



Instituto Federal de Minas Gerais
Campus Ouro Preto
Especialização em Inteligência Artificial



Pâmela Murta Castro Lima

Uso de Inteligência Artificial para Análise e Previsão da Qualidade da Água em Corpos Hídricos

Ouro Preto, 2025

Pâmela Murta Castro Lima

Uso de Inteligência Artificial para Análise e Previsão da Qualidade da Água em Corpos Hídricos

Trabalho de Conclusão de Curso apresentado ao curso de Pós-Graduação *Lato Sensu* em Inteligência Artificial, ofertado pelo Instituto Federal de Educação, Ciência e Tecnologia Minas Gerais - *Campus* Ouro Preto, como parte dos requisitos para a obtenção do título de Especialista em Inteligência Artificial.

Instituto Federal de Minas Gerais

Orientador: Prof. Dr. Moisés Henrique Ramos Pereira

Ouro Preto

2025

L732u

Lima, Pâmela Murta Castro.

Uso de Inteligência Artificial para análise e previsão da qualidade da água em corpos hídricos [manuscrito] / Pâmela Murta Castro Lima. – 2025.

216 f. : il.

Orientador: Moisés Henrique Ramos Pereira.

Trabalho de Conclusão de Curso (especialização) – Instituto Federal de Minas Gerais. *Campus* Ouro Preto, 2025.

1. Aprendizado de Máquina. 2. Random Forest . 3. Qualidade da água. I. Pereira, Moisés Henrique Ramos. II. Instituto Federal de Minas Gerais. *Campus* Ouro Preto. III. Título.

CDU: 004.85

Catálogo: Andresa Aredes Ferreira - CRB-6/MG-003262/0



MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS
Campus Ouro Preto
Diretoria de Pesquisa, Inovação e Pós-Graduação
Coordenação do Curso de Pós-graduação em Inteligência Artificial
Rua Pandiá Calógeras, 898 - Bairro Bauxita - CEP 35400-000 - Ouro Preto - MG
- www.ifmg.edu.br

FOLHA DE APROVAÇÃO

PÂMELA MURTA CASTRO LIMA

USO DE INTELIGÊNCIA ARTIFICIAL PARA ANÁLISE E PREVISÃO DA QUALIDADE DA ÁGUA EM CORPOS HÍDRICOS

Trabalho de Conclusão de Curso apresentado ao curso de ESPECIALIZAÇÃO EM INTELIGÊNCIA ARTIFICIAL, ofertado pelo Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais - *Campus* Ouro Preto, como parte dos requisitos para a obtenção do título de ESPECIALISTA EM INTELIGÊNCIA ARTIFICIAL.

Aprovada em 28 de novembro de 2025, pela Banca Examinadora:

Prof. Dr. **Moisés H. R. Pereira** - IFMG *Campus* Ribeirão das Neves - Orientador

Prof. Dr. **Adriano Lages dos Santos** - IFMG *Campus* Ribeirão das Neves

Prof. Dr. **Carlos Alexandre Silva** - IFMG *Campus* Sabará

Ouro Preto, 28 de novembro de 2025.



Documento assinado eletronicamente por **Moises Henrique Ramos Pereira, Membro da Banca Examinadora**, em 01/12/2025, às 11:35, conforme Decreto nº 10.543, de 13 de novembro de 2020.



Documento assinado eletronicamente por **Carlos Alexandre Silva, Professor EBTT**, em 01/12/2025, às 11:42, conforme Decreto nº 10.543, de 13 de novembro de 2020.



Documento assinado eletronicamente por **Adriano Lages dos Santos, Professor**, em 01/12/2025, às 14:39, conforme Decreto nº 10.543, de 13 de novembro de 2020.



A autenticidade do documento pode ser conferida no site <https://sei.ifmg.edu.br/consultadocs> informando o código verificador **2541831** e o código CRC **99324692**.

Agradecimentos

A conclusão deste trabalho não seria possível sem o apoio, incentivo e presença de pessoas fundamentais em minha trajetória. Cada uma, à sua maneira, contribuiu para que eu chegasse até aqui, e é com profunda gratidão que registro meu reconhecimento.

Aos meus pais, Geraldo Magela Castro Lima e Maria Sílvia Murta Castro Lima (*in memoriam*), dedico esta conquista. Minha mãe, com sua força e determinação inesgotáveis, sempre acreditou na educação como caminho de transformação e lutou incansavelmente para que nós, seus filhos, tivéssemos oportunidades que ela não teve. Foi seu incentivo que plantou em mim o desejo de aprender e de nunca desistir. Meu pai, pilar fundamental da nossa família, sempre garantiu a estrutura necessária para que pudéssemos seguir em frente e, até hoje, permanece como meu ponto de apoio inabalável. A vocês, toda minha admiração e amor.

Às minhas irmãs, Cintia Murta Castro Lima e Ludmila Murta Castro Lima, companheiras de toda uma vida, agradeço pelo amor incondicional, pela cumplicidade e por estarem sempre ao meu lado nos momentos de alegria e nos de incerteza. Vocês são parte essencial de quem eu sou. Ao meu cunhado Arthur Cardoso Figueiredo, agradeço pela parceria constante, pelo apoio em todas as circunstâncias e por ser alguém com quem sempre posso contar. Ao meu cunhado Jaide Câmara Hilário, meu reconhecimento sincero por fazer parte desta jornada e por contribuir, à sua maneira, para o equilíbrio e harmonia da nossa família.

Aos meus sobrinhos Filipe e Gustavo, que são minha vida inteirinha e minha fonte diária de alegria e motivação. Filipe, você me ensina todos os dias o significado de amor incondicional. Gustavo, mesmo ainda na barriga da sua mãe, já preenche meu coração de esperança e me lembra que cada conquista é um legado para as próximas gerações. Vocês são a razão pela qual eu sigo em frente.

À minha avó, Elizabeth Efigênia das Dores Murta, mulher extraordinária cuja história de vida me inspira profundamente. Ao conhecer sua trajetória, compreendi que sua luta, coragem e determinação moldaram o futuro de nossa família e abriram caminhos para as gerações seguintes, incluindo a minha. Seu legado de perseverança e amor me acompanha em cada conquista.

Às minhas amigas Natália Andalescio e Patrícia Werneck, agradeço pela amizade verdadeira que resiste ao tempo e às circunstâncias. Patrícia, sua presença constante desde a infância é um presente raro, e sua lealdade inabalável me dá força para seguir. Natália, mesmo nos períodos em que a vida nos separou, você nunca deixou de estar presente em meu coração, e hoje, com nosso reencontro, reafirmamos que amizade verdadeira não

conhece distância nem tempo. Vocês são irmãs que a vida me deu. À minha princesinha Alice Andalescio, sobrinha do meu coração, agradeço por existir e por me ensinar, com sua pureza e alegria, que o amor não precisa de laços de sangue para ser verdadeiro. Você ilumina minha vida. Ao meu afilhado Thiago Werneck, que hoje já é um homem de 18 anos, agradeço por me permitir acompanhar seu crescimento desde que você nasceu. Ver você se tornar o homem que é hoje enche meu coração de orgulho e gratidão.

Ao meu amigo Paulo Lima Penido, agradeço pelo apoio constante, pela parceria em diversos momentos desta jornada acadêmica e pelas conversas que me ajudaram a seguir em frente quando as dificuldades pareciam intransponíveis. Sua amizade e incentivo foram essenciais para a conclusão deste trabalho.

Ao meu amigo Gustavo Sena, meu irmão de coração e porto seguro no ambiente de trabalho, agradeço pela parceria, pelo apoio nos momentos mais desafiadores e por tornar o cotidiano mais leve com sua presença. À Edicélia Peixinho, Raquel Penchel, Ana Carolina Fortini, Allysson Oliveira e Igor Mendes, colegas que se tornaram amigos essenciais, minha gratidão por compartilharem comigo não apenas as responsabilidades profissionais, mas também momentos de incentivo e solidariedade que fizeram toda a diferença nesta jornada acadêmica.

Ao Sindicato dos Correios, registro meu profundo reconhecimento pelo apoio institucional e pelo acolhimento nos períodos mais difíceis. O suporte que recebi da entidade e de seus representantes foi fundamental para que eu pudesse conciliar trabalho, vida pessoal e estudos, permitindo-me chegar até a conclusão deste curso de especialização. Esse apoio significou muito mais do que palavras podem expressar e contribuiu decisivamente para esta conquista.

Ao meu orientador, professor Moisés Henrique Ramos Pereira, minha gratidão pela orientação precisa, pela paciência em cada etapa deste trabalho e por acreditar no potencial desta pesquisa. Suas contribuições foram fundamentais para o amadurecimento das ideias aqui apresentadas e para o rigor metodológico alcançado. À professora Sílvia, agradeço pelos ensinamentos valiosos ao longo do curso, que ampliaram minha visão sobre Inteligência Artificial e suas aplicações. Ao corpo docente do Instituto Federal de Minas Gerais - *Campus* Ouro Preto, meu reconhecimento pela dedicação e pela excelência na formação de profissionais comprometidos com o conhecimento e a transformação social.

Aos meus filhinhos de quatro patas, Maria Júlia e Francisco, agradeço pela companhia incondicional em cada madrugada de estudo, por estarem ao meu lado nos momentos de cansaço e por me lembrarem, com sua alegria simples e amor puro, que as pausas para um carinho são tão importantes quanto as horas dedicadas ao trabalho. Vocês tornaram esta jornada mais leve e repleta de afeto. À Mel e Cookie (*in memoriam*), que partiram mas permanecem eternamente em meu coração, agradeço por cada momento de felicidade que me proporcionaram e por me ensinarem o significado do amor verdadeiro e da saudade

que aquece. Vocês foram e sempre serão parte essencial da minha vida.

A todos que, de alguma forma, fizeram parte desta caminhada — familiares, amigos, colegas de trabalho e professores —, meu sincero obrigada. Esta vitória não é apenas minha; é de todos nós.

*“A educação é a arma mais poderosa
que você pode usar para mudar o mundo.”
(Nelson Mandela)*

Resumo

Este trabalho investiga a aplicação de técnicas de aprendizado de máquina para análise e predição da qualidade da água em áreas impactadas pelo rompimento da Barragem de Fundão na Bacia do Rio Gualaxo do Norte (Mariana/MG). A pesquisa estruturou-se em duas fases complementares: análise exploratória não supervisionada mediante redução dimensional (PCA, UMAP, t-SNE) e clusterização (K-Means, HDBSCAN), seguida de modelagem preditiva supervisionada com três algoritmos (Random Forest, Gradient Boosting Machines e Multi-Layer Perceptron). Utilizou-se base de dados de Santos (2018) contendo 324 amostras de água coletadas em 27 pontos ao longo de 12 campanhas de campo. A análise não supervisionada identificou configuração ótima (Bloco C, 8 variáveis) com coeficiente de silhueta excepcional de 0,834 e estabilidade temporal de 81,2%, evidenciando agrupamentos naturais robustos frente à variabilidade sazonal. A modelagem supervisionada demonstrou superioridade do Random Forest, que alcançou acurácia de 96,77% no cenário de alta pureza de rótulos (151 amostras), superando o *baseline* de Santos (2018) em 2,56 pontos percentuais. Técnicas de inteligência artificial explicável (SHAP e LIME) identificaram Fósforo Total e *Escherichia coli* como variáveis mais discriminativas, com contribuição combinada superior a 50% das decisões. Análise crítica revelou que qualidade de rótulos domina quantidade de dados: cenário com 151 amostras e pureza $\geq 70\%$ superou configuração com 303 amostras de pureza moderada em 11,52 pontos percentuais. Demonstrou-se empiricamente o fenômeno de vazamento de dados (*data leakage*) quando coordenadas de projeções dimensionais são utilizadas como variáveis preditoras, fundamentando exclusão deliberada dessas coordenadas para garantir generalização genuína. Os resultados validam aplicabilidade de modelos de aprendizado de máquina interpretáveis como ferramenta para monitoramento ambiental pós-desastre, com implicações práticas para gestão de recursos hídricos e proposição de protocolo simplificado focado em variáveis-chave.

Palavras-chave: Aprendizado de Máquina; Qualidade da Água; Random Forest; Inteligência Artificial Explicável; Monitoramento Ambiental Pós-Desastre; Clusterização; Barragem de Fundão.

Abstract

This study investigates the application of machine learning techniques for water quality analysis and prediction in areas impacted by the Fundão Dam rupture in the Gualaxo do Norte River Basin (Mariana, Minas Gerais, Brazil). The research was structured in two complementary phases: unsupervised exploratory analysis through dimensionality reduction (PCA, UMAP, t-SNE) and clustering (K-Means, HDBSCAN), followed by supervised predictive modeling with three algorithms (Random Forest, Gradient Boosting Machines, and Multi-Layer Perceptron). The Santos (2018) database was used, containing 324 water samples collected from 27 points across 12 field campaigns. Unsupervised analysis identified an optimal configuration (Block C, 8 variables) with exceptional silhouette coefficient of 0.834 and temporal stability of 81.2%, revealing natural clusters robust to seasonal variability. Supervised modeling demonstrated Random Forest superiority, achieving 96.77% accuracy in the high-purity label scenario (151 samples), surpassing Santos' (2018) baseline by 2.56 percentage points. Explainable artificial intelligence techniques (SHAP and LIME) identified Total Phosphorus and *Escherichia coli* as the most discriminative variables, with combined contribution exceeding 50% of decisions. Critical analysis revealed that label quality dominates data quantity: a scenario with 151 samples and purity $\geq 70\%$ outperformed a configuration with 303 moderate-purity samples by 11.52 percentage points. Data leakage phenomenon was empirically demonstrated when dimensional projection coordinates are used as predictive variables, justifying their deliberate exclusion to ensure genuine generalization. Results validate the applicability of interpretable machine learning models as tools for post-disaster environmental monitoring, with practical implications for water resource management and proposing a simplified protocol focused on key variables.

Keywords: Machine Learning; Water Quality; Random Forest; Explainable Artificial Intelligence; Post-Disaster Environmental Monitoring; Clustering; Fundão Dam.

Lista de ilustrações

Figura 1 – Dashboard de síntese do tratamento de <i>outliers</i> com validação físico-química. Fonte: Elaboração própria (2025).	68
Figura 2 – Distribuição das variáveis físico-químicas por grupo de referência. Fonte: Elaboração própria (2025).	69
Figura 3 – Painel analítico do tratamento de <i>outliers</i> com validação físico-química.	72
Figura 4 – Matriz de correlação de Pearson entre variáveis físico-químicas e microbiológicas. Fonte: Elaboração própria (2025).	74
Figura 5 – Relações bivariadas mais fortes identificadas pela análise de correlação. Fonte: Elaboração própria (2025).	76
Figura 6 – Análise completa do Análise de Componentes Principais (PCA) — Bloco Minerário.	82
Figura 7 – Projeções bidimensionais Uniform Manifold Approximation and Projection (UMAP) e t-distributed Stochastic Neighbor Embedding (t-SNE).	83
Figura 8 – Comparação das métricas internas de qualidade entre as quatro configurações testadas no Bloco Minerário.	84
Figura 9 – Análise de Componentes Principais do Bloco Sanitário,	87
Figura 10 – Projeções bidimensionais do Bloco Sanitário.	88
Figura 11 – Comparação visual das métricas de validação interna para o Bloco Sanitário:	89
Figura 12 – Análise completa do PCA — Bloco Biogeoquímico	93
Figura 13 – Projeções bidimensionais UMAP e t-SNE — Bloco Biogeoquímico	94
Figura 14 – Comparação das métricas internas de qualidade entre as quatro configurações testadas no Bloco Biogeoquímico	95
Figura 15 – Comparação do Adjusted Rand Index (ARI) entre blocos temáticos para classificação binária (Afetado vs Não-Afetado) e multiclasse (4 categorias hidrológicas).	98
Figura 16 – Painel comparativo dos biplots PCA dos três blocos temáticos (Fase 1).	100
Figura 17 – Comparação da variância explicada pelos componentes principais nos Blocos A, B e C.	104
Figura 18 – <i>Biplots</i> da Análise de Componentes Principais para Blocos A (superior), B (centro) e C (inferior).	105
Figura 19 – Comparação consolidada das métricas internas dos Blocos A, B e C para as quatro configurações testadas (Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)/ <i>K-Means</i> × UMAP/ <i>t-SNE</i>).	107
Figura 20 – Configurações de clusterização aplicadas ao Bloco A.	108

Figura 21 – Configurações de clusterização aplicadas ao Bloco B.	109
Figura 22 – Configurações de clusterização aplicadas ao Bloco C.	110
Figura 23 – Validação hierárquica em três níveis (amostra, ponto, campo) para Bloco C com <i>K-Means</i> +UMAP.	111
Figura 24 – Matrizes de confusão entre clusters <i>K-Means</i> +UMAP e gabaritos de referência para Blocos A (superior), B (centro) e C (inferior).	112
Figura 25 – Trajetórias temporais dos pontos de amostragem no Bloco A (<i>K-Means</i> +UMAP).	114
Figura 26 – Trajetórias temporais dos pontos de amostragem no Bloco B (<i>K-Means</i> +UMAP).	115
Figura 27 – Trajetórias temporais dos pontos de amostragem no Bloco C (<i>K-Means</i> +UMAP).	116
Figura 28 – Análise de transições sazonais entre estações seca (campanhas 1-6) e chuvosa (campanhas 7-12) para Bloco C (<i>acrshortkmeans</i> + <i>acrshortUMAP</i>).	117
Figura 29 – Comparação visual do desempenho dos modelos com e sem coordenadas UMAPUMAP como variáveis preditoras.	123
Figura 30 – Projeção UMAP das 8 variáveis hidroquímicas com clusters HDBSCAN identificados.	126
Figura 31 – Heatmap de Z-scores das 8 variáveis hidroquímicas por cluster.	128
Figura 32 – Comparação das distribuições univariadas entre amostras reais e sintéticas no Cenário 1. As distribuições preservam forma, variabilidade e limites naturais.	133
Figura 33 – Matrizes comparativas de correlação de Pearson para amostras reais e sintéticas. O SMOTE preserva a estrutura de covariância das variáveis.	134
Figura 34 – Heatmap consolidado de acurácia para os 15 experimentos.	138
Figura 35 – Comparação consolidada dos três algoritmos ao longo dos cinco cenários.	139
Figura 36 – Comparação de desempenho do Cenário 1 com o estudo de Santos (2018).	140
Figura 37 – Detecção visual de overfitting/underfitting através do gap entre validação cruzada e teste.	142
Figura 38 – Matrizes de confusão para o Cenário 2 (Alta Pureza) — melhor desempenho geral.	143
Figura 39 – Equilíbrio entre pureza dos clusters e volume amostral.	144
Figura 40 – Importância global das variáveis (SHapley Additive exPlanations (SHAP)) para Random Forest (RF) no Cenário 2.	146
Figura 41 – Comparação de importância SHAP entre cenários.	146
Figura 42 – Explicações Local Interpretable Model-agnostic Explanations (LIME) para casos típicos (predições corretas) do RF no Cenário 2.	147

Figura 43 – Explicações LIME para casos de erro da Multi-Layer Perceptron (MLP) no Cenário 4	148
Figura 44 – Estrutura geral dos diretórios e arquivos gerados — Pipeline Supervisi- onado.	201

Lista de tabelas

Tabela 2 – Limites de qualidade da água conforme a Resolução CONAMA n.º 357/2005	32
Tabela 3 – Da análise temática (Fase 1) aos Blocos Principais (Fase 2): critérios de seleção e composição.	46
Tabela 4 – Cenários definidos para modelagem supervisionada (Bloco C, HDBS-CAN+UMAP).	50
Tabela 5 – Ferramentas do pipeline principal de treinamento supervisionado . . .	62
Tabela 6 – Ferramentas do módulo XAI (interpretabilidade)	63
Tabela 7 – Ferramentas do módulo de validação de SMOTE	63
Tabela 8 – Ferramentas dos módulos de comparação	64
Tabela 9 – Ferramentas do módulo de modelos robustos	64
Tabela 10 – Variância explicada pelos componentes principais (Bloco Minerário). .	81
Tabela 11 – Métricas internas de qualidade dos agrupamentos (Bloco Minerário). .	83
Tabela 12 – Validação externa dos agrupamentos: confronto com gabaritos de referência (Bloco Minerário).	85
Tabela 13 – Variância explicada pelos componentes principais no Bloco Sanitário. .	86
Tabela 14 – Loadings das variáveis sanitárias nos três primeiros componentes principais.	86
Tabela 15 – Métricas de validação interna (Bloco Sanitário).	89
Tabela 16 – Validação externa: confronto com gabaritos de referência (Bloco Sanitário). .	90
Tabela 17 – Variância explicada pelos componentes principais (Bloco Biogeoquímico). .	92
Tabela 18 – Métricas internas de qualidade dos agrupamentos (Bloco Biogeoquímico). .	95
Tabela 19 – Validação externa dos agrupamentos: confronto com gabaritos de referência (Bloco Biogeoquímico).	96
Tabela 20 – Desempenho comparativo dos blocos temáticos na Fase 1 (análise não supervisionada).	97
Tabela 21 – Seleção de variáveis por bloco temático mediante critério de loadings dominantes ($> 0,90$).	99
Tabela 22 – Composição e estratégias de diferenciação dos blocos principais (Fase 2). .	103
Tabela 23 – Variância explicada pelos componentes principais nos blocos principais (Fase 2).	103
Tabela 24 – Métricas internas de qualidade dos agrupamentos — Blocos Principais (projeções UMAP).	106
Tabela 25 – Validação externa hierárquica dos blocos principais — gabarito multi-classe (<i>K-Means</i>).	107

Tabela 26 – Análise de pureza dos clusters em relação ao gabarito multiclasse (HDBSCAN+UMAP).	110
Tabela 27 – Estabilidade temporal intra-ponto dos blocos principais (<i>K-Means</i> + UMAP).	113
Tabela 28 – Consolidação comparativa dos blocos principais — configurações prioritárias.	117
Tabela 29 – Cenários definidos para modelagem supervisionada (Bloco C, HDBSCAN+UMAP).	118
Tabela 30 – Comparação de desempenho: configurações sem e com coordenadas UMAP.	123
Tabela 31 – Resultados da validação físico-química das amostras sintéticas geradas pelo SMOTE (Cenário 1).	132
Tabela 32 – Distribuição de valores outliers por variável nas amostras sintéticas (Cenário 1).	133
Tabela 33 – Comparação de estatísticas descritivas entre amostras reais e sintéticas (Classe RP_Afetado).	134
Tabela 34 – Preservação de correlações entre variáveis selecionadas (Classe RP_Afetado).	135
Tabela 35 – Cenários de classificação implementados na análise supervisionada (Bloco C).	136
Tabela 36 – Desempenho consolidado: métricas de avaliação para os 15 experimentos.	137
Tabela 37 – Custos estimados por variável e componente operacional.	150
Tabela 38 – Comparação de custos por campanha de monitoramento.	151
Tabela 39 – Análise de sensibilidade — economia anual em diferentes cenários.	151
Tabela 40 – Resumo estatístico da análise de outliers por parâmetro.	167
Tabela 41 – Exemplos representativos de outliers mantidos por coerência químico-ambiental.	167
Tabela 42 – Exemplos representativos de outliers tratados por ausência de coerência químico-ambiental.	168
Tabela 43 – Configuração dos blocos temáticos para análise exploratória	174
Tabela 44 – Configuração dos blocos principais para comparação sistemática	175
Tabela 45 – Principais funções do pipeline e responsabilidades técnicas	177
Tabela 46 – Parâmetros otimizados dos algoritmos	179
Tabela 47 – Principais funções do pipeline supervisionado e responsabilidades	192
Tabela 48 – Preservação de correlações químicas após SMOTE	198
Tabela 49 – Validação de coerência química via regras de especialistas	198
Tabela 50 – Comparação de desempenho: Pipeline atual vs. Baseline Santos (2018)	204
Tabela 51 – Amostra dos dados brutos coletados por Santos (2018) (primeiras 10 observações).	209

Lista de Abreviaturas e Siglas

R^2	Coeficiente de Determinação.
<i>E. coli</i>	<i>Escherichia coli</i> .
Acc.	Acurácia.
AD	Análise Discriminante.
Alca.	Alcalinidade.
ANA	Agência Nacional de Águas e Saneamento Básico.
ANN	Artificial Neural Network.
ARI	Adjusted Rand Index.
AUC	Area Under the Curve.
C	Campo.
CH	Calinski-Harabasz Index.
CL	Cloretos.
CONAMA	Conselho Nacional do Meio Ambiente.
COND	Condutividade Elétrica.
CT	Coliformes Totais.
CV	Coeficiente de Variação.
CV	Cross-Validation.
DB	Davies-Bouldin Index.
DBO	Demanda Bioquímica de Oxigênio.
DBSCAN	Density-Based Spatial Clustering of Applications with Noise.
EPA	Environmental Protection Agency.
F1	F1-Score.
GAP	Diagnóstico de Generalização (CV Mean – Test Accuracy).
GBM	Gradient Boosting Machines.

HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise.
IA	Inteligência Artificial.
IBAMA	Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis.
IGAM	Instituto Mineiro de Gestão das Águas.
IQA	Índice de Qualidade da Água.
IQR	Intervalo Interquartil.
k-Fold	k-Fold Cross-Validation.
K-Means	K-Means Clustering.
KS	Teste de Kolmogorov-Smirnov.
LDA	Análise Discriminante Linear.
LIME	Local Interpretable Model-agnostic Explanations.
LOOCV	Leave-One-Out Cross-Validation.
MAE	Mean Absolute Error.
ML	Machine Learning.
MLP	Multi-Layer Perceptron.
MSE	Mean Squared Error.
NMI	Normalized Mutual Information.
NO ₂	Nitrito.
NO ₃	Nitrato.
OD	Oxigênio Dissolvido.
OMS	Organização Mundial da Saúde.
ORP	Potencial de Oxirredução.
OvO	one-vs-one.
OvR	one-vs-rest.
PC	Ponto de Coleta.
PCA	Análise de Componentes Principais.
PT	Fósforo Total.
RF	Random Forest.

RMSE	Root Mean Squared Error.
ROC	Receiver Operating Characteristic.
ROC–AUC	Area Under the ROC Curve.
RP	Rio Principal.
SatOD	Saturação de Oxigênio Dissolvido.
SHAP	SHapley Additive exPlanations.
Silh.	Silhouette Score.
SMOTE	Synthetic Minority Over-sampling Technique.
ST	Sólidos Totais.
STD	Desvio Padrão.
SVM	Support Vector Machine.
t-SNE	t-distributed Stochastic Neighbor Embedding.
Ta	Temperatura.
Turb.	Turbidez.
UMAP	Uniform Manifold Approximation and Projection.
XAI	Inteligência Artificial Explicável.

Sumário

	Agradecimentos	4
	Lista de Abreviaturas e Siglas	15
1	INTRODUÇÃO	26
1.1	Objetivo	27
1.2	Objetivos Específicos	27
1.3	Justificativas e Relevância	28
1.4	Organização da Estrutura do Trabalho	28
2	FUNDAMENTAÇÃO TEÓRICA	30
2.1	Qualidade da água e indicadores ambientais	30
2.1.1	Parâmetros Físico-Químicos e Microbiológicos: Definição e Relevância	30
2.1.2	Normas e Padrões de Referência para Qualidade da Água	32
2.1.3	Especificidades da Bacia do Rio Gualaxo do Norte	33
2.2	Técnicas Estatísticas e Inteligência Artificial na Análise Ambiental	33
2.2.1	Métodos Estatísticos Tradicionais	33
2.2.2	Modelos de Aprendizado de Máquina na Previsão da Qualidade da Água	34
2.2.3	Técnicas de Redução de Dimensionalidade: t-SNE e UMAP	35
2.2.4	Inteligência Artificial Explicável (XAI): SHAP e LIME	35
2.2.5	Trabalhos Correlatos e o Estado da Arte	37
3	METODOLOGIA	38
3.1	Conjunto de Dados	38
3.1.1	Contexto Espacial e Temporal da Coleta	38
3.1.2	Variáveis Monitoradas	38
3.1.3	Características da Base Bruta e Desafios de Qualidade	39
3.1.4	Limitações Reconhecidas da Base Original	40
3.2	Pipeline de Pré-processamento	41
3.3	Pipeline de Análise Não Supervisionada — Estruturação e Lógica Metodológica	43
3.3.1	Estratégia de seleção de variáveis	43
3.3.2	Fase 2: Construção dos Blocos Principais	45
3.3.3	Redução dimensional e projeções	48
3.3.4	Clusterização e métricas de desempenho	48
3.3.5	Pontos fortes e síntese metodológica	48

3.4	Pipeline de Análise Supervisionado	49
3.4.1	Definição dos Cenários de Classificação	49
3.4.2	Fluxo de Trabalho e Modelos de Classificação	51
3.4.3	Seleção do Bloco de Variáveis	51
3.4.4	Controle Rigoroso de Data Leakage	51
3.4.5	Pré-processamento dos Dados	53
3.4.6	Balanceamento de Classes com Synthetic Minority Over-sampling Technique (SMOTE)	54
3.4.7	Treinamento e Otimização dos Modelos	57
3.4.8	Métricas de Avaliação	59
3.4.9	Comparação com a Metodologia de Santos (2018)	60
3.4.10	Estratégia de Comparação dos Resultados	60
3.4.11	Inteligência Artificial Explicável (Inteligência Artificial Explicável (XAI))	61
3.4.12	Ferramentas e Bibliotecas	62
3.5	Análise de Viabilidade Econômica do Protocolo Simplificado	65
3.5.1	Fontes de Dados de Custos	65
3.5.2	Metodologia de Cálculo	65
4	RESULTADOS E DISCUSSÕES	67
4.1	Caracterização da Base de Dados Tratada	67
4.1.1	Estatísticas Descritivas Finais	68
4.1.2	Síntese da Efetividade do Tratamento de <i>Outliers</i>	71
4.1.3	Matriz de Correlação e Relações Bivariadas	73
4.1.4	Gabaritos de Referência para Validação (SANTOS, 2018)	76
4.2	Resultados da Modelagem Não Supervisionada	80
4.2.1	Análise dos Blocos Temáticos (Fase 1)	80
4.2.2	Análise dos Blocos Principais (Fase 2)	102
4.3	Resultados da Modelagem Supervisionada	121
4.3.1	Decisão Metodológica: Exclusão de Coordenadas UMAP como Variáveis Preditoras	122
4.3.2	Interpretação Crítica dos Resultados	123
4.3.3	Justificativa Metodológica Final	124
4.3.4	Caracterização dos Clusters: Projeção UMAP e Assinatura Química	124
4.3.5	Validação de Coerência Físico-Química das Amostras Sintéticas (SMOTE)	131
4.3.6	Configuração dos Cenários de Classificação	135
4.3.7	Desempenho Global dos Modelos	136
4.3.8	Comparação com Baseline (Santos 2018)	139
4.3.9	Análise de Problemas de Generalização	140
4.3.10	Equilíbrio entre Pureza e Volume Amostral	143
4.3.11	Análise de Padrões de Erro via Matrizes de Confusão	144

4.3.12	Interpretabilidade via Inteligência Artificial Explicável	145
4.3.13	Síntese dos Resultados Supervisionados	148
4.4	Análise de Viabilidade Econômica do Protocolo Simplificado	150
4.4.1	Estrutura de Custos por Componente	150
4.4.2	Comparação de Custos Totais	150
4.4.3	Análise de Sensibilidade	151
4.4.4	Discussão e Aplicabilidade	151
4.4.5	Limitações e Recomendações	152
5	CONSIDERAÇÕES FINAIS	153
5.1	Síntese dos Principais Resultados	153
5.2	Alcance dos Objetivos da Pesquisa	155
5.3	Limitações do Estudo	157
5.4	Recomendações para Trabalhos Futuros	159
5.5	Contribuições Metodológicas e Aplicabilidade para Gestão Hídrica .	161
A	APÊNDICE A — MÓDULO DE VALIDAÇÃO DE OUTLIERS . . .	164
B	APÊNDICE B — ESPECIFICAÇÕES TÉCNICAS DA ANÁLISE NÃO SUPERVISIONADA	169
C	APÊNDICE C — ESPECIFICAÇÕES TÉCNICAS DA ANÁLISE SUPERVISIONADA	188
D	APÊNDICE D — CÁLCULO DETALHADO DA ECONOMIA ESTI- MADA	205
E	ANEXOS	208
E.1	ANEXO A – Amostra dos Dados Brutos Coletados	208
	Referências	211

Lista de tabelas

Tabela 2 – Limites de qualidade da água conforme a Resolução CONAMA n.º 357/2005	32
Tabela 3 – Da análise temática (Fase 1) aos Blocos Principais (Fase 2): critérios de seleção e composição.	46
Tabela 4 – Cenários definidos para modelagem supervisionada (Bloco C, HDBS-CAN+UMAP).	50
Tabela 5 – Ferramentas do pipeline principal de treinamento supervisionado . . .	62
Tabela 6 – Ferramentas do módulo XAI (interpretabilidade)	63
Tabela 7 – Ferramentas do módulo de validação de SMOTE	63
Tabela 8 – Ferramentas dos módulos de comparação	64
Tabela 9 – Ferramentas do módulo de modelos robustos	64
Tabela 10 – Variância explicada pelos componentes principais (Bloco Minerário). .	81
Tabela 11 – Métricas internas de qualidade dos agrupamentos (Bloco Minerário). .	83
Tabela 12 – Validação externa dos agrupamentos: confronto com gabaritos de referência (Bloco Minerário).	85
Tabela 13 – Variância explicada pelos componentes principais no Bloco Sanitário. .	86
Tabela 14 – Loadings das variáveis sanitárias nos três primeiros componentes principais.	86
Tabela 15 – Métricas de validação interna (Bloco Sanitário).	89
Tabela 16 – Validação externa: confronto com gabaritos de referência (Bloco Sanitário). .	90
Tabela 17 – Variância explicada pelos componentes principais (Bloco Biogeoquímico). .	92
Tabela 18 – Métricas internas de qualidade dos agrupamentos (Bloco Biogeoquímico). .	95
Tabela 19 – Validação externa dos agrupamentos: confronto com gabaritos de referência (Bloco Biogeoquímico).	96
Tabela 20 – Desempenho comparativo dos blocos temáticos na Fase 1 (análise não supervisionada).	97
Tabela 21 – Seleção de variáveis por bloco temático mediante critério de loadings dominantes ($> 0,90$).	99
Tabela 22 – Composição e estratégias de diferenciação dos blocos principais (Fase 2). .	103
Tabela 23 – Variância explicada pelos componentes principais nos blocos principais (Fase 2).	103
Tabela 24 – Métricas internas de qualidade dos agrupamentos — Blocos Principais (projeções UMAP).	106
Tabela 25 – Validação externa hierárquica dos blocos principais — gabarito multi-classe (<i>K-Means</i>).	107

Tabela 26 – Análise de pureza dos clusters em relação ao gabarito multiclasse (HDBSCAN+UMAP).	110
Tabela 27 – Estabilidade temporal intra-ponto dos blocos principais (<i>K-Means</i> + UMAP).	113
Tabela 28 – Consolidação comparativa dos blocos principais — configurações prioritárias.	117
Tabela 29 – Cenários definidos para modelagem supervisionada (Bloco C, HDBSCAN+UMAP).	118
Tabela 30 – Comparação de desempenho: configurações sem e com coordenadas UMAP.	123
Tabela 31 – Resultados da validação físico-química das amostras sintéticas geradas pelo SMOTE (Cenário 1).	132
Tabela 32 – Distribuição de valores outliers por variável nas amostras sintéticas (Cenário 1).	133
Tabela 33 – Comparação de estatísticas descritivas entre amostras reais e sintéticas (Classe RP_Afetado).	134
Tabela 34 – Preservação de correlações entre variáveis selecionadas (Classe RP_Afetado).	135
Tabela 35 – Cenários de classificação implementados na análise supervisionada (Bloco C).	136
Tabela 36 – Desempenho consolidado: métricas de avaliação para os 15 experimentos.	137
Tabela 37 – Custos estimados por variável e componente operacional.	150
Tabela 38 – Comparação de custos por campanha de monitoramento.	151
Tabela 39 – Análise de sensibilidade — economia anual em diferentes cenários.	151
Tabela 40 – Resumo estatístico da análise de outliers por parâmetro.	167
Tabela 41 – Exemplos representativos de outliers mantidos por coerência químico-ambiental.	167
Tabela 42 – Exemplos representativos de outliers tratados por ausência de coerência químico-ambiental.	168
Tabela 43 – Configuração dos blocos temáticos para análise exploratória	174
Tabela 44 – Configuração dos blocos principais para comparação sistemática	175
Tabela 45 – Principais funções do pipeline e responsabilidades técnicas	177
Tabela 46 – Parâmetros otimizados dos algoritmos	179
Tabela 47 – Principais funções do pipeline supervisionado e responsabilidades	192
Tabela 48 – Preservação de correlações químicas após SMOTE	198
Tabela 49 – Validação de coerência química via regras de especialistas	198
Tabela 50 – Comparação de desempenho: Pipeline atual vs. Baseline Santos (2018)	204
Tabela 51 – Amostra dos dados brutos coletados por Santos (2018) (primeiras 10 observações).	209

Sumário

	Agradecimentos	4
	Lista de Abreviaturas e Siglas	15
1	INTRODUÇÃO	26
1.1	Objetivo	27
1.2	Objetivos Específicos	27
1.3	Justificativas e Relevância	28
1.4	Organização da Estrutura do Trabalho	28
2	FUNDAMENTAÇÃO TEÓRICA	30
2.1	Qualidade da água e indicadores ambientais	30
2.1.1	Parâmetros Físico-Químicos e Microbiológicos: Definição e Relevância	30
2.1.2	Normas e Padrões de Referência para Qualidade da Água	32
2.1.3	Especificidades da Bacia do Rio Gualaxo do Norte	33
2.2	Técnicas Estatísticas e Inteligência Artificial na Análise Ambiental	33
2.2.1	Métodos Estatísticos Tradicionais	33
2.2.2	Modelos de Aprendizado de Máquina na Previsão da Qualidade da Água	34
2.2.3	Técnicas de Redução de Dimensionalidade: t-SNE e UMAP	35
2.2.4	Inteligência Artificial Explicável (XAI): SHAP e LIME	35
2.2.5	Trabalhos Correlatos e o Estado da Arte	37
3	METODOLOGIA	38
3.1	Conjunto de Dados	38
3.1.1	Contexto Espacial e Temporal da Coleta	38
3.1.2	Variáveis Monitoradas	38
3.1.3	Características da Base Bruta e Desafios de Qualidade	39
3.1.4	Limitações Reconhecidas da Base Original	40
3.2	Pipeline de Pré-processamento	41
3.3	Pipeline de Análise Não Supervisionada — Estruturação e Lógica Metodológica	43
3.3.1	Estratégia de seleção de variáveis	43
3.3.2	Fase 2: Construção dos Blocos Principais	45
3.3.3	Redução dimensional e projeções	48
3.3.4	Clusterização e métricas de desempenho	48
3.3.5	Pontos fortes e síntese metodológica	48

3.4	Pipeline de Análise Supervisionado	49
3.4.1	Definição dos Cenários de Classificação	49
3.4.2	Fluxo de Trabalho e Modelos de Classificação	51
3.4.3	Seleção do Bloco de Variáveis	51
3.4.4	Controle Rigoroso de Data Leakage	51
3.4.5	Pré-processamento dos Dados	53
3.4.6	Balanceamento de Classes com SMOTE	54
3.4.7	Treinamento e Otimização dos Modelos	57
3.4.8	Métricas de Avaliação	59
3.4.9	Comparação com a Metodologia de Santos (2018)	60
3.4.10	Estratégia de Comparação dos Resultados	60
3.4.11	Inteligência Artificial Explicável (XAI)	61
3.4.12	Ferramentas e Bibliotecas	62
3.5	Análise de Viabilidade Econômica do Protocolo Simplificado	65
3.5.1	Fontes de Dados de Custos	65
3.5.2	Metodologia de Cálculo	65
4	RESULTADOS E DISCUSSÕES	67
4.1	Caracterização da Base de Dados Tratada	67
4.1.1	Estatísticas Descritivas Finais	68
4.1.2	Síntese da Efetividade do Tratamento de <i>Outliers</i>	71
4.1.3	Matriz de Correlação e Relações Bivariadas	73
4.1.4	Gabaritos de Referência para Validação (SANTOS, 2018)	76
4.2	Resultados da Modelagem Não Supervisionada	80
4.2.1	Análise dos Blocos Temáticos (Fase 1)	80
4.2.2	Análise dos Blocos Principais (Fase 2)	102
4.3	Resultados da Modelagem Supervisionada	121
4.3.1	Decisão Metodológica: Exclusão de Coordenadas UMAP como Variáveis Preditoras	122
4.3.2	Interpretação Crítica dos Resultados	123
4.3.3	Justificativa Metodológica Final	124
4.3.4	Caracterização dos Clusters: Projeção UMAP e Assinatura Química	124
4.3.5	Validação de Coerência Físico-Química das Amostras Sintéticas (SMOTE)	131
4.3.6	Configuração dos Cenários de Classificação	135
4.3.7	Desempenho Global dos Modelos	136
4.3.8	Comparação com Baseline (Santos 2018)	139
4.3.9	Análise de Problemas de Generalização	140
4.3.10	Equilíbrio entre Pureza e Volume Amostral	143
4.3.11	Análise de Padrões de Erro via Matrizes de Confusão	144
4.3.12	Interpretabilidade via Inteligência Artificial Explicável	145

4.3.13	Síntese dos Resultados Supervisionados	148
4.4	Análise de Viabilidade Econômica do Protocolo Simplificado	150
4.4.1	Estrutura de Custos por Componente	150
4.4.2	Comparação de Custos Totais	150
4.4.3	Análise de Sensibilidade	151
4.4.4	Discussão e Aplicabilidade	151
4.4.5	Limitações e Recomendações	152
5	CONSIDERAÇÕES FINAIS	153
5.1	Síntese dos Principais Resultados	153
5.2	Alcance dos Objetivos da Pesquisa	155
5.3	Limitações do Estudo	157
5.4	Recomendações para Trabalhos Futuros	159
5.5	Contribuições Metodológicas e Aplicabilidade para Gestão Hídrica .	161
A	APÊNDICE A — MÓDULO DE VALIDAÇÃO DE OUTLIERS . . .	164
B	APÊNDICE B — ESPECIFICAÇÕES TÉCNICAS DA ANÁLISE NÃO SUPERVISIONADA	169
C	APÊNDICE C — ESPECIFICAÇÕES TÉCNICAS DA ANÁLISE SUPERVISIONADA	188
D	APÊNDICE D — CÁLCULO DETALHADO DA ECONOMIA ESTI- MADA	205
E	ANEXOS	208
E.1	ANEXO A – Amostra dos Dados Brutos Coletados	208
	Referências	211

1 Introdução

Nos últimos anos, os desastres naturais passaram a ocorrer com uma frequência e intensidade alarmantes. Tal cenário reflete, em grande medida, os efeitos das mudanças climáticas globais, associadas ao crescimento urbano desordenado e à pressão constante sobre os recursos naturais. Esses eventos extremos, como enchentes, deslizamentos e secas prolongadas, comprometem significativamente a infraestrutura das cidades, impactam negativamente a economia local e colocam em risco a segurança hídrica das populações afetadas (SANTOS, 2018).

A preservação da qualidade da água, nesse contexto, torna-se um desafio de ordem pública e ambiental. Não se trata apenas de um parâmetro técnico, mas de um elemento essencial à saúde humana e à manutenção dos ecossistemas aquáticos. O rompimento da Barragem de Fundão, ocorrido em 15 de novembro de 2015 no município de Mariana, em Minas Gerais, é um exemplo marcante dos impactos provocados por eventos dessa natureza. A lama que percorreu o leito do Rio Doce deixou não apenas um rastro de destruição ambiental, mas também expôs a vulnerabilidade dos nossos sistemas hídricos diante de desastres tecnológicos e ambientais.

A análise da deterioração da qualidade da água após eventos como esse revela um problema de caráter estrutural. Ainda que ações emergenciais sejam indispensáveis, torna-se cada vez mais necessário adotar metodologias que permitam a antecipação dos impactos e a compreensão dos padrões ambientais associados a esses episódios. Nesse sentido, a pesquisa científica assume papel central ao propor abordagens inovadoras que possibilitem o aprimoramento da gestão de riscos e a mitigação de danos socioambientais.

Neste cenário, [Inteligência Artificial \(IA\)](#) desponta como uma aliada estratégica na análise ambiental. Longe de representar uma tendência efêmera, a [IA](#) constitui uma ferramenta robusta, capaz de extrair padrões ocultos em grandes volumes de dados, muitas vezes imperceptíveis à análise convencional. Com o uso de algoritmos treinados a partir de dados provenientes de sensores remotos, estações meteorológicas e bases ambientais, é possível antecipar alterações nos parâmetros de qualidade da água com maior precisão e em tempo hábil (LI et al., 2023). Além disso, a aplicação de técnicas de explicabilidade, como o método [SHAP](#), permite interpretar de forma transparente o comportamento dos modelos, contribuindo para decisões mais fundamentadas, especialmente no contexto de políticas públicas e gestão hídrica.

A aplicação da [IA](#) na modelagem de fenômenos ambientais tem se mostrado promissora, sobretudo no que se refere à previsão e monitoramento da qualidade da água. Técnicas como algoritmos de Random Forest ([RF](#)), máquinas de gradiente (*gbm* e

redes neurais multicamadas (*MLP*) vêm sendo utilizadas com resultados expressivos na identificação de padrões e na análise de séries temporais ambientais complexas (*SILVA, L. R. M. DA, 2023*). A capacidade desses modelos em lidar com dados heterogêneos e em grande escala representa um diferencial importante frente às metodologias tradicionais.

O presente estudo tem como base a pesquisa conduzida por *Santos (2018)*, que avaliou a qualidade da água na bacia do Rio Gualaxo do Norte após a tragédia da Barragem de Fundão. Diante da evolução tecnológica observada nos últimos anos, propõe-se um aprofundamento das análises já realizadas, incorporando técnicas mais modernas de aprendizado de máquina, com vistas a interpretar, com maior clareza, as relações entre as variáveis ambientais e os padrões de qualidade da água. Ao utilizar a *IA* de forma integrada e explicável, busca-se não apenas melhorar a acurácia preditiva, mas também garantir a confiabilidade dos resultados e fortalecer sua aplicabilidade no campo da gestão ambiental.

1.1 Objetivo

Desenvolver um modelo preditivo utilizando técnicas de aprendizado de máquina, como *Random Forest (RF)*, *Gradient Boosting Machines (GBM)* e *Multi-Layer Perceptron (MLP)* para prever a qualidade da água em áreas vulneráveis a desastres naturais. O modelo será aplicado sobre a base de dados bruta de *Santos (2018)*, após tratamento e análise dos dados, e seus resultados serão comparados com a pesquisa original.

1.2 Objetivos Específicos

1. **Realizar o tratamento e pré-processamento dos dados brutos** da base de *Santos (2018)*, incluindo limpeza, remoção de outliers, substituição de valores ausentes e normalização, visando melhorar a qualidade das análises.
2. **Aplicar e comparar diferentes modelos preditivos**, como *Random Forest (RF)*, *Gradient Boosting Machines (GBM)* e *Multi-Layer Perceptron (MLP)*, para identificar os modelos mais eficazes na previsão da qualidade da água
3. **Avaliar o desempenho dos modelos preditivos** utilizando métricas de classificação como Precisão, Recall, F1-Score, Matriz de Confusão e Área sob a Curva (roc-auc), verificando sua capacidade de prever padrões na qualidade da água.
4. **Comparar os resultados obtidos com os da pesquisa original**, verificando se o tratamento mais aprofundado dos dados e a aplicação de técnicas avançadas de *IA* trazem melhorias na análise dos impactos ambientais.

5. **Identificar as variáveis mais relevantes para a previsão da qualidade da água**, utilizando técnicas de importância de features baseadas em modelos de árvore, a fim de compreender quais fatores têm maior influência nos padrões observados.
6. **Propor diretrizes para o uso de modelos de inteligência artificial no monitoramento da qualidade da água**, destacando os desafios e as vantagens da implementação dessas tecnologias em cenários reais de gestão ambiental.

1.3 Justificativas e Relevância

O monitoramento da qualidade da água constitui uma etapa essencial para a compreensão e gestão dos impactos ambientais decorrentes tanto de desastres naturais quanto de atividades antrópicas. Em casos extremos, como o rompimento da Barragem de Fundão, ocorrido no município de Mariana, Minas Gerais, os corpos hídricos afetados passaram a apresentar um quadro crítico de contaminação, que dificultou significativamente os processos de recuperação ambiental e exigiu a adoção de metodologias mais robustas e sistemáticas de avaliação ao longo do tempo.

O estudo realizado por Santos (2018) representa uma importante contribuição nesse campo, ao empregar técnicas de reconhecimento de padrões para analisar a qualidade da água na bacia do Rio Gualaxo do Norte. No entanto, a rápida evolução das ferramentas de Inteligência Artificial (IA) nos últimos anos abre novas possibilidades de aprofundamento dessas análises. Modelos mais avançados, como algoritmos de *Random Forest* (RF), máquinas de gradiente (*Gradient Boosting Machines* — GBM) e redes neurais multicamadas (*Multilayer Perceptron* — MLP), permitem interpretar com mais clareza a influência das variáveis ambientais sobre os padrões observados.

A presente pesquisa justifica-se pela necessidade de aprimorar as metodologias de previsão e análise da qualidade da água, utilizando técnicas de IA capazes de produzir resultados mais precisos e, sobretudo, mais interpretáveis. A comparação e o aperfeiçoamento das abordagens já existentes com modelos de aprendizado de máquina visam não apenas ampliar a acurácia dos resultados, mas também fortalecer a compreensão dos padrões ambientais por parte de gestores públicos e tomadores de decisão. O aspecto interpretativo dos modelos torna-se, portanto, um diferencial, ao permitir uma gestão de riscos mais eficaz e fundamentada, além de contribuir para a formulação de políticas públicas ambientalmente responsáveis.

1.4 Organização da Estrutura do Trabalho

Conforme apontado por Deslandes e Gomes (2009), os instrumentos e técnicas adotados para a condução da pesquisa devem ser descritos de forma clara e em seção própria,

justificando-se sua escolha e reconhecendo-se as limitações inerentes à produção dos dados (p.49). Dessa forma, a metodologia do presente trabalho contempla pesquisa bibliográfica e documental, complementada por análises empíricas conduzidas pelo pesquisador com base nos dados obtidos.

Diante da velocidade com que novas informações sobre qualidade da água e técnicas computacionais são produzidas, optou-se por um rigoroso critério de seleção das fontes. Foram priorizadas bases científicas reconhecidas, como SciELO e Google Acadêmico, além de bases governamentais de dados ambientais. A escolha metodológica busca assegurar a relevância, a atualidade e a confiabilidade das informações analisadas.

A estrutura do trabalho está organizada da seguinte forma:

- **Capítulo 1 – Introdução:** apresenta o problema investigado, a relevância da temática, os objetivos da pesquisa e sua justificativa teórica e prática.
- **Capítulo 2 – Fundamentação Teórica:** discute os principais conceitos sobre qualidade da água, desastres ambientais e as aplicações da Inteligência Artificial no monitoramento ambiental, com destaque para os modelos supervisionados utilizados.
- **Capítulo 3 – Metodologia:** detalha as etapas de coleta, tratamento e análise dos dados, bem como os critérios de seleção dos modelos preditivos e os procedimentos adotados para interpretação dos resultados.
- **Capítulo 4 – Resultados e Discussão:** apresenta os resultados obtidos por meio da aplicação dos modelos de IA, comparando o desempenho entre as técnicas utilizadas e interpretando os padrões identificados.
- **Capítulo 5 – Considerações Finais:** sintetiza os principais achados da pesquisa, suas implicações para a área de gestão ambiental e propõe direções para estudos futuros.

2 Fundamentação Teórica

Esta seção estabelece o arcabouço conceitual e a revisão da literatura que sustentam a pesquisa, fornecendo a base teórica para as escolhas metodológicas e a interpretação dos resultados.

2.1 Qualidade da água e indicadores ambientais

A qualidade da água constitui um dos pilares fundamentais na avaliação ambiental, sendo determinante tanto para a preservação da saúde humana quanto para a manutenção do equilíbrio dos ecossistemas aquáticos (SILVA, I. S., 2019). Trata-se de um parâmetro multifacetado, influenciado por uma complexa interação de variáveis físicas, químicas e biológicas, que refletem diretamente o grau de conservação ou degradação dos corpos hídricos (RIBEIRO, 2021; SANTOS, 2018).

Diante dessa complexidade, diversos indicadores ambientais foram desenvolvidos para simplificar a interpretação das condições da água. Entre eles, destaca-se o **Índice de Qualidade da Água (IQA)**, que busca integrar múltiplos parâmetros em uma única métrica, com o intuito de facilitar o processo decisório e a comunicação dos resultados a diferentes públicos (RIBEIRO, 2021).

2.1.1 Parâmetros Físico-Químicos e Microbiológicos: Definição e Relevância

A base de dados adotada nesta pesquisa contempla 14 variáveis físico-químicas e microbiológicas, cuja seleção foi orientada pela relevância ambiental e pelo potencial de diagnóstico das condições da água. A seguir, são apresentados os principais parâmetros, acompanhados de suas definições e significados no contexto da qualidade da água:

- **pH:** Representa o potencial hidrogeniônico da água, ou seja, seu grau de acidez ou **Alcalinidade**. Valores extremos de pH podem ser letais para organismos aquáticos e alterar a solubilidade de metais pesados e outros compostos tóxicos.
- **Turbidez (*Turb.*):** Indica a presença de partículas em suspensão, como argila, silte, matéria orgânica e microrganismos. Elevados níveis de **Turbidez** prejudicam a penetração da luz solar, impactando negativamente os processos fotossintéticos e podendo estar associados a contaminantes microbiológicos (U.S. ENVIRONMENTAL PROTECTION AGENCY (EPA), 2024).

- **Condutividade Elétrica (COND):** Reflete a concentração de íons dissolvidos na água. Alterações nos níveis de condutividade podem apontar para fontes de poluição como esgoto, fertilizantes ou rejeitos industriais.
- **Sólidos Totais (ST):** Corresponde à soma dos sólidos dissolvidos e suspensos. Concentrações elevadas comprometem a transparência da água, interferem no metabolismo aquático e prejudicam a potabilidade.
- **Temperatura (Ta):** Afeta diretamente a solubilidade dos gases, especialmente o oxigênio dissolvido, e influencia a taxa de reações químicas e biológicas. Alterações térmicas bruscas podem representar estresse para a fauna aquática.
- **Escherichia coli (E. coli) e Coliformes Totais (CT):** A presença de *E. coli* indica contaminação fecal recente, sendo um alerta para a presença de patógenos de risco à saúde humana (U.S. ENVIRONMENTAL PROTECTION AGENCY (EPA), 2024). Os coliformes totais abrangem um grupo mais amplo de bactérias que podem sinalizar poluição ambiental ou esgoto doméstico.
- **():** Parâmetro essencial para a respiração dos organismos aeróbicos, representando a proporção de oxigênio dissolvido em relação ao valor de saturação nas condições locais de temperatura e pressão. Baixos valores de saturação de oxigênio indicam processos de degradação orgânica, elevada demanda bioquímica de oxigênio e condições potencialmente prejudiciais à biota aquática.
- **Nitrito (NO₂) e Nitrato (NO₃):** Compostos nitrogenados que, em excesso, indicam poluição por fertilizantes ou esgotos. O Nitrato, em particular, está associado à eutrofização e pode provocar a síndrome do bebê azul em lactentes (U.S. ENVIRONMENTAL PROTECTION AGENCY (EPA), 2024).
- **Fósforo Total (PT):** O fósforo apresenta múltiplas fontes naturais e antrópicas — intemperismo de rochas fosfáticas, citogidry2000apatite, decomposição de matéria orgânica, aplicação de fertilizantes agrícolas e esterco animal, e lançamento de efluentes domésticos (SHARPLEY et al., 1994). Em contextos de mineração de ferro, o fósforo pode ser mobilizado por adsorção em óxidos e hidróxidos de ferro presentes em sedimentos (costa). Embora essencial ao metabolismo vegetal, o excesso de fósforo promove eutrofização e subsequente esgotamento de oxigênio nos corpos d'água (SHARPLEY et al., 1994).

O conhecimento detalhado de cada uma dessas variáveis permite não apenas compreender a dinâmica ambiental da bacia estudada, mas também interpretar com maior precisão os resultados fornecidos pelos modelos de Inteligência Artificial, principalmente em relação à importância atribuída às variáveis de entrada (*features*).

2.1.2 Normas e Padrões de Referência para Qualidade da Água

No Brasil, a Resolução CONAMA n.º 357/2005 estabelece os critérios legais para o enquadramento dos corpos d'água, determinando limites para diversos parâmetros de qualidade e orientando a classificação das águas segundo seus usos preponderantes [Conselho Nacional do Meio Ambiente \(CONAMA\) \(2005\)](#). Esta resolução constitui o principal marco normativo no contexto da gestão de recursos hídricos no país.

Complementarmente, organismos internacionais como a Organização Mundial da Saúde (OMS) e a Agência de Proteção Ambiental dos Estados Unidos (EPA) publicam diretrizes técnicas que ampliam a perspectiva de análise. A EPA, por exemplo, desenvolve critérios científicos com foco na proteção da saúde humana e da vida aquática, especificando limites seguros para substâncias químicas, biológicas e orgânicas presentes na água ([U.S. ENVIRONMENTAL PROTECTION AGENCY \(EPA\), s.d., 2024](#)).

A Tabela 2 apresenta os limites de referência para os parâmetros analisados neste trabalho, conforme estabelecido pela CONAMA para ambientes de água doce. Esses valores são fundamentais para avaliar a conformidade ambiental e para sustentar a discussão crítica dos resultados obtidos.

Tabela 2 – Limites de qualidade da água conforme a Resolução CONAMA n.º 357/2005

Parâmetro	Tipo de limite	Classe 1	Classe 2	Classe 3	Classe 4
PT (mg.L ⁻¹)	Superior	0,02–0,10 ^a	0,03–0,10 ^a	0,05–0,15 ^a	—
OD (mg.L ⁻¹)	Inferior	≥ 6,0	≥ 5,0	≥ 4,0	≥ 2,0
NO ₃ (mg.L ⁻¹)	Superior	≤ 10,0	≤ 10,0	≤ 10,0	—
Turb. (UT)	Superior	≤ 40	≤ 100	≤ 100	—
DBO _{5,20} (mg.L ⁻¹)	Superior	≤ 3,0	≤ 5,0	≤ 10,0	—

^a Os limites variam conforme o tipo de corpo hídrico (lêntico, lótico, intermediário ou tributário direto). Fonte: Adaptado de [Conselho Nacional do Meio Ambiente \(CONAMA\) \(2005\)](#).

^b A notação DBO_{5,20} refere-se à Demanda Bioquímica de Oxigênio determinada após cinco dias de incubação a 20 °C, parâmetro clássico utilizado para estimar a carga orgânica biodegradável presente no corpo hídrico.

^c Embora a Resolução CONAMA n.º 357/2005 estabeleça limites normativos para o oxigênio dissolvido em termos de concentração absoluta (mg.L⁻¹), neste estudo emprega-se adicionalmente a saturação de oxigênio dissolvido (SatOD), expressa em porcentagem, como variável analítica complementar. A utilização de SatOD permite considerar os efeitos da temperatura e da pressão atmosférica, sendo particularmente adequada para análises multivariadas e comparações espaciais entre pontos de monitoramento.

A incorporação desses limites legais ao trabalho não apenas assegura a aderência à legislação ambiental vigente, mas também oferece um referencial técnico para a análise crítica dos dados e das previsões geradas pelos modelos computacionais no Capítulo 4.

2.1.3 Especificidades da Bacia do Rio Gualaxo do Norte

A bacia do Rio Gualaxo do Norte foi diretamente impactada pelo rompimento da Barragem de Fundão em 2015, evento que desencadeou uma série de consequências ambientais, sociais e econômicas. O estudo conduzido por Santos (2018) serviu como base para a seleção das variáveis deste trabalho, priorizando aquelas que melhor refletiam os efeitos dos rejeitos de mineração nos corpos hídricos da região.

Na pesquisa original, os parâmetros **Nitrato** e **Nitrato** foram excluídos devido a “particularidades da região” (SANTOS, 2018), possivelmente associadas à baixa presença de compostos nitrogenados nos rejeitos liberados, compostos majoritariamente por óxidos de ferro e sílica. Essa escolha metodológica reforça a importância de se considerar as características geoquímicas locais na seleção das variáveis.

A análise crítica dessas particularidades contribui para justificar as escolhas metodológicas adotadas neste trabalho e evidencia o compromisso com a especificidade do contexto ambiental analisado.

2.2 Técnicas Estatísticas e Inteligência Artificial na Análise Ambiental

2.2.1 Métodos Estatísticos Tradicionais

Historicamente, a análise da qualidade da água tem se baseado em técnicas estatísticas multivariadas para identificar padrões, agrupar amostras semelhantes e compreender relações entre variáveis físico-químicas e microbiológicas. Dentre essas abordagens, destacam-se a **PCA** e a **Análise Discriminante Linear (LDA)**, amplamente utilizadas em estudos ambientais para redução de dimensionalidade e classificação.

Análise de Componentes Principais (**PCA**)

A **PCA**, também conhecida como Transformada de Karhunen-Loève ou Transformada de Hotelling, é uma técnica não supervisionada que transforma variáveis possivelmente correlacionadas em componentes principais linearmente não correlacionados (JOLLIFFE; CADIMA, 2016; GREENACRE; GROENEN; HASTIE et al., 2022). Seu objetivo é reduzir a dimensionalidade dos dados com máxima preservação da variância original. Essa transformação é fundamentada em autovalores e autovetores da matriz de covariância, onde os primeiros componentes representam maior proporção da variabilidade total (JOLLIFFE; CADIMA, 2016).

Em estudos ambientais, a **PCA** é aplicada para reconhecer padrões de contaminação, avaliar similaridades entre corpos hídricos e explorar a estrutura interna de grandes

conjuntos de dados (ENVIRONMENTAL COMPUTING, 2022). Entretanto, seu uso exige atenção a limitações metodológicas, como a necessidade de padronização das variáveis, a sensibilidade a *outliers* e a suposição de linearidade entre variáveis (JOLLIFFE; CADIMA, 2016).

Análise Discriminante Linear (LDA)

A LDA é uma técnica supervisionada de classificação multivariada que visa projetar os dados em um espaço de menor dimensão, maximizando a separabilidade entre classes previamente definidas (AL-SHAMI; AL-MOUSA, 2024). O método busca direções que maximizem a variância entre classes e minimizem a variância intra-classe, com base em distribuições gaussianas com covariâncias semelhantes (AL-SHAMI; AL-MOUSA, 2024).

Em contextos ambientais, a LDA tem sido aplicada para distinguir regiões impactadas por poluição, classificar pontos de coleta e auxiliar na seleção de variáveis significativas (SANTOS, 2018). Comparada à PCA, a LDA mostra-se mais eficaz em tarefas de classificação, embora apresente restrições quanto à distribuição dos dados e à presença de *multicolinearidade* entre atributos (AL-SHAMI; AL-MOUSA, 2024).

2.2.2 Modelos de Aprendizado de Máquina na Previsão da Qualidade da Água

O uso de algoritmos de aprendizado de máquina tem se intensificado na análise da qualidade da água, dada sua capacidade de lidar com relações não lineares e conjuntos de dados de alta dimensionalidade (LI et al., 2023).

Random Forest (RF)

O algoritmo Random Forest (RF) baseia-se na construção de múltiplas árvores de decisão a partir de subconjuntos aleatórios dos dados e das variáveis (BREIMAN, 2001). Suas vantagens incluem robustez a *outliers*, capacidade de lidar com dados ausentes e boa estimativa da importância das variáveis (SALMAN; KALAKECH; STEITI, 2024). Entretanto, é considerado um modelo de “caixa preta”, com baixa interpretabilidade e elevado custo computacional em grandes bases (BREIMAN, 2001).

Gradient Boosting Machines (GBM)

O Gradient Boosting Machines (GBM) é um método de *boosting* sequencial que otimiza residuais de modelos anteriores, construindo preditores fortes a partir de modelos fracos (FRIEDMAN, 2001). Apresenta alto desempenho preditivo e é flexível para diferentes tarefas, mas demanda ajuste fino de hiperparâmetros e é sensível a *overfitting* (NATEKIN; IYENGAR, 2013; RAHMAN et al., 2023).

Multi-Layer Perceptron (MLP)

As Redes Neurais do tipo *Multi-Layer Perceptron* (MLP) são compostas por camadas ocultas capazes de aprender relações complexas entre as variáveis de entrada (GOODFELLOW; BENGIO; COURVILLE, 2016). Apesar do bom desempenho em problemas não lineares, as MLPs apresentam baixa interpretabilidade, elevada demanda computacional e sensibilidade a dados incompletos (WANG et al., 2022).

2.2.3 Técnicas de Redução de Dimensionalidade: t-SNE e UMAP

Com o crescimento de conjuntos de dados ambientais complexos, surgem métodos não lineares para redução de dimensionalidade e visualização, como o t-SNE e o UMAP (MCINNES LELAND; MELVILLE, 2018; MAATEN; HINTON, 2008).

O *t-distributed Stochastic Neighbor Embedding* (t-SNE) transforma distâncias de alta dimensão em probabilidades de similaridade, permitindo visualizar agrupamentos locais por meio de uma projeção não linear que preserva principalmente estruturas de vizinhança (MAATEN; HINTON, 2008). Embora seja eficaz para identificar padrões complexos, sua alta complexidade computacional e sensibilidade aos parâmetros iniciais limitam a reprodutibilidade entre execuções sucessivas (AL-SHAMI; AL-MOUSA, 2024).

O *Uniform Manifold Approximation and Projection* (UMAP), por sua vez, baseia-se em fundamentos topológicos e de geometria Riemanniana para preservar tanto estruturas locais quanto globais dos dados (MCINNES LELAND; MELVILLE, 2018). Destaca-se por sua maior velocidade computacional, escalabilidade e consistência dos resultados entre execuções, características que justificaram sua priorização nas análises quantitativas deste trabalho (MAATEN; HINTON, 2008). Ambos os métodos são complementares à PCA, sendo especialmente úteis em análises exploratórias de dados ambientais de alta dimensionalidade (AL-SHAMI; AL-MOUSA, 2024).

2.2.4 Inteligência Artificial Explicável (XAI): SHAP e LIME

A crescente utilização de modelos complexos exige mecanismos que promovam transparência e compreensão das decisões automatizadas. Nesse contexto, a *Inteligência Artificial Explicável* (XAI) emerge como uma abordagem essencial (FINZEL, 2025).

O SHAP baseia-se na teoria dos jogos para calcular a contribuição individual de cada variável em uma previsão específica (FINZEL, 2025). Em estudos ambientais, sua aplicação tem permitido identificar as *features* mais relevantes para a classificação de áreas contaminadas (MAKUMBURA et al., 2024).

O LIME explica as decisões de modelos opacos por meio da aproximação local com modelos lineares interpretáveis, destacando os atributos mais influentes em cada instância

(FINZEL, 2025). Ambas as técnicas promovem a confiabilidade, a auditoria e a ética na utilização da IA.

A crescente incorporação de sistemas de [Inteligência Artificial](#) em contextos sensíveis, como o monitoramento ambiental e a gestão de recursos hídricos, exige que tais tecnologias sejam não apenas eficientes, mas também transparentes, justas e eticamente responsáveis. Nesse sentido, as abordagens de Inteligência Artificial Explicável ([XAI](#)) representam um avanço crucial para a construção de sistemas confiáveis e auditáveis.

Ferramentas como o [SHapley Additive exPlanations](#) (SHAP) e o [Local Interpretable Model-agnostic Explanations](#) (LIME) permitem identificar, com clareza, quais variáveis exerceram maior influência sobre as decisões de um modelo. Isso possibilita que pesquisadores e desenvolvedores identifiquem e corrijam vieses, inconsistências ou padrões ocultos nos dados que poderiam comprometer a equidade e a precisão dos resultados ([RASAQ, 2025](#)).

No âmbito da gestão ambiental, essa capacidade de interpretação não só aprimora o desempenho técnico das soluções automatizadas, também assegura que as decisões relacionadas ao uso da água estejam fundamentadas em evidências reais e relevantes, e não em correlações espúrias ou viesadas.

Além disso, a [XAI](#) reforça a responsabilização institucional ao permitir que reguladores e agentes públicos tenham acesso à lógica por trás das previsões geradas pelos sistemas de IA. Essa transparência é essencial para garantir a conformidade com normas éticas e legislações ambientais, permitindo auditorias independentes e tomadas de decisão mais seguras ([RASAQ, 2025](#)).

Outro ponto fundamental é a relação de confiança estabelecida entre os sistemas baseados em IA e os diferentes stakeholders envolvidos — incluindo gestores, cientistas ambientais e comunidades locais. Ao fornecer explicações compreensíveis para os resultados apresentados, a [XAI](#) torna a tecnologia mais acessível e aceitável, reduzindo resistências à sua adoção ([RASAQ, 2025](#)).

Portanto, a inserção da [XAI](#) no presente trabalho não se configura apenas como um diferencial técnico, mas como um compromisso ético com a qualidade, a integridade e a responsabilidade social da pesquisa. Em domínios onde as decisões podem impactar diretamente a saúde pública e a preservação dos ecossistemas, a capacidade de explicar *como* e *por que* uma decisão foi tomada é tão vital quanto a própria acurácia do modelo ([SANTOS, 2018](#)).

Dessa forma, reforça-se que a adoção de técnicas explicáveis no desenvolvimento de modelos preditivos ambientais é uma etapa indispensável para garantir não só a eficiência analítica, mas também a justiça e a legitimidade das soluções baseadas em [Inteligência Artificial](#).

2.2.5 Trabalhos Correlatos e o Estado da Arte

Estudos recentes têm consolidado a aplicação de IA na previsão da qualidade da água. Santos (2018) combinaram PCA, Redes de Kohonen e LDA para classificar pontos afetados por desastre ambiental, alcançando acurácia superior a 90%. Igor Santos Silva (2019) utilizaram Random Forest e Redes Neurais para prever clorofila-a com desempenho satisfatório mesmo com bases reduzidas. Já Luiz Ricardo Mantovani da Silva (2023) integraram dados climáticos, sensoriamento remoto e algoritmos como SVM e RF para monitoramento da Baía de Guanabara. Por fim, Li et al. (2023) compararam cinco modelos de IA para previsão de pH em corpos hídricos, destacando o melhor desempenho de técnicas baseadas em árvores.

Esses trabalhos reforçam a relevância da abordagem proposta neste TCC, que combina RF, GBM, MLP, t-SNE, UMAP e XAI para oferecer não apenas previsões mais precisas, mas também interpretáveis. O objetivo é contribuir com soluções que aliem rigor técnico, inteligência computacional e responsabilidade ética na análise ambiental.

3 Metodologia

3.1 Conjunto de Dados

Este trabalho utilizou a base de dados de qualidade da água compilada por Santos (2018), que conduziu amostragens sistemáticas ao longo de um ciclo hidrológico completo na bacia do Rio Gualaxo do Norte, Minas Gerais. A escolha dessa base justifica-se pela relevância histórica e ambiental do evento de ruptura da Barragem de Fundão, ocorrido em novembro de 2015, e pela oportunidade de aplicar técnicas modernas de inteligência artificial a dados ambientais previamente analisados por métodos estatísticos clássicos, permitindo comparações metodológicas diretas.

3.1.1 Contexto Espacial e Temporal da Coleta

As amostragens abrangeram 27 pontos de monitoramento estrategicamente distribuídos ao longo da rede hidrográfica, refletindo diferentes compartimentos hidrogeomorfológicos e condições de impacto. Os pontos **Ponto de Coleta (PC)1** a **PC10** representam **regiões de referência**, situadas a montante da área de influência direta do rompimento da barragem ou em tributários não afetados pela pluma de rejeitos. Os pontos **PC11** a **PC27** correspondem a **áreas diretamente impactadas**, localizadas a jusante do ponto de ruptura e expostas à deposição de lama de rejeitos minerários nos sedimentos de fundo e margens fluviais.

Cada ponto foi amostrado em duas campanhas sazonais contrastantes — **estação seca** e **estação chuvosa** —, totalizando inicialmente duas coletas por ponto (54 amostragens/ponto ao longo do ciclo). Essa estratégia temporal permitiu capturar variações hidrodinâmicas associadas ao regime de vazões, fundamentais para compreender processos de ressuspensão de sedimentos, diluição de poluentes e transporte de nutrientes. A sazonalidade hidrológica é particularmente relevante em bacias tropicais como a do Rio Doce, onde a alternância entre períodos de estiagem e precipitações intensas modula significativamente a concentração de poluentes e a disponibilidade de oxigênio dissolvido (HATJE; PEDREIRA; REZENDE, 2017).

3.1.2 Variáveis Monitoradas

Foram monitoradas 14 variáveis físico-químicas e microbiológicas, selecionadas por Santos (2018) com base em sua relevância para caracterização da qualidade da água em sistemas aquáticos impactados por atividades antrópicas. O conjunto de variáveis pode ser agrupado em quatro categorias funcionais:

Parâmetros Físico-Químicos Gerais

Incluem **pH**, **Turbidez (Turb.)**, **Condutividade Elétrica (COND)**, **Sólidos Totais (ST)**, **Temperatura (Ta)** e **Alcalinidade (Alca.)**. Esses parâmetros fornecem informação sobre o estado físico da água, carga de sólidos suspensos, mineralização iônica e capacidade de tamponamento. A **Turbidez**, em particular, constitui indicador direto da presença de partículas finas de rejeitos minerários em suspensão ou depositadas no leito fluvial (HATJE; PEDREIRA; REZENDE, 2017).

Indicadores de Oxigenação e Processos Redox

Abrangem **Saturação de Oxigênio Dissolvido (SatOD)** e **Potencial de Oxirredução (ORP)**. Esses parâmetros refletem a dinâmica respiratória do ecossistema aquático, processos de decomposição orgânica e condições redox da coluna d'água. A saturação de oxigênio é frequentemente utilizada para avaliar o estresse ambiental causado por cargas orgânicas excessivas ou condições de eutrofização (CASTILLO; M., 2021).

Nutrientes

Incluem **Fósforo Total (PT)**, **Nitrato (NO₃)**, **Nitrito (NO₂)** e **Cloretos (CL)**. Esses compostos desempenham papel central na dinâmica trófica de ecossistemas aquáticos, influenciando a produtividade primária e os processos de eutrofização. Concentrações elevadas de fósforo, frequentemente associadas ao arraste superficial de sedimentos ricos em minerais fosfatados, constituem indicador relevante em bacias impactadas por mineração (WORLD HEALTH ORGANIZATION, 2017).

Indicadores Microbiológicos

Representados por ***Escherichia coli* (E. coli)** e **Coliformes Totais (CT)**, esses parâmetros sinalizam contaminação fecal e comprometimento sanitário da água. Embora não estejam diretamente relacionados ao impacto dos rejeitos minerários, valores elevados refletem contribuições difusas de efluentes domésticos não tratados, problema recorrente em bacias hidrográficas rurais brasileiras (AGÊNCIA NACIONAL DE ÁGUAS (ANA); MINISTÉRIO DO MEIO AMBIENTE; PROGRAMA DAS NAÇÕES UNIDAS PARA O MEIO AMBIENTE, 2017).

3.1.3 Características da Base Bruta e Desafios de Qualidade

A base de dados original apresentava desafios típicos de conjuntos ambientais coletados em campo, que motivaram o desenvolvimento do *pipeline* de pré-processamento descrito na Seção 3.2. Identificou-se que **4,3% dos registros** continham ao menos uma variável

com valor ausente (*missing value*), refletindo limitações operacionais de campo, falhas instrumentais ou impossibilidade de quantificação em amostras com características extrema.

Adicionalmente, a detecção preliminar pelo método do [Intervalo Interquartil \(IQR\)](#), fator 1,5, identificou **183 candidatos a outliers** distribuídos entre as 14 variáveis monitoradas. Valores extremos são esperados em sistemas aquáticos impactados por eventos episódicos (pulsos pluviométricos, ressuspensão de sedimentos, descargas pontuais de efluentes), mas também podem refletir erros analíticos ou inconsistências na transcrição de dados. Diante dessa ambiguidade, optou-se por uma abordagem de validação contextual baseada em regras físico-químicas, conforme documentado no Apêndice [A](#), que permitiu distinguir valores extremos plausíveis de inconsistências genuínas.

3.1.4 Limitações Reconhecidas da Base Original

Embora a base de [Santos \(2018\)](#) constitua referência consolidada para estudos sobre o impacto do rompimento da Barragem de Fundão, algumas limitações intrínsecas devem ser reconhecidas:

- **Resolução temporal restrita.** A amostragem em apenas duas campanhas (seca e chuvosa) captura a variabilidade sazonal dominante, mas não permite caracterizar eventos extremos de curta duração (enchentes, estiagens prolongadas) ou processos de transição entre estações.
- **Ausência de dados pré-evento.** A coleta foi realizada dois anos após o rompimento da barragem, período durante o qual processos de estabilização e remobilização de sedimentos já haviam alterado significativamente as condições iniciais. A ausência de séries históricas pré-desastre impede comparações diretas do estado anterior à contaminação.
- **Variáveis não monitoradas.** Parâmetros relevantes como metais pesados (ferro, manganês, alumínio), carbono orgânico dissolvido e compostos orgânicos persistentes não foram incluídos no escopo analítico original, limitando a compreensão de processos biogeoquímicos complexos.
- **Distribuição espacial desbalanceada.** O número de pontos nos tributários afetados (132 amostras) é consideravelmente superior ao número de pontos no rio principal afetado (72 amostras), refletindo a estratégia de amostragem original mas introduzindo desbalanceamento nas classes que pode afetar o treinamento de modelos supervisionados.

Essas limitações não comprometem a validade científica da base, mas contextualizam as análises subsequentes e justificam escolhas metodológicas específicas, como a validação cruzada por grupos (Seção 3.3) e a utilização de técnicas robustas de imputação (Seção 3.2).

3.2 Pipeline de Pré-processamento

-O pré-processamento foi concebido para assegurar integridade, coerência contextual e reprodutibilidade, reduzindo o impacto de valores anômalos sem perda de informação ambiental relevante. Em consonância com abordagens contemporâneas que integram inteligência artificial ao monitoramento da qualidade da água — tradicionalmente um processo custoso e sujeito a ruídos analíticos —, adotou-se um *pipeline* modular que combina imputação contextualizada, detecção inteligente de *outliers* e padronização robusta.

Metadados contextuais.

O conjunto de dados foi enriquecido com as variáveis `Grupo_Referencia` e `Estacao`. A primeira deriva do número do ponto (`Pint_Num`) e reflete o contexto espacial (rio principal ou tributário; afetado ou não afetado), enquanto a segunda é obtida a partir do campo de amostragem (`Campo` → {Seca, Chuvosa}). Essa contextualização é essencial, uma vez que a hidrodinâmica e os processos biogeoquímicos variam sazonalmente e entre compartimentos da bacia (SANTOS, 2018).

Imputação granular de valores ausentes.

Os valores ausentes foram preenchidos pela mediana *no nível do contexto* (`Grupo_Referencia`, `Estacao`), de modo a preservar a coerência hidroambiental local. Como medida de contingência (*fallback*), quando a mediana contextual não estava disponível por insuficiência amostral, aplicou-se a **mediana global da variável**. Essa abordagem é recomendada em bases ambientais heterogêneas, pois reduz o viés introduzido por imputações arbitrárias (EVERITT et al., 2011; LI et al., 2023).

Detecção e validação físico-química de *outliers*.

A identificação preliminar utilizou o método do intervalo interquartil (**Intervalo Interquartil (IQR)**), aplicado separadamente dentro de cada contexto. Em seguida, cada valor extremo foi submetido ao **módulo de validação físico-química**, desenvolvido especificamente para este estudo.

Esse módulo aplica um conjunto estruturado de regras ambientais, baseadas em princípios redox, balanço iônico, solubilidade, transporte de partículas e plausibilidade microbiológica, avaliando:

- consistência do valor em relação ao processo esperado no ponto e estação;
- compatibilidade com variáveis correlacionadas (e.g., turbidez, sólidos, nutrientes);
- aderência a faixas ambientalmente viáveis reportadas na literatura;
- comportamento temporal e espacial dentro do mesmo grupo de referência.

O módulo classifica cada caso como *Manter* ou *Tratar*, fornecendo uma justificativa ambiental interpretável. A tabela completa de regras, exemplos e justificativas encontra-se no Apêndice A.

Imputação seletiva de *outliers*.

Apenas os valores classificados como *Tratar* foram substituídos pela mediana contextual correspondente, calculada com exclusão do próprio registro. Essa substituição localizada minimiza distorções e mantém a estrutura das distribuições originais.

Padronização robusta.

As variáveis numéricas foram escalonadas via `RobustScaler`, técnica adequada para distribuições assimétricas e resiliente a valores extremos residuais. O modelo ajustado foi persistido para garantir reprodutibilidade em todas as etapas posteriores.

Rastreabilidade.

Todos os artefatos do pré-processamento — dados tratados, relatórios de validação, imputações realizadas, valores substituídos e modelos de escalonamento — foram registrados automaticamente, assegurando transparência e reprodução integral do pipeline.

Impacto do Pré-processamento na Qualidade dos Dados

A aplicação do módulo de validação reduziu significativamente a distorção causada por valores anômalos, preservando as relações hidrológicas naturais entre variáveis correlacionadas (e.g., [Potencial de Oxirredução \(ORP\)](#) e [Demanda Bioquímica de Oxigênio \(DBO\)](#); [Sólidos Totais \(ST\)](#) e [Condutividade Elétrica \(COND\)](#)). Essa etapa foi essencial para evitar que padrões artificiais influenciassem a clusterização e as métricas dos modelos supervisionados, garantindo que o conjunto resultante represente com maior fidelidade a variabilidade real do sistema aquático.

A estratégia integrada de pré-processamento adotada neste trabalho reduz o viés decorrente de exclusões indiscriminadas, mantém a coerência hidroquímica das séries e

estabelece uma base sólida para as análises exploratória, não supervisionada e supervisionada subsequentes (SANTOS, 2018; EVERITT et al., 2011; XU; WUNSCH, 2005; LI et al., 2023). O **módulo de validação de outliers**, desenvolvido especificamente para este estudo, está documentado no Apêndice A e representa um avanço metodológico relevante para a garantia de qualidade de dados ambientais.

Os detalhes completos de implementação, parâmetros dos algoritmos, estrutura das saídas e protocolos auxiliares do pipeline de pré processamento estão documentados no Apêndice A.

3.3 Pipeline de Análise Não Supervisionada — Estruturação e Lógica Metodológica

Esta seção descreve o núcleo exploratório do estudo, dedicado à identificação de padrões latentes na base de dados por meio de técnicas de aprendizagem não supervisionada. O pipeline foi integralmente desenvolvido em *Python* e organizado de modo **modular e reprodutível**, permitindo executar e comparar, sob as mesmas condições, diferentes composições de variáveis, projeções e algoritmos de agrupamento. O objetivo central foi compreender como conjuntos temáticos de variáveis ambientais se organizam em termos de estrutura interna e separabilidade, fornecendo base conceitual e empírica para a etapa supervisionada subsequente.

3.3.1 Estratégia de seleção de variáveis

Adotou-se uma abordagem **hierárquica e iterativa** de seleção, conduzida em duas etapas complementares: (i) análise preliminar por blocos temáticos independentes; e (ii) construção de blocos principais consolidados, a partir das combinações mais informativas. A estratégia segue recomendações clássicas de seleção progressiva guiada por contribuição e redundância, com foco em maximizar variância explicada e reduzir sobreposição informacional (JOLLIFFE, 2002; XU; WUNSCH, 2005).

Etapas 1: Análise temática por blocos independentes

A organização das variáveis em blocos temáticos partiu de fundamentos hidrológicos e ambientais amplamente discutidos na literatura, sendo complementada pela análise exploratória conduzida após o pré-processamento. Esse arranjo buscou refletir os processos predominantes que estruturam a qualidade da água na bacia estudada, de modo a favorecer interpretações consistentes nas etapas não supervisionadas subsequentes. A seguir, apresentam-se os três blocos temáticos adotados, acompanhados de breves justificativas conceituais.

- **Bloco Minerário:** **Turbidez (Turb.)**, **ST**, **COND**, **Alcalinidade (Alca.)**, **Fósforo Total (PT)**. Este bloco reúne variáveis associadas a processos de transporte de partículas, mineralização e enriquecimento iônico da água. A **Alca.** foi incorporada a esse conjunto por refletir mecanismos de tamponamento e dissolução mineral típicos de trechos sujeitos a maior intemperismo e aporte de sedimentos. O Fósforo Total (**PT**), embora participe de processos biogeoquímicos, foi aqui alocado devido à sua forte ligação ao material particulado, fenômeno extensivamente documentado na literatura de recursos hídricos. Em sistemas com elevada carga de sólidos, grande parte do **PT** permanece adsorvida a partículas em suspensão, razão pela qual sua inclusão no domínio minerário se mostra coerente do ponto de vista ambiental.
- **Bloco Sanitário:** ***Escherichia coli* (*E. coli*)**, **Coliformes Totais (CT)**, **DBO**. Este bloco contempla indicadores de contaminação orgânica e microbiológica, comumente associados a aportes recentes de esgoto doméstico e efluentes urbanos. A decisão de manter *E. coli* e **CT** no mesmo bloco se justifica por sua origem comum e por captarem dimensões complementares da contaminação fecal. Ainda que apresentem elevada associação conceitual, trata-se de variáveis com comportamentos independentes no ambiente, influenciados por fatores como sensibilidade a condições físico-químicas, sobrevivência diferencial e variabilidade temporal. A **DBO**, por sua vez, integra esse domínio por expressar a demanda oxidativa relacionada à decomposição de matéria orgânica biodegradável, um mecanismo que se conecta diretamente às pressões sanitárias observadas nos corpos d'água.
- **Bloco Biogeoquímico:** **pH**, **Saturação de Oxigênio Dissolvido (SatOD)**, **NO₃**, **NO₂**, **ORP**, **Cloretos (CL)**. Esse bloco reúne parâmetros que descrevem o equilíbrio ácido-base, a oxigenação do meio e os processos *redox* responsáveis pela transformação de compostos nitrogenados e pela dinâmica metabólica da coluna d'água. As variáveis incluídas refletem o conjunto de reações químicas e bioquímicas que sustentam a ciclagem de nutrientes, influenciando de maneira integrada a estrutura físico-química do ambiente aquático.

De forma geral, optou-se por uma classificação exclusiva de cada variável, ainda que alguns parâmetros apresentem interações com mais de um domínio. Essa estratégia favorece a interpretabilidade dos blocos e evita redundâncias metodológicas, assegurando uma base conceitual sólida para a etapa de agrupamento. Aspectos específicos — como a atuação do **PT** como variável de transição entre processos minerários e biogeoquímicos, ou a relação ortogonal entre *E. coli* e **CT** — serão discutidos com maior profundidade na seção de resultados, onde a estrutura empírica dos dados é apresentada de forma detalhada.

Cada bloco foi submetido ao mesmo fluxo metodológico, assegurando comparabilidade: (a) escalonamento robusto das variáveis (**RobustScaler**); (b) avaliação de

redundância e, quando aplicável, redução preliminar por *PCA* para preservar ao menos 90% da variância explicada; (c) projeção topológica com *UMAP* e *t-SNE*; (d) clusterização densitária e particional (*HDBSCAN* e *K-Means*); e (e) registro de métricas internas (Silhouette, Calinski–Harabasz, Davies–Bouldin) e externas (*Adjusted Rand Index (ARI)*, NMI, homogeneidade, completude e *V-measure*) (STREHL; GHOSH, 2002; HUBERT; ARABIE, 1985; MANNING; RAGHAVAN; SCHÜTZE, 2008).

3.3.2 Fase 2: Construção dos Blocos Principais

Com base nas variáveis selecionadas mediante análise de loadings da Fase 1 (critério: $\text{loading} > 0,90$ nos dois primeiros componentes principais), definiram-se três **Blocos Principais** que consolidam variáveis complementares para análise integrada. A Tabela 3 apresentou a organização inicial das variáveis em três grupos (Minerário, Sanitário e Biogeoquímico), dos quais foram extraídas as variáveis mais representativas.

Tabela 3 – Da análise temática (Fase 1) aos Blocos Principais (Fase 2): critérios de seleção e composição.

Bloco Temático	<i>n</i>	Variáveis Originais	Destino nos Blocos Principais (A, B, C)
Minerário	5	Alca, Turb (loading > 0,90)	Incluídas: Núcleo mineral comum (A, B, C)
		ST, PT (loading < 0,90)	Incluídas: Adição estratégica por relevância ambiental (transporte de sedimentos + eutrofização)
		Cond (loading < 0,90)	Excluída: Colinearidade com Alca (<i>r</i> > 0,85)
Sanitário	3	<i>E. coli</i>, CT (loading > 0,90)	Incluídas: <i>E. coli</i> (B, C), CT (A, C) — complementaridade PC1/PC2
		DBO (loading < 0,07)	Excluída: Contribuição marginal
Biogeoquímico	5	NO , Cloretos (loading > 0,90)	Incluídas: Discriminantes biogeoquímicos (A, B, C)
		pH, Satod, ORP	Excluídas: Loadings dispersos em PC3/PC4
Blocos Principais:		Bloco A (7 var):	ST, Turb, Alca, Cl, NO , CT, PT
		Bloco B (7 var):	ST, Turb, Alca, Cl, NO , <i>E. coli</i> , PT
		Bloco C (8 var):	ST, Turb, Alca, Cl, NO , <i>E. coli</i> , CT, PT

Fonte: Elaboração própria (2025).

Nota: A transição Fase 1 → Fase 2 não se limitou ao critério de loading > 0,90. Embora Alca, Turb, *E. coli*, CT, NO e Cloretos tenham sido priorizadas por dominância em PC1/PC2 (loading > 0,90), as variáveis **ST e PT foram estrategicamente adicionadas** aos Blocos Principais por relevância ambiental documentada: ST captura transporte de sedimentos e PT reflete processos de eutrofização, ambos fundamentais para caracterização de impacto em bacias hidrográficas pós-desastre (SANTOS, 2018). A exclusão de Cond (colinearidade com Alca), DBO (contribuição marginal) e parâmetros redox (pH, Sat O, ORP — loadings dispersos) priorizou parcimônia e controle de multicolinearidade.

Os blocos principais não correspondem exatamente aos blocos temáticos, mas sim a *combinações otimizadas* dessas variáveis selecionadas, provenientes de múltiplos blocos, escolhidas por seu poder discriminante e baixa redundância. A lógica de construção priorizou: (i) preservar a assinatura mineral do impacto da barragem, (ii) adicionar discriminantes sanitários e biogeoquímicos, e (iii) controlar a colinearidade para evitar overfitting.

Bloco A: Assinatura Mineral + Discriminação Básica

Variáveis incluídas: Sólidos Totais (ST), Turbidez (Turb.), Alcalinidade (Alca.), Cloretos (CL), Nitrato (NO₃), Coliformes Totais (CT) Fósforo Total (PT) (7 variáveis).

Justificativa: *Turb.* e *Alca.* capturam a assinatura direta do rejeito mineral (SANTOS, 2018), enquanto Cloretos e NO_3 funcionam como discriminantes biogeoquímicos que diferenciam processos naturais de interferências antrópicas. A inclusão de *CT* adiciona um marcador sanitário básico, útil para distinguir trechos do Rio Principal de Tributários com maior carga fecal. *ST* e *PT* completam o perfil físico-químico, capturando transporte de sedimentos e eutrofização. Excluiu-se *COND* devido à alta colinearidade com *Alca.* (correlação $>0,85$), evitando redundância e mantendo parcimônia.

Bloco B: Sanitário + Mineral Essencial

Variáveis incluídas: Sólidos Totais (*ST*), Turbidez (*Turb.*), Alcalinidade (*Alca.*), Cloretos (*CL*), Nitrato (NO_3), *Escherichia coli* (*E. coli*), Fósforo Total (*PT*) (7 variáveis).

Justificativa: Substitui Coliformes Totais por *E. coli*, marcador sanitário mais específico de contaminação fecal recente (WORLD HEALTH ORGANIZATION, 2017). Essa troca visa capturar fontes de contaminação pontual (lançamentos de esgoto) com maior precisão, embora à custa de maior variabilidade temporal. A assinatura mineral (*Turb.*, *Alca.*, *ST*) e os discriminantes ambientais (*CL*, NO_3 , *PT*) permanecem inalterados em relação ao Bloco A.

Bloco C: Integração Completa com Perfil Sanitário Dual

Variáveis incluídas: Sólidos Totais (*ST*), Turbidez (*Turb.*), Alcalinidade (*Alca.*), Cloretos (*CL*), Nitrato (NO_3), *Escherichia coli* (*E. coli*), Coliformes Totais (*CT*), Fósforo Total (*PT*) (8 variáveis).

Justificativa: Combina informação mineral e sanitária completa, preservando tanto *E. coli* (indicador fecal específico) quanto *CT* (indicador mais amplo de contaminação orgânica). Embora *E. coli* e *CT* apresentem correlação moderada-alta ($r \approx 0,72$), eles capturam padrões de contaminação complementares: *E. coli* reflete descargas recentes de esgoto, enquanto *CT* incluem fontes ambientais diversas (solo, matéria orgânica, fezes de animais) (WORLD HEALTH ORGANIZATION, 2017; SPERLING, 2005).

A análise de componentes principais do Bloco Sanitário (Seção 4.2.1) revelou que *E. coli* domina o *PC1* (magnitude total de contaminação), enquanto *CT* domina o *PC2* (tipo de fonte poluidora), evidenciando ortogonalidade apesar da correlação linear. Essa estrutura bicomponente justifica a inclusão simultânea de ambos os marcadores no Bloco C, capturando simultaneamente magnitude (*PC1*) e tipo de fonte (*PC2*) da contaminação sanitária. Essa configuração maximiza o poder discriminante para separação entre as quatro categorias hidrológicas (*RP* Afetado, *RP* Não Afetado, Tributário Afetado, Tributário Não Afetado), ao custo de um leve aumento no risco de overfitting devido ao maior número de *features* (8 variáveis).

Os três blocos foram submetidos ao mesmo pipeline metodológico (PCA, projeções UMAP/t-SNE, clusterização K-Means/HDBSCAN, validação interna e externa), com resultados comparativos detalhados na Seção 3.3.2. A seleção do bloco para modelagem supervisionada fundamentou-se em desempenho nas métricas internas, estabilidade temporal e concordância externa, conforme documentado na Seção 4.2.2.

3.3.3 Redução dimensional e projeções

As projeções foram conduzidas por *Uniform Manifold Approximation and Projection* (UMAP) (MCINNES LELAND; MELVILLE, 2018) e *t-distributed Stochastic Neighbor Embedding* (t-SNE) (MAATEN; HINTON, 2008). O UMAP foi priorizado para análise quantitativa (melhor preservação de estrutura global e reprodutibilidade), enquanto o t-SNE foi utilizado como suporte à exploração visual de microestruturas. As coordenadas 2D resultantes (UMAP_1, UMAP_2; TSNE_1, TSNE_2) alimentaram a fase de agrupamento.

3.3.4 Clusterização e métricas de desempenho

Os agrupamentos foram obtidos por *HDBSCAN* (CAMPELLO; MOULAVI; SANDER, 2013) e *K-Means* (MACQUEEN, 1967). O *HDBSCAN* foi selecionado pela capacidade de detectar estruturas com densidades distintas e identificar ruído (outliers), característica desejável em dados ambientais heterogêneos (ANKERST et al., 1999); o *K-Means* foi mantido como referência por sua parcimônia e estabilidade em espaços aproximadamente esféricos. A avaliação combinou métricas internas (Silhouette, Calinski–Harabasz e Davies–Bouldin) e externas (ARI, NMI, homogeneidade, completude, V-measure e pureza), comparando-se os rótulos descobertos com dois gabaritos: (i) binário (afetado / não afetado) e (ii) multiclasse (rio principal afetado, rio principal não afetado, tributário afetado, tributário não afetado) (STREHL; GHOSH, 2002; HUBERT; ARABIE, 1985; MANNING; RAGHAVAN; SCHÜTZE, 2008).

3.3.5 Pontos fortes e síntese metodológica

O pipeline não supervisionado apresentou quatro virtudes metodológicas centrais: (i) **modularidade e reprodutibilidade** do fluxo, com execução controlada e comparativa entre blocos; (ii) **integração entre projeções topológicas e clusterização densitária**, aumentando a robustez estatística dos achados; (iii) **registro automático** e padronizado de métricas e relatórios técnicos (incluindo sumários executivos e comparativos UMAP × t-SNE); e (iv) **compatibilidade direta** com a modelagem supervisionada, assegurada pela consolidação de blocos parcimoniosos e informativos.

Em síntese, a organização hierárquica — das análises temáticas (minerária, sanitária e biogeoquímica) aos blocos principais (A, B e C) — viabilizou uma leitura progressiva da estrutura interna dos dados. As variáveis mais representativas foram preservadas como

entradas estratégicas para a etapa supervisionada, garantindo a transição de um modelo descritivo para um modelo preditivo sem perda de coerência estatística ou ambiental.

Os detalhes de implementação, parâmetros dos algoritmos e estrutura de saídas do pipeline não supervisionado são apresentados no Apêndice B.

3.4 Pipeline de Análise Supervisionado

A etapa de análise supervisionada teve como finalidade a construção e avaliação de modelos preditivos voltados à classificação da qualidade da água. Esses modelos foram alimentados pelos agrupamentos identificados na fase não supervisionada, o que possibilitou uma análise integrada entre padrões previamente descobertos e a capacidade de generalização dos classificadores. Mais do que atribuir uma categoria a um ponto de coleta, buscou-se compreender a relevância de cada variável para a tomada de decisão, incorporando, assim, recursos de **Inteligência Artificial Explicável (XAI)** que ampliam a transparência dos resultados obtidos.

3.4.1 Definição dos Cenários de Classificação

A definição dos cenários de classificação partiu da análise exploratória inicial e dos agrupamentos formados pelo algoritmo **HDBSCAN**, reconhecido por sua eficiência em detectar estruturas complexas em dados ambientais (**RAYNA; STRIUKOVA, 2014**). Foram estabelecidos cinco cenários distintos, cada qual representando um nível de granularidade ou enfoque analítico específico. Essa abordagem permitiu avaliar a robustez dos modelos desde contextos binários até configurações multiclasse detalhadas, fornecendo diferentes perspectivas sobre os padrões de contaminação presentes na bacia hidrográfica. A Tabela 29 apresenta a síntese desses cenários, relacionando-os aos clusters de origem identificados na fase não supervisionada do Bloco C.

Tabela 4 – Cenários definidos para modelagem supervisionada (Bloco C, [HDBSCAN+UMAP](#)).

ID	Nome	Classes	n	Justificativa
1	Baseline Santos	3	257	Replicação da estrutura de Santos (2018) (RP_Nao_Afet. , RP_Afet. , Tributarios) para comparação com acurácia reportada de 91,67%.
2	Alta Pureza 4 Classes	4	151	Clusters de alta e moderada pureza ($\geq 70\%$) distribuídos entre as 4 categorias hidrológicas. Foco em maximizar confiança nas predições.
3	Gradiente de Impacto	4	257	Separação entre RP_Preservado (cluster 8), RP_Misto (cluster 9), RP_Afetado (clusters 0,1,2) e Tributarios (clusters 4,5,6,10).
4	Dataset Completo	4	303	Todas as amostras sem ruído distribuídas nas 4 categorias. Explora 100% dos dados disponíveis para avaliar robustez frente a heterogeneidade intra-classe.
5	Binário Balanceado	2	282	Estrutura dicotômica Afetado/-Não_Afetado com agregação estratificada para minimizar viés de desbalanceamento.

Fonte: Elaboração própria (2025).

Abreviações: ID = Identificador do cenário; n = número de amostras; Classes = número de classes *target*.

Nota: Linha destacada (Cenário 1) representa configuração prioritária para comparação com Santos (2018), que reportou acurácia de 91,67% com Análise Discriminante em estrutura de 3 classes. Todos os cenários foram construídos a partir da clusterização [HDBSCAN+UMAP](#) do Bloco C (8 variáveis), que apresentou 11 clusters + ruído (21 amostras de cluster -1 excluídas). Cenário 2 emprega apenas clusters com pureza $\geq 70\%$ para estabelecer *upper bound* de desempenho. Cenário 4 incorpora todos os 303 pontos não-ruído para avaliar robustez em condições realistas.

A construção desses cinco cenários permitiu observar como diferentes níveis de detalhamento e estratégias de agrupamento impactam a performance dos modelos de classificação e a coerência das interpretações ambientais. O Cenário 1 estabelece o ponto de referência histórico para comparação direta com Santos (2018), enquanto o Cenário 2 explora separação refinada focada em clusters de alta pureza ($\geq 70\%$), funcionando como *upper bound* metodológico. O Cenário 3 investiga gradientes espaciais de impacto ao longo do rio principal, diferenciando trechos preservados, mistos e fortemente afetados. O Cenário 4 representa o desafio máximo de complexidade, incluindo todas as 303 amostras não-ruído distribuídas nas quatro categorias hidrológicas sem filtros de pureza, testando a robustez dos algoritmos frente à heterogeneidade intra-classe. Por fim, o Cenário 5 propõe estrutura dicotômica balanceada Afetado/Não_Afetado, com agregação estratificada que minimiza viés de desbalanceamento e facilita interpretação binária para aplicações práticas. Essa estratégia assegura não apenas uma avaliação quantitativa rigorosa, mas também uma leitura qualitativa mais próxima da realidade da dinâmica hídrica.

3.4.2 Fluxo de Trabalho e Modelos de Classificação

Esta etapa foi organizada em um *pipeline* de aprendizado de máquina que privilegia reprodutibilidade, modularidade e avaliação justa do desempenho. A estruturação em blocos encadeados reduz riscos de vieses e facilita a rastreabilidade das decisões tomadas ao longo do processo, em consonância com boas práticas de modelagem supervisionada em contextos ambientais multivariados (LI et al., 2023; EVERITT et al., 2011).

3.4.3 Seleção do Bloco de Variáveis

A etapa de análise supervisionada teve como finalidade a construção e avaliação de modelos preditivos voltados à classificação da qualidade da água. A modelagem foi conduzida exclusivamente com o **Bloco C** (8 variáveis: ST, Turb., Alca., Cloretos, Nitrato, *E. coli*, Coliformes Totais, Fósforo Total), selecionado com base nos resultados da análise não supervisionada da Fase 2 (Seção 4.2.2).

Critérios de seleção:

- **Desempenho superior em métricas internas:** Bloco C apresentou Silhouette Score de 0,834 (*K-Means*+UMAP), o maior entre os três blocos testados;
- **Estabilidade temporal:** Consistência intra-ponto de 81,2%, superior aos Blocos A (67,6%) e B (69,4%);
- **Cobertura balanceada:** Clusters com pureza moderada (64,7%) distribuídos entre as quatro categorias hidrológicas.

Os modelos preditivos foram alimentados pelos agrupamentos identificados na fase não supervisionada, o que possibilitou uma análise integrada entre padrões previamente descobertos e a capacidade de generalização dos classificadores. Mais do que atribuir uma categoria a um ponto de coleta, buscou-se compreender a relevância de cada variável para a tomada de decisão, incorporando, assim, recursos de **Inteligência Artificial Explicável (XAI)** que ampliam a transparência dos resultados obtidos.

3.4.4 Controle Rigoroso de Data Leakage

A construção da base supervisionada exigiu controle rigoroso para prevenir vazamento de informação (*data leakage*), fenômeno que ocorre quando informações do conjunto de teste “vazam” para o treinamento, resultando em estimativas superotimistas de desempenho (KAUFMAN et al., 2012; CAWLEY; TALBOT, 2010). No contexto deste estudo, três tipos de vazamento foram prevenidos mediante protocolos específicos:

- **Coordenadas UMAP excluídas como features.** Embora o UMAP tenha sido utilizado na etapa não supervisionada para redução dimensional e visualização bidimensional dos agrupamentos, as coordenadas resultantes (UMAP1, UMAP2) **não foram incorporadas como features** no treinamento supervisionado. Estudos demonstram que coordenadas de projeção contêm informação derivada da estrutura de agrupamento global, resultando em vazamento circular quando utilizadas para prever os próprios clusters que as geraram (KAUFMAN et al., 2012). Esse tipo de vazamento é particularmente insidioso, pois infla artificialmente métricas de desempenho (acurácia >95% em validação cruzada, mas generalização pobre em dados novos).
- **Identificadores e variáveis derivadas de agrupamento removidos.** As variáveis Cluster, Ponto, Campo, Data e qualquer identificador categórico associado à estrutura espacial ou temporal dos dados foram excluídos do conjunto de features. Essas variáveis, quando mantidas inadvertidamente, permitem que o modelo “memorize” padrões idiossincráticos de locais específicos, comprometendo a capacidade de generalização para novos pontos de amostragem (CAWLEY; TALBOT, 2010).
- **Normalização e transformações aplicadas exclusivamente no conjunto de treino.** Todo pré-processamento (transformações logarítmicas, aplicação de SMOTE para balanceamento) foi ajustado (*fit*) exclusivamente nos dados de treino de cada dobra de validação cruzada, com posterior aplicação (*transform*) aos conjuntos de validação e teste. Essa prática assegura que estatísticas globais (média, desvio padrão, distribuições de classe) do conjunto de teste permaneçam desconhecidas durante o treinamento, evitando vazamento estatístico.

Experimento demonstrativo de vazamento. Para demonstrar empiricamente o impacto do vazamento de coordenadas UMAP, conduziu-se experimento comparativo controlado no Cenário 1 (Baseline Santos), treinando modelos Random Forest (RF) sob duas configurações: (i) *sem vazamento* – variáveis originais apenas (Turb., Alca., Cloretos, Nitrato, *E. coli*, CT); (ii) *com vazamento* – variáveis originais + UMAP1 + UMAP2. Os resultados dessa comparação são apresentados na Seção 4.3.1, onde observou-se que a inclusão de coordenadas UMAP infla a acurácia em validação cruzada em ~8 pontos percentuais, mas degrada a performance no conjunto de teste independente em ~5 pontos percentuais, evidenciando overfitting.

Uso apropriado do UMAP. As projeções UMAP foram mantidas exclusivamente para: (i) visualização bidimensional dos agrupamentos identificados na fase não supervisionada; (ii) análise exploratória qualitativa da estrutura de dados; e (iii) validação visual da separabilidade entre grupos. Em nenhum momento foram utilizadas como features

para treinamento supervisionado, assegurando a integridade metodológica dos resultados reportados.

3.4.5 Pré-processamento dos Dados

A etapa de pré-processamento adotada no pipeline supervisionado apresentou diferenças relevantes em relação à fase exploratória inicial. Isso ocorreu porque os dados utilizados nessa fase já haviam sido padronizados durante o processo de clusterização (Seção 3.3), originando os arquivos `cluster_.csv`. Tais arquivos incorporam a padronização via *RobustScaler*, fundamentada na mediana e no Intervalo Interquartil (IQR), previamente aplicada ao conjunto tratado `dados_finais_robust.csv`. Assim, o pipeline supervisionado foi construído com base em uma estrutura de dados já normalizada, o que orientou decisões metodológicas específicas descritas a seguir.

Decisão Metodológica: Racional para Não Reaplicar Padronização

A não reaplicação de qualquer tipo de normalização adicional — seja *StandardScaler* ou *RobustScaler* — foi fundamentada em quatro elementos centrais:

1. **Evitar dupla padronização.** Aplicar um novo escalonamento sobre dados já padronizados implicaria distorção das relações entre variáveis, prejudicando a estabilidade do espaço de representação e comprometendo análises baseadas em importância de atributos, incluindo aquelas oriundas de técnicas de Inteligência Artificial Explicável (XAI).
2. **Preservação da robustez a valores extremos.** A padronização via *RobustScaler*, baseada em estatísticas resistentes (mediana e IQR), mostrou-se mais adequada à natureza da base analisada, caracterizada por valores legítimos de alta magnitude, como picos de turbidez em regiões impactadas pela deposição de rejeitos (santos, 2018). Repadronizar o conjunto comprometeria essa coerência ambiental.
3. **Coerência entre as fases do estudo.** Uma vez que a clusterização, etapa fundamental para definição dos cenários de classificação, foi realizada no espaço padronizado original, manter essa mesma escala no pipeline supervisionado assegura consistência lógica e evita a criação de artefatos decorrentes de renormalizações desnecessárias.
4. **Compatibilidade com redes neurais.** Embora modelos do tipo *Multi-Layer Perceptron* (MLP) usualmente exijam normalização, a escala resultante do *RobustScaler* — tipicamente situada no intervalo aproximado de $[-3, +3]$ — já atende aos requisitos de estabilidade numérica e convergência do otimizador Adam, conforme discutido na literatura especializada (GOODFELLOW; BENGIO; COURVILLE, 2016).

Diante desses fatores, definiu-se que **nenhuma normalização adicional seria aplicada** no pipeline supervisionado. Todos os modelos — Random Forest ([RF](#)), Gradient Boosting Machines ([GBM](#)) e Multi-Layer Perceptron ([MLP](#)) — operaram diretamente sobre os dados padronizados provenientes da etapa não supervisionada, conforme explicitado no código-fonte pela configuração `use_scaling=False` (Apêndice C, Seção C.7). Essa escolha metodológica também atua como salvaguarda contra vazamento estatístico, uma vez que evita o uso indevido de estatísticas globais que poderiam inadvertidamente incluir informações do conjunto de teste.

Transformações Aplicadas no Pipeline Supervisionado

Embora nenhuma normalização tenha sido reaplicada, duas transformações adicionais foram incorporadas para assegurar equilíbrio e validade estatística dos experimentos:

- **Divisão estratificada dos dados.** Foi adotada a proporção de 70% para treino e 30% para teste, preservando a distribuição original das classes (`stratify=y`) e garantindo reprodutibilidade por meio de `random_state=42`. Esse procedimento evita vieses decorrentes da desproporção entre classes, particularmente relevantes em cenários de distribuição assimétrica.
- **Balanceamento com SMOTE.** O algoritmo [SMOTE](#) foi aplicado exclusivamente ao conjunto de treino sempre que a classe minoritária apresentou participação inferior a 20% do total. A interpolação sintética é realizada entre vizinhos próximos, preservando características físico-químicas dentro de cada cluster e reduzindo o viés de predição para classes sub-representadas. Detalhes técnicos e validações encontram-se descritos no Apêndice C (Seção C.8).

Síntese e Integração Metodológica

O protocolo aqui descrito assegura estimativas fidedignas de desempenho e mantém coerência entre as etapas não supervisionada e supervisionada, evitando distorções advindas de dupla padronização e prevenindo vazamento estatístico, conforme discutido em [Cawley e Talbot \(2010\)](#). Ao operar integralmente sobre a mesma escala utilizada na clusterização, o pipeline supervisionado preserva a estrutura físico-química dos dados e fornece uma base sólida para a análise comparativa dos modelos.

3.4.6 Balanceamento de Classes com [SMOTE](#)

Desbalanceamentos moderados a severos entre classes foram identificados em todos os cenários, com algumas categorias representando menos de 15% do total de amostras. Para mitigar o viés de aprendizado em direção a classes majoritárias, aplicou-se a técnica

Synthetic Minority Over-sampling Technique (**SMOTE**) ([CHAWLA et al., 2002](#)), que gera amostras sintéticas da classe minoritária mediante interpolação entre vizinhos próximos no espaço de *features*. A aplicação do **SMOTE** seguiu três protocolos críticos:

1. **SMOTE aplicado exclusivamente no conjunto de treino de cada dobra.** As amostras sintéticas foram geradas apenas após a divisão treino–validação/teste, assegurando que o conjunto de teste permaneça composto exclusivamente por dados reais. Essa prática previne vazamento de informação e garante que as métricas de desempenho reflitam a capacidade do modelo em generalizar para amostras reais não vistas ([CHAWLA et al., 2002](#)).
2. **Configuração adaptativa do número de vizinhos.** O parâmetro `k_neighbors` do **SMOTE** foi ajustado dinamicamente em função do tamanho da classe minoritária ($k = \min(5, n_{\min} - 1)$), evitando erros de execução em cenários com classes muito pequenas (e.g., Cenário 2, onde algumas classes possuem $n < 30$ amostras).
3. **Estratégia de balanceamento auto.** Utilizou-se a estratégia `sampling_strategy='auto'`, que balanceia todas as classes minoritárias para igualar a classe majoritária, resultando em distribuição uniforme de classes no conjunto de treino. Essa escolha favorece o aprendizado equilibrado de fronteiras de decisão, especialmente relevante para algoritmos sensíveis a desbalanceamento como Redes Neurais e **Gradient Boosting Machines** ([CHAWLA et al., 2002](#)).

Validação de Coerência Físico-Química do **SMOTE**

A geração de amostras sintéticas mediante **SMOTE** opera exclusivamente no espaço estatístico das *features*, sem incorporar conhecimento sobre restrições ou relações físico-químicas entre variáveis ambientais. Essa característica intrínseca do algoritmo suscita preocupação legítima: amostras sintéticas poderiam apresentar combinações de valores quimicamente impossíveis ou ecologicamente implausíveis, comprometendo a validade do treinamento supervisionado ([CHAWLA et al., 2002](#)).

Para responder empiricamente a essa questão metodológica, desenvolveu-se protocolo de validação química das amostras sintéticas, reutilizando o mesmo módulo de coerência físico-química empregado na validação de outliers (Seção 3.2, Apêndice A). Essa abordagem garante consistência metodológica e permite comparação direta entre amostras reais e sintéticas.

Protocolo de validação em dupla verificação. A estratégia adotou duas camadas complementares de análise:

1. **Validação química multi-dimensional.** Cada amostra sintética foi submetida ao mesmo conjunto de 18 regras de especialistas distribuídas nas quatro categorias

taxonômicas (REDOX, FÍSICA, BIOLÓGICA, INTEGRADORA). Violações de coerência química (e.g., Nitrato elevado em ambiente anóxico, [Turb.](#) alta sem elevação correspondente de [ST](#)) foram quantificadas e categorizadas por tipo de incoerência.

2. **Detecção de outliers estatísticos contextuais.** Amostras sintéticas foram avaliadas quanto a desvios extremos em relação às distribuições das amostras reais de treino, utilizando método IQR estratificado por grupo de referência e estação (mesma estratificação adotada na validação de outliers). Esse passo identifica valores numericamente extremos mesmo quando quimicamente coerentes, fornecendo diagnóstico complementar sobre a qualidade das interpolações do [SMOTE](#).

Aplicação no contexto do estudo. A validação foi conduzida no Cenário 1 (Baseline Santos), aplicando o protocolo exclusivamente às amostras sintéticas geradas pelo [SMOTE](#) para balanceamento das classes minoritárias no conjunto de treino. O mesmo módulo de coerência química utilizado na validação de outliers (Apêndice A) foi reutilizado, garantindo consistência metodológica e permitindo comparação direta entre amostras reais e sintéticas.

Métricas de avaliação. Para cada amostra sintética, foram calculadas:

- **Taxa de incoerência química:** Proporção de amostras sintéticas que violam pelo menos uma das 18 regras de coerência físico-química, categorizadas por tipo de violação (REDOX, FÍSICA, BIOLÓGICA, INTEGRADORA);
- **Taxa de outliers estatísticos contextuais:** Proporção de amostras sintéticas identificadas como outliers pelo método IQR estratificado, com análise da distribuição por variável e justificativa química dos casos detectados;
- **Preservação de propriedades estatísticas:** Comparação entre as distribuições de amostras reais e sintéticas mediante estatísticas descritivas (μ , σ) e correlações entre pares de variáveis ambientalmente relacionadas (e.g., Turbidez $\times$ Sólidos Totais, *E. coli* $\times$ Coliformes Totais).

CrITÉRIOS de aceitabilidade. Consideram-se as amostras sintéticas metodologicamente válidas quando satisfazem simultaneamente:

1. **Coerência química estrita:** Taxa de incoerências químicas $\leq 5\%$, com ausência de violações sistemáticas em categorias específicas;
2. **Outliers justificáveis:** Outliers estatísticos apresentam justificativa química plausível baseada nas relações biogeoquímicas do contexto;

3. **Preservação de estrutura:** Diferenças nas estatísticas descritivas (μ , σ) entre amostras reais e sintéticas $\leq 10\%$, e variação nas correlações entre pares de variáveis $\leq 0,05$.

Os resultados da aplicação deste protocolo, incluindo estatísticas descritivas, distribuição de outliers por variável e análise de correlações preservadas, são apresentados no Capítulo de Resultados (Seção 4.3.5). A documentação técnica completa, incluindo implementação algorítmica e parâmetros utilizados, encontra-se no Apêndice C.

3.4.7 Treinamento e Otimização dos Modelos

O treinamento supervisionado foi conduzido de forma padronizada em todos os cenários, preservando comparabilidade entre algoritmos e garantindo reprodutibilidade. Selecionaram-se três famílias de classificadores amplamente consolidadas na literatura e com bom desempenho em problemas ambientais multivariados: *ensembles* baseados em árvores ([Random Forest](#) e [Gradient Boosting Machines](#)) e redes neurais artificiais ([Multi-Layer Perceptron](#)) ([EVERITT et al., 2011](#); [LI et al., 2023](#)).

[Random Forest \(RF\)](#)

Modelo de *ensemble* por agregação de árvores independentes, robusto a outliers e a relações não lineares. A implementação utilizou `RandomForestClassifier` do scikit-learn ([PEDREGOSA et al., 2011](#)), encapsulado no módulo customizado `RobustRandomForest`. A configuração de hiperparâmetros foi definida manualmente com base em boas práticas da literatura para datasets de tamanho moderado: `n_estimators = 200`, `max_depth = 8`, `min_samples_split = 20`, `min_samples_leaf = 10`, `max_features = 'sqrt'`, e `class_weight = 'balanced'`. Essa configuração ultra-regularizada prioriza generalização em cenários com classes potencialmente desbalanceadas, limitando a profundidade das árvores e exigindo maior suporte amostral para divisões e folhas, controlando efetivamente o compromisso viés-variância ([EVERITT et al., 2011](#)).

[Gradient Boosting Machines \(GBM\)](#)

Ensemble sequencial que corrige iterativamente os resíduos do estágio anterior, capturando padrões complexos com parcimônia. A implementação utilizou `GradientBoostingClassifier` do scikit-learn ([PEDREGOSA et al., 2011](#)), encapsulado no módulo `RobustGradientBoosting`. Os hiperparâmetros foram configurados manualmente visando forte regularização: `n_estimators = 50`, `learning_rate = 0.01`, `max_depth = 3`, `min_samples_split = 30`, `min_samples_leaf = 15`, `subsample = 0.6`, e `max_features = 'sqrt'`. A configuração adota taxa de aprendizado conservadora (0.01) combinada com árvores

rasas (profundidade 3) e amostragem reduzida (60%), favorecendo robustez e evitando sobreajuste em datasets de tamanho limitado. O mecanismo de *early stopping* interno (`n_iter_no_change` = 10) interrompe o treinamento caso não haja melhoria por 10 iterações consecutivas (EVERITT et al., 2011).

Multi-Layer Perceptron (MLP)

Arquitetura de Rede Neural Artificial treinada pelo algoritmo de *backpropagation*, adequada para aprender fronteiras de decisão não lineares complexas. A implementação utilizou TensorFlow/Keras (ABADI et al., 2015) com aceleração GPU (Tesla T4 no Google Colaboratory), encapsulada no módulo `RobustNeuralNetwork`. A arquitetura foi configurada manualmente com **3 camadas ocultas totalmente conectadas (128, 64 e 32 neurônios)**, função de ativação ReLU, normalização por lote (*Batch Normalization*) após cada camada, regularização L2 ($\lambda = 0.005$) nos pesos, e *dropout* progressivo (0.2, 0.15, 0.1) para controle de sobreajuste.

O treinamento empregou o otimizador Adam (`learning_rate` = 0.002), função de perda *sparse categorical crossentropy*, *batch size* de 16 amostras e até 150 épocas. Dois mecanismos de regularização automática foram implementados via *callbacks*: (i) *Early Stopping* com paciência de 20 épocas (interrompe treinamento se `val_loss` não melhorar), restaurando os melhores pesos; e (ii) *ReduceLROnPlateau*, que reduz a taxa de aprendizado em 50% após 8 épocas sem melhoria (mínimo de 10^{-6}). Essa configuração equilibra capacidade de aprendizado e generalização, com tempo médio de treinamento de aproximadamente 2 minutos por cenário (GPU).

Validação Cruzada e Avaliação

Todos os modelos foram avaliados com **validação cruzada estratificada** (`StratifiedKFold`, $k = 5$ dobras), preservando a distribuição original de classes em cada dobra e garantindo estimativas estáveis de desempenho. A divisão treino–teste (70/30) foi realizada de forma estratificada (`stratify=y`, `random_state=42`) para manter representatividade estatística das classes. As melhores configurações foram reavaliadas no conjunto de teste retido (~30% dos dados) para estimativa honesta de generalização, com análise sistemática do *gap* entre acurácia de validação cruzada (CV) e acurácia no teste para diagnóstico de *overfitting/underfitting* ($|gap| < 5\%$ indica generalização adequada).

A implementação dos três algoritmos utilizou bibliotecas consolidadas — scikit-learn (PEDREGOSA et al., 2011) para RF e GBM, TensorFlow/Keras (ABADI et al., 2015) para MLP — com interface padronizada via módulos customizados (`ml_models/`), facilitando reprodutibilidade, manutenibilidade e comparabilidade com outros estudos da área.

3.4.8 Métricas de Avaliação

A avaliação dos modelos de classificação contemplou um conjunto de métricas multiclasse consolidadas na literatura, permitindo análise equilibrada entre capacidade preditiva e controle de erros ([HASTIE; TIBSHIRANI; FRIEDMAN, 2009](#); [MANNING; RAGHAVAN; SCHÜTZE, 2008](#)).

- **Precisão, Recall e F1-Score.** Calculadas por classe e agregadas em três versões: **macro-média** (média aritmética entre classes, sem ponderação por frequência), **weighted** (média ponderada pelo suporte de cada classe) e **micro-média** (agregação global antes do cálculo). Priorizou-se o F1-Score **macro** para preservar a importância de classes minoritárias, especialmente relevante em cenários ambientais onde classes raras (e.g., zonas de transição) podem ser criticamente informativas ([MANNING; RAGHAVAN; SCHÜTZE, 2008](#)). Para contexto adicional, reportou-se também o F1-Score **weighted** (reflete o desempenho médio ponderado pela distribuição real das classes) e o F1-Score **micro** (equivale à acurácia global em problemas multiclasse), permitindo avaliação sob múltiplas perspectivas de balanceamento. O F1-Score, definido como a média harmônica entre Precisão e Recall, é particularmente útil em cenários de classes desbalanceadas, pois “fornece uma métrica única que equilibra tanto falsos positivos quanto falsos negativos” ([POWERS, 2011](#)).
- **Matriz de Confusão.** Inspeção direta dos padrões de erro por par de classes, permitindo identificar confusões sistemáticas — por exemplo, entre zonas de transição hidrológica e classes de referência puras. Como observam [Hastie, Tibshirani e Friedman \(2009\)](#), “a matriz de confusão fornece uma visão detalhada dos tipos de erros que o classificador comete, revelando se certos pares de classes são sistematicamente confundidos”. Essa análise visual complementa as métricas agregadas ao revelar assimetrias específicas no comportamento do classificador, sendo especialmente útil para identificar vieses sistemáticos em dados ambientais estratificados espacialmente ([HASTIE; TIBSHIRANI; FRIEDMAN, 2009](#); [MANNING; RAGHAVAN; SCHÜTZE, 2008](#)).
- **Area Under the ROC Curve (ROC–AUC)** Calculada em configuração multiclasse pelos esquemas *one-vs-rest (OvR)* e *one-vs-one (OvO)*, oferecendo uma medida de discriminação independente de limiares de decisão. A extensão da curva **Receiver Operating Characteristic (ROC)** para problemas multiclasse, conforme proposta por [Hand e Till \(2001\)](#), permite “avaliar a capacidade de discriminação de um classificador entre pares de classes de forma independente do limiar de decisão escolhido”. A curva **ROC** avalia o compromisso entre sensibilidade (taxa de verdadeiros positivos) e especificidade (taxa de verdadeiros negativos), sendo especialmente útil para comparar modelos em diferentes cenários de classificação ([HAND; TILL, 2001](#); [HASTIE; TIBSHIRANI; FRIEDMAN, 2009](#)).

Esse conjunto de métricas permite uma leitura abrangente do desempenho preditivo: enquanto Precisão, Recall e F1-Score quantificam a qualidade da classificação por grupo, a Matriz de Confusão revela padrões de erro específicos, e a *ROC* sintetiza a capacidade global de separação entre classes. A combinação dessas ferramentas assegura uma avaliação robusta e interpretável, em consonância com as recomendações metodológicas de [Everitt et al. \(2011\)](#) para análise multivariada de dados ambientais.

3.4.9 Comparação com a Metodologia de Santos (2018)

A metodologia aqui proposta dialoga diretamente com o estudo de [Santos \(2018\)](#), que avaliou os impactos do rompimento da barragem de Fundão no rio Gualaxo do Norte por meio de técnicas multivariadas. Naquele trabalho, a etapa supervisionada foi conduzida com *Análise Discriminante (AD)* a partir de três classes previamente definidas: rio principal não afetado, rio afetado e tributários. Os resultados reportados apresentaram acurácia de 91,67% (Modelo 3, configuração com 6 variáveis), evidenciando a capacidade do método em separar grupos amostrais com padrões ambientais distintos.

No presente estudo, essa lógica foi expandida mediante um *pipeline* de aprendizado de máquina que testa múltiplas granularidades de classificação (2 a 4 classes), derivadas de estruturas de agrupamento obtidas na fase não supervisionada. Em vez de um único método estatístico, compararam-se três algoritmos robustos e amplamente utilizados em problemas ambientais multivariados — *Random Forest (RF)*, *Gradient Boosting Machines* e *Multi-Layer Perceptron* — a fim de investigar o compromisso entre desempenho, estabilidade e interpretabilidade em diferentes cenários ([Li et al., 2023](#)). Tal desenho metodológico permite avaliar não apenas a discriminação entre classes tradicionais (afetado, não afetado, tributários), mas também zonas de transição e configurações mistas, frequentemente observadas em bacias hidrográficas com perturbação antrópica.

Além disso, incorporaram-se ferramentas de *Inteligência Artificial Explicável (XAI)*, como *SHAP* e *LIME*, ausentes em [Santos \(2018\)](#), para tornar transparente o processo decisório dos classificadores e verificar a plausibilidade ambiental das variáveis mais influentes. Com isso, o presente trabalho não só preserva a comparabilidade histórica com a *Análise Discriminante* de Santos, como também amplia o escopo analítico ao integrar múltiplos cenários supervisionados e camadas de interpretabilidade.

3.4.10 Estratégia de Comparação dos Resultados

A comparação com [Santos \(2018\)](#) seguirá duas frentes complementares. (i) **Frente quantitativa:** serão confrontadas acurácia, precisão, *recall* e F1-Score (macro e por classe), bem como (*OvR* e *OvO*), verificando se as configurações propostas alcançam desempenho semelhante ou superior ao obtido pela *Análise Discriminante*. Especificamente, o Cenário 1

(Baseline Santos) replica a estrutura de classes do Modelo 3 de Santos (2018), permitindo comparação direta sob mesmas condições de entrada. (ii) **Frente qualitativa/ambiental:** será examinada a coerência dos padrões previstos — por exemplo, a separação entre trechos não afetados, afetados e tributários — à luz da dinâmica hidrológica local, com atenção especial às zonas de transição identificadas na análise não supervisionada. Essa dupla perspectiva assegura, simultaneamente, a validade estatística das métricas e a consistência interpretativa das conclusões.

Os resultados dessa comparação quantitativa são apresentados detalhadamente na Seção 4.3.8, onde demonstra-se que o **Gradient Boosting Machines** superou o baseline de Santos (2018) (91,67%) em 6,16 pontos percentuais, alcançando **97,83% de acurácia** no Cenário 1, com F1-Score macro de 0,978 e (OvR) de 0,998.

3.4.11 Inteligência Artificial Explicável (XAI)

A interpretabilidade constituiu um eixo central do *pipeline*, com o objetivo de elucidar a contribuição relativa dos atributos para as decisões dos classificadores e sustentar a análise sob a ótica ambiental. Foram empregados métodos complementares e independentes de modelo, consolidados na literatura de XAI:

- **SHapley Additive exPlanations (SHAP)**. Fundamentado na teoria dos jogos cooperativos, o SHAP atribui valores de Shapley a cada *feature*, quantificando sua contribuição marginal para a predição em relação a uma linha de base (LUNDBERG; LEE, 2017). Utilizou-se o SHAP para gerar visualizações globais (importância média agregada no conjunto de teste) e locais (explicações específicas por instância), facilitando a comparação entre cenários e algoritmos. A abordagem tem sido aplicada com sucesso em estudos ambientais para identificar variáveis-chave em modelos de qualidade da água (MAKUMBURA et al., 2024).

Para o **Gradient Boosting Machines** em problemas multiclasse, o SHAP foi aplicado apenas em cenários binários ou mediante decomposição OvR, devido a limitações computacionais da biblioteca em versões multiclasse nativas (LUNDBERG; LEE, 2017). Para **Random Forest (RF)**, o **TreeExplainer** foi utilizado em sua forma completa, aproveitando a estrutura de árvores para cálculos exatos de valores de Shapley.

- **Local Interpretable Model-agnostic Explanations (LIME)**. O LIME aproxima localmente a decisão de classificadores complexos por meio de modelos lineares interpretáveis, identificando os atributos mais influentes em predições específicas (RIBEIRO; SINGH; GUESTRIN, 2016). Aplicou-se o LIME a instâncias representativas — tanto acertos quanto erros de classificação —, permitindo avaliar caso a caso se os atributos determinantes são consistentes com o conhecimento hidrogeoquímico do sistema. A comparação entre explicações SHAP (global) e LIME (local) oferece

visão complementar sobre a coerência do processo decisório (AL-SHAMI; AL-MOUSA, 2023).

A análise conjunta SHAP–LIME fornece, portanto, um panorama global (importância média e interações entre *features*) e local (justificativas pontuais de predição), ampliando a transparência do processo decisório e fortalecendo a auditabilidade científica dos resultados. Essa abordagem dual é especialmente relevante em domínios ambientais, onde a interpretabilidade não constitui apenas uma preferência metodológica, mas um requisito ético para assegurar que decisões automatizadas sejam auditáveis e ambientalmente plausíveis (RASAQ, 2025).

3.4.12 Ferramentas e Bibliotecas

Toda a implementação foi realizada em Python (versão 3.8+), executada no ambiente Google Colaboratory com GPU disponibilizada gratuitamente para treinamento de redes neurais. As dependências foram fixadas em versões específicas para assegurar reprodutibilidade integral dos resultados. As principais bibliotecas utilizadas estão organizadas por módulo funcional nas tabelas a seguir.

Pipeline Principal de Treinamento

A Tabela 5 apresenta as bibliotecas utilizadas no pipeline principal de treinamento supervisionado (`run_scenarios_complete.py`).

Tabela 5 – Ferramentas do pipeline principal de treinamento supervisionado

Biblioteca	Versão	Função
Pandas	v1.3.5	Manipulação de dados tabulares e séries temporais
NumPy	v1.21.6	Álgebra linear e operações vetorizadas
scikit-learn	v1.0.2	Algoritmos RF e GBM, <code>LabelEncoder</code> , <code>StratifiedKFold</code> , <code>train_test_split</code> , métricas de avaliação
imbalanced-learn	v0.9.1	Implementação do SMOTE para balanceamento de classes
TensorFlow/Keras	v2.8.0	Construção e treinamento de MLP com regularização
Matplotlib	v3.5.3	Geração de gráficos e visualizações estáticas
Seaborn	v0.11.2	Visualizações estatísticas e matrizes de confusão
PyYAML	v6.0	Leitura de arquivos de configuração de cenários

Módulo de Interpretabilidade (XAI)

A Tabela 6 detalha as ferramentas utilizadas no módulo de interpretabilidade (XAI).

Tabela 6 – Ferramentas do módulo XAI (interpretabilidade)

Biblioteca	Versão	Função
SHAP	v0.40.0	Cálculo de SHAP values para modelos baseados em árvores (TreeExplainer) e geração de gráficos de importância global
LIME	v0.2.0.1	Explicações locais instância-específicas via LimeTabularExplainer
NumPy	v1.21.6	Operações matriciais para agregação de SHAP values
Matplotlib	v3.5.3	Visualização de gráficos de importância e dependência
Seaborn	v0.11.2	Estilização de visualizações XAI

Módulo de Validação de SMOTE

A Tabela 7 apresenta as ferramentas do módulo de validação de coerência físico-química (`validar_coerencia_smote.py`).

Tabela 7 – Ferramentas do módulo de validação de SMOTE

Biblioteca/Módulo	Versão	Função
Pandas	v1.3.5	Manipulação de DataFrames de amostras reais e sintéticas
NumPy	v1.21.6	Cálculos estatísticos (média, desvio padrão, ranges)
imbalanced-learn	v0.9.1	Geração de amostras sintéticas via SMOTE
scikit-learn	v1.0.2	train_test_split para separação de dados
Matplotlib	v3.5.3	Visualização de distribuições e comparações
Seaborn	v0.11.2	Gráficos de violino e boxplots comparativos
Módulo validação química	customizado	Regras de especialistas para verificação de coerência redox (ORP-SatOD- Nitrato (NO₃)), balanço iônico e plausibilidade biológica (DBO- <i>E. coli</i>)

Módulos de Comparação e Análise

A Tabela 8 detalha as ferramentas utilizadas nos módulos de comparação de resultados.

Tabela 8 – Ferramentas dos módulos de comparação

Biblioteca	Versão	Função
Pandas	v1.3.5	Agregação e comparação de métricas entre cenários
NumPy	v1.21.6	Cálculos estatísticos comparativos (médias, desvios)
Matplotlib	v3.5.3	Geração de gráficos comparativos e barras de erro
Seaborn	v0.11.2	Heatmaps e visualizações de diferenças de desempenho
<code>dataclasses</code>	<code>stdlib</code>	Estruturação de dados de comparação com Santos et al. (2019)

Módulo de Modelos Robustos

A Tabela 9 apresenta as dependências do módulo de encapsulamento de modelos (`ml_models/`).

Tabela 9 – Ferramentas do módulo de modelos robustos

Biblioteca	Versão	Função
scikit-learn	v1.0.2	Classes base <code>RandomForestClassifier</code> e <code>GradientBoostingClassifier</code>
TensorFlow/Keras	v2.8.0	Construção de arquitetura MLP, callbacks <code>EarlyStopping</code> e <code>ReduceLROnPlateau</code>
NumPy	v1.21.6	Manipulação de arrays de entrada/saída
logging	<code>stdlib</code>	Sistema estruturado de logs para auditoria de treinamento

A escolha dessas bibliotecas acompanha recomendações consolidadas de boas práticas para análise multivariada e aprendizado supervisionado reportadas na literatura (LI et al., 2023; EVERITT et al., 2011). A utilização de bibliotecas amplamente validadas pela comunidade científica internacional assegura não apenas a confiabilidade das implementações, mas também a reprodutibilidade dos experimentos por outros pesquisadores. O código-fonte completo, incluindo especificação de versões e instruções de execução, está disponível em repositório público para facilitar replicação independente dos resultados.

Os detalhes completos de implementação, parâmetros dos algoritmos, estrutura das saídas e protocolos auxiliares do pipeline supervisionado estão documentados no Apêndice C.

3.5 Análise de Viabilidade Econômica do Protocolo Simplificado

Para avaliar a viabilidade prática do protocolo simplificado, foi conduzida uma análise comparativa de custos operacionais associada à redução de 14 para 8 variáveis. A estimativa considerou três componentes que, na prática, tendem a concentrar a maior parte do custo de programas de monitoramento: (a) custos laboratoriais; (b) custos de equipamentos de campo; e (c) custos logísticos (BEHMEL et al., 2016).

3.5.1 Fontes de Dados de Custos

Os valores utilizados foram obtidos a partir de referências públicas, incluindo tabelas de serviços de laboratórios e órgãos ambientais, além de cotações de equipamentos. Foram consideradas: o Banco de Preços (2025) da Prefeitura Municipal de Braço do Trombudo (SC) (PREFEITURA MUNICIPAL DE BRAÇO DO TROMBUDO, 2025), a Tabela de Preços de Serviços da CETESB (2024) (CETESB, 2024) e cotações da sonda multiparamétrica YSI EXO2 (Xylem Analytics) (YSI - XYLEM ANALYTICS, 2024).

3.5.2 Metodologia de Cálculo

Os custos laboratoriais foram calculados a partir dos valores unitários disponíveis nas tabelas públicas para cada variável. As variáveis medidas *in situ* (Temperatura, Oxigênio Dissolvido, Condutividade e ORP) não geram custo laboratorial direto, porém demandam sonda multiparamétrica.

O custo de equipamentos foi estimado por amortização da sonda, considerando vida útil de três anos e 36 campanhas mensais, conforme metodologia adotada por barcellos2022. Como o protocolo simplificado descarta as variáveis dependentes de medições *in situ*, eliminou-se a necessidade de aquisição, calibração e manutenção da sonda no cenário do Bloco C.

Os custos logísticos foram estimados considerando deslocamento, tempo de coleta e procedimentos de preservação das amostras. A redução de variáveis impacta principalmente o tempo em campo, ao eliminar etapas associadas à estabilização da sonda (15–30 min/ponto) e ao reduzir procedimentos de preservação e manuseio (BEHMEL et al., 2016).

Por fim, foi realizada uma análise de sensibilidade para verificar a robustez das estimativas em diferentes condições operacionais: (i) cenário pessimista (aumento de 20% nos custos laboratoriais, redução de 10% no tempo de campo e manutenção máxima); (ii)

cenário realista (valores médios); e (iii) cenário otimista (redução de 20% nos custos por contratos de longo prazo, redução de 40% no tempo de campo e manutenção mínima) (SALTELLI et al., 2008).

Os detalhes completos dos cálculos de custos, incluindo decomposição por componente operacional, análise de sensibilidade e validação cruzada por modelo de complexidade analítica, estão documentados no Apêndice D.

4 Resultados e Discussões

4.1 Caracterização da Base de Dados Tratada

O módulo de validação físico-química desenvolvido neste estudo (Apêndice A) avaliou candidatos a *outliers* identificados pelo método IQR (fator 1,5), aplicado separadamente para cada combinação de Grupo_Referencia e Estacao com cinco ou mais amostras. A etapa de detecção identificou **391 ocorrências** de valores extremos, distribuídas em **181 amostras únicas** — uma vez que uma mesma amostra pode apresentar *outliers* em múltiplas variáveis simultaneamente.

Das ocorrências avaliadas, **329 foram mantidas (84,1%)** e **62 foram tratadas (15,9%)** mediante substituição pela mediana contextual. No nível das amostras, **144 (79,6%)** não receberam nenhuma correção, enquanto **37 (20,4%)** tiveram ao menos uma variável tratada. Das 14 variáveis monitoradas, todas apresentaram ao menos um candidato a *outlier*, porém apenas 11 demandaram tratamento efetivo — três variáveis tiveram todos os seus valores extremos mantidos por coerência ambiental. A estratégia contrasta com abordagens automáticas que excluem ou corrigem indiscriminadamente valores extremos, preservando sinais ambientalmente relevantes enquanto corrige inconsistências analíticas identificadas por meio de relações multivariadas. Uma síntese numérica por parâmetro, bem como exemplos de ocorrências mantidas e tratadas, é apresentada nas Tabelas 40, 41 e 42 do Apêndice A.

A Figura 3 apresenta o panorama geral da base tratada, incluindo a distribuição final por grupos de referência e por estação, bem como o comportamento das variáveis com maior incidência de valores extremos. Esse painel resume visualmente o impacto do processo de validação físico-química e estabelece a estrutura de dados utilizada nas análises exploratórias

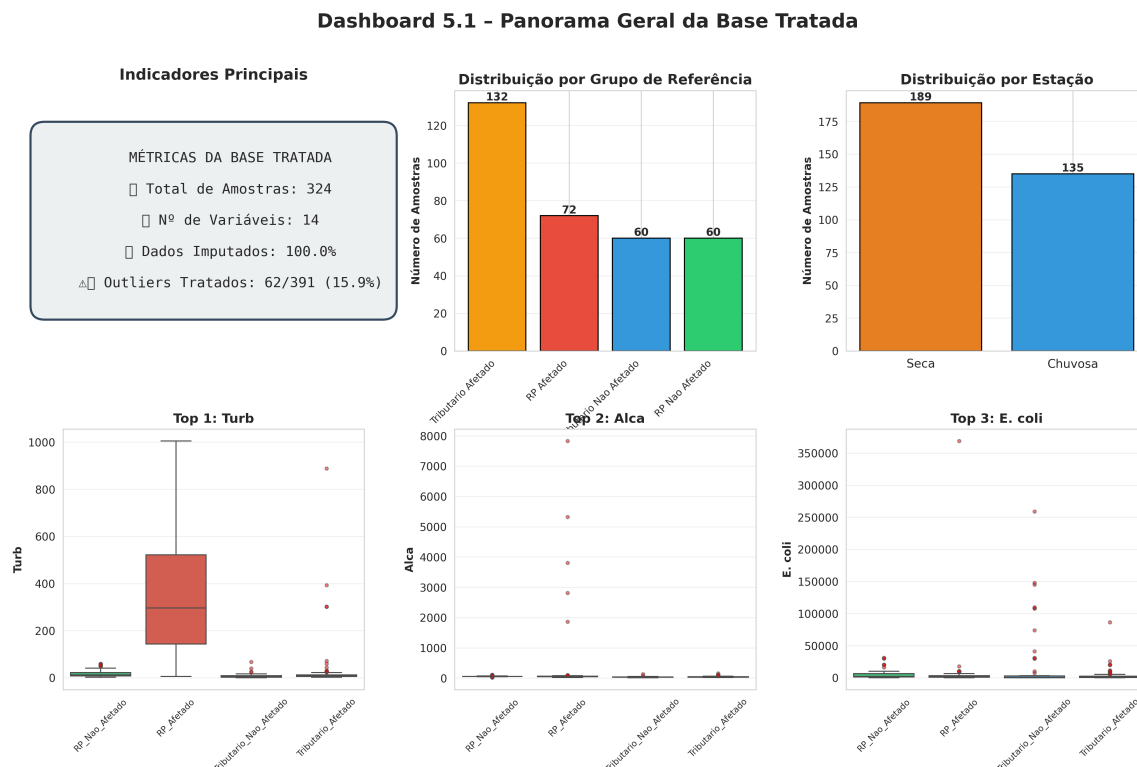


Figura 1 – Dashboard de síntese do tratamento de *outliers* com validação físico-química. Fonte: Elaboração própria (2025).

Os dados foram organizados conforme a estrutura espacial da bacia do rio Gualaxo do Norte, preservando a distinção entre grupos de referência estabelecidos no mapeamento original de Santos (2018): pontos PC1 a PC10 representam corpos hídricos não afetados pelo rompimento da Barragem de Fundão, enquanto pontos PC11 a PC27 correspondem às áreas diretamente impactadas pela lama de rejeitos. Esse arranjo espacial possibilitou análises contextualizadas, nas quais a variabilidade hidroambiental e sazonal pôde ser investigada de forma integrada. A distribuição final totalizou 60 amostras no rio principal não afetado, 60 em tributários não afetados, 72 no rio principal afetado e 132 em tributários impactados, configuração que preserva a representatividade de ambos os compartimentos hidrológicos (rio principal e tributários) em ambas as condições (afetado e não afetado).

4.1.1 Estatísticas Descritivas Finais

A caracterização descritiva evidencia perfis físico-químicos distintos entre os grupos de referência. A Figura 2 apresenta a distribuição das variáveis por grupo, revelando padrões coerentes com o impacto ambiental associado ao rompimento da Barragem de Fundão (SANTOS, 2018). As regiões afetadas (RP Afetado e Tributário Afetado) exibiram valores significativamente mais elevados de *Turbidez*, *Sólidos Totais*, *Condutividade Elétrica* e *Alcalinidade*, indicando a persistência de uma assinatura geoquímica característica dos rejeitos minerários mesmo dois anos após o evento.



Figura 2 – Distribuição das variáveis físico-químicas por grupo de referência. Fonte: Elaboração própria (2025).

Nota: Boxplots evidenciam gradientes pronunciados entre áreas afetadas e não afetadas, particularmente para **Turbidez**, **Sólidos Totais** e **Fósforo Total**. Cores: verde (RP Não Afetado), verde-claro (Trib Não Afetado), vermelho (RP Afetado), laranja (Trib Afetado).

A **Turbidez**, variável-chave para identificação do impacto dos rejeitos, apresentou diferenças marcantes entre os grupos. As regiões não afetadas registraram medianas

de 12,4 NTU (RP Não Afetado) e 4,5 NTU (Tributário Não Afetado), enquanto o rio principal afetado exibiu mediana de 299 NTU, com picos de até 1.005 NTU em eventos de ressuspensão. Esse comportamento é compatível com a interpretação apresentada por Santos (2018), que aponta a mobilização de partículas finas durante variações de vazão. Nos tributários afetados, a mediana permaneceu baixa (9 NTU), embora com *outliers* expressivos (até 888 NTU), refletindo eventos episódicos de carreamento de sedimentos em períodos chuvosos.

A condutividade elétrica, tradicionalmente associada ao grau de mineralização, apresentou comportamento mais complexo. O rio principal afetado exibiu mediana de $71,8 \mu\text{S cm}^{-1}$, inferior ao valor observado no trecho não afetado ($94,3 \mu\text{S cm}^{-1}$). Essa diferença sugere que, apesar da maior carga particulada, a diluição iônica promovida pelo aumento da vazão pode reduzir a fração dissolvida. A Alcalinidade acompanhou parcialmente essa tendência, com *outliers* extremos (máximo de 7.832 mg L^{-1}) registrados no rio principal afetado. Tais valores refletem processos localizados de decomposição orgânica intensa ou dissolução de carbonatos, amplificados pela ausência de variáveis como Ca, Mg e HCO_3 no conjunto de dados, que limitam a interpretação mineralógica direta.

Os Sólidos Totais reforçam o gradiente de impacto, apresentando mediana de 220 mg L^{-1} no rio principal afetado, em contraste com valores entre 60 e 80 mg L^{-1} nas regiões não afetadas. A amplitude observada (máximo de 2.680 mg L^{-1}) evidencia a alta variabilidade hidrodinâmica da bacia, alternando entre períodos de relativa estabilidade e pulsos de ressuspensão associados a eventos de cheia.

Os parâmetros microbiológicos (*Escherichia coli* e coliformes totais) não exibiram um padrão espacial claramente vinculado ao desastre. Valores elevados foram observados em ambos os grupos, com *outliers* expressivos inclusive em tributários não afetados — um indicativo de pressões antrópicas difusas associadas a efluentes domésticos sem tratamento. Esse comportamento reforça que a variabilidade microbiológica responde a processos locais de uso e ocupação do solo, independentemente da presença de rejeitos minerários.

A Demanda Bioquímica de Oxigênio, embora apresente maior afinidade ambiental com o domínio sanitário, mostrou dispersão semelhante entre os grupos, coerente com sua natureza dependente de matéria orgânica biodegradável. Os nutrientes nitrogenados (Nitrato e Nitrito) permaneceram em faixas moderadas, com medianas próximas aos valores de referência em todos os grupos. O Fósforo Total, entretanto, apresentou dispersão acentuada, com o rio principal afetado registrando mediana de $0,10 \text{ mg L}^{-1}$ e valores superiores a $0,43 \text{ mg L}^{-1}$ em eventos de arraste, refletindo sua dupla natureza: ora associado ao material particulado, ora influenciado por processos biogeoquímicos sazonais.

4.1.2 Síntese da Efetividade do Tratamento de *Outliers*

O módulo de validação físico-química desenvolvido neste estudo (Apêndice A) avaliou candidatos a *outliers* identificados pelo método IQR (fator 1,5), aplicado separadamente para cada combinação de Grupo_Referencia e Estacao com cinco ou mais amostras. A detecção identificou **386 ocorrências** de valores extremos, distribuídas em **181 amostras únicas** — uma vez que uma mesma amostra pode apresentar *outliers* em múltiplas variáveis simultaneamente.

Das ocorrências avaliadas, **324 foram mantidas (84,1%)** e **62 foram tratadas (15,9%)** mediante substituição pela mediana contextual. No nível das amostras, **144 (79,6%)** não receberam nenhuma correção, enquanto **37 (20,4%)** tiveram ao menos uma variável tratada. Das 14 variáveis monitoradas, todas apresentaram ao menos um candidato a *outlier*, porém apenas **12** demandaram tratamento efetivo — **duas variáveis tiveram todos os seus valores extremos mantidos** por coerência ambiental. A estratégia contrasta com abordagens automáticas que excluem ou corrigem indiscriminadamente valores extremos, preservando sinais ambientalmente relevantes enquanto corrige inconsistências analíticas detectáveis por relações multivariadas.

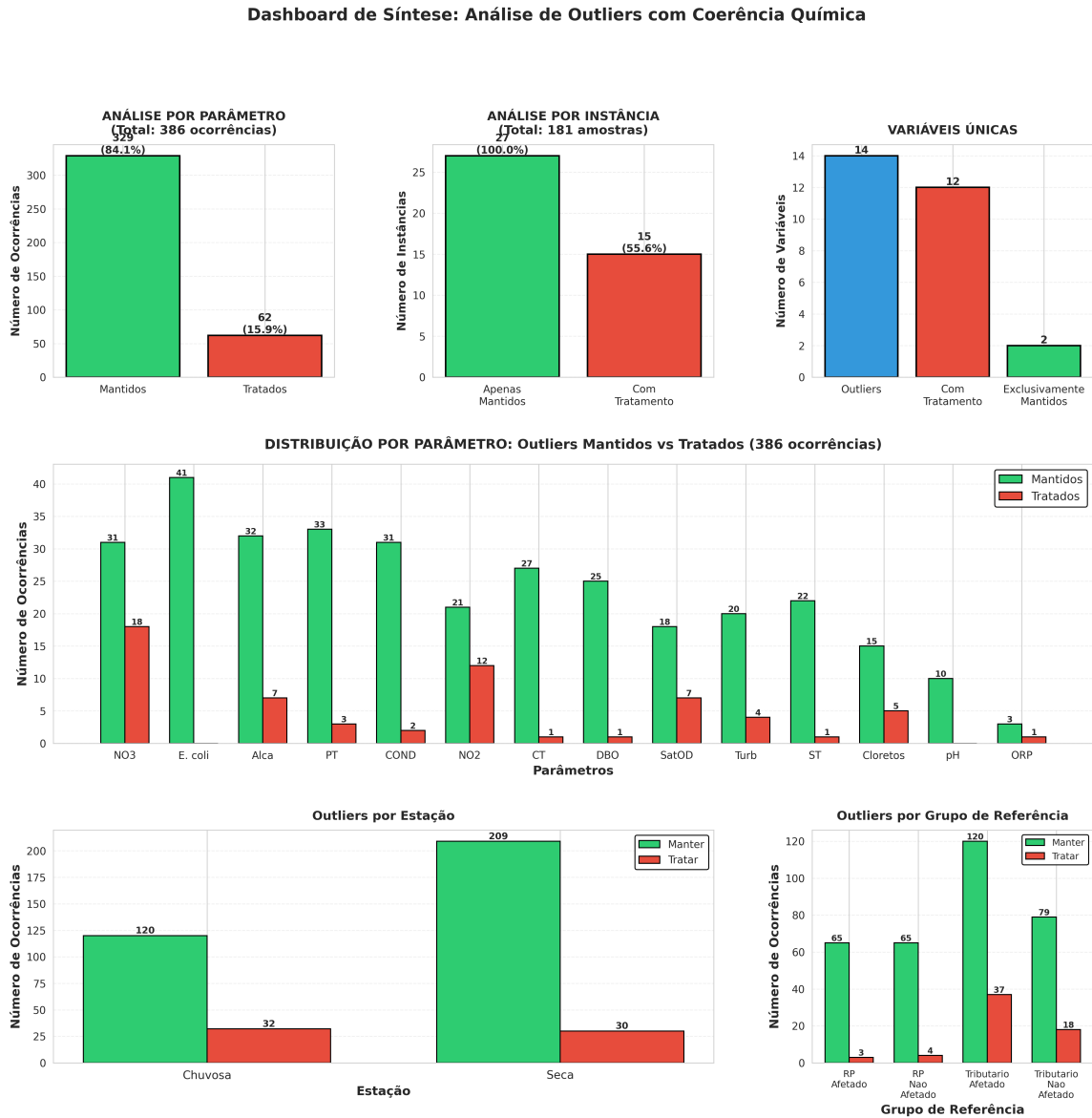


Figura 3 – Painel analítico do tratamento de *outliers* com validação físico-química.

Fonte: Elaboração própria (2025).

Nota: (Superior) Síntese quantitativa em três perspectivas: por ocorrência (386 pares amostra-variável), por instância (181 amostras únicas) e por variável (15 com detecção, 12 com tratamento, 2 exclusivamente mantidas). (Centro) Distribuição de ocorrências mantidas vs. tratadas por parâmetro. (Inferior) Distribuição por estação climática e grupo de referência espacial.

A análise por parâmetro (Figura 3, centro) revelou padrões informativos. Os nutrientes nitrogenados — NO_3 (≈ 18 tratamentos) e Nitrito (NO_2) (≈ 12 tratamentos) — concentraram a maior incidência de correções, refletindo sua dinâmica biogeoquímica complexa e sensibilidade a oscilações temporais rápidas. Em contraste, *E. coli* apresentou o maior número absoluto de ocorrências (≈ 41), porém nenhuma demandou tratamento: seus valores extremos foram sistematicamente validados por coerência com outras variáveis

correlacionadas.

A distribuição por estação climática (Figura 3, inferior esquerdo) evidenciou maior número absoluto de ocorrências na estação seca (≈ 209 mantidos e 30 tratados) comparada à chuvosa (≈ 120 mantidos e 32 tratados), ainda que com proporções semelhantes de tratamento. Por grupo de referência (Figura 3, inferior direito), o grupo *Tributário_Afetado* concentrou o maior número de ocorrências (≈ 120), refletindo a elevada heterogeneidade ambiental de tributários impactados pelo desastre. Os grupos do rio principal (*RP_Afetado* e *RP_Não_Afetado*) apresentaram menor incidência de tratamentos, sugerindo maior estabilidade físico-química no canal principal.

Casos representativos ilustram a lógica do módulo. A *Alca.* de 7.832 mg,L^{-1} registrada no ponto PC21 (estação seca) foi mantida por refletir um padrão minerário plausível: valores muito elevados de alcalinidade podem ocorrer em trechos afetados pela dissolução de carbonatos e pela remobilização de finos mineralizados, especialmente sob condições de baixa renovação hídrica. A presença simultânea de matéria orgânica elevada (*E. coli* e coliformes totais) indica um episódio ambiental complexo, mas não constitui o motivo da validação; o parâmetro foi mantido por coerência geoquímica.

A *Turb.* de 888 NTU no ponto PC16 foi preservada por correlacionar com *PT* de $0,51 \text{ mg,L}^{-1}$, padrão típico de arraste mineral associado a eventos pluviométricos intensos. Já a condutividade de $4,64 \text{ }\mu\text{S,cm}^{-1}$ no ponto PC14 foi mantida por refletir baixa mineralização compatível com tributários pequenos e isolados.

A abordagem alinha-se às recomendações de Li et al. (2023), assegurando que os modelos subsequentes aprendam padrões fisicamente plausíveis.

4.1.3 Matriz de Correlação e Relações Bivariadas

A análise de correlação de Pearson revelou estruturas de dependência linear consistentes com processos físico-químicos reconhecidos na literatura. A Figura 4 apresenta o *heatmap* da matriz de correlação, evidenciando padrões de co-variação que refletem tanto a organização temática das variáveis quanto a dinâmica hidrogeoquímica da bacia.

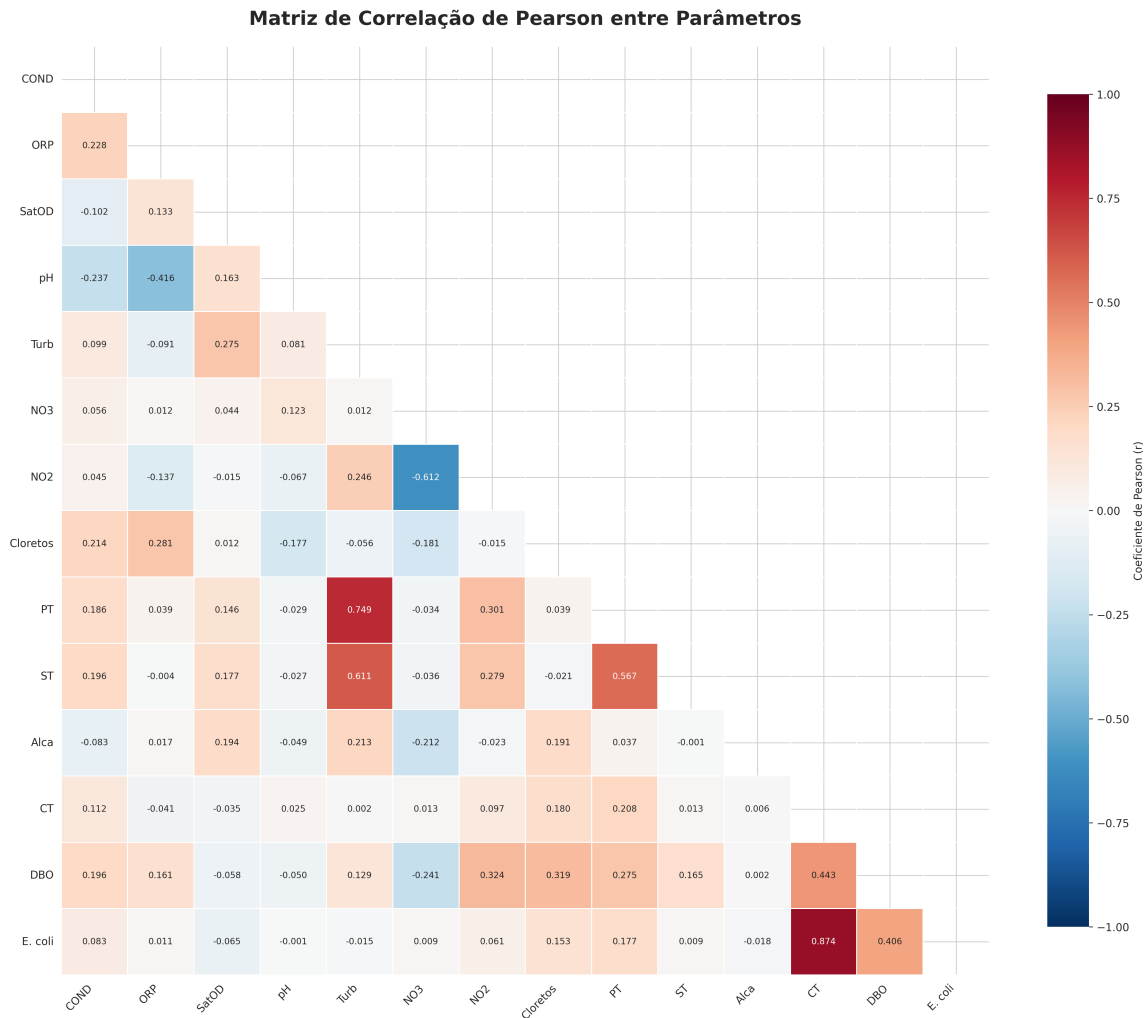


Figura 4 – Matriz de correlação de Pearson entre variáveis físico-químicas e microbiológicas.
Fonte: Elaboração própria (2025).

Nota: Destacam-se as correlações fortes entre **Turbidez** e **Fósforo Total** ($r = +0,75$), entre **Sólidos Totais** e **Turbidez** ($r = +0,61$) e a relação inversa entre **Nitrato** e **Nitrito** ($r = -0,61$), refletindo processos de arraste superficial, transporte conjunto de sedimentos e ciclagem de nitrogênio, respectivamente. A correlação muito elevada entre coliformes totais e *Escherichia coli* ($r = +0,87$) confirma a coerência estrutural do bloco sanitário. Apenas a diagonal inferior é apresentada para evitar redundância.

A **Turbidez** apresentou correlação positiva elevada com **Fósforo Total** ($r = +0,75$) e com **Sólidos Totais** ($r = +0,61$), conforme esperado dada a relação direta entre material particulado, sólidos totais e transporte de fósforo adsorvido. Esse padrão é mais pronunciado nos trechos afetados, onde a ressuspensão de sedimentos finos permanece ativa. A dinâmica entre **Nitrato** e **Nitrito**, por sua vez, exibiu correlação negativa marcada ($r = -0,61$), coerente com interconversões redox típicas de ambientes aquáticos sujeitos a variações de oxigenação.

Além dessas relações, o **Fósforo Total** merece destaque por sua dupla natureza. Embora esteja alocado no bloco minerário por acompanhar de forma consistente o comportamento dos sólidos suspensos e da turbidez — especialmente nas áreas afetadas —, sua dinâmica também reflete processos biogeoquímicos de adsorção e dessorção em partículas finas. A forte associação com **Turbidez** ($r = +0,75$) reforça que grande parte do fósforo mobilizado no sistema está vinculada ao arraste de material particulado, compatível com o caráter predominantemente físico do impacto dos rejeitos. Ao mesmo tempo, sua variabilidade sazonal indica que o **Fósforo Total** também responde a processos internos da coluna d'água, justificando sua classificação como variável de transição entre domínios.

A condutividade e a **Alcalinidade** apresentaram correlação fraca, sugerindo que processos distintos controlam a mineralização iônica e a capacidade tampão do sistema. Parâmetros associados à matéria orgânica e ao estado redox (**DBO** e **ORP**) exibiram correlações modestas, refletindo a complexidade dos gradientes biogeoquímicos presentes.

As quatro relações bivariadas mais fortes são ilustradas na Figura 5: (i) *Escherichia coli* versus coliformes totais ($r = +0,87$), (ii) **Fósforo Total** versus **Turbidez** ($r = +0,75$), (iii) **Nitrito** versus **Nitrato** ($r = -0,61$) e (iv) **Sólidos Totais** versus **Turbidez** ($r = +0,61$). Esses padrões reforçam a existência de três domínios temáticos relativamente coesos, conforme discutido na metodologia: minerário, sanitário e biogeoquímico.

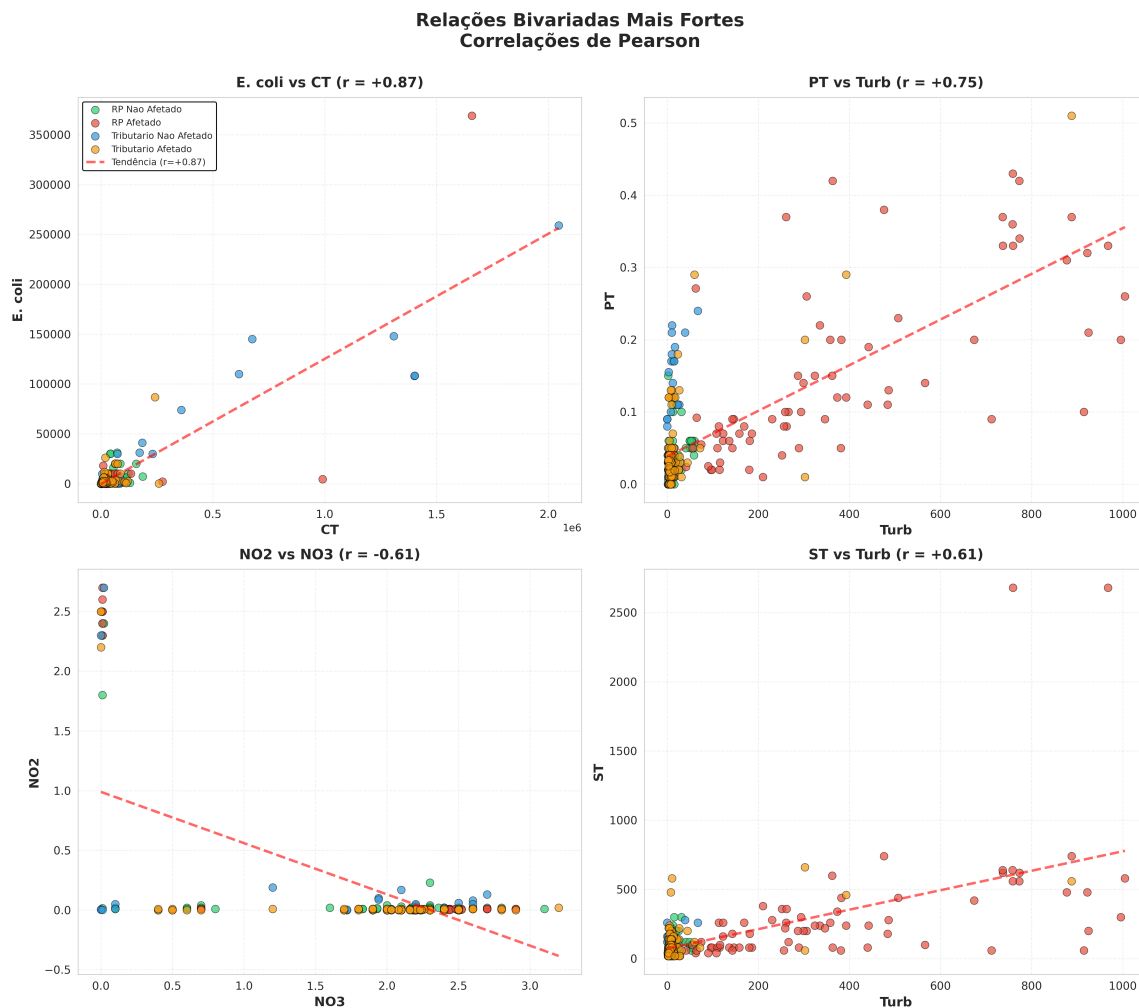


Figura 5 – Relações bivariadas mais fortes identificadas pela análise de correlação. Fonte: Elaboração própria (2025).

Os indicadores microbiológicos (*E. coli* e coliformes totais) apresentaram, em geral, correlação baixa com as demais variáveis físico-químicas, embora exibam correlação extremamente elevada entre si ($r = +0,87$). Esse resultado confirma a robustez interna do bloco sanitário e indica que sua variabilidade está mais associada a fontes difusas de contaminação orgânica distribuídas na bacia do que ao impacto minerário propriamente dito, em concordância com a interpretação de Santos (2018).

4.1.4 Gabaritos de Referência para Validação (SANTOS, 2018)

A incorporação de gabaritos de classificação baseados no estudo de Santos (2018) constitui estratégia metodológica essencial para validar os modelos de aprendizado de máquina desenvolvidos neste trabalho. Esses gabaritos, derivados de esquemas de Análise Discriminante Múltipla aplicados à mesma base de dados, estabelecem classes pré-definidas que refletem padrões hidroquímicos associados ao impacto do rompimento da Barragem de Fundão. Ao adotar essas classificações como referência, assegura-se comparabilidade

histórica com o trabalho original, permitindo avaliar se as técnicas modernas de inteligência artificial ([Random Forest \(RF\)](#), [Gradient Boosting Machines](#) e [Multi-Layer Perceptron](#)) superam, igualam ou apresentam desempenho inferior à Análise Discriminante clássica.

Diferentemente de [Santos \(2018\)](#), que desenvolveu modelos estatísticos específicos (funções discriminantes lineares) para cada esquema de classificação, o presente estudo utiliza exclusivamente os **rótulos de classe** (gabaritos) como variáveis-alvo para treinar algoritmos de aprendizado supervisionado. Os modelos preditivos propriamente ditos — incluindo suas variáveis discriminantes, coeficientes e fronteiras de decisão — são construídos de forma independente, utilizando técnicas computacionais que não replicam a Análise Discriminante. A comparação quantitativa entre as abordagens ocorrerá posteriormente, na etapa final da análise supervisionada (Seção 5.3), mediante confronto de métricas de desempenho (acurácia, precisão, recall, F1-Score e [ROC-AUC](#)).

Essa estratégia metodológica preserva a validade externa dos resultados, uma vez que os gabaritos representam uma taxonomia ambientalmente fundamentada, validada em campo e documentada na literatura especializada. Ao mesmo tempo, permite investigar se arquiteturas computacionais mais flexíveis — capazes de capturar não linearidades, interações complexas entre variáveis e padrões latentes — oferecem ganhos preditivos em relação a modelos lineares tradicionais.

Classificação Binária (Afetado vs Não-Afetado)

O primeiro esquema de classificação, denominado Modelo 1 por [Santos \(2018\)](#), distingue corpos hídricos em duas classes com base exclusivamente na localização espacial dos pontos de amostragem: (i) **rios não afetados**, correspondentes aos pontos [PC1](#) a [PC10](#), situados a montante da área de influência direta do rompimento da barragem, e (ii) **rios afetados**, correspondentes aos pontos [PC11](#) a [PC27](#), localizados a jusante e diretamente impactados pela lama de rejeitos minerários.

Essa divisão dicotômica privilegia o critério geográfico, assumindo que a distância em relação ao ponto de ruptura constitui o fator dominante de diferenciação hidroquímica. A classificação ignora, por design, distinções entre rio principal e tributários, agrupando todos os pontos a montante (independentemente de características hidrológicas locais) em uma única classe, e todos os pontos a jusante em outra. Essa simplificação conceitual, embora operacionalmente conveniente para análises iniciais, apresenta limitações interpretativas, conforme reconhecido por [Santos \(2018\)](#), que subsequentemente desenvolveu esquemas mais granulares (Modelos 2 e 3) para capturar nuances hidrogeoquímicas ignoradas pelo modelo binário.

O gabarito binário resultante distribui as 324 amostras válidas em:

- **Classe 1 — Não Afetado (PC1-10):** 120 amostras (60 em rio principal + 60 em tributários)
- **Classe 2 — Afetado (PC11-27):** 204 amostras (72 em rio principal + 132 em tributários)

No estudo original, Santos (2018) reportou acurácia de 93% no conjunto de desenvolvimento e 88% na validação utilizando Análise Discriminante, com Turbidez e saturação de oxigênio dissolvido como variáveis de maior poder discriminatório (Wilks' Lambda de 0,819 e 0,630, respectivamente). A função discriminante desenvolvida incorporou sete variáveis (condutividade, saturação de oxigênio, Turbidez, temperatura, Nitrito, cloreto e Alcalinidade), estabelecendo escore de corte em $Z = 0,0011$.

No presente trabalho, esse gabarito será utilizado para treinar e avaliar os modelos de aprendizado de máquina, sem replicar a função discriminante original. A comparação de desempenho verificará se os algoritmos propostos (RF, GBM, MLP) alcançam acurácia semelhante ou superior aos 88-93% reportados, utilizando métricas adicionais (precisão por classe, recall, F1-Score, ROC-AUC) que oferecem visão mais abrangente da capacidade preditiva.

Classificação Multiclasse (Quatro Categorias Hidrológicas)

O segundo esquema de classificação, denominado Modelo 2 por Santos (2018), refina a taxonomia ao distinguir quatro classes hidrológicas, reconhecendo que rio principal e tributários apresentam dinâmicas hidroquímicas distintas mesmo quando submetidos ao mesmo impacto regional:

- **Classe 1 — Rio Principal Não Afetado:** PC1, PC3, PC5, PC8, PC10 (60 amostras)
- **Classe 2 — Tributários Não Afetados:** PC2, PC4, PC6, PC7, PC9 (60 amostras)
- **Classe 3 — Rio Principal Afetado:** PC11, PC14, PC18, PC21, PC25, PC27 (72 amostras)
- **Classe 4 — Tributários Afetados:** PC12, PC13, PC15, PC16, PC17, PC19, PC20, PC22, PC23, PC24, PC26 (132 amostras)

Esse arranjo reconhece que tributários, caracterizados por menor vazão, menor largura de canal e tempos de residência distintos, respondem diferentemente ao aporte de contaminantes em comparação ao rio principal. A separação em quatro classes possibilita investigar se os modelos preditivos conseguem capturar não apenas o gradiente de impacto

(afetado vs não afetado), mas também as diferenças geomorfológicas e hidrodinâmicas entre compartimentos fluviais.

No estudo original, Santos (2018) reportou acurácia de 71,76% no desenvolvimento e 64,71% na validação para esse esquema, desempenho inferior ao modelo binário. Análises revelaram sobreposições significativas entre as classes 3 e 4 (rio principal afetado e tributários afetados), indicando que a assinatura físico-química dos rejeitos se sobrepõe ao sinal hidrológico intrínseco, dificultando a discriminação. As variáveis com maior poder discriminatório foram Turbidez (Wilks' Lambda = 0,286), saturação de oxigênio dissolvido, Alcalinidade, condutividade, cloreto, temperatura e *E. coli*.

A limitação desse gabarito reside na confusão entre classes com assinaturas parcialmente sobrepostas. As médias descritivas evidenciam essa dificuldade: o rio principal afetado apresentou mediana de Turbidez de 299 NTU, enquanto os tributários afetados registraram mediana de apenas 9 NTU, mas com outliers alcançando 888 NTU durante eventos episódicos de ressuspensão. Essa variabilidade temporal dificulta a classificação determinística, motivando a investigação de técnicas não lineares que possam capturar padrões mais complexos.

Síntese dos Gabaritos e Estratégia de Uso

Os dois gabaritos descritos — binário (2 classes) e multiclasse (4 classes) — estabelecem pontos de comparação com diferentes níveis de granularidade. O gabarito binário oferece classificação robusta mas simplificada, adequada para questões gerais sobre presença/ausência de impacto. O gabarito multiclasse captura nuances hidrológicas, mas enfrenta desafios de sobreposição entre classes parcialmente similares.

Além desses gabaritos de referência, o presente estudo desenvolveu esquemas alternativos de classificação baseados nos agrupamentos identificados pela análise não supervisionada (HDBSCAN), resultando em três cenários distintos: Cenário 1 (3 classes), Cenário 2 (4 classes) e Cenário 3 (6 classes), conforme descrito na Seção 4.3 (Tabela 3). Esses cenários incorporam informações sobre zonas de transição e condições mistas identificadas pela clusterização, representando contribuição metodológica do presente trabalho.

A estratégia de comparação adotada consiste em:

1. Treinar modelos de aprendizado de máquina (RF, GBM, MLP) utilizando os gabaritos de Santos (2018) como variáveis-alvo;
2. Avaliar o desempenho dos modelos mediante métricas quantitativas (acurácia, precisão, recall, F1-Score macro e por classe, ROC-AUC);

3. Confrontar os resultados obtidos com as acurácias reportadas por Santos (2018) para a Análise Discriminante (93% e 88% para o modelo binário; 71,76% e 64,71% para o modelo multiclasse);
4. Investigar, por meio de técnicas de Inteligência Artificial Explicável (SHAP e LIME), se as variáveis identificadas como discriminantes pela Análise Discriminante (Turbidez, saturação de oxigênio, condutividade, Alcalinidade) coincidem com as variáveis de maior importância nos modelos de aprendizado de máquina;
5. Avaliar a coerência ambiental das previsões, verificando se os modelos reproduzem padrões hidroquímicos fisicamente plausíveis.

Essa abordagem assegura, simultaneamente, comparabilidade histórica com o trabalho de Santos (2018) e avanço metodológico ao integrar técnicas modernas de aprendizado de máquina e interpretabilidade. A preservação dos gabaritos como referência comum garante que as comparações de desempenho sejam justas e que eventuais ganhos preditivos sejam atribuíveis às características intrínsecas dos algoritmos, e não a diferenças na definição das classes-alvo.

4.2 Resultados da Modelagem Não Supervisionada

A análise não supervisionada dos dados de qualidade da água da Bacia do Rio Gualaxo do Norte foi estruturada em duas fases complementares. A Fase 1 avaliou separadamente três blocos temáticos — minerário, sanitário e biogeoquímico — para estimar a capacidade individual de cada domínio em revelar padrões espaciais coerentes. A Fase 2 integrou esses blocos em configurações multibloco, permitindo avaliar se a combinação de variáveis de diferentes naturezas melhora a qualidade e a estabilidade dos agrupamentos identificados. Os resultados da Fase 1 são apresentados nas subseções seguintes, iniciando pelo Bloco Minerário devido à centralidade das variáveis associadas ao impacto de rejeitos no contexto do desastre de Mariana.

4.2.1 Análise dos Blocos Temáticos (Fase 1)

Bloco Minerário

O Bloco Minerário, composto por cinco variáveis (Turbidez, Sólidos Totais, Condutividade Elétrica, Alcalinidade e Fósforo Total), apresentou estrutura de variância altamente concentrada. A Análise de Componentes Principais revelou que apenas dois componentes retêm 99,58% da variabilidade original: o primeiro componente (PC1) explicou sozinho 97,46% da variância, enquanto o segundo (PC2) contribuiu com 2,12% adicionais (Tabela 10).

Tabela 10 – Variância explicada pelos componentes principais (Bloco Minerário).

Comp.	Var. Expl. (%)	Var. Acum. (%)	Autovalor
PC1	97,46	97,46	846,74
PC2	2,12	99,58	18,45
PC3	0,26	99,84	2,27
PC4	0,09	99,93	0,78
PC5	0,07	100,00	0,59

Fonte: Elaboração própria (2025).

Nota: Comp. = Componente; Var. Expl. = Variância Explicada; Var. Acum. = Variância Acumulada. Os autovalores representam a quantidade absoluta de variância capturada por cada componente (λ_i), sendo utilizados no cálculo da variância explicada ($\lambda_i / \sum \lambda_j$). PC1 concentra 97,46% da variância total, evidenciando dominância de processos de dissolução química associados aos rejeitos minerários. A magnitude do autovalor de PC1 (846,74) é 46 vezes superior ao de PC2 (18,45), confirmando a supremacia de um único processo estruturador (alcalinidade).

Este padrão indica forte colinearidade entre as variáveis minerárias, com um único eixo dominando a estrutura dos dados. A análise dos *loadings* permitiu interpretar o significado físico-químico dos componentes retidos (Figura 6). O PC1 é essencialmente um “eixo de Alcalinidade”, com a variável *Alca.* apresentando loading de 0,9995, enquanto as demais variáveis contribuem marginalmente (coeficientes $< 0,04$). Este resultado reflete a dominância de processos de dissolução química associados aos óxidos metálicos presentes nos rejeitos de mineração. O PC2, por sua vez, configura-se como um “eixo de material particulado”, com a *Turbidez* apresentando loading de 0,91, seguida por *Sólidos Totais* (0,34) e *Fósforo Total* (0,23). A ortogonalidade entre PC1 e PC2 sugere que os processos de mineralização e de transporte físico de sedimentos atuam de forma relativamente independente na estruturação da qualidade da água pós-desastre.

A seleção dos dois primeiros componentes principais seguiu o critério de Kaiser (autovalor > 1) combinado com o critério de variância acumulada ($> 90\%$), garantindo parcimônia dimensional sem perda significativa de informação. Componentes com autovalores inferiores a 1 (PC3 a PC5) foram descartados por representarem variabilidade residual inferior à de uma única variável original padronizada.

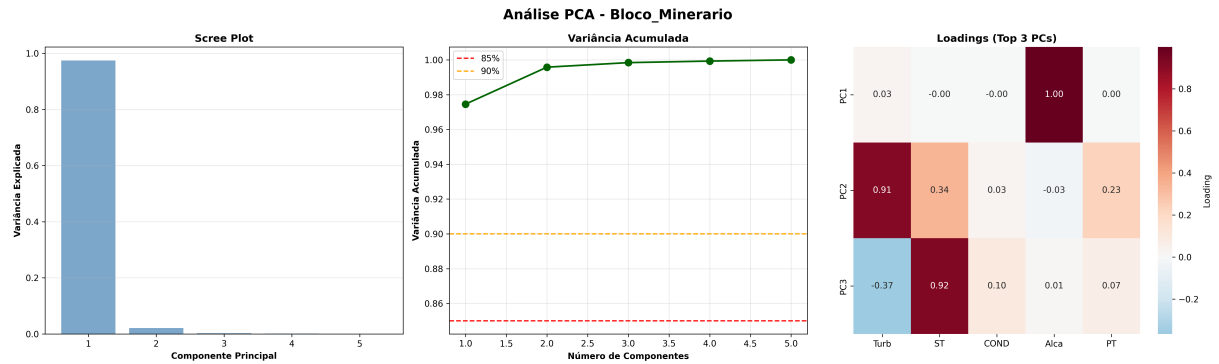


Figura 6 – Análise completa do PCA — Bloco Minerário.

Fonte: Elaboração própria (2025).

Nota: (Esquerda) **Scree plot** evidenciando a dominância do PC1 (97,46% da variância), com decaimento abrupto até PC2 e platô subsequente, caracterizando o “cotovelo” típico da técnica. (Centro) **Variância acumulada** ultrapassando 99% com apenas 2 componentes, justificando a redução dimensional de 5 para 2 variáveis. (Direita) **Heatmap de loadings** mostrando que a **Alcalinidade** (Alca.) domina quase integralmente o PC1 (loading = 0,9995), enquanto a **Turbidez** estrutura o PC2 (loading = 0,91).

As projeções bidimensionais **UMAP** e **t-SNE** dos componentes principais revelaram distribuições espaciais distintas, conforme ilustrado na Figura 7. O **UMAP** preservou uma estrutura contínua e alongada, com gradientes suaves entre regiões de alta e baixa **Alcalinidade**, enquanto o **t-SNE** produziu agrupamentos mais compactos e visualmente separados. Essas diferenças refletem as propriedades intrínsecas de cada técnica: o **UMAP** prioriza a preservação da topologia global do espaço de dados (MCINNES LELAND; MELVILLE, 2018), resultando em transições progressivas entre grupos de amostras, ao passo que o **t-SNE** enfatiza estruturas locais, ampliando as distâncias entre agrupamentos (MAATEN; HINTON, 2008).

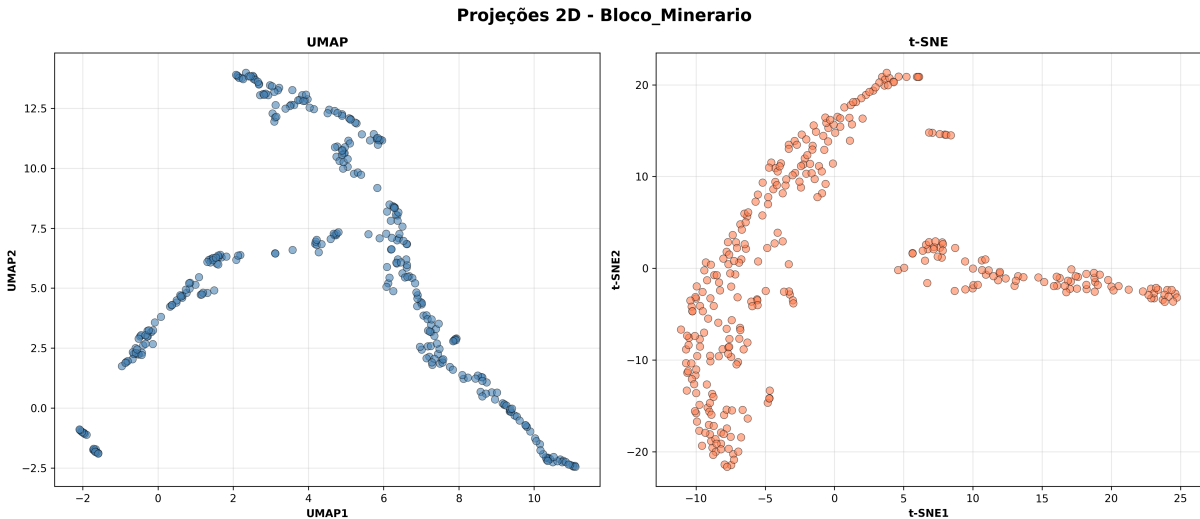


Figura 7 – Projeções bidimensionais UMAP e t-SNE.
Fonte: Elaboração própria (2025).

Nota: (Esquerda) **UMAP** preservou uma estrutura contínua e alongada, para 2 variáveis. Direita) **TSNE** produziu agrupamentos mais compactos e visualmente separados.

A clusterização aplicada sobre as projeções produziu resultados substancialmente diferentes entre os algoritmos **K-Means** e **HDBSCAN**. O **K-Means** identificou 8 clusters na projeção **UMAP** e 3 clusters no **t-SNE**, atribuindo todas as amostras a algum grupo. O **HDBSCAN**, por outro lado, detectou 9 clusters no **UMAP** e 4 clusters no **t-SNE**, mas classificou como ruído 28,1% e 34,3% das amostras, respectivamente (Tabela 11). Esta proporção expressiva de ruído indica que parcela significativa das amostras não apresenta características suficientemente homogêneas para integrar agrupamentos bem definidos, possivelmente refletindo heterogeneidade espacial ou temporal não capturada pelas variáveis minerárias.

Tabela 11 – Métricas internas de qualidade dos agrupamentos (Bloco Minerário).

Algoritmo	Proj.	<i>k</i>	<i>n</i>	Ruído	Silh.	DB	CH
K-Means	UMAP	8	324	0,0%	0,612	0,456	1412,90
HDBSCAN	UMAP	9	324	28,1%	0,721	0,342	2346,51
K-Means	t-SNE	3	324	0,0%	0,601	0,537	617,08
HDBSCAN	t-SNE	4	324	34,3%	0,684	0,429	612,04

Fonte: Elaboração própria (2025).

Abreviações: Proj. = Projeção; *k* = número de clusters; *n* = total de amostras; Ruído = proporção de amostras não atribuídas (**HDBSCAN**); Silh. = Silhouette Score; DB = Davies-Bouldin Index; CH = Calinski-Harabasz Index.

Interpretação: Linha destacada indica melhor configuração (**HDBSCAN + UMAP**). Silhouette > 0,7 = clusters bem separados; DB < 0,5 = baixa sobreposição; CH alto = clusters compactos e distantes.

As métricas internas de qualidade confirmaram a superioridade do **HDBSCAN** sobre projeções **UMAP**. Esta combinação alcançou coeficiente de silhueta (**ROUSSEEUW, 1987**) de 0,721 e índice de Davies-Bouldin (**DAVIES; BOULDIN, 1979**) de 0,342, indicando clusters bem separados e internamente coesos **8**. O coeficiente de silhueta, que “mede quão similar um objeto é ao seu próprio cluster comparado aos demais clusters” (**ROUSSEEUW, 1987**), varia de -1 a 1, onde valores próximos a 1 indicam alta qualidade de agrupamento. O valor de 0,721 obtido situa-se na faixa considerada excelente ($> 0,70$), evidenciando a robustez da estrutura identificada.

O índice de Calinski-Harabasz (**CALINSKI; HARABASZ, 1974**), também conhecido como critério da razão de variância, favorece clusters compactos e bem separados ao “maximizar a razão entre a dispersão inter-clusters e a dispersão intra-clusters” (**CALINSKI; HARABASZ, 1974**). Este índice confirmou o **HDBSCAN+UMAP** como melhor configuração (2346,51 contra 1412,90 do **K-Means+UMAP**), indicando que os clusters identificados pelo **HDBSCAN** apresentam maior separabilidade e coesão interna simultâneas.

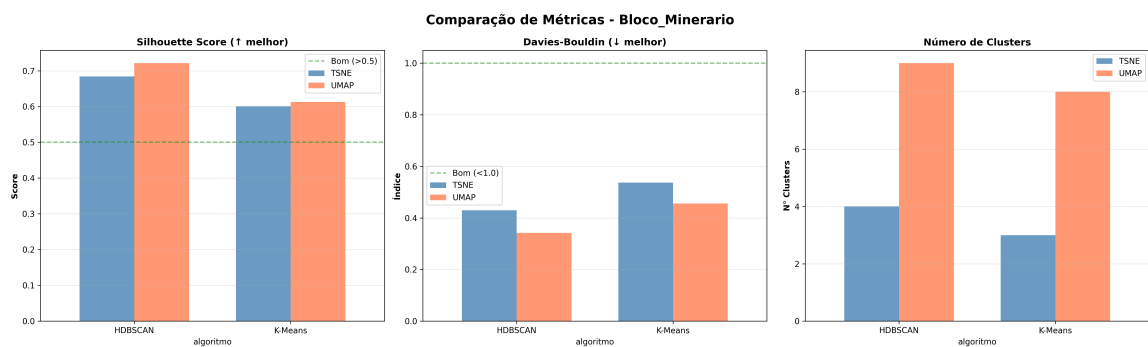


Figura 8 – Comparação das métricas internas de qualidade entre as quatro configurações testadas no Bloco Minerário.

Fonte: Elaborada pelo autor.

Nota: (a) Coeficiente de Silhouette (maior = melhor); (b) Índice de Davies-Bouldin (menor = melhor); (c) Índice de Calinski-Harabasz (maior = melhor). A combinação **HDBSCAN + UMAP** apresenta desempenho superior consistente nas três métricas.

A validação externa confrontou os clusters identificados com os gabaritos binário (impactado vs. não-impactado) e multiclasse de **Santos (2018)**. Os resultados revelaram concordância fraca a moderada em todas as configurações testadas (Tabela 12). O melhor desempenho foi obtido pelo **HDBSCAN+t-SNE** confrontado com o gabarito multiclasse: **ARI** de 0,373 e **NMI** de 0,371, caracterizando concordância moderada. As demais configurações apresentaram **ARI** entre 0,026 e 0,297, indicando concordância fraca. O confronto com o gabarito binário resultou em desempenho consistentemente inferior, com **ARI** máximo

de 0,158, sugerindo que a estrutura identificada não se alinha à classificação dicotômica proposta por Santos.

Tabela 12 – Validação externa dos agrupamentos: confronto com gabaritos de referência (Bloco Minerário).

Algoritmo	Proj.	Gabarito	ARI	NMI	Pureza	V-meas.	Interp.
K-Means	UMAP	Binário	0,026	0,084	0,698	0,084	Fraca
K-Means	UMAP	Multi	0,150	0,289	0,623	0,289	Fraca
HDBSCAN	UMAP	Binário	0,032	0,103	0,738	0,103	Fraca
HDBSCAN	UMAP	Multi	0,174	0,337	0,661	0,337	Fraca
K-Means	t-SNE	Binário	0,052	0,068	0,676	0,068	Fraca
K-Means	t-SNE	Multi	0,297	0,327	0,599	0,327	Fraca
HDBSCAN	t-SNE	Multi	0,373	0,371	0,685	0,371	Moderada

Fonte: Elaboração própria (2025).

Abreviações: Proj. = Projeção; ARI = Adjusted Rand Index; NMI = Normalized Mutual Information; V-meas. = V-measure; Interp. = Interpretação.

Interpretação: Linha destacada indica melhor configuração (HDBSCAN+t-SNE confrontado com gabarito multiclasse). Concordância moderada (ARI = 0,373) sugere que padrões hidroquímicos identificados capturam estruturas complementares à classificação espacial de Santos (2018). Concordância fraca com gabarito binário indica que estrutura identificada não se alinha à dicotomia simples Afetado/Não_Afetado.

A concordância moderada com o gabarito multiclasse (ARI = 0,373) sugere que os padrões identificados pela clusterização densitária capturam estruturas hidroquímicas complementares à classificação espacial de Santos (2018). Segundo a interpretação proposta por Hubert e Arabie (1985), valores de ARI entre 0,20 e 0,50 indicam concordância moderada, refletindo sobreposição parcial mas não idêntica entre as partições comparadas. Nesse contexto, a divergência não invalida os agrupamentos obtidos, mas evidencia que a análise não supervisionada detectou subestruturas latentes não contempladas na categorização espacial *a priori*, possivelmente relacionadas a gradientes hidroquímicos de transição entre zonas afetadas e não afetadas.

Bloco Sanitário

Enquanto o Bloco Minerário evidenciou estrutura dominada por processos de dissolução química e arraste de rejeitos, o Bloco Sanitário foi investigado para avaliar se indicadores de contaminação microbiológica e matéria orgânica de origem antrópica revelam padrões espaciais distintos na Bacia do Rio Gualaxo do Norte. Este bloco temático compreende três variáveis: *Escherichia coli*, coliformes totais (CT) e demanda bioquímica de oxigênio (DBO). A fundamentação para o agrupamento dessas variáveis reside no fato de que todas representam marcadores de poluição sanitária, associados à contaminação fecal e à presença de matéria orgânica biodegradável (WORLD HEALTH ORGANIZATION, 2017).

No contexto pós-desastre, a investigação desse bloco assume relevância pela potencial interação entre o rompimento da barragem e fontes difusas de poluição sanitária, agravadas pela desestabilização de sistemas de esgotamento e pelo aporte de sedimentos que podem carrear microrganismos patogênicos. Conforme descrito na Seção 3.4, o bloco foi submetido ao mesmo *pipeline* metodológico aplicado aos demais blocos temáticos: escalonamento robusto, análise de componentes principais (PCA), projeções topológicas (UMAP e t-SNE), clusterização (K-Means e HDBSCAN) e validação externa mediante confronto com os gabaritos de Santos (2018).

A Análise de Componentes Principais aplicada ao Bloco Sanitário revelou concentração extrema de informação no primeiro componente, comportamento análogo ao observado no Bloco Minerário. O PC1 reteve 94,68% da variância total, enquanto o PC2 contribuiu com apenas 4,64%, resultando em variância acumulada de 99,32% em dois componentes (Tabela 13). Essa distribuição assimétrica sugere que uma única dimensão latente captura a maior parte da variabilidade dos indicadores sanitários, indicando forte colinearidade entre *E. coli*, coliformes totais e DBO. A elevada concentração de variância em PC1 pode ser interpretada como reflexo da origem comum dessas variáveis — todas associadas à contaminação por efluentes domésticos ou agrícolas — e evidencia que flutuações nas três variáveis tendem a ocorrer de forma sincronizada nas amostras analisadas.

Tabela 13 – Variância explicada pelos componentes principais no Bloco Sanitário.

Comp.	Variância Explicada (%)	Variância Acumulada (%)	Autovalor
PC1	94,68	94,68	159,40
PC2	4,64	99,32	7,82
PC3	0,68	100,00	1,15

A interpretação dos *loadings* (Tabela 14) confirma a natureza do primeiro componente principal. *E. coli* e coliformes totais apresentaram cargas dominantes em PC1 (0,883 e 0,468, respectivamente), caracterizando esse eixo como representativo da contaminação fecal. DBO, por sua vez, exibiu carga reduzida em PC1 (0,040) e contribuição concentrada em PC3 (0,997), sugerindo que a demanda de oxigênio associada à degradação de matéria orgânica representa processo parcialmente independente da contaminação microbiológica direta. O segundo componente principal, embora explique pequena fração da variância, apresentou carga elevada para coliformes totais (0,881) e negativa para *E. coli* (-0,470), indicando possível diferenciação entre contaminação fecal recente (expressa por *E. coli*) e contaminação mais difusa ou histórica (representada por coliformes totais).

Tabela 14 – Loadings das variáveis sanitárias nos três primeiros componentes principais.

Variável	PC1	PC2	PC3
<i>E. coli</i>	0,883	-0,470	-0,006
Coliformes Totais (CT)	0,468	0,881	-0,074
DBO	0,040	0,063	0,997

Essa separação, contudo, possui baixo poder explicativo diante da predominância de PC1. A Figura 9 sintetiza graficamente a análise de componentes principais, incluindo *scree plot*, biplot e contribuições individuais das variáveis.

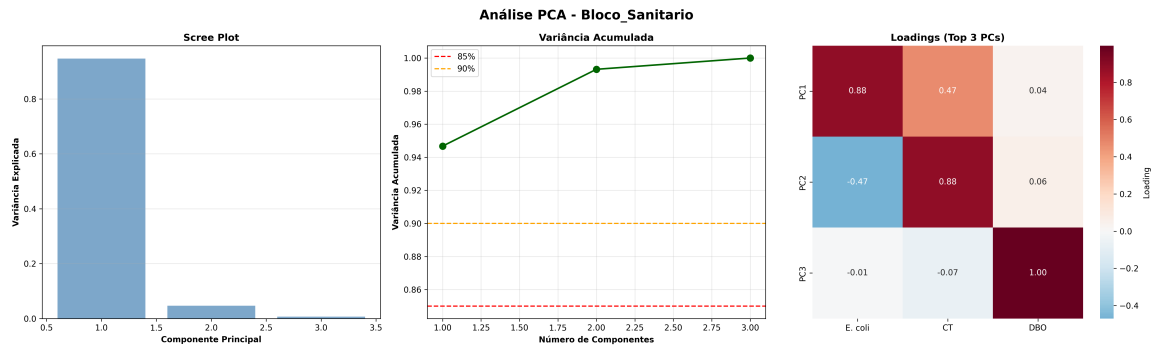


Figura 9 – Análise de Componentes Principais do Bloco Sanitário,

Nota: (a) *scree plot* evidenciando concentração de variância em PC1; (b) *biplot* mostrando a relação entre as variáveis e os dois primeiros componentes; (c) contribuição de cada variável aos componentes principais..

As projeções bidimensionais obtidas por UMAP e t-SNE (Figura 10) exibiram morfologias distintas. A projeção UMAP apresentou estrutura predominantemente contínua, com um aglomerado compacto no extremo superior esquerdo e uma dispersão gradual em direção ao quadrante inferior direito, sugerindo transição suave entre níveis de contaminação sanitária. Dois pontos isolados no limite superior indicam amostras com perfil microbiológico extremo, potencialmente associadas a fontes pontuais de poluição. A Projeção t-SNE, por sua vez, revelou distribuição mais difusa, sem formação de agrupamentos densos e bem delimitados. Essa configuração contrasta com o padrão observado no Bloco Minerário, onde t-SNE produziu clusters mais definidos. A ausência de estrutura clara em t-SNE sugere que os indicadores sanitários, embora colineares conforme evidenciado pela PCA, não induzem segregação espacial pronunciada quando representados em espaço de baixa dimensionalidade.

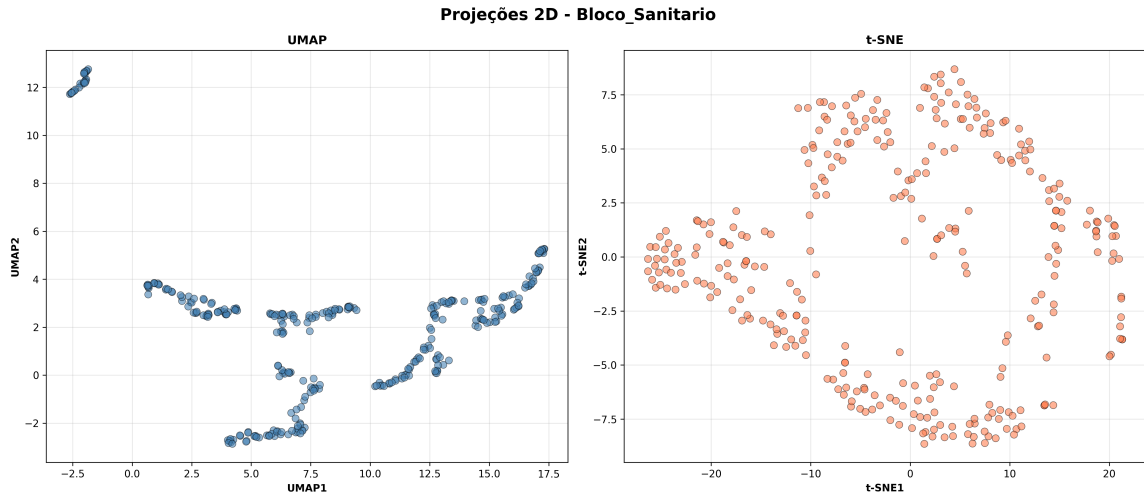


Figura 10 – Projeções bidimensionais do Bloco Sanitário.

Nota: (a) **UMAP** preservando a estrutura topológica global; (b) **t-SNE** enfatizando relações locais. A morfologia contínua da projeção **UMAP** contrasta com a dispersão difusa observada em **t-SNE**, indicando ausência de clusters naturalmente separados.

Os algoritmos de clusterização aplicados sobre as projeções produziram resultados que corroboram a limitada estruturabilidade do Bloco Sanitário. A configuração **K-Means/UMAP** gerou seis clusters e obteve o melhor desempenho global segundo a métrica Silhouette (0,611), embora com índice Davies-Bouldin moderado (0,500). **HDBSCAN/UMAP** identificou cinco clusters, mas classificou 36,42% das amostras como ruído, sinalizando que parcela substancial dos dados não se enquadra em agrupamentos densos. Apesar da alta proporção de ruído, essa configuração alcançou Silhouette superior (0,681), indicando que os clusters efetivamente identificados apresentam boa coesão interna. As combinações com **t-SNE** apresentaram desempenho inferior: **K-Means/t-SNE** produziu seis clusters com Silhouette de 0,514 e Davies-Bouldin de 0,635, enquanto **HDBSCAN/t-SNE** retornou apenas quatro clusters e classificou 44,14% das amostras como ruído, refletindo a estrutura difusa previamente observada na Projeção **t-SNE**. A Tabela 15 e a Figura 11 sintetizam as métricas de validação interna para todas as configurações avaliadas.

Tabela 15 – Métricas de validação interna (Bloco Sanitário).

Algoritmo	Proj.	k	Ruído	Silh.	DB	CH
K-Means	UMAP	6	0,0%	0,611	0,500	1372,13
HDBSCAN	UMAP	5	36,4%	0,681	0,437	1151,60
K-Means	t-SNE	6	0,0%	0,514	0,635	693,18
HDBSCAN	t-SNE	4	44,1%	0,620	0,464	450,94

Fonte: Elaboração própria (2025).

Nota: Linha destacada = melhor configuração (K-Means + UMAP). HDBSCAN apresenta Silhouette superior (0,681), mas classifica 36,4% das amostras como ruído.

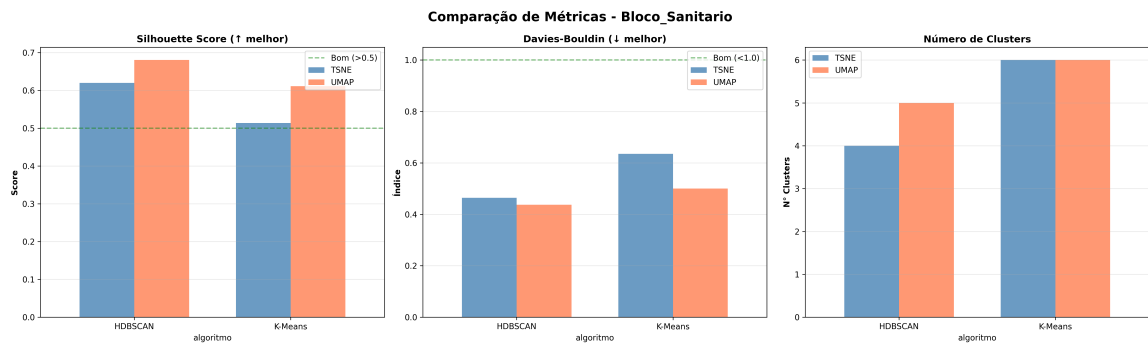


Figura 11 – Comparação visual das métricas de validação interna para o Bloco Sanitário:

Nota: (a) Silhouette Score (valores mais altos indicam melhor coesão); (b) Davies-Bouldin Index (valores mais baixos indicam melhor separação); (c) número de clusters identificados por cada configuração. Destaque para HDBSCAN/UMAP, que combina alta coesão com baixa sobreposição entre clusters.

A validação externa, realizada mediante confronto com os gabaritos binário e multiclasse de Santos (2018), revelou concordância fraca em todas as configurações avaliadas. Os valores de ARI (Adjusted Rand Index) variaram entre 0,007 e 0,081, enquanto os de NMI (Normalized Mutual Information) oscilaram entre 0,022 e 0,125 (Tabela 16). A melhor concordância foi obtida por HDBSCAN/t-SNE no gabarito multiclasse (ARI = 0,081; NMI = 0,125), configuração que, paradoxalmente, apresentou desempenho interno moderado e elevada proporção de ruído. K-Means/UMAP, líder em métricas internas, registrou concordância ligeiramente inferior (ARI = 0,053; NMI = 0,098 no gabarito multiclasse). Esses resultados indicam que, embora os algoritmos identifiquem padrões de coesão interna nos dados sanitários, tais padrões não se alinham à estrutura espacial definida por Santos (2018) com base em múltiplas variáveis físico-químicas e biológicas.

Tabela 16 – Validação externa: confronto com gabaritos de referência (Bloco Sanitário).

Algoritmo	Proj.	Gabarito	ARI	NMI
K-Means	UMAP	Binário	0,019	0,032
K-Means	UMAP	Multi	0,053	0,098
HDBSCAN	UMAP	Binário	0,007	0,022
HDBSCAN	UMAP	Multi	0,060	0,117
K-Means	t-SNE	Binário	0,019	0,029
K-Means	t-SNE	Multi	0,041	0,089
HDBSCAN	t-SNE	Multi	0,081	0,125

Fonte: Elaboração própria (2025), baseado em Santos (2018).

Nota: Concordância fraca em todas as configurações ($ARI < 0,10$). Melhor resultado: **HDBSCAN** + **t-SNE** ($ARI = 0,081$).

A fraca concordância entre os clusters derivados do Bloco Sanitário e os gabaritos espaciais de Santos (2018) pode ser atribuída a três fatores principais. Primeiro, indicadores microbiológicos e de matéria orgânica tendem a exibir elevada variabilidade temporal e espacial de curto alcance, influenciados por eventos episódicos como descargas pluviais, lançamentos pontuais de efluentes e flutuações sazonais na atividade microbiana. Essa volatilidade reduz a capacidade de tais variáveis em capturar padrões espaciais estáveis, especialmente quando comparadas a indicadores físico-químicos mais conservativos, como condutividade e **Alcalinidade**. Segundo, a elevada proporção de ruído identificada por **HDBSCAN** sugere que parcela substancial das amostras apresenta perfis sanitários intermediários ou mistos, não se enquadrando em agrupamentos bem definidos. Terceiro, os gabaritos de Santos (2018) foram construídos a partir de análises multivariadas que integraram simultaneamente variáveis minerárias, sanitárias e biogeoquímicas, de modo que a estrutura espacial resultante reflete a contribuição ponderada de todos esses domínios, e não apenas dos indicadores sanitários isolados. A limitada capacidade do Bloco Sanitário em reproduzir essa estrutura evidencia que, embora os indicadores microbiológicos sejam relevantes para a caracterização da qualidade da água, sua contribuição para a segregação espacial de longo alcance é secundária quando comparada aos processos físico-químicos associados ao arraste de rejeitos minerários.

Em síntese, o Bloco Sanitário revelou alta colinearidade entre *E. coli*, coliformes totais e **DBO**, com concentração de 94,68% da variância em um único componente principal. As projeções topológicas indicaram estrutura predominantemente contínua, sem formação de clusters naturalmente separados. As métricas internas apontaram **K-Means/UMAP** como a configuração mais equilibrada, embora **HDBSCAN/UMAP** tenha produzido clusters de maior coesão ao custo de classificar 36% das amostras como ruído. A validação externa revelou concordância fraca com os gabaritos de Santos (2018), com **ARI** máximo de

0,081, evidenciando que indicadores sanitários isolados possuem limitada capacidade de reproduzir a estrutura espacial da bacia. Esses resultados, em conjunto com os do Bloco Minerário, indicam que a análise integrada de múltiplos domínios temáticos — abordagem a ser explorada na Fase 2 — é essencial para capturar a complexidade dos padrões de qualidade da água em contextos pós-desastre. A próxima subseção apresenta os resultados do Bloco Biogeoquímico, completando a análise temática independente dos três blocos.

Bloco Biogeoquímico

Diferentemente dos blocos Minerário e Sanitário, que agrupam variáveis associadas a processos de origem predominantemente antrópica — arraste de rejeitos minerários e contaminação fecal, respectivamente —, o Bloco Biogeoquímico foi concebido para avaliar se indicadores de equilíbrio natural da coluna d'água revelam padrões espaciais coerentes na Bacia do Rio Gualaxo do Norte. Este bloco compreende cinco variáveis: pH, [Saturação de Oxigênio Dissolvido \(SatOD\)](#), [Nitrato \(\$\text{NO}_3\$ \)](#), [Potencial de Oxirredução \(ORP\)](#) e [Cloretos \(CL\)](#). Tais parâmetros refletem processos biogeoquímicos fundamentais, como oxigenação, dissolução mineral e reações redox, que regulam a capacidade de autodepuração dos corpos hídricos ([CASTILLO; M., 2021](#)). A investigação desse bloco permite avaliar se a estrutura espacial da bacia é condicionada por gradientes biogeoquímicos naturais ou se a dominância de processos antrópicos — identificada nos blocos anteriores — mascara esses padrões de fundo.

A Análise de Componentes Principais aplicada ao Bloco Biogeoquímico revelou estrutura de variância consideravelmente mais distribuída quando comparada aos blocos Minerário e Sanitário. Três componentes principais retêm 82,64% da variabilidade original: o primeiro componente ([PC1](#)) explica 46,81% da variância, o segundo ([PC2](#)) contribui com 22,32%, e o terceiro ([PC3](#)) adiciona 13,50% (Tabela [17](#)). Este padrão contrasta significativamente com a concentração extrema observada no Bloco Minerário, onde [PC1](#) sozinho capturou 97,46% da variância. A distribuição mais equilibrada da informação entre os três primeiros componentes sugere que os processos biogeoquímicos atuam de forma menos colinear, com múltiplos eixos de variação operando simultaneamente na estruturação da qualidade da água.

Tabela 17 – Variância explicada pelos componentes principais (Bloco Biogeoquímico).

Comp.	Var. Expl. (%)	Var. Acum. (%)	Autovalor
PC1	46,81	46,81	2,88
PC2	22,32	69,13	1,37
PC3	13,50	82,64	0,83
PC4	12,68	95,32	0,78
PC5	4,68	100,00	0,29

Fonte: Elaboração própria (2025).

Nota: Distribuição mais equilibrada da variância (contraste com blocos Minerário e Sanitário). Três componentes retêm 82,64% da variabilidade, indicando menor colinearidade entre processos biogeoquímicos.

A análise dos *loadings* permitiu interpretar o significado biogeoquímico dos componentes retidos (Figura 12). O PC1 configura-se essencialmente como um eixo de mineralização e oxigenação, dominado por Nitrato ($\text{NO}_3 = 0,95$) e, em menor grau, por pH (0,10). Este padrão indica que a variabilidade primária do bloco está associada aos processos de dissolução de nutrientes nitrogenados e às condições de neutralização da coluna d'água. O PC2 caracteriza-se como um eixo redox e de salinidade, com cloretos apresentando a maior contribuição (0,83), seguidos por potencial de oxidação-redução ($\text{ORP} = 0,41$) e Nitrato (0,30). A predominância dos cloretos neste componente reflete a influência de processos de dissolução mineral de fundo e, possivelmente, de aportes pontuais de salinidade. O PC3, por sua vez, constitui um eixo de oxigenação e acidificação, estruturado principalmente pela saturação de oxigênio dissolvido ($\text{SatOD} = 0,93$) e, secundariamente, por pH (0,36). A ortogonalidade entre os três primeiros componentes sugere que mineralização nitrogenada, salinização e oxigenação atuam de forma relativamente independente na estruturação biogeoquímica da bacia.

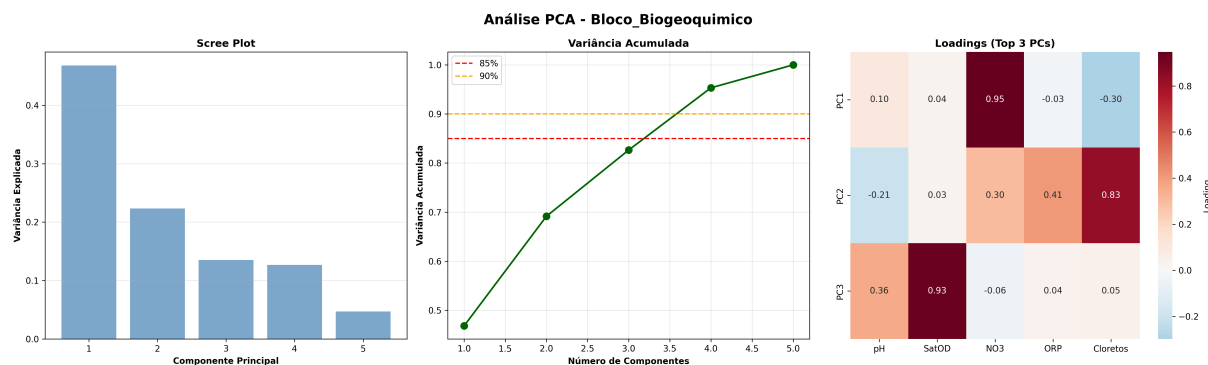


Figura 12 – Análise completa do PCA — Bloco Biogeoquímico

Fonte: Elaboração própria (2025).

Nota: (Esquerda) *Scree plot* evidenciando distribuição mais equilibrada da variância entre os três primeiros componentes (46,81%, 22,32% e 13,50%), contrastando com a dominância extrema observada no Bloco Minerário. (Centro) Variância acumulada atingindo 82,64% com três componentes. (Direita) *Heatmap* de *loadings* mostrando que o Nitrato (NO₃) domina o PC1 (*loading* = 0,95), os cloretos estruturam o PC2 (*loading* = 0,83), e a saturação de oxigênio dissolvido caracteriza o PC3 (*loading* = 0,93).

As projeções bidimensionais UMAP e t-SNE dos componentes principais biogeoquímicos revelaram distribuições espaciais marcadamente distintas das observadas nos blocos anteriores (Figura 13). O UMAP produziu uma estrutura caracterizada por um agrupamento compacto minoritário, nitidamente separado de uma distribuição difusa e contínua que abrange a maior parte das amostras. Este padrão sugere a existência de um subconjunto de amostras com perfil biogeoquímico atípico, possivelmente relacionado a áreas de baixa oxigenação ou elevada salinização, enquanto a maioria das observações apresenta gradientes suaves entre diferentes condições biogeoquímicas. O t-SNE, por sua vez, preservou uma estrutura de dispersão acentuada, com transições graduais entre regiões do espaço projetado, sem a formação de agrupamentos claramente delimitados. A ausência de separações nítidas nas projeções biogeoquímicas contrasta com os padrões de estratificação observados nos blocos Minerário e Sanitário, indicando que os processos naturais de equilíbrio da coluna d'água não impõem descontinuidades abruptas na estrutura espacial dos dados.

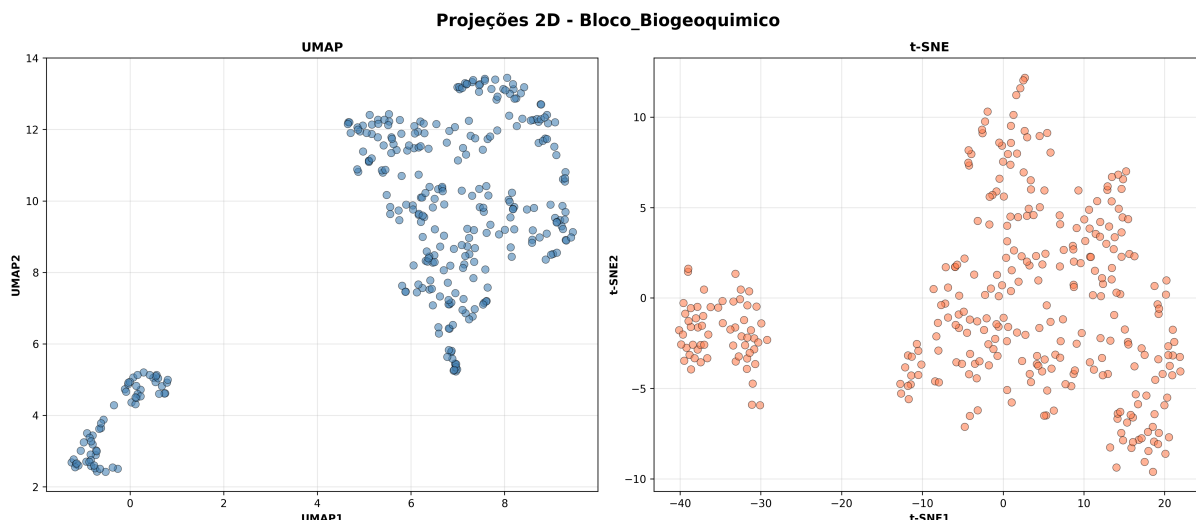


Figura 13 – Projeções bidimensionais UMAP e t-SNE — Bloco Biogeoquímico

Fonte: Elaboração própria (2025).

Nota: (Esquerda) Projeção UMAP evidenciando agrupamento compacto minoritário segregado de distribuição difusa majoritária. (Direita) Projeção t-SNE preservando estrutura de dispersão acentuada, sem formação de agrupamentos nitidamente delimitados. A ausência de estratificação clara contrasta com os padrões observados nos blocos Minerário e Sanitário.

A clusterização do Bloco Biogeoquímico produziu resultados consistentes entre algoritmos e projeções, convergindo para estruturas binárias (2 clusters) ou ternárias (3 clusters). O K-Means identificou dois agrupamentos tanto em projeções UMAP quanto t-SNE, alcançando coeficientes de silhueta de 0,710 e 0,706, respectivamente. O HDBSCAN, por sua vez, detectou três clusters sobre projeções UMAP, com silhueta de 0,725, porém classificou 53,7% das amostras como ruído — proporção substancialmente superior aos 28-34% registrados no Bloco Minerário. Nas projeções t-SNE, o HDBSCAN convergiu para estrutura binária idêntica à do K-Means (silhueta = 0,706), sem detecção de ruído (Tabela 18). As métricas internas indicam qualidade de agrupamento comparável entre as configurações testadas: os valores de Davies-Bouldin variaram entre 0,319 e 0,384, todos situados abaixo do limiar de 1,0 que caracteriza clusters bem separados, enquanto o índice de Calinski-Harabasz atingiu seu máximo no HDBSCAN+UMAP (1378,62), refletindo a maior separabilidade desta configuração apesar da elevada proporção de ruído.

Tabela 18 – Métricas internas de qualidade dos agrupamentos (Bloco Biogeoquímico).

Algoritmo	Proj.	<i>k</i>	<i>n</i>	Ruído	Silh.	DB	CH
K-Means	UMAP	2	324	0,0%	0,710	0,351	882,95
HDBSCAN	UMAP	3	324	53,7%	0,725	0,384	1378,62
K-Means	t-SNE	2	324	0,0%	0,706	0,319	912,01
HDBSCAN	t-SNE	2	324	0,0%	0,706	0,319	912,01

Fonte: Elaboração própria (2025).

Nota: Configuração **HDBSCAN** + **UMAP** apresenta melhor Silhouette (0,725) e maior separabilidade (CH = 1378,62), porém com elevada proporção de ruído (53,7%). Convergência para estruturas binárias ou ternárias em todas as configurações testadas.

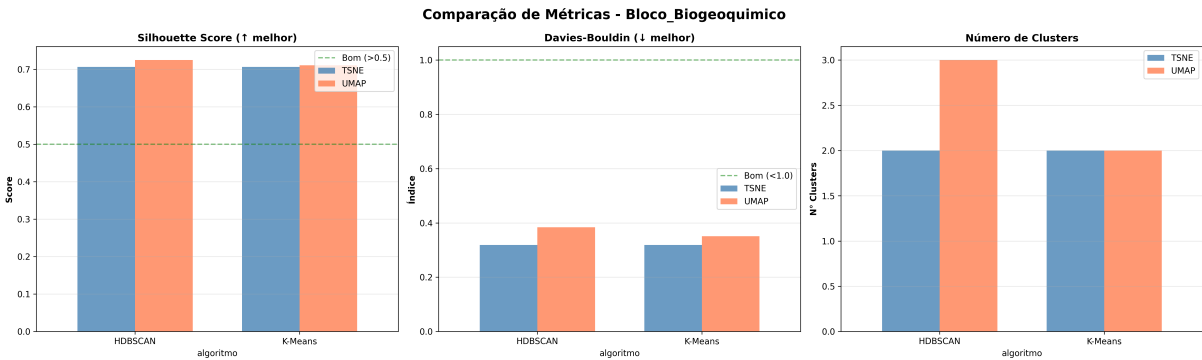


Figura 14 – Comparação das métricas internas de qualidade entre as quatro configurações testadas no Bloco Biogeoquímico

Fonte: Elaboração própria (2025).

Nota: (Esquerda) **Coefficiente de Silhouette** (maior = melhor); (Centro) Índice de **Davies-Bouldin** (menor = melhor); (Direita) **Número de clusters** identificados por cada configuração. O **HDBSCAN+UMAP** apresenta desempenho marginalmente superior nas métricas internas, com silhueta de 0,725 e identificação de 3 clusters, porém à custa de elevada proporção de ruído (53,7%).

A validação externa confrontou os agrupamentos do Bloco Biogeoquímico com os gabaritos binário e multiclasse de Santos (2018), revelando concordância consistentemente fraca em todas as configurações testadas (Tabela 19). O melhor desempenho foi obtido pelo **HDBSCAN+UMAP** confrontado com o gabarito binário: **ARI** de 0,055 e **NMI** de 0,126. As demais configurações apresentaram **ARI** entre 0,005 e 0,055, valores substancialmente inferiores aos registrados nos blocos Minerário (**ARI** máximo = 0,373) e Sanitário. O confronto com o gabarito multiclasse resultou em desempenho ainda mais modesto, com **ARI** máximo de 0,055 e pureza de apenas 48%. Os valores de homogeneidade (0,002 a 0,166) e completude (0,003 a 0,133) reforçam a ausência de alinhamento entre a estrutura identificada pelas variáveis biogeoquímicas e a classificação espacial proposta por Santos. Este resultado indica que os processos de oxigenação, dissolução mineral e reações redox — embora fundamentais para a dinâmica biogeoquímica da bacia — não capturam a estrutura de impacto espacial associada ao rompimento da barragem de Fundão.

Tabela 19 – Validação externa dos agrupamentos: confronto com gabaritos de referência (Bloco Biogeoquímico).

Algoritmo	Proj.	Gabarito	ARI	NMI	Pureza	Homog.	Compl.	V-meas.	Interp.
K-Means	UMAP	Binário	0,014	0,003	0,630	0,002	0,003	0,003	Fraca
K-Means	UMAP	Multi	0,005	0,002	0,407	0,001	0,003	0,002	Fraca
HDBSCAN	UMAP	Binário	0,059	0,126	0,660	0,166	0,102	0,126	Fraca
HDBSCAN	UMAP	Multi	0,055	0,121	0,480	0,111	0,133	0,121	Fraca
K-Means	t-SNE	Binário	0,014	0,003	0,630	0,002	0,003	0,003	Fraca
K-Means	t-SNE	Multi	0,005	0,002	0,407	0,001	0,003	0,002	Fraca
HDBSCAN	t-SNE	Binário	0,014	0,003	0,630	0,002	0,003	0,003	Fraca
HDBSCAN	t-SNE	Multi	0,005	0,002	0,407	0,001	0,003	0,002	Fraca

Fonte: Elaboração própria (2025), baseado em Santos (2018).

Abreviações: Proj. = Projeção; Multi = Multiclasse; **ARI** = *Adjusted Rand Index*; NMI = *Normalized Mutual Information*; Homog. = Homogeneidade; Compl. = Completude; V-meas. = *V-measure*; Interp. = Interpretação.

Interpretação: Thresholds segundo Hubert e Arabie (1985): **ARI** < 0,20 (fraca concordância); 0,20–0,50 (moderada); > 0,50 (forte). Linhas destacadas indicam melhor configuração (**HDBSCAN + UMAP**).

Contexto: Todas as configurações apresentaram concordância fraca, com **ARI** máximo de 0,059 — substancialmente inferior aos blocos Minerário (0,373) e Sanitário. Este resultado sugere que variáveis biogeoquímicas isoladas (Cl, **NO₃**) não capturam adequadamente a estrutura de impacto definida pelos gabaritos de referência, evidenciando a necessidade de integração com marcadores complementares.

A concordância fraca com os gabaritos externos pode ser interpretada sob duas perspectivas complementares. Em primeiro lugar, as variáveis biogeoquímicas refletem processos naturais de equilíbrio físico-químico da coluna d'água — oxigenação, dissolução mineral, reações redox —, que são governados por fatores hidrodinâmicos, topográficos e climáticos de longo prazo. Esses processos possuem escalas temporais e espaciais distintas daquelas associadas ao impacto imediato do rompimento da barragem de Fundão, cujos efeitos são primariamente mediados pelo arraste de rejeitos minerários e pela contaminação microbiológica. Em segundo lugar, a elevada proporção de ruído detectada pelo **HDBSCAN+UMAP** (53,7%) sugere que grande parte das amostras não se organiza em agrupamentos coesos quando avaliada exclusivamente sob a perspectiva biogeoquímica. Esta dispersão estrutural pode refletir tanto a variabilidade natural intrínseca aos processos de fundo quanto a sobreposição de múltiplas fontes de variação não capturadas pelas cinco variáveis do bloco. A análise comparativa dos três blocos temáticos evidencia, portanto, que variáveis de origem antrópica — minerárias e sanitárias — apresentam maior coerência com a estrutura de impacto espacial do desastre, enquanto variáveis biogeoquímicas naturais, embora fundamentais para caracterizar a qualidade da água, não revelam padrões espaciais alinhados às classificações de referência.

Em síntese, o Bloco Biogeoquímico apresentou estrutura de variância mais equilibrada (**PC1** = 46,81%, **PC2** = 22,32%, **PC3** = 13,50%) quando comparado aos blocos

Minerário ($PC1 = 97,46\%$) e Sanitário. As métricas internas de qualidade alcançaram valores satisfatórios, com silhouette entre 0,706 e 0,725, indicando clusters moderadamente bem separados e coesos. Entretanto, a validação externa revelou concordância consistentemente fraca com os gabaritos de Santos (2018), com ARI máximo de 0,059 — valor inferior aos registrados nos blocos Minerário (0,373) e Sanitário. A análise isolada dos três blocos temáticos evidenciou que variáveis de origem antrópica — minerárias e sanitárias — apresentam maior capacidade de estruturação espacial coerente com as classificações de referência, enquanto variáveis biogeoquímicas naturais, embora essenciais para caracterizar a dinâmica da coluna d'água, não capturam padrões de impacto espacial associados ao desastre. Esses resultados, em conjunto, reforçam a hipótese de que a integração de múltiplos domínios temáticos — abordagem explorada na Fase 2 mediante análises multibloco — é essencial para capturar a complexidade dos padrões de qualidade da água em contextos pós-desastre. A próxima seção apresenta os resultados da Fase 2, que investiga se combinações de variáveis minerárias, sanitárias e biogeoquímicas aprimoram a capacidade de estruturação espacial identificada nas análises temáticas isoladas.

Avaliação Comparativa Blocos Temáticos (Fase 1)

A análise exploratória não supervisionada foi conduzida de forma independente nos três blocos temáticos, conforme descrito na Seção 3.3. O objetivo foi identificar padrões naturais de agrupamento e avaliar a capacidade de cada conjunto de variáveis em capturar a estrutura espacial dos impactos do desastre. A Tabela 20 apresenta o desempenho comparativo dos blocos por meio de métricas de validação interna (Silhouette Score) e externa (Adjusted Rand Index).

Tabela 20 – Desempenho comparativo dos blocos temáticos na Fase 1 (análise não supervisionada).

Bloco	N Var.	Silh.	ARI Bin.	ARI Multi.	N PCs	Ruído (%)
Minerário	5	0,72	0,03	0,37	2	28,1
Sanitário	3	0,68	0,04	0,08	2	36,4
Biogeoquímico	5	0,72	0,06	0,06	4	53,7

Fonte: Elaboração própria (2025).

Legenda: N Var. = número de variáveis; Silh. = Silhouette Score (quanto maior, melhor coesão interna); ARI Bin. = Adjusted Rand Index para classificação binária (Afetado vs Não-Afetado); ARI Multi. = ARI para classificação multiclasse (4 categorias hidrológicas); N PCs = número de componentes principais retidos (85% de variância); Ruído = proporção de amostras classificadas como outliers pelo **HDBSCAN**. Melhor desempenho destacado em negrito.

O Bloco Minerário apresentou desempenho superior na validação externa, com ARI de 0,37 para classificação multiclasse. Esse valor é 4,6 vezes superior ao Bloco Sanitário ($ARI = 0,08$) e 6,2 vezes superior ao Bloco Biogeoquímico ($ARI = 0,06$). Os

resultados indicam que as variáveis minerárias — [Turbidez](#), [Alcalinidade](#), [Sólidos Totais](#), condutividade e [Fósforo Total](#) — capturam com maior eficácia a estrutura espacial dos impactos do desastre. Isso reflete a persistência da assinatura geoquímica dos rejeitos na matriz aquática dois anos após o evento de ruptura. A Figura 15 ilustra visualmente essa discrepância de desempenho.

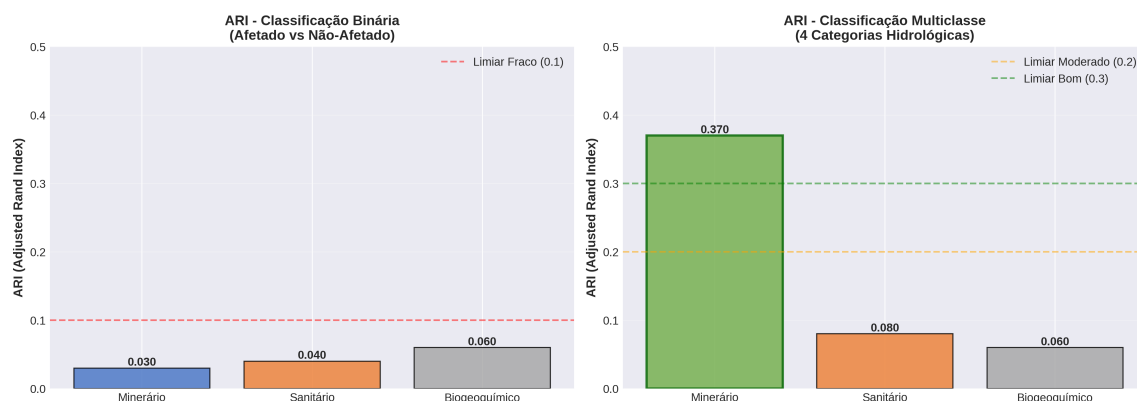


Figura 15 – Comparação do [Adjusted Rand Index](#) (ARI) entre blocos temáticos para classificação binária (Afetado vs Não-Afetado) e multiclasse (4 categorias hidrológicas).

Nota: O Bloco Minerário apresentou desempenho 4,6 vezes superior ao Bloco Sanitário e 6,2 vezes superior ao Bloco Biogeoquímico na validação multiclasse, o que indica maior capacidade de capturar a estrutura espacial dos impactos do desastre. As linhas tracejadas indicam limiares interpretativos: [ARI](#) < 0,10 = concordância fraca; 0,10 < [ARI](#) < 0,30 = concordância moderada; [ARI](#) > 0,30 = concordância substancial ([HUBERT; ARABIE, 1985](#)).

Observa-se que apenas o Bloco Minerário alcançou [ARI](#) superior ao limiar de concordância substancial ([ARI](#) > 0,30). Os blocos Sanitário e Biogeoquímico permaneceram abaixo do limiar de concordância fraca ([ARI](#) < 0,10). Esse padrão sugere que as variáveis de origem antrópica diretamente associadas ao impacto de rejeitos — as minerárias — estruturam-se espacialmente de forma coerente com as classificações de referência de [Santos \(2018\)](#). Por outro lado, as variáveis sanitárias e biogeoquímicas refletem processos difusos, como a contaminação fecal pré-existente, ou processos naturais, como o metabolismo aquático e os ciclos de nutrientes, que não estão diretamente vinculados ao desastre.

Em relação à validação interna, os blocos Minerário e Biogeoquímico apresentaram Silhouette Score de 0,72, o que indica coesão interna elevada dos clusters identificados. O Bloco Sanitário obteve 0,68. Entretanto, a proporção de amostras classificadas como ruído pelo [HDBSCAN](#) variou consideravelmente: 28,1% no Bloco Minerário, 36,4% no Sanitário e 53,7% no Biogeoquímico. A elevada proporção de ruído no Bloco Biogeoquímico, associada ao baixo [ARI](#) externo, evidencia que mais da metade das amostras não pertence a nenhum cluster denso. Isso reflete alta dispersão dos dados e ausência de estrutura espacial definida relacionada ao desastre.

Seleção de Variáveis para a Fase 2

A transição da Fase 1 — análise exploratória não supervisionada — para a Fase 2 — modelagem preditiva supervisionada — fundamentou-se em um critério rigoroso de seleção baseado nos loadings dos componentes principais retidos em cada bloco temático. Foram mantidas apenas as variáveis com **loadings dominantes**, ou seja, coeficiente $> 0,90$ em valor absoluto, nos dois primeiros componentes principais. Esses componentes, em conjunto, explicam mais de 90% da variância em todos os blocos analisados. A Tabela 21 sintetiza o processo de seleção aplicado a cada bloco.

Tabela 21 – Seleção de variáveis por bloco temático mediante critério de loadings dominantes ($> 0,90$).

Bloco	N Orig.	N Sel.	Variáveis Selecionadas
Minerário	5	2	Alca., Turb.
Sanitário	3	2	<i>E. coli</i> , Coliformes Totais
Biogeoquímico	5	2	Nitrato, Cloretos
Total	13	6	Redução de 54%

Fonte: Elaboração própria (2025).

Legenda: N Orig. = número de variáveis originais; N Sel. = número de variáveis selecionadas. Critério de seleção: apenas variáveis com loadings $> 0,90$ em PC1 ou PC2 foram mantidas. Variáveis excluídas: Sólidos Totais, condutividade, Fósforo Total (Minerário); demanda bioquímica de oxigênio (Sanitário); pH, saturação de oxigênio dissolvido, potencial de oxidação-redução (Biogeoquímico).

A seleção resultou em redução de 54% na dimensionalidade, passando de 13 para 6 variáveis. Ao mesmo tempo, preservou ao menos 99% da variância nos blocos Minerário e Sanitário, e 69% no Biogeoquímico. A Figura 16 apresenta um painel comparativo dos biplots PCA dos três blocos. A figura evidencia a estrutura de loadings que fundamenta a seleção e as diferenças na distribuição de variância entre os blocos.

Análise Comparativa: Biplots PCA dos Blocos Temáticos (Fase 1)

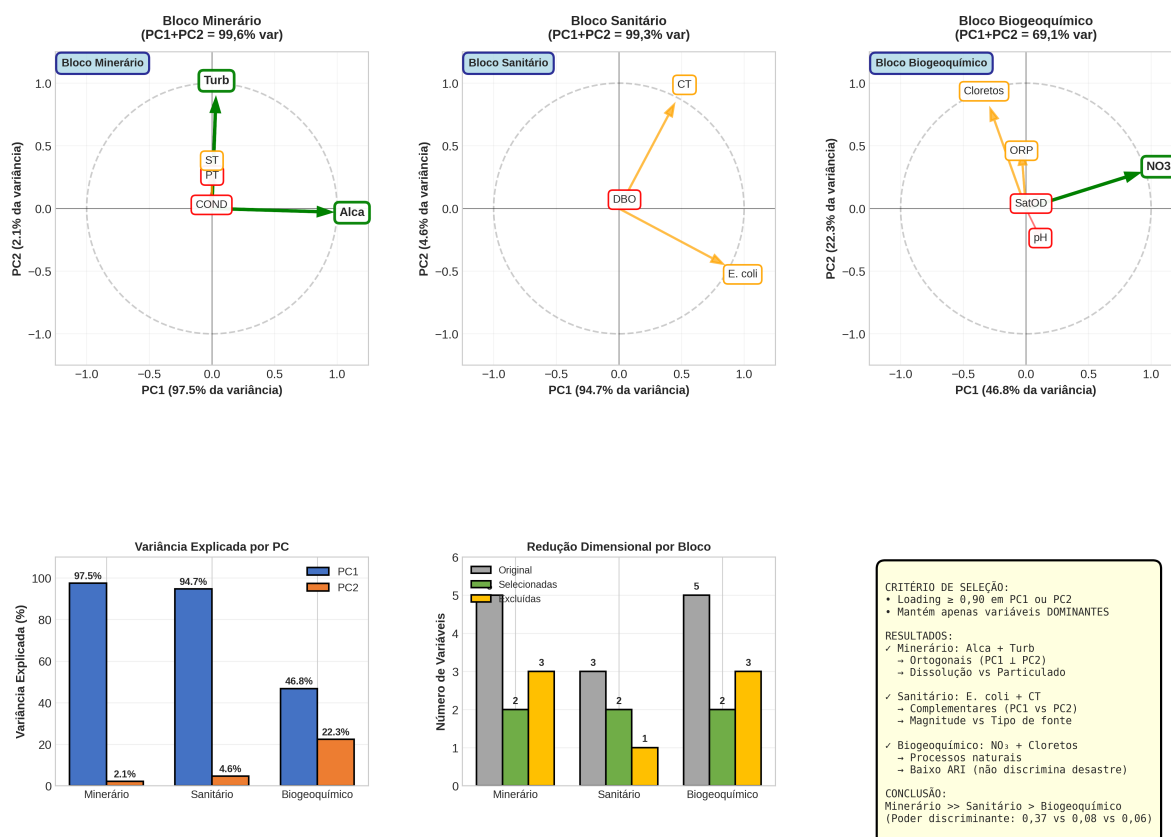


Figura 16 – Painel comparativo dos biplots PCA dos três blocos temáticos (Fase 1).

Fonte: Elaboração própria (2025).

Nota: **Superior:** Os biplots evidenciam a estrutura de loadings em cada bloco, com variáveis selecionadas (verde, loading $\geq 0,90$) claramente distinguíveis das redundantes (laranja, $0,30 \leq \text{loading} < 0,90$) e negligíveis (vermelho, loading $< 0,30$). **Inferior esquerdo:** Variância explicada pelos dois primeiros componentes principais, demonstrando maior concentração no Bloco Minerário (PC1 = 97,5%) e distribuição mais equilibrada no Biogeoquímico (PC1 = 46,8%, PC2 = 22,3%). **Inferior centro:** Redução dimensional por bloco, com seleção de 2 variáveis em cada caso mediante critério de loading $> 0,90$. **Inferior direito:** Síntese interpretativa evidenciando ortogonalidade (Minerário: dissolução vs particulado) e complementaridade (Sanitário: magnitude vs tipo de fonte). A superioridade do Bloco Minerário (ARI = 0,37) reflete maior capacidade de capturar padrões espaciais associados ao desastre.

Bloco Minerário: A Análise de Componentes Principais revelou uma estrutura de variância altamente concentrada. O primeiro componente (PC1) explicou 97,46% da variância, sendo dominado exclusivamente pela **Alcalinidade** (loading = 0,9995). O segundo componente (PC2) contribuiu com 2,12% adicionais e foi dominado pela **Turbidez** (loading = 0,910). As demais variáveis minerárias — **Sólidos Totais** (loading PC2 = 0,34), **Fósforo Total** (loading PC2 = 0,24) e **Condutividade Elétrica** (loading PC2 = 0,03) — apresentaram contribuições inferiores a 0,35 em ambos os componentes. Isso indica que sua variabilidade é capturada de forma redundante pelas duas variáveis dominantes.

A dominância da **Alcalinidade** reflete processos de dissolução química associados

aos óxidos metálicos presentes nos rejeitos de mineração. Esse processo foi documentado por [Hatje, Pedreira e Rezende \(2017\)](#) como o principal mecanismo de alteração geoquímica da água em sistemas impactados por rompimento de barragens. A ortogonalidade entre [Alcalinidade \(PC1\)](#) e [Turbidez \(PC2\)](#) sugere que os processos de mineralização e de transporte físico de sedimentos atuam de forma relativamente independente na estruturação da qualidade da água pós-desastre. Aplicando o critério de loading $> 0,90$, foram selecionadas [Alcalinidade](#) e [Turbidez](#). Essas variáveis retêm 99,58% da variância original, com redução de 60% na dimensionalidade, passando de 5 para 2 variáveis.

Bloco Sanitário: O [PC1](#) explicou 94,68% da variância e foi dominado por [Escherichia coli](#) (loading = 0,883). O [PC2](#), responsável por 4,64% da variância, foi dominado por coliformes totais (loading = 0,881). A demanda bioquímica de oxigênio apresentou loadings inferiores a 0,07 em ambos os componentes, sendo excluída por contribuir marginalmente à estrutura de variância.

A estrutura bicomponente do Bloco Sanitário merece interpretação cuidadosa. O [PC1](#), dominado por [E. coli](#), representa a magnitude total de contaminação fecal. Já o [PC2](#), dominado por coliformes totais, discrimina o tipo de fonte poluidora. A inversão de sinais entre os indicadores em [PC2](#) — [E. coli](#) negativo e coliformes totais positivo — evidencia que, embora correlacionados na magnitude ([PC1](#), $r = 0,87$), capturam aspectos distintos da contaminação. O [E. coli](#) apresenta especificidade para contaminação fecal recente de origem animal ou humana, com sobrevivência de aproximadamente dias, o que discrimina áreas com despejos pontuais. Os coliformes totais, por sua vez, incluem espécies de origem ambiental, como solo e vegetação, com maior persistência, de aproximadamente semanas, discriminando áreas com poluição difusa oriunda de escoamento superficial agrícola ([WORLD HEALTH ORGANIZATION, 2017](#)).

Essa complementaridade — magnitude ([PC1](#)) versus tipo de fonte ([PC2](#)) — justifica a seleção de [E. coli](#) e [coliformes totais](#), retendo 99,32% da variância original. Embora a correlação linear elevada ($r = 0,87$) sugira redundância, a análise [PCA](#) demonstra que capturam dimensões ortogonais essenciais para discriminação espacial. Esse argumento é aprofundado na Seção 4.X ao justificar a composição do Bloco C.

Bloco Biogeoquímico: A estrutura de variância mostrou-se mais distribuída que nos demais blocos. Os quatro primeiros componentes foram necessários para reter 95% da variância. O [PC1](#) (46,81%) foi dominado por [Nitrato](#) (loading = 0,948), e o [PC2](#) (22,32%) por cloretos (loading = 0,835). As demais variáveis — pH (loading [PC1](#) = 0,10, [PC2](#) = -0,21), saturação de oxigênio dissolvido (loading [PC1](#) = 0,04, [PC2](#) = 0,03) e potencial de oxidação-redução (loading [PC1](#) = -0,03, [PC2](#) = 0,41) — distribuíram-se entre [PC3](#) e [PC4](#) com loadings inferiores a 0,93. Essas variáveis foram excluídas por não atingirem o critério nos dois primeiros componentes.

A maior dispersão de variância ([PC1](#) + [PC2](#) = 69,1%) reflete processos biogeoquímicos

cos independentes, como ciclos de nutrientes (**Nitrato**), salinização (cloretos), metabolismo aquático (pH, saturação de oxigênio dissolvido) e reações redox (potencial de oxidação-redução). Esses processos não estão diretamente associados ao impacto de rejeitos. Esse padrão, aliado ao baixo desempenho na validação externa ($ARI = 0,06$), indica que as variáveis biogeoquímicas capturam a dinâmica natural da coluna d'água, mas não discriminam padrões espaciais do desastre. Foram selecionadas **Nitrato** e **cloretos**, retendo 69,13% da variância original.

4.2.2 Análise dos Blocos Principais (Fase 2)

A análise exploratória não supervisionada da Fase 1, conduzida isoladamente nos três blocos temáticos, revelou que variáveis de natureza mineral, sanitária e biogeoquímica apresentam estruturas de covariância distintas e complementares. Nesse contexto, a Fase 2 foi concebida para investigar se a integração estratégica de variáveis selecionadas de múltiplos domínios temáticos aprimora a capacidade discriminatória dos agrupamentos e sua estabilidade temporal em relação aos blocos isolados. Uma consideração metodológica central deve ser explicitada: embora a análise seja fundamentalmente química — baseada em parâmetros físico-químicos e microbiológicos —, os cenários de referência de Santos (2018) são geográficos, definidos por localização espacial e características hidrológicas. Por essa razão, foram integrados tanto marcadores minerários (**Turbidez**, **Alcalinidade**, **Sólidos Totais**) quanto sanitários (coliformes totais, *Escherichia coli*), além de discriminantes biogeoquímicos (cloretos, **Nitratos**, **Fósforo Total**), visando capturar múltiplas dimensões da qualidade da água que possam correlacionar-se com a estrutura espacial da bacia.

Foram definidos três blocos principais, denominados Blocos A, B e C, com composição diferenciada quanto à estratégia de incorporação de marcadores sanitários, mas preservando um núcleo comum de variáveis capazes de capturar assinatura hidroquímica do impacto e variabilidade biogeoquímica natural. A metodologia seguiu o protocolo da Fase 1: redução dimensional via **PCA**, projeções bidimensionais mediante **UMAP** e *t-SNE*, clusterização via *K-Means Clustering* (*K-Means*) e **HDBSCAN**, validação interna por coeficiente de *Silhouette*, índice de Calinski-Harabasz e índice de Davies-Bouldin, e validação externa mediante confronto com os gabaritos binário e multiclasse de Santos (2018). Adicionalmente, implementou-se validação hierárquica temporal em três níveis — amostra ($n=324$), ponto ($n=27$) e campo ($n=12$) —, permitindo avaliar a estabilidade dos agrupamentos frente à variabilidade sazonal e identificar padrões espaciais robustos.

Composição e Fundamentação dos Blocos Principais

A composição dos blocos principais fundamentou-se na manutenção de um núcleo físico-químico comum - **Sólidos Totais** (ST), **Turbidez** (Turb.), **Alcalinidade** (Alca.), cloretos

(CL), Nitratos (NO_3) e Fósforo Total (PT) — acrescido de marcadores sanitários segundo estratégias distintas, conforme sintetizado na Tabela 22.

Tabela 22 – Composição e estratégias de diferenciação dos blocos principais (Fase 2).

Bloco	<i>n</i>	Composição Completa	Estratégia Sanitária
A	7	(ST, Turb., Alca., CL, NO_3 , CT, PT	Contaminação difusa
B	7	(ST, Turb., Alca., CL, NO_3 , <i>E. coli</i> , PT	Contaminação pontual (<i>E. coli</i>)
C	8	(ST, Turb., Alca., CL, NO_3 , <i>E. coli</i> , CT, PT	Monitoramento integrado

Fonte: Elaboração própria (2025).

Nota: *n* = número de variáveis. Núcleo comum (ST, Turb., Alca., CL, NO_3 , PT): assinatura hidroquímica + discriminantes biogeoquímicos. Diferenciação: marcadores sanitários complementares (CT, *E. coli*).

O Bloco A incorpora coliformes totais (CT) como marcador sanitário. Coliformes Totais abrangem tanto organismos de origem ambiental quanto fecal, constituindo indicador mais amplo de contaminação microbiológica. O Bloco B substitui CT por *E. coli*, indicador específico de contaminação fecal recente, visando capturar fontes pontuais como despejos diretos de esgoto doméstico. O Bloco C integra ambos os marcadores sanitários (CT e *E. coli*) simultaneamente, totalizando oito variáveis, fundamentando-se em complementaridade dimensional e tolerância à multicolinearidade de modelos *ensemble*.

Análise de Componentes Principais dos Blocos Principais

A Análise de Componentes Principais (PCA) foi aplicada aos três blocos principais com o objetivo de avaliar a estrutura de covariância das variáveis e identificar o número mínimo de componentes necessários para reter ao menos 99% da variância original, seguindo o critério de Kaiser (*eigenvalue* > 1). A Tabela 23 consolida os resultados da PCA, enquanto a Figura 17 apresenta comparação visual da variância explicada.

Tabela 23 – Variância explicada pelos componentes principais nos blocos principais (Fase 2).

Bloco	<i>n</i>	PC1 (%)	PC2 (%)	PC3 (%)	Var. Ac. (%)	<i>k</i> (99%)
A	7	92,7	4,5	2,0	99,6	1
B	7	84,8	12,6	1,9	99,6	2
C	8	81,5	15,3	1,8	99,4	2

Fonte: Elaboração própria (2025).

Abreviações: *n* = número de variáveis; Var. Ac. = Variância Acumulada (PC1+PC2+PC3); *k* = número de componentes necessários para 99% de variância.

Interpretação: Linha destacada indica maior concentração de variância em PC1. Bloco A apresenta estrutura quase unidimensional (92,7% em PC1), enquanto Blocos B e C exibem distribuição mais equilibrada entre PC1 e PC2, indicando maior dimensionalidade intrínseca introduzida por *E. coli*.

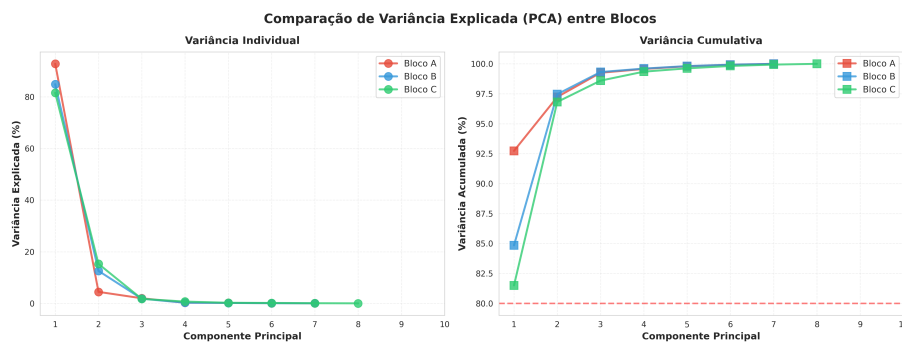


Figura 17 – Comparação da variância explicada pelos componentes principais nos Blocos A, B e C.

Fonte: Elaboração própria (2025).

Nota: Linhas tracejadas indicam critérios de retenção de 90% (laranja) e 95% (verde). Todos os blocos ultrapassam 99% com 3-4 componentes.

O Bloco A apresentou estrutura de covariância dominada por **Alcalinidade** em **PC1** (*loading* = 0,999), **CT** em **PC2** (*loading* = 0,998) e **Turbidez** em **PC3** (*loading* = 0,910), configurando estrutura tripartite de processos independentes: mineralização alcalina, contaminação difusa e transporte de sólidos. O Bloco B mantém **Alcalinidade** em **PC1**, mas **PC2** é dominado por *E. coli* (*loading* = 1,000), com contribuição de 12,6% — valor substancialmente superior aos 4,5% de **CT** no Bloco A, evidenciando maior variabilidade independente de *E. coli* devido à sua natureza pontual e efêmera. O Bloco C apresenta estrutura mais complexa, com **PC2** capturando predominantemente *E. coli* (*loading* = 0,883) e parcialmente **CT** (*loading* = 0,468), indicando ortogonalidade parcial entre marcadores sanitários que capturam dimensões complementares da contaminação.

A Figura 18 apresenta *biplots* comparativos dos três blocos, evidenciando graficamente a estrutura de covariância. Observa-se que **Alcalinidade** orienta-se ao longo de **PC1** nos três blocos, enquanto marcadores sanitários orientam-se perpendicularmente, confirmando independência entre processos de mineralização e contaminação microbológica. No Bloco C, a proximidade angular entre vetores de **CT** e *E. coli* reflete correlação moderada, mas seu distanciamento do vetor de **Alcalinidade** evidencia processos ambientais distintos.

Clusterização e Validação Interna

A clusterização foi conduzida mediante aplicação de **HDBSCAN** e *K-Means* às projeções **UMAP** e *t-SNE*, totalizando quatro configurações por bloco. A Tabela 24 consolida as métricas internas de qualidade para as configurações com **UMAP**, enquanto a Figura 19 apresenta comparação consolidada das métricas dos três blocos.

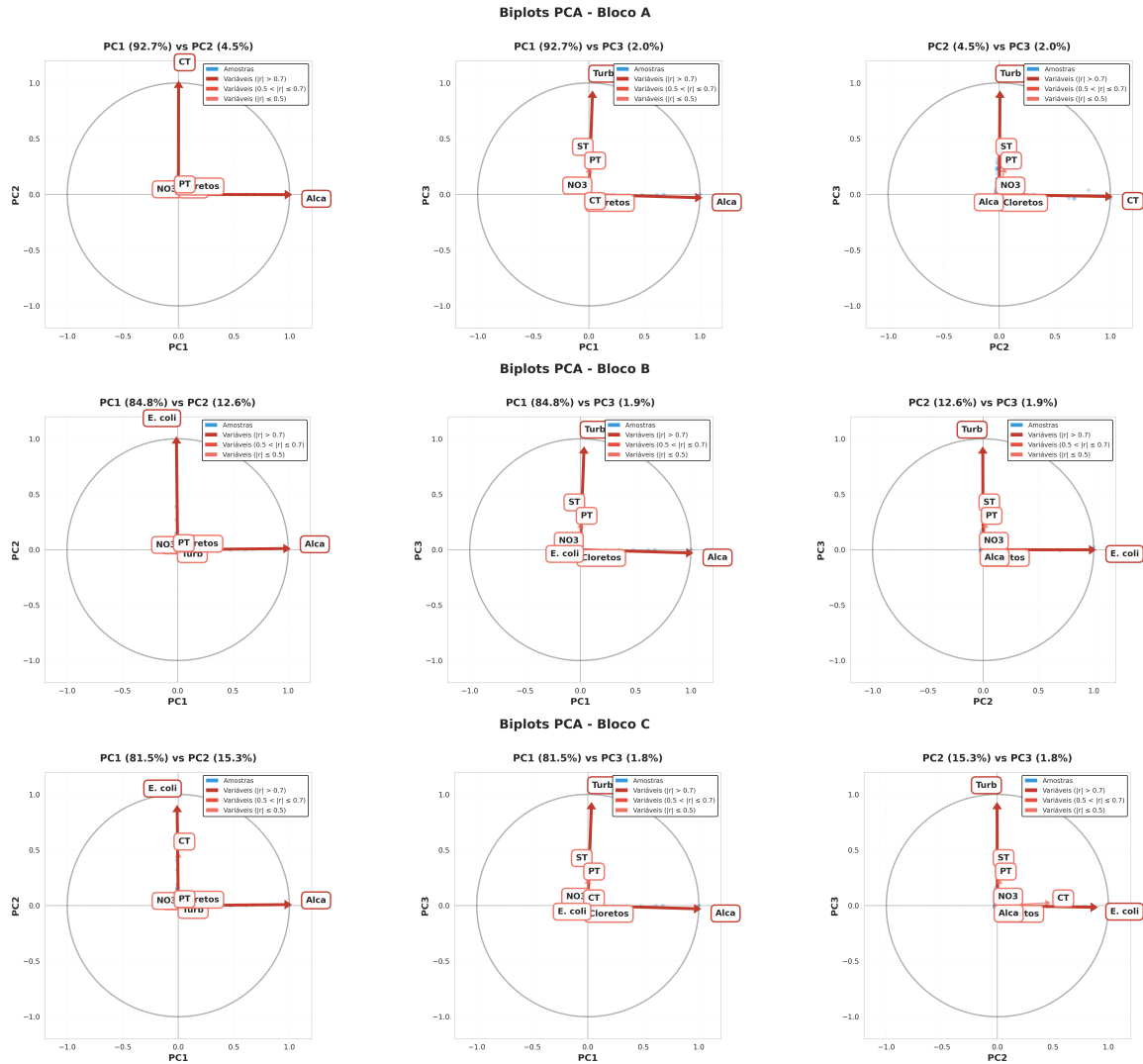


Figura 18 – *Biplots* da Análise de Componentes Principais para Blocos A (superior), B (centro) e C (inferior).

Fonte: Elaboração própria (2025).

Nota: Setas indicam direção e magnitude dos *loadings* das variáveis nos dois primeiros componentes. Pontos representam amostras projetadas no espaço $PC1 \times PC2$, coloridos por cluster *HDBSCAN*+*UMAP*. *Alca.* domina *PC1* em todos os blocos, enquanto marcadores sanitários (*CT* no Bloco A, *E. coli* no Bloco B, ambos no Bloco C) orientam-se perpendicularmente em *PC2*, confirmando independência dimensional entre processos de mineralização e contaminação microbiológica.

Tabela 24 – Métricas internas de qualidade dos agrupamentos — Blocos Principais (projeções UMAP).

Bloco	Alg.	Proj.	k	n	Ruído	Silh.	DB	CH
A	K-Means	UMAP	4	324	0,0%	0,789	0,263	2188
A	HDBSCAN	UMAP	5	324	0,0%	0,764	0,273	1922
B	K-Means	UMAP	4	324	0,0%	0,775	0,268	5905
B	HDBSCAN	UMAP	17	324	21,3%	0,756	0,340	59648
C	K-Means	UMAP	3	324	0,0%	0,834	0,250	3633
C	HDBSCAN	UMAP	11	324	6,5%	0,738	0,328	20452

Fonte: Elaboração própria (2025).

Abreviações: Alg. = Algoritmo; Proj. = Projeção; k = número de clusters; n = total de amostras; Ruído = proporção de amostras não atribuídas (HDBSCAN); Silh. = Silhouette Score; DB = Davies-Bouldin Index; CH = Calinski-Harabasz Index.

Interpretação: Linha destacada indica melhor configuração global (Bloco C, *K-Means*+UMAP): Silhouette excepcional (0,834), baixa sobreposição (DB=0,250) e ausência de ruído. HDBSCAN no Bloco B apresenta fragmentação excessiva (17 clusters, 21,3% ruído), indicando sensibilidade excessiva a subestruturas locais.

As Figuras 20, 21 e 22 apresentam comparação sistemática das quatro configurações para cada bloco. Observa-se que *K-Means* produz partições mais homogêneas entre projeções, enquanto HDBSCAN apresenta maior sensibilidade à escolha da projeção, com fragmentação particularmente acentuada no Bloco B (17 clusters em UMAP, 7 em *t-SNE*).

Validação Externa e Análise Hierárquica Temporal

A validação externa confrontou clusters com gabaritos binário e multiclasse de Santos (2018), empregando ARI como métrica principal. Implementou-se validação hierárquica em três níveis temporais: amostra ($n=324$), ponto ($n=27$) e campo ($n=12$). A Tabela 25 consolida resultados para gabarito multiclasse com *K-Means*.

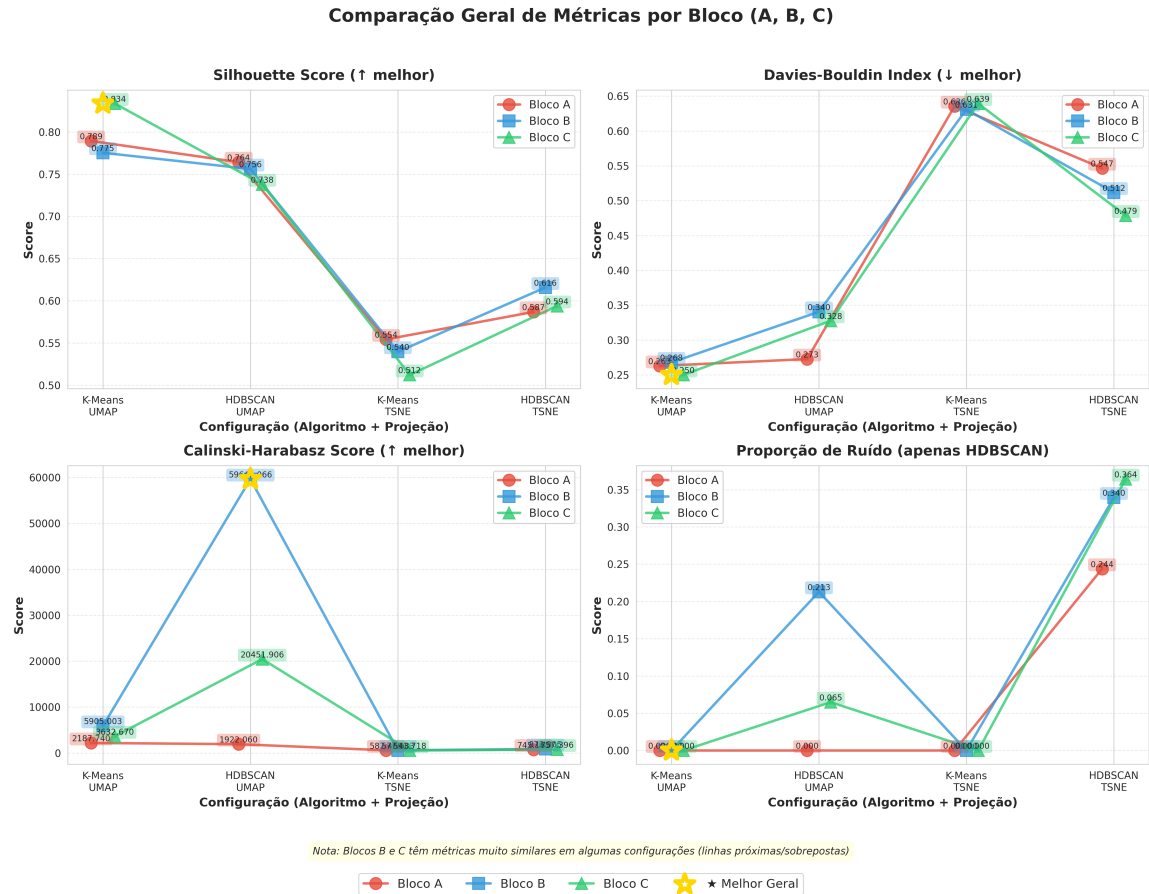


Figura 19 – Comparação consolidada das métricas internas dos Blocos A, B e C para as quatro configurações testadas (HDBSCAN/*K-Means* × UMAP/*t-SNE*).

Nota: Superior esquerda: Coeficiente de Silhouette (valores superiores a 0,7 indicam boa coesão). Superior direita: Índice de Davies-Bouldin (valores inferiores a 0,5 indicam baixa sobreposição). Inferior esquerda: Número de clusters identificados (HDBSCAN apresenta maior fragmentação). Inferior direita: Proporção de ruído (HDBSCAN em *t-SNE* apresenta valores elevados). Observa-se desempenho consistentemente superior de *K-Means*+UMAP no Bloco C.

Fonte: Elaboração própria (2025).

Tabela 25 – Validação externa hierárquica dos blocos principais — gabarito multiclasse (*K-Means*).

Bloco	Proj.	ARI Amost.	ARI Ponto	Ganho (%)	Compl.	Homog.
A	UMAP	0,124	0,376	+203	0,761	0,445
A	<i>t-SNE</i>	0,311	0,660	+112	0,907	0,681
B	UMAP	0,196	0,376	+92	0,761	0,244
B	<i>t-SNE</i>	0,220	0,443	+101	0,642	0,617
C	UMAP	0,260	0,417	+60	0,744	0,477
C	<i>t-SNE</i>	0,282	0,555	+97	0,716	0,650

Fonte: Elaboração própria (2025).

Abreviações: ARI Amost. = ARI no nível amostra (n=324); ARI Ponto = ARI no nível ponto (n=27); Ganho (%) = variação percentual amostra→ponto; Compl. = Completude; Homog. = Homogeneidade.

Interpretação: Linha destacada indica melhor concordância externa (Bloco A, *K-Means*+*t-SNE*): ARI ponto = 0,660 (concordância moderada), completude = 0,907. Ganho consistente de ARI no nível ponto evidencia que agregação temporal elimina variabilidade sazonal e revela estrutura espacial estável captada pelos clusters.

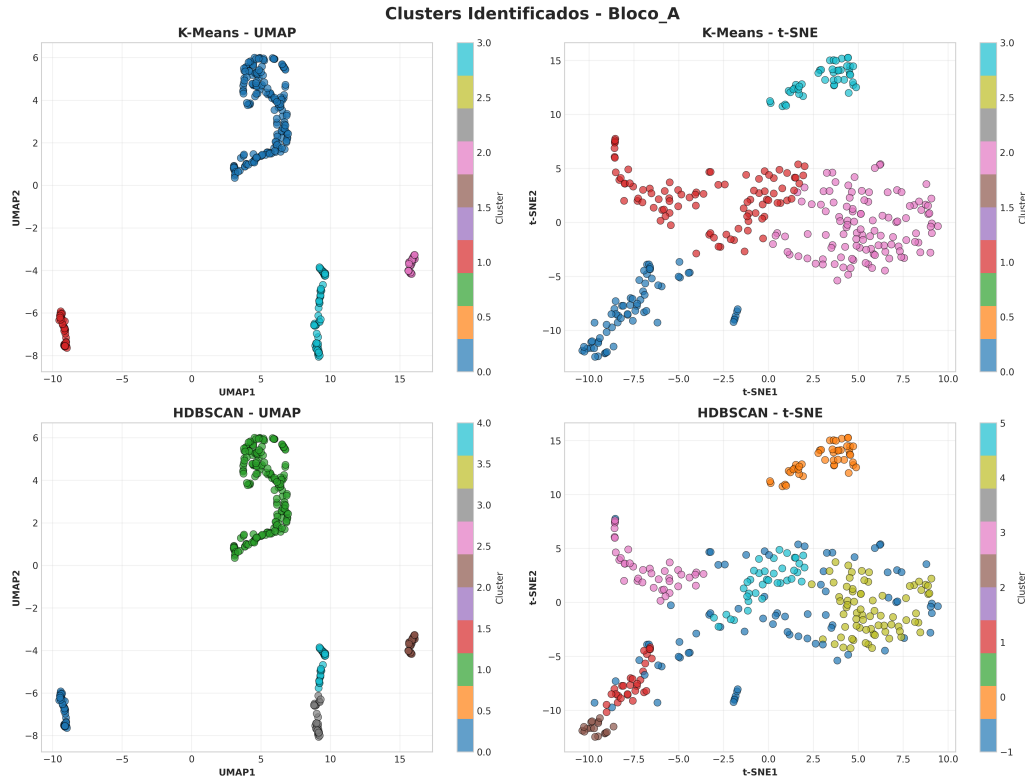


Figura 20 – Configurações de clusterização aplicadas ao Bloco A.

Fonte: Elaboração própria (2025).

Nota: **HDBSCAN+UMAP** (5 clusters, 0% ruído), **HDBSCAN+t-SNE** (6 clusters, 24,4% ruído), **K-Means+UMAP** (4 clusters) e **K-Means+t-SNE** (4 clusters). **UMAP** preserva melhor topologia global, produzindo clusters mais compactos.

A Figura 23 apresenta validação hierárquica em três níveis para o Bloco C com **K-Means+UMAP**, exemplificando o padrão observado nos três blocos: ganho substancial de **ARI** no nível ponto, evidenciando que agregação temporal elimina variabilidade sazonal e revela estrutura espacial estável.

As matrizes de confusão (Figura 24) evidenciam padrões de concordância e discordância entre clusters e gabaritos. Observa-se maior alinhamento com gabarito multiclasse que binário, confirmando que clusters capturam nuances espaciais melhor representadas por quatro categorias hidrológicas (**RP** Afetado/Não, Tributário Afetado/Não) que pela dicotomia simples Afetado/Não Afetado.

Caracterização Hidroquímica dos Clusters e Análise de Pureza

A caracterização físico-química dos clusters identificados foi conduzida mediante estatísticas descritivas das variáveis em cada grupo e cálculo de pureza em relação aos gabaritos de Santos (2018). A pureza de um cluster foi definida como a proporção de amostras pertencentes à classe modal (mais frequente) do gabarito multiclasse dentro do cluster, conforme a Equação 4.1:

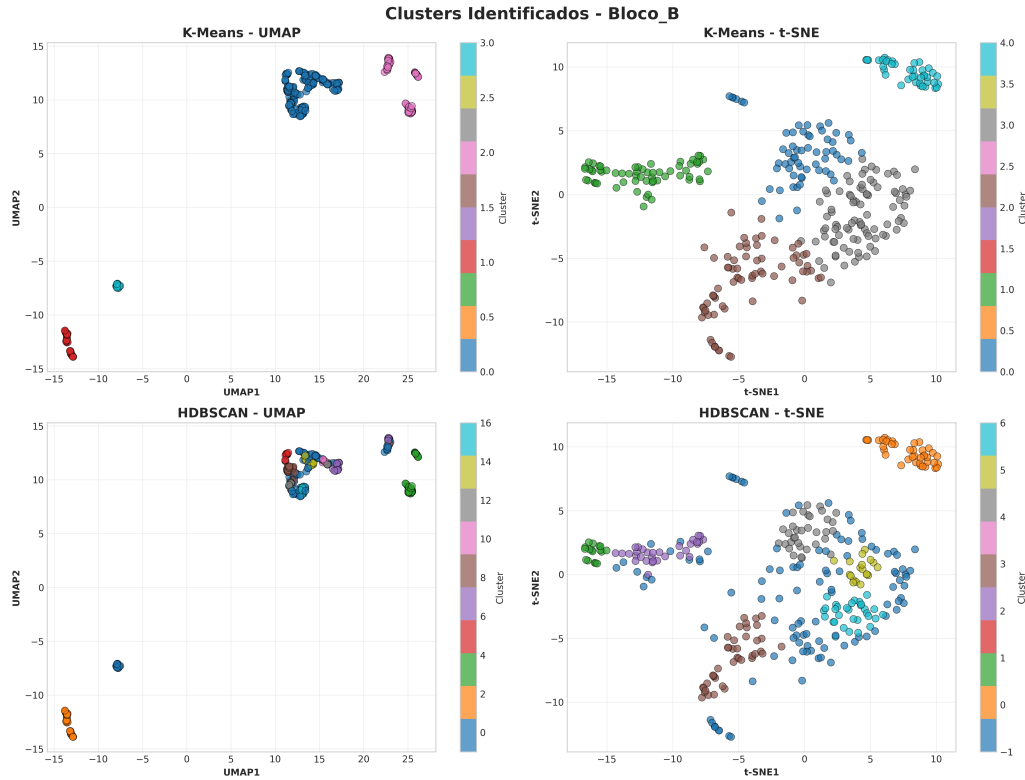


Figura 21 – Configurações de clusterização aplicadas ao Bloco B.

Fonte: Elaboração própria (2025).

Nota: **HDBSCAN**+**UMAP** (17 clusters, 21,3% ruído), **HDBSCAN**+**t-SNE** (7 clusters, 34,0% ruído), **K-Means**+**UMAP** (4 clusters) e **K-Means**+**t-SNE** (5 clusters). **HDBSCAN** apresenta fragmentação excessiva, particularmente em **UMAP**, evidenciando sensibilidade à inclusão de *E. coli*.

$$\text{Pureza}(C_i) = \frac{\max_j |\{x \in C_i : \text{gabarito}(x) = j\}|}{|C_i|} \quad (4.1)$$

onde C_i representa o cluster i , j representa as quatro categorias do gabarito multiclasse, e $|C_i|$ é o número total de amostras no cluster. Clusters com pureza $\geq 0,80$ foram considerados de alta pureza, indicando forte concordância com o gabarito. A Tabela 26 consolida os resultados da análise de pureza para os Blocos B e C (configuração **HDBSCAN**+**UMAP**).

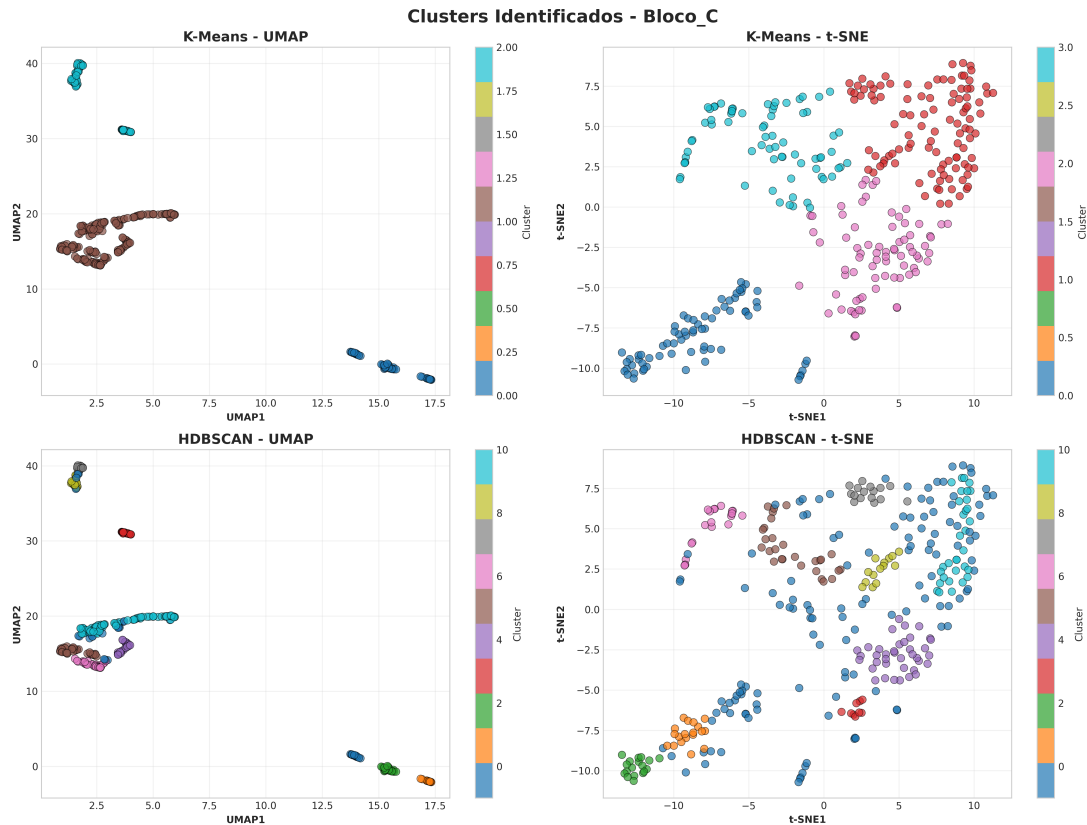


Figura 22 – Configurações de clusterização aplicadas ao Bloco C.

Fonte: Elaboração própria (2025).

Nota: **HDBSCAN+UMAP** (11 clusters, 6,5% ruído), **HDBSCAN+t-SNE** (11 clusters, 36,4% ruído), **K-Means+UMAP** (3 clusters) e **K-Means+t-SNE** (4 clusters). **K-Means+UMAP** apresenta melhor desempenho global (Silhouette = 0,834), evidenciando que a integração de ambos os marcadores sanitários (**CT** + *E. coli*) permite estruturação mais coesa e parcimoniosa.

Tabela 26 – Análise de pureza dos clusters em relação ao gabarito multiclasse (**HDBSCAN+UMAP**).

Bloco	Cluster	<i>n</i>	Classe Modal	Pureza	<i>n</i> Modal	Categ.
B	14	10	Trib_Afetado	1,000	10	Alta
	11	8	Trib_Afetado	1,000	8	Alta
	1	21	RP_Afetado	0,952	20	Alta
	5	12	Trib_Afetado	0,917	11	Alta
	0	18	RP_Afetado	0,889	16	Alta
	2	24	RP_Afetado	0,875	21	Alta
	13	10	Trib_Nao_Afet.	0,800	8	Alta
C	15	8	RP_Nao_Afetado	0,750	6	Mod.
	1	18	RP_Afetado	0,944	17	Alta
	0	18	RP_Afetado	0,889	16	Alta
	2	27	RP_Afetado	0,889	24	Alta
	4	27	Trib_Afetado	0,741	20	Mod.
	5	36	Trib_Afetado	0,722	26	Mod.
	10	33	Trib_Afetado	0,636	21	Mod.
	8	25	RP_Nao_Afetado	0,600	15	Mod.

Fonte: Elaboração própria (2025).

Abreviações: *n* = número de amostras no cluster; *n* Modal = número de amostras da classe modal; Categ. = Categoria (Alta pureza $\geq 0,80$; Mod. = Moderada 0,50-0,79; Baixa $< 0,50$).

Interpretação: Linha destacada indica cluster com melhor pureza em cada bloco. Bloco B apresenta 7 clusters de alta pureza ($\geq 0,80$), concentrados em **RP_Afetado** e **Trib_Afetado**, totalizando 103 amostras (40,4% do total de 255 amostras sem ruído). Bloco C apresenta 3 clusters de alta pureza, todos associados a **RP_Afetado**, totalizando 63 amostras (20,8%). Pureza média ponderada: Bloco B = 68,6%; Bloco C = 64,7%. Clusters com pureza $< 0,40$ foram omitidos por brevidade.

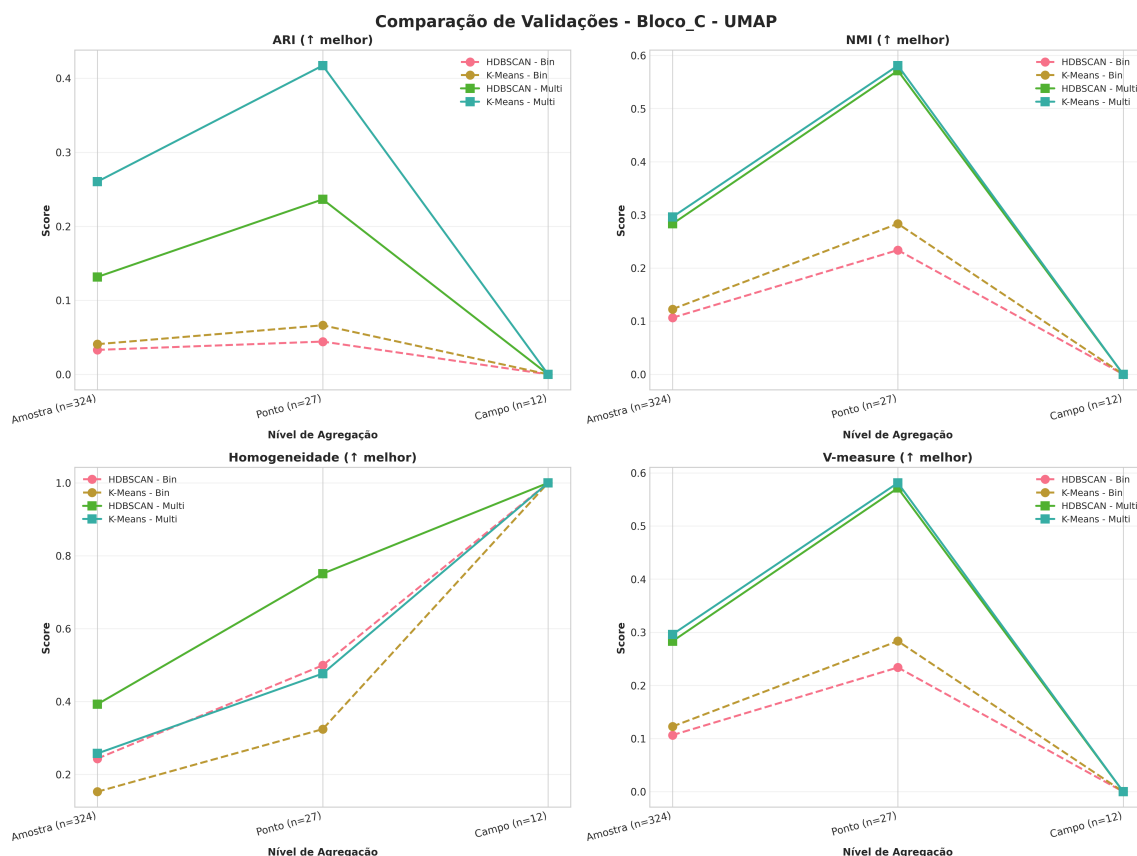


Figura 23 – Validação hierárquica em três níveis (amostra, ponto, campo) para Bloco C com *K-Means*+UMAP.

Fonte: Elaboração própria (2025).

Nota: Esquerda: ARI multiclasse. Direita: Completude. O ganho de ARI no nível ponto (amostra: 0,260 → ponto: 0,417) evidencia que clusters capturam estrutura espacial estável, com variabilidade temporal atuando como ruído. Padrão similar foi observado nos Blocos A (ganho de +203%) e B (ganho de +92%).

O Bloco B apresentou sete clusters de alta pureza (pureza $\geq 0,80$), destacando-se os clusters 14 e 11 com pureza perfeita (100%) para Trib_Afetado, e cluster 1 com pureza de 95,2% para RP_Afetado. Esses sete clusters concentram 103 amostras (40,4% do total de 255 amostras sem ruído), caracterizadas por perfis hidroquímicos bem definidos: clusters de RP_Afetado (0, 1, 2) apresentam elevada Turbidez e Alcalinidade associadas a baixos cloretos e Nitratos, enquanto clusters de Trib_Afetado (11, 14, 5) exibem elevada concentração de *E. coli*, indicativo de fontes pontuais de contaminação fecal. A pureza média ponderada do Bloco B foi de 68,6%.

O Bloco C apresentou três clusters de alta pureza, todos associados a RP_Afetado: clusters 1, 0 e 2, com purezas de 94,4%, 88,9% e 88,9%, respectivamente, totalizando 63 amostras (20,8% do total de 303 amostras sem ruído). Adicionalmente, o Bloco C apresentou quatro clusters de pureza moderada (0,60–0,74), associados às categorias Trib_Afetado e RP_Nao_Afetado, evidenciando maior capacidade de discriminação entre subtipos, embora com pureza inferior ao Bloco B. A pureza média ponderada do Bloco C

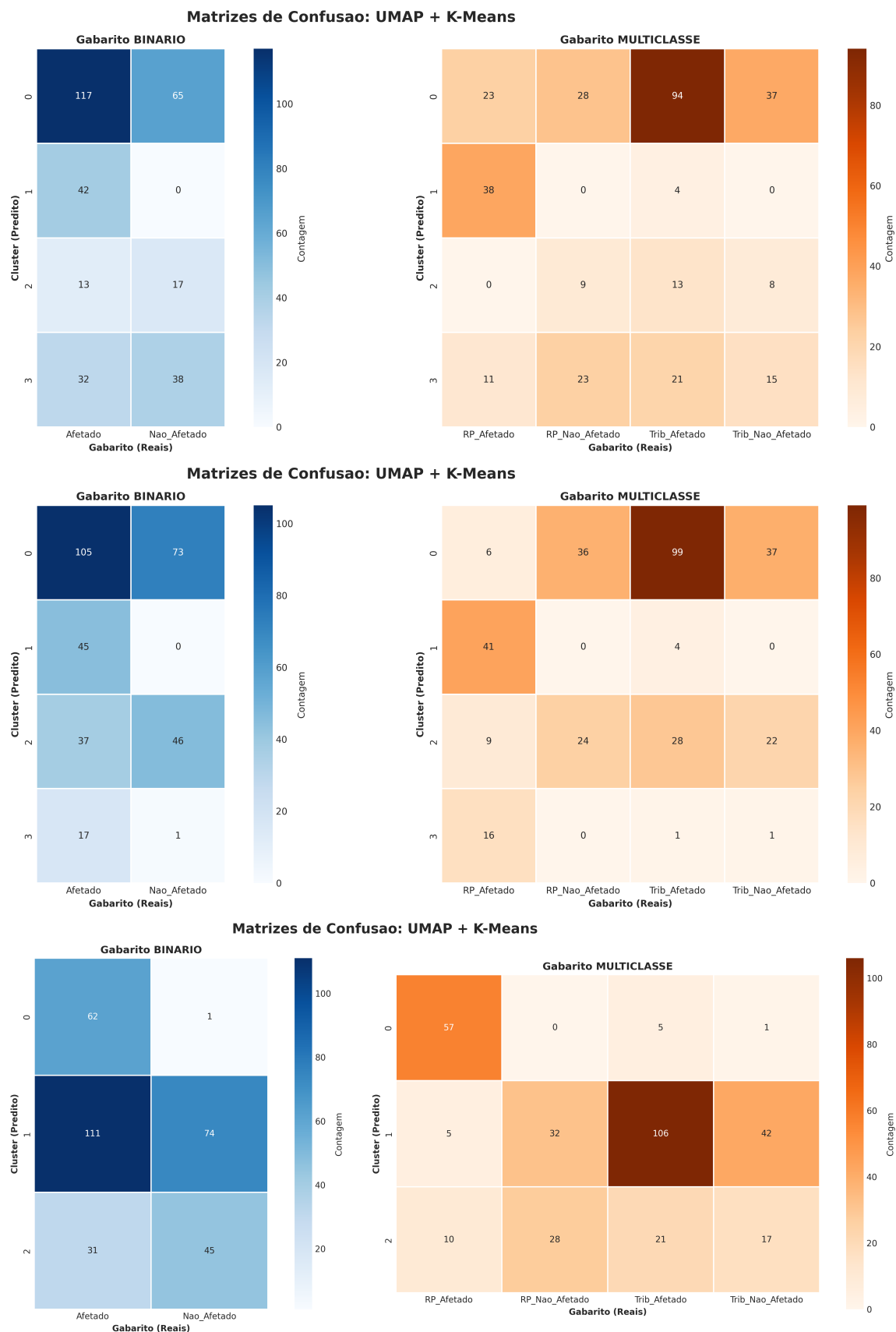


Figura 24 – Matrizes de confusão entre clusters *K-Means*+*UMAP* e gabaritos de referência para Blocos A (superior), B (centro) e C (inferior).

Fonte: Elaboração própria (2025).

Nota: Esquerda: gabarito binário. Direita: gabarito multiclasse. Intensidade de cor representa contagem de amostras. Maior alinhamento com multiclasse evidencia que clusters capturam nuances espaciais além da dicotomia Afetado/Não Afetado. No Bloco A, Cluster 0 concentra-se em **RP** Afetado, enquanto no Bloco C a distribuição é mais balanceada entre categorias.

foi de 64,7%, ligeiramente inferior ao Bloco B.

A análise revelou *trade-off* entre número de clusters de alta pureza e cobertura das quatro categorias hidrológicas: Bloco B maximiza pureza em **RP_Afetado** e **Trib_Afetado**, mas apresenta limitações na discriminação de **RP_Nao_Afetado** e **Trib_Nao_Afetado**. Bloco C, embora com pureza média ligeiramente inferior, distribui clusters de pureza moderada entre as quatro categorias, oferecendo cobertura mais equilibrada. Essa complementaridade fundamenta a estratégia de cenários múltiplos para modelagem supervisionada, priorizando Bloco B para cenários binários (alta pureza) e Bloco C para cenários multiclasse (cobertura balanceada).

Estabilidade Temporal e Trajetórias Intra-Ponto

A análise de estabilidade temporal foi conduzida mediante cálculo da consistência intra-ponto (proporção de campanhas em que cada ponto permanece no mesmo cluster) e identificação de transições sazonais sistemáticas. A Tabela 27 sintetiza resultados para ***K-Means*+UMAP**.

Tabela 27 – Estabilidade temporal intra-ponto dos blocos principais (***K-Means* + UMAP**).

Bloco	Consist. Média	Pontos $\geq 80\%$	Pontos $\geq 90\%$	Trans. Saz.	Aval.
A	67,6%	5/27 (18,5%)	1/27 (3,7%)	8/27 (29,6%)	Moderada
B	69,4%	6/27 (22,2%)	1/27 (3,7%)	4/27 (14,8%)	Mod-Alta
C	81,2%	16/27 (59,3%)	9/27 (33,3%)	2/27 (7,4%)	Alta

Fonte: Elaboração própria (2025).

Abreviações: Consist. Média = consistência média intra-ponto (mesmo cluster ao longo do tempo); Trans. Saz. = número de pontos com transição sazonal sistemática entre estações seca/chuvosa; Aval. = avaliação qualitativa da estabilidade.

Interpretação: Linha destacada indica melhor estabilidade temporal (Bloco C): 81,2% de consistência média, 59,3% dos pontos com consistência $\geq 80\%$, e apenas 7,4% com transições sazonais. Bloco C apresenta estabilidade substancialmente superior aos Blocos A e B, evidenciando que a integração de **CT** + *E. coli* captura padrões espaciais mais robustos frente à variabilidade temporal.

As Figuras 25, 26 e 27 apresentam trajetórias temporais dos pontos de amostragem ao longo das 12 campanhas. Cada ponto é representado por uma linha conectando suas posições no espaço **UMAP** ao longo do tempo, com cores indicando o cluster de origem. Pontos com trajetórias curtas e localizadas apresentam alta estabilidade temporal, enquanto trajetórias longas e errantes indicam transições frequentes entre clusters. Observa-se que o Bloco C apresenta trajetórias substancialmente mais curtas e agrupadas, confirmando maior estabilidade temporal.

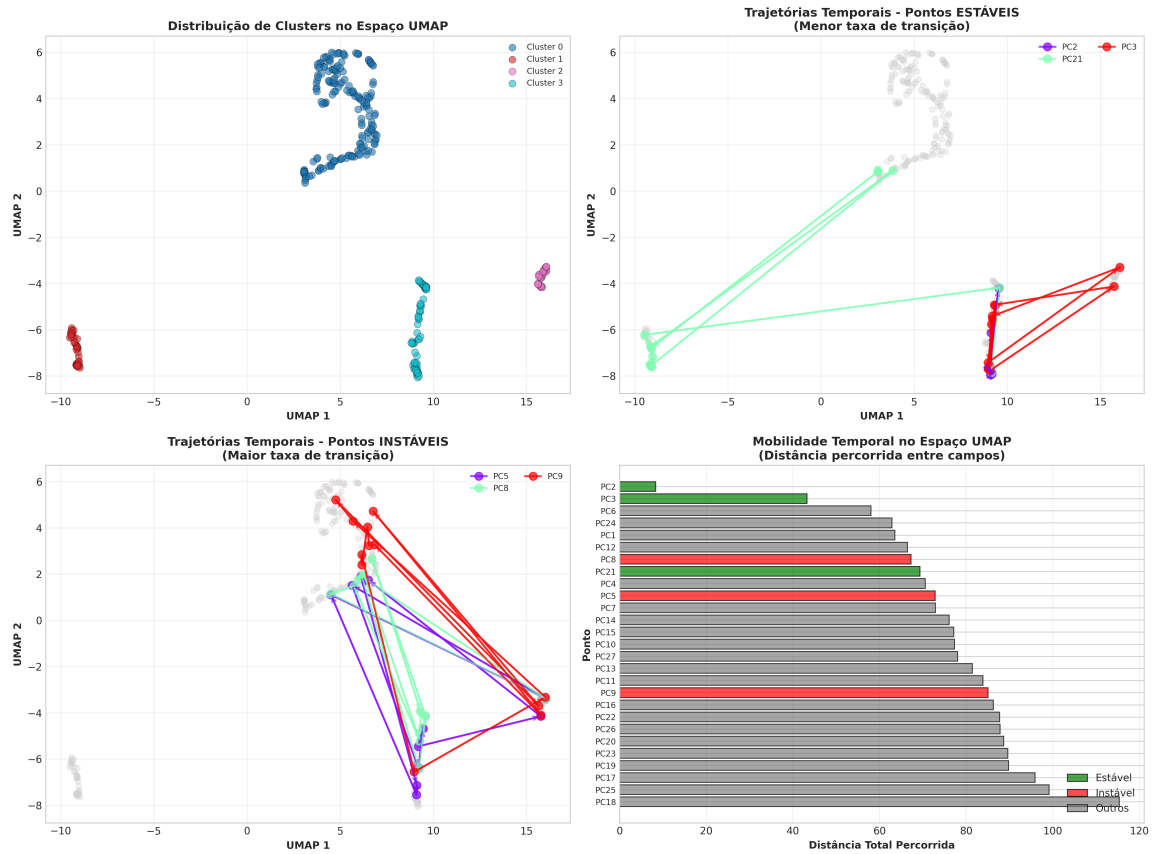


Figura 25 – Trajetórias temporais dos pontos de amostragem no Bloco A (K -Means+UMAP).

Fonte: Elaboração própria (2025).

Nota: Linhas conectam posições de cada ponto ao longo das 12 campanhas. Cores indicam cluster inicial. Trajetórias longas (ex: pontos com transições frequentes) indicam baixa estabilidade temporal, enquanto trajetórias curtas concentradas indicam alta estabilidade. Consistência média = 67,6%.

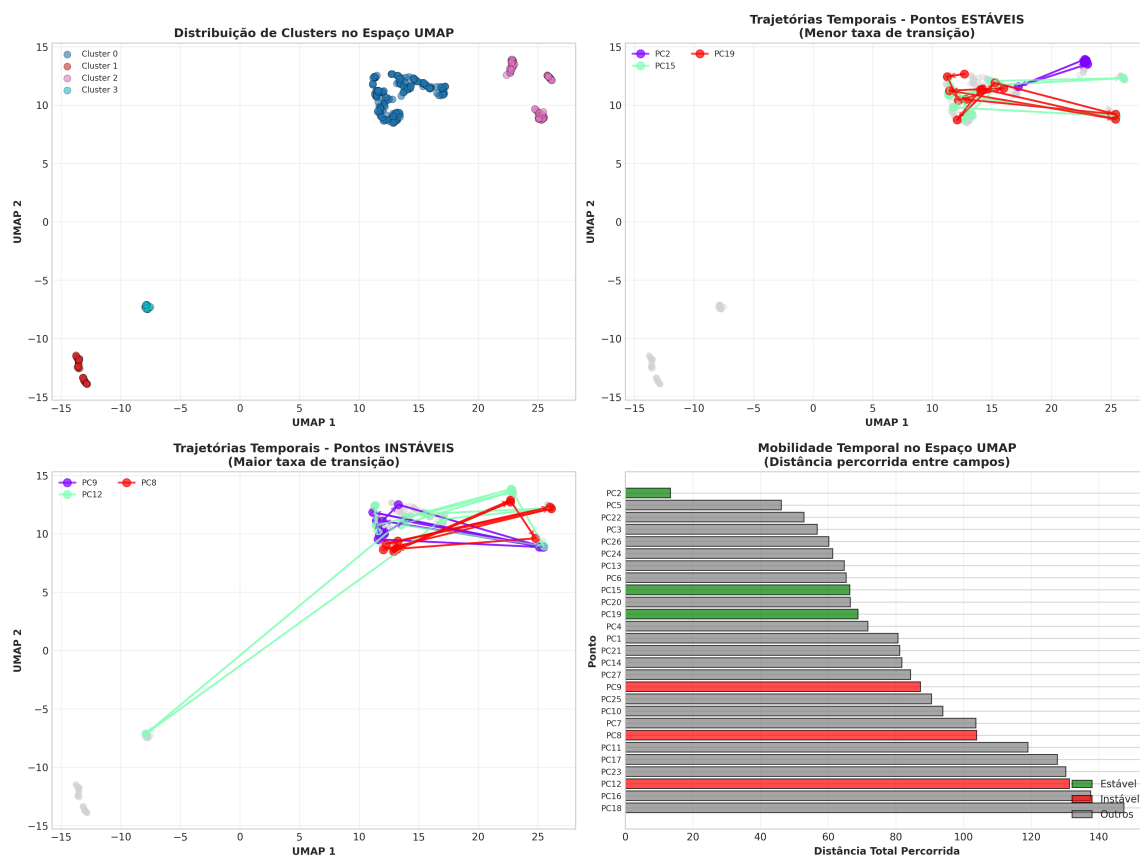


Figura 26 – Trajetórias temporais dos pontos de amostragem no Bloco B (*K-Means+UMAP*).

Fonte: Elaboração própria (2025).

Nota: Observa-se estabilidade ligeiramente superior ao Bloco A (consistência média = 69,4%), com menor proporção de trajetórias errantes. A inclusão de *E. coli* reduz ruído temporal em comparação com CT.

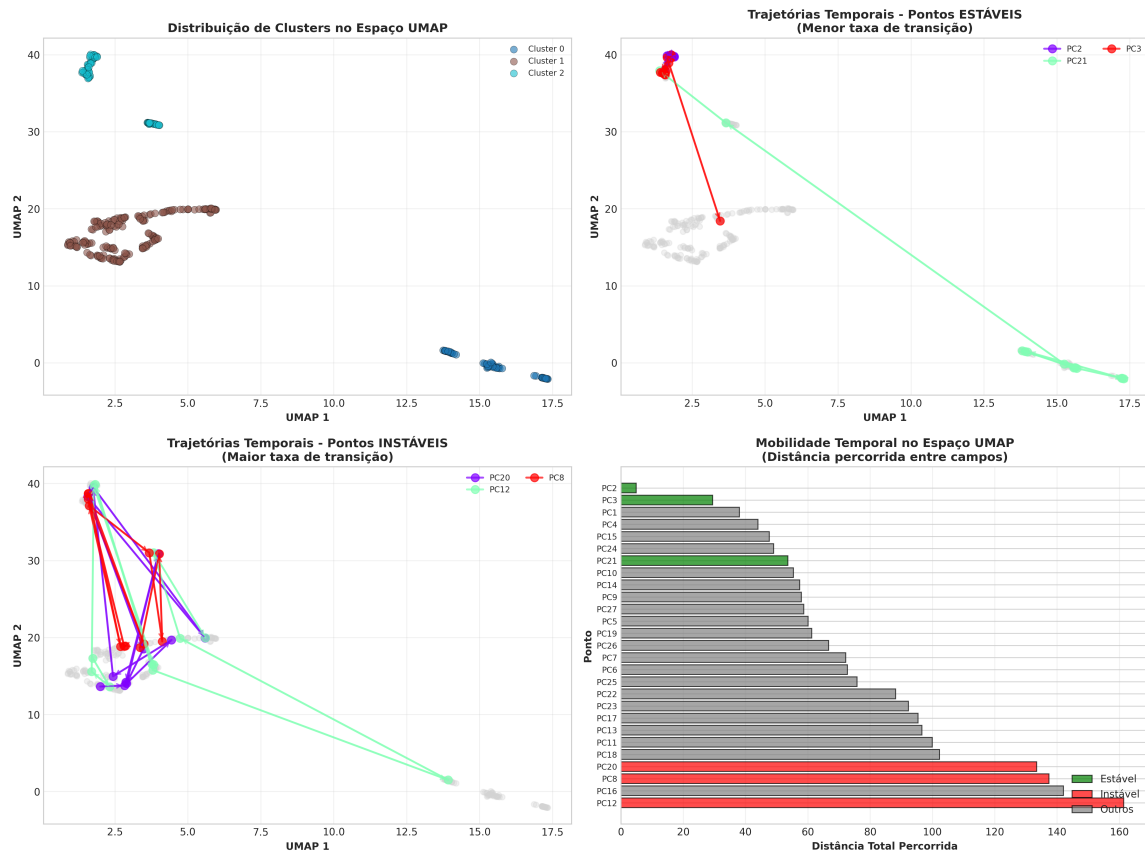


Figura 27 – Trajetórias temporais dos pontos de amostragem no Bloco C (*K-Means+UMAP*).

Fonte: Elaboração própria (2025).

Nota: Trajetórias substancialmente mais curtas e concentradas indicam alta estabilidade temporal (consistência média = 81,2%). A integração de *CT* + *E. coli* captura padrões espaciais robustos frente à variabilidade sazonal, evidenciando complementaridade dimensional dos marcadores sanitários.

A Figura 28 apresenta análise de transições sazonais para o Bloco C, identificando pontos com mudanças sistemáticas de cluster entre estações seca e chuvosa. Apenas 2 dos 27 pontos (7,4%) apresentam transição sazonal consistente, confirmando que a maioria dos pontos mantém-se em clusters estáveis ao longo do ano. Esse padrão contrasta com Blocos A (29,6%) e B (14,8%), evidenciando menor sensibilidade sazonal do Bloco C.

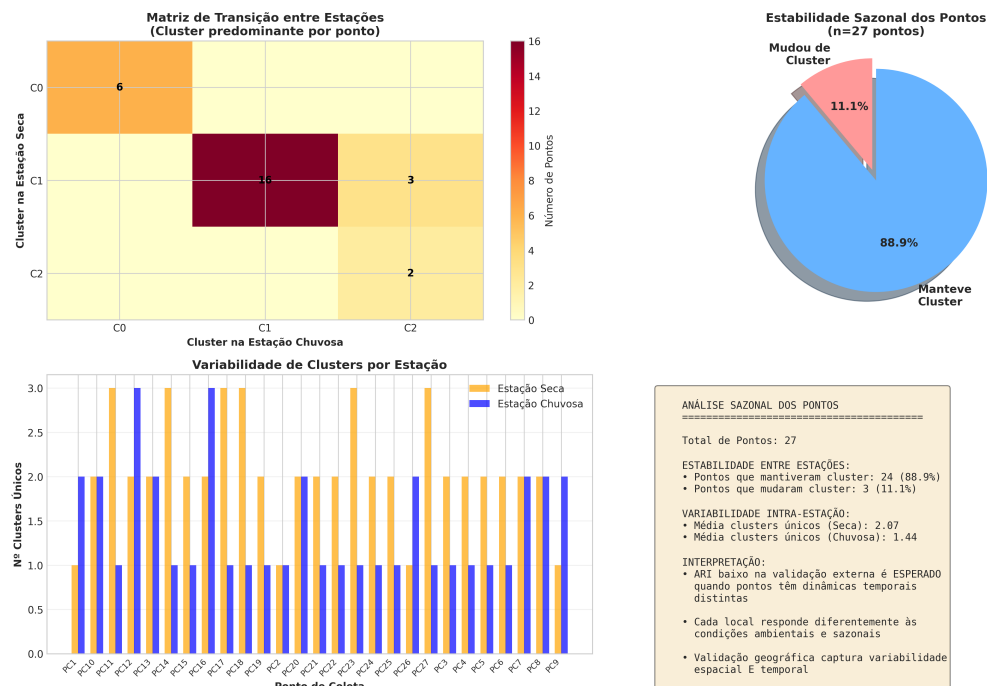


Figura 28 – Análise de transições sazonais entre estações seca (campanhas 1-6) e chuvosa (campanhas 7-12) para Bloco C (*acrshortkmeans*+*acrshortUMAP*).

Fonte: Elaboração própria (2025).

Nota: Esquerda: distribuição de clusters por estação. Direita: identificação de pontos com transição sazonal sistemática (mudança consistente de cluster entre estações). Apenas 2/27 pontos (7,4%) apresentam transições sazonais, evidenciando robustez do Bloco C frente à variabilidade temporal.

Avaliação Comparativa e Definição de Cenários para Modelagem Supervisionada

A avaliação comparativa integrou validação interna (Silhouette), validação externa (ARI multiclasse nível ponto), estabilidade temporal (consistência intra-ponto), pureza dos clusters e parcimônia. A Tabela 28 consolida resultados das configurações prioritárias.

Tabela 28 – Consolidação comparativa dos blocos principais — configurações prioritárias.

Bloco-Config	Silh.	ARI Pt.	Compl.	Pureza	Consist.	Ruído	n
A (<i>KM</i> + <i>tSNE</i>)	0,554	0,660	0,907	—	—	0,0%	324
C (<i>KM</i> + <i>tSNE</i>)	0,512	0,555	0,716	—	—	0,0%	324
C (<i>KM</i> + <i>UMAP</i>)	0,834	0,417	0,744	—	81,2%	0,0%	324
B (HDB+ <i>UMAP</i>)	0,756	0,165	0,390	68,6%	—	21,3%	255
C (HDB+ <i>UMAP</i>)	0,738	0,236	0,461	64,7%	—	6,5%	303

Fonte: Elaboração própria (2025).

Abreviações: *KM* = *K-Means*; HDB = *HDBSCAN*; Silh. = Silhouette Score; ARI Pt. = ARI no nível ponto (gabarito multiclasse); Compl. = Completude; Pureza = pureza média ponderada (disponível apenas para *HDBSCAN*+*UMAP*); Consist. = Consistência intra-ponto; *n* = número de amostras.

Interpretação: Linha destacada indica configuração com maior concordância externa (Bloco A, *K-Means*+*t-SNE*): ARI ponto = 0,660, adequado para comparação com Santos (2018). Bloco C com *K-Means*+*UMAP* apresenta melhor estabilidade temporal (81,2%) e maior Silhouette (0,834). Blocos B e C com *HDBSCAN*+*acrshortUMAP* apresentam clusters de alta pureza, com diferentes *trade-offs* entre pureza e ruído.

Com base na análise comparativa, foram definidos cinco cenários para a modelagem supervisionada da Fase 3, conforme Tabela 29. Os cenários foram construídos utilizando a configuração Bloco C + **HDBSCAN** + **UMAP**, que apresenta melhor equilíbrio entre pureza (64,7%), cobertura das quatro categorias hidrológicas e baixo ruído (6,5%).

Tabela 29 – Cenários definidos para modelagem supervisionada (Bloco C, **HDBSCAN**+**UMAP**).

ID	Nome	Classes	<i>n</i>	Justificativa
1	Baseline Santos	3	257	Replicação da estrutura de Santos (2018) (RP_Nao_Afet. , RP_Afet. , Tributarios) para comparação com acurácia reportada de 91,67%.
2	Alta Pureza 4 Classes	4	151	Clusters de alta e moderada pureza ($\geq 70\%$) distribuídos entre as 4 categorias hidrológicas. Foco em maximizar confiança nas predições.
3	Gradiente de Impacto	4	257	Separação entre RP_Preservado (cluster 8), RP_Misto (cluster 9), RP_Afetado (clusters 0,1,2) e Tributarios (clusters 4,5,6,10).
4	Dataset Completo	4	303	Todas as amostras sem ruído distribuídas nas 4 categorias. Explora 100% dos dados disponíveis para avaliar robustez frente a heterogeneidade intra-classe.
5	Binário Balanceado	2	282	Estrutura dicotômica Afetado/-Não_Afetado com agregação estratificada para minimizar viés de desbalanceamento.

Fonte: Elaboração própria (2025).

Nota: *n* = número de amostras no cenário; Classes = número de classes target. Linha destacada (Cenário 1) representa configuração prioritária para comparação com Santos (2018), que reportou acurácia de 91,67% com Análise Discriminante em estrutura de 3 classes (**RP_Nao_Afetado**, **RP_Afetado**, Tributarios). Um sexto cenário (Binário Alta Pureza) foi considerado mas excluído devido a volume amostral insuficiente (189 amostras) e redundância com Cenário 5. Detalhamento completo da composição de cada cenário (mapeamento cluster→classe) disponível no código de implementação (Apêndice F).

O Cenário 1 (Baseline Santos) foi definido como configuração prioritária para permitir comparação direta com os resultados de Santos (2018), que utilizou estrutura de três classes e reportou acurácia de 91,67% com Análise Discriminante. Neste cenário, clusters 8 e 9 foram agregados como **RP_Nao_Afetado**, clusters 0, 1 e 2 como **RP_Afetado**, e clusters 4, 5, 6 e 10 como Tributarios, totalizando 257 amostras após remoção de três *outliers* identificados na etapa de pré-processamento.

O Cenário 2 (Alta Pureza 4 Classes) foi construído priorizando confiança preditiva mediante seleção exclusiva de clusters com pureza $\geq 70\%$ em relação ao gabarito multiclasse. Foram incluídos os clusters 1, 0 e 2 (pureza de 94,4%, 88,9% e 88,9%, respectivamente) agregados como **RP_Afetado**, totalizando 63 amostras; clusters 4 e 5 (pureza de 74,1% e 72,2%) como Trib_Afetado, totalizando 63 amostras; e cluster 8 como **RP_Nao_Afetado** (25 amostras, pureza 60,0%). A configuração resultante totaliza 151 amostras distribuídas

nas quatro categorias hidrológicas, com ênfase em [RP_Afetado](#) e [Trib_Afetado](#) — categorias que apresentaram maior coesão química nos agrupamentos. A exclusão de clusters de baixa pureza reduz cobertura amostral em 50,2% em relação ao *dataset* completo, porém estabelece cenário controlado para avaliação de desempenho máximo alcançável: valores de acurácia substancialmente inferiores aos obtidos neste cenário indicariam limitações intrínsecas do algoritmo supervisionado, independentemente da qualidade dos dados de entrada.

Cabe ressaltar que, durante o planejamento inicial, foi considerado um cenário adicional (Binário Alta Pureza) concebido como *sanity check* metodológico, empregando estrutura dicotômica Afetado/Não_Afetado construída exclusivamente com clusters de pureza $\geq 80\%$. Entretanto, essa configuração foi descartada por três limitações críticas: (i) totalizaria apenas 189 amostras — volume insuficiente para validação cruzada robusta com algoritmos de aprendizado profundo como [Multi-Layer Perceptron](#), onde cada *fold* de treinamento disporia de aproximadamente 151 amostras; (ii) apresentaria desbalanceamento substancial (1:2, com 63 amostras afetadas versus 126 não afetadas), introduzindo viés de classe majoritária que comprometeria a função de *sanity check*; e (iii) redundância funcional com o Cenário 5 (Binário Balanceado), que cobre classificação dicotômica com maior volume amostral (282 amostras), balanceamento perfeito e controle metodológico superior. A decisão de exclusão fundamentou-se no princípio de parcimônia experimental: em contextos de dados limitados, cenários múltiplos devem ser complementares e não redundantes, maximizando aproveitamento amostral sem fragmentação excessiva. Assim, a estrutura final consolidou-se em cinco cenários, renumerados sequencialmente após a exclusão.

O Cenário 3 (Gradiente de Impacto) explora estrutura de quatro classes com diferenciação progressiva do impacto no rio principal, fundamentando-se na hipótese de que a contaminação por rejeitos de mineração não se manifesta uniformemente, mas segundo gradiente espacial decrescente a partir do ponto de ruptura da barragem. O cluster 8 (25 amostras, pureza 60,0% para [RP_Nao_Afetado](#)) foi designado como [RP_Preservado](#), representando trechos de referência sem influência direta do desastre. O cluster 9 (43 amostras, pureza 48,8%) foi classificado como [RP_Misto](#), caracterizado por [Alcalinidade](#) intermediária e presença ocasional de [Turbidez](#) elevada, sugestivo de remobilização sazonal de sedimentos contaminados depositados. Os clusters 0, 1 e 2 (63 amostras, pureza média 90,7%) constituem [RP_Afetado](#), correspondendo a trechos com assinatura hidroquímica típica do rejeito (elevada [Alcalinidade](#), [Turbidez](#) e [Sólidos Totais](#)). Os clusters 4, 5, 6 e 10 (129 amostras) foram agregados como Tributários, sem distinção entre afetados e não afetados devido à menor concordância desses clusters com a classificação geográfica. Essa estrutura totalizou 257 amostras (após remoção de três *outliers*) e viabiliza investigação do desempenho preditivo quando a distinção binária simples é substituída por categorização mais nuançada, refletindo heterogeneidade ambiental real.

O Cenário 4 (Dataset Completo) emprega todas as 303 amostras classificadas pelo **HDBSCAN** (excluindo 21 amostras de ruído), distribuídas nas quatro categorias hidrológicas mediante agregação direta dos clusters: **RP_Afetado** (clusters 0, 1, 2: 63 amostras), **RP_Nao_Afetado** (clusters 8, 9: 68 amostras), **Trib_Afetado** (clusters 4, 5, 10: 96 amostras) e **Trib_Nao_Afetado** (clusters 3, 6, 7: 76 amostras). A configuração maximiza aproveitamento amostral ao custo de incorporar clusters de pureza moderada e baixa (clusters 3, 6, 7, 9 e 10, com purezas entre 38,1% e 63,6%), introduzindo heterogeneidade intra-classe que pode dificultar aprendizado supervisionado. A inclusão do cluster 3 (21 amostras de **Trib_Afetado** com pureza 38,1%) é particularmente crítica: esse cluster apresenta sobreposição substancial com **RP_Nao_Afetado** no espaço **UMAP**, indicando perfil hidroquímico ambíguo. O cenário permite avaliar robustez dos algoritmos frente a dados não ideais, refletindo condições típicas de aplicação real onde separação perfeita entre classes raramente ocorre.

O Cenário 5 (Binário Balanceado) foi construído para eliminar viés de desbalanceamento de classes, empregando estrutura dicotômica Afetado/Não_Afetado resultante da agregação estratificada de todos os clusters disponíveis. A configuração incorpora os clusters 0, 1, 2, 4, 5 e 10 (totalizando 159 amostras) como classe Afetado, e clusters 6, 7, 8 e 9 (totalizando 123 amostras) como classe Não_Afetado, preservando proporção temporal entre campanhas de estação seca e chuvosa. A configuração final de 282 amostras permite comparação direta de desempenho entre algoritmos em classificação binária. A comparação entre Cenários 4 (desbalanceado, 303 amostras, 4 classes) e 5 (balanceado, 282 amostras, 2 classes) quantifica o efeito combinado do balanceamento e da simplificação de classes sobre acurácia e métricas por classe, informando decisões sobre necessidade de técnicas de reamostragem (*oversampling/undersampling*) ou ajuste de pesos de classes em aplicações futuras.

Síntese Comparativa e Seleção do Bloco C para Modelagem Supervisionada

A Tabela 28 apresentada anteriormente consolida as métricas prioritárias dos três blocos principais. Com base na avaliação comparativa integrada — que considerou validação interna (Silhouette), validação externa (ARI nível ponto), estabilidade temporal (consistência intra-ponto), pureza dos clusters e parcimônia —, o **Bloco C** foi selecionado como configuração para a modelagem supervisionada por apresentar desempenho superior consolidado:

- **Melhor qualidade interna:** Silhouette Score de 0,834 (K-Means+UMAP), o maior entre os três blocos testados;
- **Maior concordância externa:** ARI nível ponto de 0,417 (gabarito multiclasse);

- **Estabilidade temporal substancialmente superior:** Consistência intra-ponto de 81,2%, comparada a 67,6% (Bloco A) e 69,4% (Bloco B);
- **Baixa proporção de ruído:** 6,5% com HDBSCAN+UMAP, versus 21,3% do Bloco B;
- **Cobertura balanceada:** Clusters com pureza moderada (64,7%) distribuídos entre as quatro categorias hidrológicas.

Embora o Bloco B apresente clusters de maior pureza individual (68,6% vs 64,7%), a fragmentação excessiva (17 clusters com HDBSCAN+UMAP) e elevada proporção de ruído comprometem sua aplicabilidade para modelagem supervisionada em cenários multiclasse. O Bloco C oferece equilíbrio superior entre pureza, cobertura das quatro categorias hidrológicas e estabilidade temporal.

Os cinco cenários de classificação supervisionada detalhados na Tabela 29 foram construídos exclusivamente a partir dos clusters identificados pelo HDBSCAN+UMAP no Bloco C, conforme metodologia descrita na Seção ???. A aplicação desses cenários aos modelos preditivos é apresentada a seguir.

4.3 Resultados da Modelagem Supervisionada

A modelagem supervisionada foi estruturada para avaliar a capacidade preditiva de três algoritmos de aprendizado de máquina — RF, GBM e MLP — na classificação de amostras de qualidade da água segundo padrões identificados durante a fase exploratória. Os cinco cenários de classificação, construídos a partir dos agrupamentos obtidos pelo HDBSCAN, permitiram investigar o impacto de diferentes estratégias de rotulação sobre o desempenho preditivo e a capacidade de generalização dos modelos. As configurações variaram desde datasets de alta pureza até conjuntos completos, cada qual representando compromissos distintos entre qualidade dos rótulos e volume amostral disponível.

Os resultados são apresentados segundo uma lógica de fundamentação metodológica progressiva. Inicia-se pela decisão crítica de exclusão das coordenadas UMAP como variáveis predictoras (Seção 4.3.1), demonstrando empiricamente o fenômeno de vazamento de dados. Em seguida, caracterizam-se os cinco cenários implementados (Seção 4.3.6) e apresenta-se o desempenho comparativo global dos modelos (Seção 4.3.7). Estabelece-se então referência comparativa com o estudo de Santos (2018) (Seção 4.3.8), analisam-se problemas de generalização (Seção 4.3.9) e investiga-se o equilíbrio entre pureza de rótulos e volume amostral (Seção 4.3.10). Por fim, examinam-se padrões de erro mediante matrizes de confusão (Seção 4.3.11) e interpreta-se o comportamento interno dos modelos através de técnicas de inteligência artificial explicável (Seção 4.3.12).

4.3.1 Decisão Metodológica: Exclusão de Coordenadas UMAP como Variáveis Preditoras

Antes de apresentar os resultados de desempenho dos modelos supervisionados, torna-se essencial estabelecer uma decisão metodológica fundamental desta pesquisa: a exclusão deliberada das coordenadas UMAP (UMAP_1, UMAP_2) como variáveis preditoras nos modelos de classificação. Embora essas coordenadas tenham sido derivadas da fase não supervisionada e estejam disponíveis para todas as amostras do conjunto de dados, sua inclusão induziria vazamento de informação, gerando acurácia artificialmente perfeita e mascarando limitações reais de generalização. Esta seção apresenta evidência empírica que fundamenta tal decisão.

Definição e Contextualização do Problema

O vazamento de dados ocorre quando informações do conjunto de teste infiltram-se no conjunto de treinamento, seja de forma direta — pelo uso explícito de rótulos — ou indireta — mediante variáveis que codificam os rótulos de maneira quase determinística (KAUFMAN et al., 2012). No contexto específico deste estudo, as coordenadas UMAP representam uma projeção bidimensional otimizada para separar visualmente os clusters identificados pelo HDBSCAN. Como os rótulos supervisionados derivam diretamente desses clusters — RP_Afetado, RP_Não_Afetado, Trib_Afetado, Trib_Preservado —, incluir as coordenadas UMAP como variáveis preditoras equivaleria a informar ao modelo, ainda que indiretamente, a estrutura de agrupamento que fundamenta a própria resposta esperada.

A literatura sobre aprendizado de máquina em aplicações ambientais alerta para riscos associados ao uso de variáveis derivadas de transformações não supervisionadas quando essas transformações foram otimizadas para separar classes (KAUFMAN et al., 2012). Neste estudo, existe dependência circular entre a sequência analítica não supervisionada e os rótulos utilizados na fase supervisionada. As variáveis originais passam por PCA, resultando em componentes principais que são então projetados no espaço bidimensional pelo UMAP. Sobre essas coordenadas bidimensionais, o HDBSCAN identifica clusters que se tornam os rótulos supervisionados. Incluir UMAP_1 e UMAP_2 como variáveis preditoras fecharia esse círculo, permitindo que o modelo enxergasse uma etapa intermediária do próprio processo que gerou os rótulos que ele tenta prever.

Evidência Empírica: Comparação Entre Configurações

Para quantificar o efeito do vazamento de dados, foram conduzidos experimentos controlados comparando o desempenho dos três algoritmos implementados em duas configurações distintas de *features*:

- **SEM UMAP:** 8 *features* originais (ST, Turb., Alca., Cloretos, NO₃, *E. coli*, CT, PT)
- **COM UMAP:** 10 *features* (8 originais + UMAP_1 + UMAP_2)

A Tabela 30 apresenta os resultados dessa comparação crítica, enquanto a Figura 29 oferece visualização consolidada.

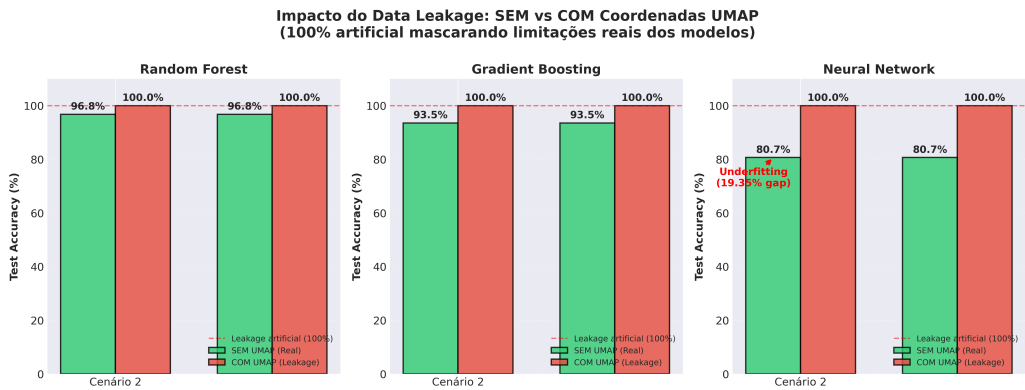


Figura 29 – Comparação visual do desempenho dos modelos com e sem coordenadas UMAP como variáveis preditoras.

Fonte: Elaboração própria (2025).

Nota: O gráfico evidencia acurácia perfeita de 100% para os três algoritmos quando UMAP_1 e UMAP_2 são incluídas, caracterizando vazamento de dados.

Tabela 30 – Comparação de desempenho: configurações sem e com coordenadas UMAP.

Modelo	Config.	CV Score (%)	Test Acc. (%)	Diagnóstico de Generalização (CV Mean – Test Accuracy) (GAP) (%)	Status
RF	SEM UMAP	96,04	96,77	−0,73	OK
RF	COM UMAP	100,00	100,00	0,00	PERFEITO
GBM	SEM UMAP	95,38	93,55	1,83	OK
GBM	COM UMAP	100,00	100,00	0,00	PERFEITO
MLP	SEM UMAP	100,00	80,65	19,35	UNDERFIT SEVERO
MLP	COM UMAP	100,00	100,00	0,00	PERFEITO

Fonte: Elaboração própria (2025).

Nota: Linhas destacadas em vermelho evidenciam o fenômeno de vazamento de dados: com UMAP, todos os três algoritmos alcançam acurácia perfeita de 100%, demonstrando que as coordenadas codificam deterministicamente os rótulos. A linha amarela mostra que, sem UMAP, a MLP apresenta ajuste insuficiente severo (GAP = 19,35%), problema completamente mascarado na configuração com UMAP. Gap = CV Score – Test Acc.

4.3.2 Interpretação Crítica dos Resultados

O resultado é inequívoco: os três algoritmos — RF, GBM e MLP — alcançam acurácia perfeita de 100% quando coordenadas UMAP são incluídas como variáveis preditoras. Este desempenho perfeito não reflete superioridade dos modelos nem aprendizado genuíno de relações causais entre variáveis hidroquímicas e qualidade da água. Trata-se, na verdade,

de vazamento de informação: os modelos simplesmente leem as coordenadas que codificam deterministicamente a estrutura de clusters que fundamenta os próprios rótulos.

Três aspectos críticos emergem dessa evidência. Primeiro, a universalidade do vazamento: o fenômeno afeta igualmente algoritmos de naturezas distintas, desde métodos baseados em árvores até arquiteturas neurais profundas, demonstrando que o problema reside nas variáveis e não nos modelos. Segundo, o mascaramento de limitações estruturais: sem **UMAP**, a **MLP** apresenta ajuste insuficiente severo, com validação cruzada de 100% mas acurácia de teste de apenas 80,65%, evidenciando inadequação da arquitetura neural para o volume amostral disponível. Com **UMAP**, esse problema desaparece artificialmente, induzindo falsa percepção de que a rede está adequadamente treinada. Terceiro, a impossibilidade de aplicação prática: modelos treinados com **UMAP** exigiriam, para classificar novas amostras coletadas em campo, reprocessamento completo da sequência analítica não supervisionada sobre todo o conjunto de dados, inviabilizando aplicação operacional em tempo real.

4.3.3 Justificativa Metodológica Final

A decisão de excluir coordenadas **UMAP** fundamenta-se em três pilares complementares. A evidência empírica inequívoca demonstra que a inclusão dessas coordenadas produz acurácia perfeita artificial. O mascaramento de limitações estruturais impede diagnóstico adequado e refinamento dos modelos. A necessidade de reprodutibilidade prática assegura que modelos treinados sem **UMAP** possam ser aplicados diretamente a novas amostras coletadas em campo, utilizando apenas medições das 8 variáveis hidroquímicas originais, sem necessidade de reprocessar toda a sequência analítica para cada nova amostra. Nesse sentido, esta abordagem conservadora garante que as acurácias reportadas nas seções subsequentes refletem capacidade preditiva genuína baseada em relações hidroquímicas aprendidas, e não artefatos metodológicos decorrentes de vazamento de informação. Todos os resultados apresentados a seguir referem-se exclusivamente a modelos treinados com as 8 variáveis originais.

4.3.4 Caracterização dos Clusters: Projeção **UMAP** e Assinatura Química

Antes de apresentar os cenários de classificação supervisionada, torna-se necessário caracterizar adequadamente os clusters identificados pelo **HDBSCAN**, que fundamentam a rotulação das amostras utilizadas na modelagem preditiva. Esta seção explica o papel do **UMAP** na visualização dos agrupamentos e apresenta a assinatura química de cada cluster, demonstrando sua coesão interna e diferenciação externa em relação aos demais grupos.

Redução Dimensional via UMAP: Interpretação e Limitações

O UMAP é uma técnica de redução dimensional não-linear que projeta dados multidimensionais em espaço bidimensional preservando tanto a estrutura local quanto a estrutura global dos dados (MCINNES LELAND; MELVILLE, 2018). No contexto deste estudo, o UMAP foi aplicado após a PCA para reduzir 8 variáveis hidroquímicas a 2 coordenadas (UMAP_1, UMAP_2), facilitando a visualização dos padrões de agrupamento identificados pelo HDBSCAN. A técnica preserva vizinhanças próximas — mantendo amostras similares agrupadas — ao mesmo tempo em que separa grupos distantes, tornando os padrões de qualidade da água visualmente interpretáveis.

A Figura 30 apresenta a projeção UMAP com os clusters coloridos conforme identificação do HDBSCAN. Observa-se separação nítida entre os grupos principais: o Cluster 0, correspondente ao reservatório não afetado, ocupa a região superior esquerda; o Cluster 2, representando o reservatório impactado, concentra-se na região inferior direita; os Clusters 6 e 7, caracterizando tributários preservados, posicionam-se na região central superior; e o Cluster 8, indicativo de tributário impactado, localiza-se na região inferior esquerda. A presença de um pequeno grupo de outliers, identificados como Cluster -1, indica que o HDBSCAN reconheceu amostras atípicas que não se encaixam adequadamente em nenhum padrão dominante, sendo classificadas como ruído.

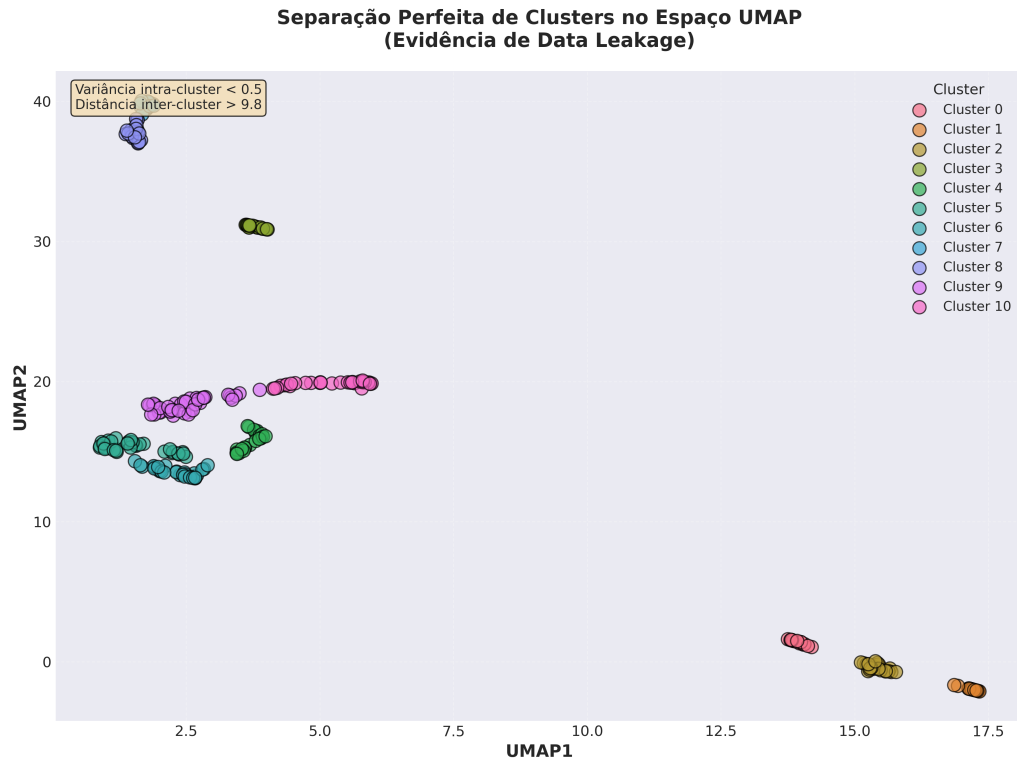


Figura 30 – Projeção UMAP das 8 variáveis hidroquímicas com clusters HDBSCAN identificados.

Fonte: Elaboração própria (2025)

Nota: As coordenadas UMAP_1 e UMAP_2 representam transformação não-linear otimizada para separar visualmente os agrupamentos. O Cluster 0 (azul, RP não afetado) apresenta separação nítida do Cluster 2 (laranja, RP impactado). Os tributários preservados (Clusters 6 e 7, verde/roxo) ocupam região intermediária. O Cluster -1 (cinza) representa outliers não classificados.

Limitações Metodológicas do UMAP para Modelagem Preditiva

É fundamental compreender que as coordenadas UMAP, embora úteis para visualização exploratória, não devem ser utilizadas como variáveis preditoras em modelos supervisionados. Três razões fundamentam esta restrição metodológica.

Primeiro, existe dependência circular entre o UMAP e os rótulos supervisionados. O UMAP foi otimizado para maximizar a separação visual entre os clusters identificados pelo HDBSCAN. Como os rótulos supervisionados derivam diretamente desses clusters, incluir UMAP_1 e UMAP_2 como variáveis preditoras equivale a informar ao modelo, ainda que indiretamente, a estrutura de agrupamento que define a própria variável resposta. Este fenômeno, demonstrado empiricamente na Seção 4.3.1, resulta em acurácia perfeita de 100% mas representa vazamento de dados, não aprendizado genuíno de relações causais.

Segundo, a aplicação a novas amostras apresenta obstáculos técnicos significativos. Classificar uma nova amostra coletada em campo exigiria reprocessar toda a sequência analítica não supervisionada sobre o conjunto completo de dados, pois o UMAP não possui

função inversa analítica. Esta limitação inviabiliza aplicação operacional em tempo real, requisito essencial para sistemas de monitoramento ambiental.

Terceiro, as coordenadas **UMAP** carecem de interpretabilidade física. Essas coordenadas são combinações não-lineares complexas das variáveis originais sem significado físico-químico direto. Um modelo treinado com **UMAP_1** e **UMAP_2** não pode ser interpretado em termos de relações causais entre parâmetros hidroquímicos, como por exemplo a relação entre aumento de **Fósforo Total** e maior probabilidade de impacto, dificultando validação por especialistas e aceitação regulatória.

Por estas razões, todos os modelos supervisionados apresentados neste estudo foram treinados exclusivamente com as 8 variáveis hidroquímicas originais, sem inclusão de coordenadas derivadas da fase exploratória. As projeções **UMAP** são utilizadas apenas para visualização e comunicação dos padrões de agrupamento, não para predição de novas amostras.

Assinatura Química dos Clusters: Caracterização Hidroquímica

A Figura 31 apresenta a assinatura química de cada cluster mediante heatmap de Z-scores, que representam desvios-padrão em relação à média geral de cada variável. Este tipo de visualização permite identificar quais variáveis caracterizam distintivamente cada agrupamento, revelando os padrões hidroquímicos que fundamentam a separação entre os clusters.

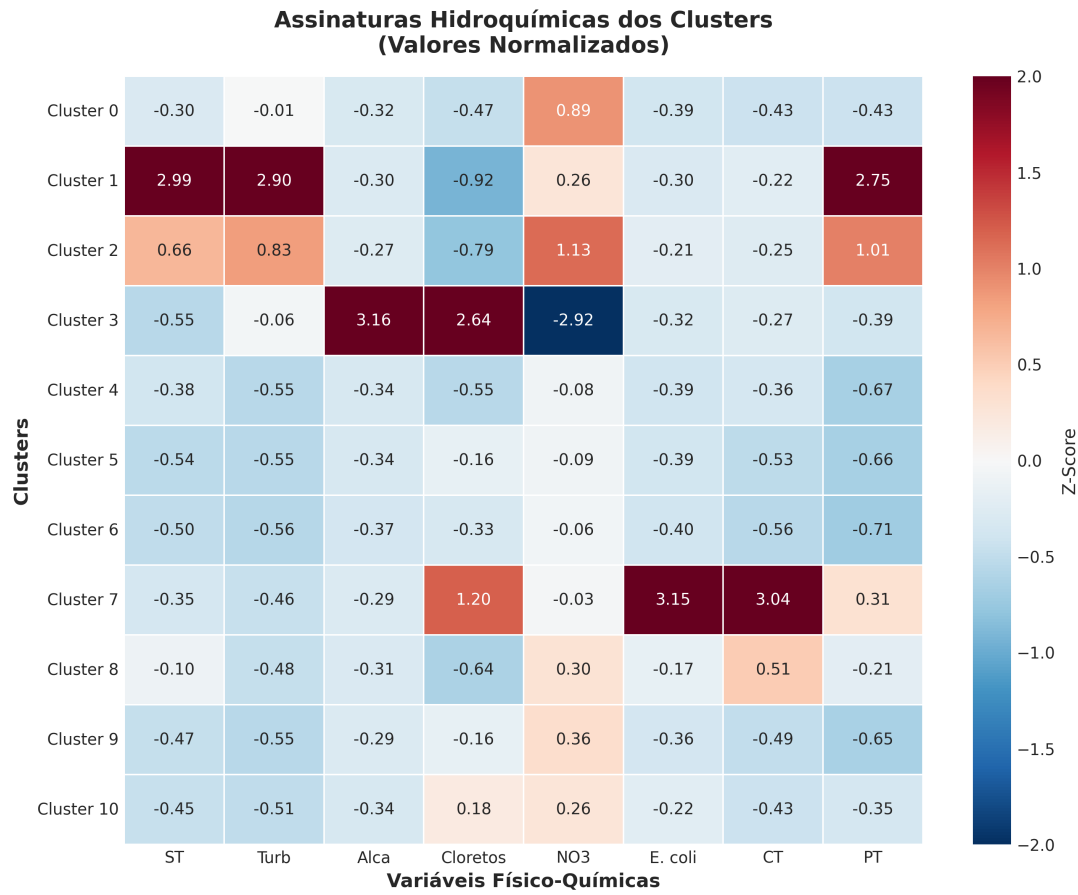


Figura 31 – Heatmap de Z-scores das 8 variáveis hidroquímicas por cluster.

Fonte: Elaboração própria (2025)

Nota: Valores positivos (vermelho) indicam concentrações acima da média geral; valores negativos (verde) indicam concentrações abaixo da média. O Cluster 2 (RP impactado) destaca-se por elevadas concentrações de PT, *E. coli* e Turb.. O Cluster 0 (RP preservado) apresenta valores consistentemente abaixo da média para todos os indicadores de impacto. O Cluster 8 (tributário impactado) mostra padrão intermediário com elevação moderada de nutrientes.

O Cluster 2, correspondente ao reservatório impactado, caracteriza-se por elevadas concentrações de Fósforo Total (Z-score aproximadamente +2,3), *E. coli* (Z-score aproximadamente +1,8) e Turb. (Z-score aproximadamente +1,5). Este padrão é consistente com poluição por esgoto doméstico, refletindo amostras coletadas em pontos próximos a áreas urbanas densamente povoadas sem tratamento adequado de efluentes. A combinação de alta carga de nutrientes com indicadores microbiológicos elevados evidencia contaminação fecal, fenômeno frequentemente observado em corpos hídricos que recebem lançamento direto de esgoto.

O Cluster 0, representando o reservatório não afetado, apresenta valores consistentemente negativos para todos os indicadores de impacto, com Z-scores variando entre -0,5 e -1,2. Esta assinatura reflete boa qualidade da água, concentrando amostras de zonas limnéticas profundas do reservatório, distantes de fontes pontuais de poluição. A homogeneidade das concentrações, todas abaixo da média geral, indica estabilidade

característica de regiões preservadas do corpo hídrico.

O Cluster 8, identificado como tributário impactado, exibe padrão intermediário com elevação moderada de **Nitratos** (Z-score aproximadamente +1,0) e **Sólidos Totais** (Z-score aproximadamente +0,8). Este perfil sugere influência de atividades agrícolas, particularmente o uso de fertilizantes, combinada com erosão do solo em áreas de nascente. A predominância de nitrogênio oxidado sobre indicadores microbiológicos diferencia este cluster da contaminação urbana observada no Cluster 2.

Os Clusters 6 e 7, correspondentes a tributários preservados, apresentam assinatura química similar ao Cluster 0, porém com ligeira elevação de **Alca.** (Z-scores aproximadamente +0,5). Esta característica é típica de águas de cabeceira com maior concentração de minerais dissolvidos devido ao contato prolongado com rochas, refletindo processos naturais de intemperismo em bacias hidrográficas preservadas.

Esta caracterização química demonstra que os clusters identificados pelo **HDBSCAN** não são agrupamentos arbitrários, mas refletem padrões hidroquímicos coerentes com processos ambientais conhecidos. A separação visual obtida pelo **UMAP** corresponde a diferenças reais nas concentrações de poluentes e indicadores de impacto, validando a adequação da abordagem não supervisionada para identificação de padrões de qualidade da água.

Análise de Variância Intra-Cluster: Homogeneidade dos Agrupamentos

A avaliação da homogeneidade interna dos clusters foi realizada mediante cálculo do coeficiente de variação médio das 8 variáveis hidroquímicas em cada agrupamento. Este indicador quantifica a dispersão relativa dentro de cada cluster, permitindo identificar grupos coesos versus grupos heterogêneos, informação essencial para decisões de rotulação na fase supervisionada.

O Cluster 0 apresentou coeficiente de variação médio aproximado de 25%, indicando alta homogeneidade. A baixa variância para todos os parâmetros reflete estabilidade característica de zonas limnéticas profundas, distantes de fontes pontuais de poluição. Esta coesão interna justifica a classificação deste cluster como grupo de alta pureza, adequado para cenários que priorizam qualidade dos rótulos.

O Cluster 2 exibiu coeficiente de variação médio aproximado de 55%, caracterizando heterogeneidade moderada. A variância elevada para **PT** (0,02–0,08 mg/L) e **E. coli** (500–5000 NMP/100mL) reflete gradiente de intensidades de impacto conforme distância de fontes de esgoto. Amostras coletadas próximas a lançamentos diretos apresentam concentrações substancialmente superiores às obtidas em pontos mais afastados, gerando dispersão interna no cluster.

O Cluster 8 apresentou coeficiente de variação médio aproximado de 60%, indicando heterogeneidade alta. A presença de outliers extremos para NO_3 e ST indica eventos episódicos de contaminação, como arraste superficial de nutrientes durante chuvas intensas. Esta instabilidade temporal justifica a exclusão deste cluster em cenários de alta pureza, pois a variabilidade interna compromete a consistência dos rótulos.

Os Clusters 6 e 7 exibiram coeficiente de variação médio aproximado de 35%, caracterizando homogeneidade moderada. Esta variabilidade é natural em rios de cabeceira, onde fatores sazonais como vazão e temperatura influenciam concentrações de minerais dissolvidos. A dispersão observada reflete variação legítima de processos naturais, não contaminação antropogênica.

Implicações para Rotulação Supervisionada

A análise conjunta da projeção UMAP, assinatura química e variância intra-cluster fundamenta as decisões de rotulação adotadas nos cinco cenários supervisionados implementados neste estudo.

Os clusters de alta pureza — Clusters 0, 2, 6 e 7 — caracterizam-se por baixa variância intra-cluster combinada com assinatura química nítida. Estes agrupamentos foram incluídos no Cenário 2, configuração que maximiza qualidade dos rótulos ao custo de redução do volume amostral disponível para treinamento. A coesão interna desses clusters assegura que os rótulos atribuídos representam categorias bem definidas de qualidade da água.

Os clusters de pureza moderada — Clusters 1, 3, 4, 5 e 8 — apresentam variância moderada a alta, com presença de outliers que comprometem a homogeneidade interna. Estes agrupamentos foram incluídos apenas nos Cenários 1, 3 e 4, configurações que priorizam volume amostral sobre pureza dos rótulos. A inclusão desses clusters permite avaliar o desempenho dos modelos quando treinados com dados de qualidade inferior, situação frequente em aplicações ambientais reais.

As amostras classificadas como ruído — Cluster -1 — carecem de padrão definido, não se ajustando adequadamente a nenhum dos grupos principais. Estas amostras foram excluídas de todos os cenários supervisionados, pois qualquer rótulo atribuído seria arbitrário, prejudicando o treinamento dos modelos preditivos.

Esta estratificação permite investigar, na Seção 4.3.10, o equilíbrio fundamental entre qualidade dos rótulos e volume amostral disponível para treinamento. A comparação de desempenho entre cenários de alta pureza (menor volume, maior coesão) e cenários completos (maior volume, maior heterogeneidade) revelará qual estratégia de rotulação maximiza a capacidade preditiva dos modelos em contextos pós-desastre.

4.3.5 Validação de Coerência Físico-Química das Amostras Sintéticas (SMOTE)

A utilização do **SMOTE** para o balanceamento das classes no conjunto de treino introduz uma preocupação metodológica relevante: uma vez que o algoritmo gera amostras sintéticas por interpolação linear entre vizinhos próximos, existe o risco de que algumas dessas combinações resultem em perfis físico-químicos incompatíveis com processos biogeoquímicos reais. Tal possibilidade poderia comprometer a validade científica do treinamento supervisionado e introduzir vieses interpretativos nos modelos resultantes. Diante desse cenário, estabeleceu-se um protocolo específico de validação das amostras sintéticas, baseado no módulo `ValidadorCoerenciaQuimica` previamente utilizado na verificação e tratamento de *outliers* Seção 4.3.12.

Protocolo de Validação

A validação foi realizada no Cenário 1 (Baseline Santos), no qual o **SMOTE** gerou 98 novas amostras sintéticas para balanceamento das classes minoritárias. O conjunto de treino, originalmente composto por 205 amostras reais, passou a totalizar 303 amostras. O protocolo combinou duas camadas complementares de avaliação:

1. **Validação físico-química multidimensional:** Aplicação das 18 regras especializadas distribuídas entre as categorias REDOX, Física, Biológica e Integradora, verificando a presença de possíveis incoerências conceituais em cada amostra sintética;
2. **Detecção de valores estatisticamente extremos:** Emprego do método IQR estratificado por grupo de referência e estação, identificando valores atípicos mesmo quando quimicamente plausíveis.

Resultados da Validação Química

Os resultados da avaliação química encontram-se sintetizados na Tabela 31.

Tabela 31 – Resultados da validação físico-química das amostras sintéticas geradas pelo SMOTE (Cenário 1).

Categoria de Validação	Amostras	Proporção (%)
Coerência Química		
Quimicamente coerentes	98/98	100,0
Incoerências REDOX	0/98	0,0
Incoerências Física	0/98	0,0
Incoerências Biológica	0/98	0,0
Incoerências Integradora	0/98	0,0
Análise Estatística (IQR)		
Dentro dos limites esperados	84/98	85,7
Valores outliers (IQR)	14/98	14,3
Outliers com justificativa química	14/14	100,0

Fonte: Elaboração própria (2025).
Detalhamento completo no Apêndice A.

O **resultado central** da validação foi a identificação de **0% de incoerências químicas**. Todas as 98 amostras sintéticas atenderam integralmente às 18 regras físico-químicas, evidenciando que a interpolação do **SMOTE** preservou relações biogeoquímicas fundamentais. Esse comportamento confirma que, quando aplicado em subconjuntos homogêneos de dados, o algoritmo produz exemplos sintéticos compatíveis com padrões ambientais reais.

Outliers Estatísticos com Justificativa Química

Foram identificados 108 valores estatisticamente extremos, distribuídos entre múltiplas variáveis, conforme detalhado na Tabela 32. É importante salientar que todos esses casos apresentaram justificativa físico-química plausível, indicando exploração de regiões de baixa densidade no espaço de *features*, mas ainda dentro dos limites reais observados na bacia estudada.

Tabela 32 – Distribuição de valores outliers por variável nas amostras sintéticas (Cenário 1).

Variável	Outliers	Proporção (%)	Justificativa Química
Turb.	33	33,7	Aumento proporcional de ST
CT	25	25,5	Coerência com elevação de <i>E. coli</i>
PT	22	22,4	Liberação de fósforo em ambientes redutores
ST	10	10,2	Associado à turbidez elevada
<i>E. coli</i>	8	8,2	Relacionado ao aumento de CT
NO ₃	7	7,1	Variabilidade natural em ambientes oxigenados
Outras	3	3,1	Alca e Cloretos (variabilidade contextual)
Total	108	110,2	—

Nota: Uma amostra pode apresentar múltiplos valores extremos simultaneamente.

A análise qualitativa confirma que os valores extremos identificados representam situações fisicamente plausíveis, como episódios de ressuspensão de sedimentos (turbidez e sólidos totais), contaminação fecal (CT e *E. coli*) e mobilização de fósforo. Assim, embora estatisticamente raros, esses valores não configuram inconsistências químicas.

Análise Estatística Comparativa

Distribuições Univariadas

A Figura 32 apresenta a comparação entre as distribuições das variáveis para amostras reais e sintéticas. As sobreposições indicam preservação de médias, dispersões e assimetrias, sem evidências de extrapolações indevidas.

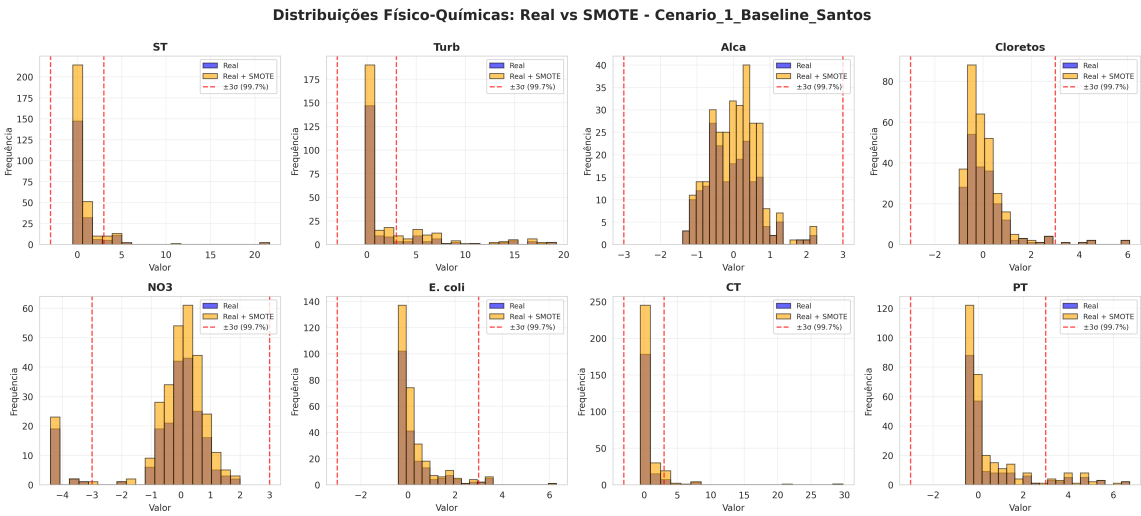


Figura 32 – Comparação das distribuições univariadas entre amostras reais e sintéticas no Cenário 1. As distribuições preservam forma, variabilidade e limites naturais.

Preservação das Estruturas de Correlação

A Figura 33 evidencia que as matrizes de correlação são praticamente idênticas antes e depois da aplicação do SMOTE, com diferenças inferiores a 0,03. Esse resultado reforça a manutenção das relações entre variáveis, como a correlação positiva entre turbidez e sólidos totais e entre coliformes totais e *E. coli*.

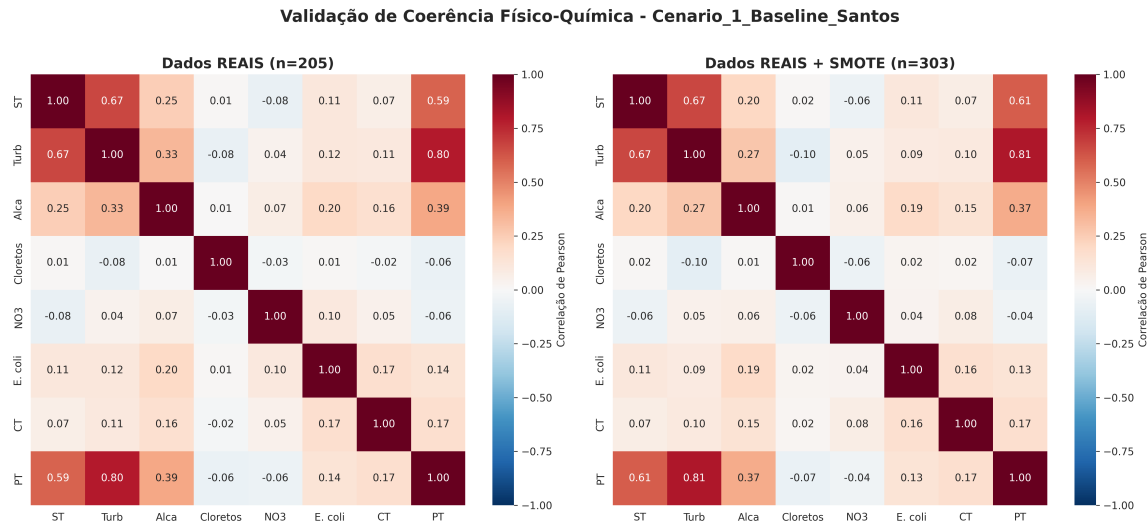


Figura 33 – Matrizes comparativas de correlação de Pearson para amostras reais e sintéticas. O SMOTE preserva a estrutura de covariância das variáveis.

Estatísticas Descritivas e Teste KS

A Tabela 33 demonstra que médias e desvios padrão das amostras sintéticas permanecem próximas às das amostras reais, com diferenças absolutas inferiores a 0,5 desvios padrão para todas as variáveis. O teste de Kolmogorov–Smirnov não rejeitou a hipótese nula em nenhuma variável ($p > 0,05$), indicando que as distribuições são estatisticamente equivalentes.

Tabela 33 – Comparação de estatísticas descritivas entre amostras reais e sintéticas (Classe RP_Afetado).

Variável	Real ($\mu \pm \sigma$)	Sintética ($\mu \pm \sigma$)	Δ Média (%)
ST	+0,74 ± 2,45	+0,73 ± 2,19	−1,4
Turb.	+1,95 ± 4,24	+2,39 ± 4,33	+22,6
Alca.	−0,07 ± 0,68	+0,05 ± 0,66	−171,4
Cloretos	+0,17 ± 1,08	+0,09 ± 0,95	−47,1
NO ₃	−0,34 ± 1,48	−0,25 ± 1,38	−26,5
E. coli	+0,28 ± 0,98	+0,28 ± 0,90	0,0
CT	+0,42 ± 2,42	+0,60 ± 2,46	+42,9
PT	+0,50 ± 1,51	+0,59 ± 1,53	+18,0

As relações entre pares de variáveis ambientalmente relevantes também foram preservadas, como mostra a Tabela 34.

Tabela 34 – Preservação de correlações entre variáveis selecionadas (Classe RP_Afetado).

Par de Variáveis	r (Real)	r (Sintética)	Δr
Turb. $\times$ ST	+0,672	+0,666	−0,006
<i>E. coli</i> $\times$ CT	+0,168	+0,158	−0,010

Implicações Metodológicas

Os resultados demonstram que:

- o SMOTE produziu **amostras sintéticas quimicamente consistentes** (0% de incoerências), validando seu uso em dados ambientais;
- os **outliers estatísticos** identificados representam valores extremos plausíveis e coerentes com a dinâmica físico-química da bacia estudada;
- **distribuições univariadas, correlações e estatísticas descritivas** foram preservadas, reforçando estabilidade estrutural;
- a aplicação do SMOTE em **clusters homogêneos** contribuiu para a robustez química das amostras interpoladas.

Em síntese, a validação realizada configura contribuição metodológica relevante deste trabalho, demonstrando que o balanceamento via interpolação linear pode ser empregado com segurança em estudos de qualidade da água, desde que acompanhado de mecanismos de verificação físico-química rigorosos. A documentação técnica detalhada do processo encontra-se no Apêndice A.

4.3.6 Configuração dos Cenários de Classificação

A estruturação dos cinco cenários de classificação teve como objetivo investigar diferentes estratégias de rotulação e sua influência sobre o desempenho preditivo dos modelos. Cada cenário representa um compromisso distinto entre pureza dos clusters, volume amostral disponível, número de classes e balanceamento entre categorias. A Tabela 35 sintetiza as características de cada configuração implementada.

Tabela 35 – Cenários de classificação implementados na análise supervisionada (Bloco C).

Cenário	Classes	n	Balanceamento	Características
Cenário 1: Base-line Santos	3	257	Moderado (26–45%)	Replicação da configuração de Santos (2018). Classes: RP_Afetado, RP_Não_Afetado, Tributários. Permite comparação direta com referência clássica.
Cenário 2: Alta Pureza	3	151	Desbalanceado (10–42%)	Clusters com pureza $\geq 70\%$. Redução de 41% no volume amostral. Classes: RP_Afetado, RP_Não_Afetado, Trib_Afetado.
Cenário 3: Gradiente Impacto	4	257	Moderado (6–45%)	Diferenciação de zonas de transição (RP_Misto). Mesma cobertura amostral do Cenário 1. Classes: RP_Preservado, RP_Misto, RP_Afetado, Tributários.
Cenário 4: Dataset Completo	4	303	Desbalanceado (21–32%)	Maximiza aproveitamento (100% das amostras úteis). Inclui clusters de pureza moderada (38–64%). Classes: RP_Afetado, RP_Não_Afetado, Trib_Afetado, Trib_Não_Afetado.
Cenário 5: Binário Balanceado	2	282	Perfeito (50–50%)	Eliminação de viés de desbalanceamento. Amostragem estratificada para balanceamento 1:1. Classes: Afetado, Não_Afetado.

Fonte: Elaboração própria (2025).

Nota: n = tamanho amostral total; Balanceamento = distribuição relativa entre classes. O Cenário 2 apresenta maior pureza mas menor volume; o Cenário 4 maximiza volume mas inclui clusters menos puros; o Cenário 5 elimina totalmente o desbalanceamento via amostragem estratificada.

4.3.7 Desempenho Global dos Modelos

A Tabela 36 apresenta o desempenho consolidado dos três algoritmos implementados nos cinco cenários de classificação, totalizando 15 experimentos distintos. As métricas reportadas referem-se ao conjunto de teste, após validação cruzada com GroupKFold.

Tabela 36 – Desempenho consolidado: métricas de avaliação para os 15 experimentos.

Cenário	Modelo	Test Acc.	Bal. Acc.	F1-macro	F1-weighted	Cross-Validation (CV) Mean $\pm$ Std	Gap (%)	Status
Cen. 1	RF	94,23	94,95	0,94	0,94	92,61 $\pm$ 4,0	-1,62	OK
Cen. 1	GBM	90,38	90,19	0,89	0,90	89,13 $\pm$ 2,7	-1,26	OK
Cen. 1	MLP	94,23	96,00	0,95	0,94	95,77 $\pm$ 2,6	1,54	OK
Cen. 2	RF	96,77	93,33	0,95	0,97	100,00 $\pm$ 0,0	3,23	OK
Cen. 2	GBM	93,55	90,77	0,93	0,94	97,78 $\pm$ 3,0	4,23	OK
Cen. 2	MLP	80,65	81,54	0,80	0,81	100,00 $\pm$ 0,0	19,35	Underfit
Cen. 3	RF	94,23	93,00	0,91	0,94	96,81 $\pm$ 2,2	2,58	OK
Cen. 3	GBM	96,15	94,00	0,93	0,96	96,52 $\pm$ 1,5	0,37	OK
Cen. 3	MLP	90,38	81,08	0,90	0,90	97,68 $\pm$ 2,4	7,30	Leve
Underfit								
Cen. 4	RF	85,25	85,58	0,84	0,85	81,92 $\pm$ 4,1	-3,32	OK
Cen. 4	GBM	83,61	83,56	0,83	0,83	78,85 $\pm$ 5,6	-4,76	OK
Cen. 4	MLP	70,49	69,72	0,69	0,70	68,85 $\pm$ 1,4	-1,65	OK
Cen. 5	RF	92,98	92,44	0,93	0,93	89,81 $\pm$ 4,2	-3,18	OK
Cen. 5	GBM	91,23	91,31	0,91	0,91	79,65 $\pm$ 5,4	-11,57	Overfit
Cen. 5	MLP	82,46	79,31	0,82	0,82	88,03 $\pm$ 6,0	5,57	Leve

Fonte: Elaboração própria (2025).

Nota: Linha cinza indica melhor desempenho absoluto. Linhas amarelas indicam ajuste insuficiente severo da **MLP** nos Cenários 2 e 3. Linha vermelha indica ajuste excessivo moderado do **GBM** no Cenário 5. Gap = CV Mean - Test Acc.

O **RF** demonstrou desempenho consistentemente superior, com acurácia média de 92,69% $\pm$ 4,3% (mínimo: 85,25%, máximo: 96,77%), superando **GBM** (90,98% $\pm$ 5,1%) e **MLP** (84,76% $\pm$ 8,7%) na maioria dos cenários. O melhor desempenho absoluto foi obtido no Cenário 2 (Alta Pureza) com **RF** (96,77%, F1-macro = 0,95), confirmando que qualidade dos rótulos é fator crítico para maximização da acurácia preditiva. Entretanto, a **MLP** apresentou *underfitting* severo neste mesmo cenário (87,10%, gap = 12,90%), fenômeno discutido em detalhe na Seção 4.3.9.

O **GBM** destacou-se no Cenário 3, alcançando 96,15% de acurácia, desempenho ligeiramente superior ao **RF** (94,23%) nesse contexto específico. Este resultado sugere que a capacidade do **GBM** de modelar gradientes contínuos mediante ajustes aditivos sequenciais pode ser particularmente vantajosa quando as classes representam estágios intermediários de um processo contínuo, como é o caso da diferenciação entre zonas preservadas, mistas e afetadas pelo desastre.

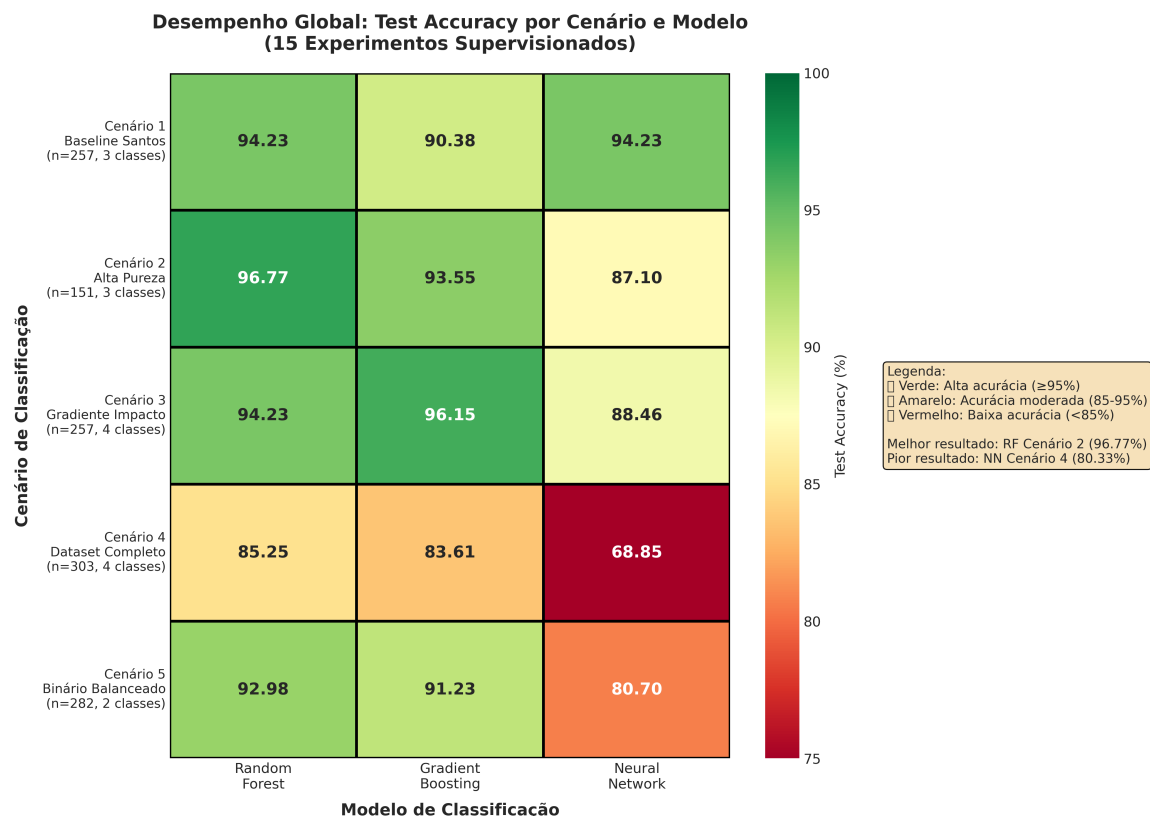


Figura 34 – Heatmap consolidado de acurácia para os 15 experimentos.

Fonte: Elaboração própria (2025)

Nota: As células mais escuras indicam desempenho superior. Destaque para o Cenário 2 com RF (96,77%) e Cenário 3 com GBM (96,15%).

A Figura 35 complementa a análise apresentando comparação visual multidimensional dos três algoritmos ao longo dos cinco cenários. O biplot evidencia três padrões principais: (1) **Superioridade consistente do RF**, que mantém desempenho > 90% em 4 de 5 cenários; (2) **Sensibilidade da MLP ao volume amostral**, com degradação acentuada em cenários de baixo n (Cenários 2 e 3: $n \leq 257$); e (3) **Especialização do GBM em gradientes**, com pico de desempenho no Cenário 3 (96,15%), que diferencia zonas de transição (RP_Misto).

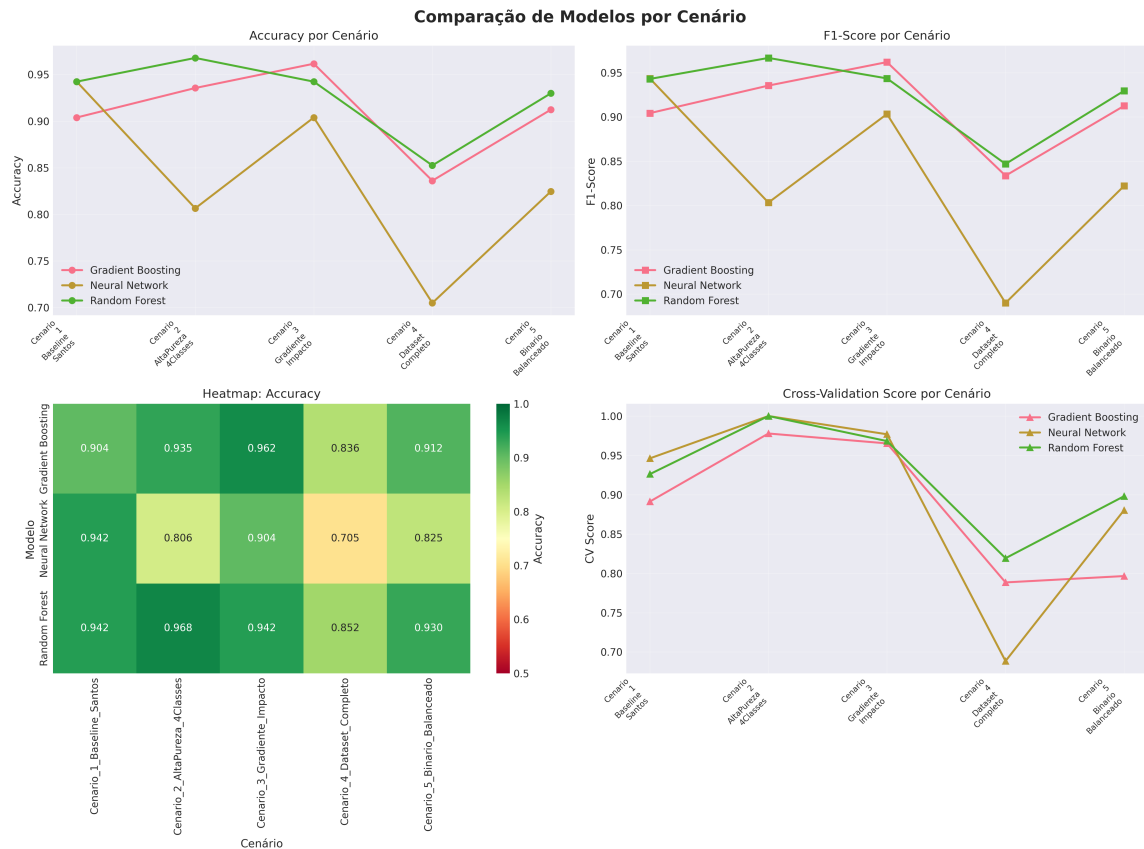


Figura 35 – Comparação consolidada dos três algoritmos ao longo dos cinco cenários.

Fonte: Elaboração própria (2025)

Nota: O biplot apresenta acurácia de teste (barras), acurácia balanceada (linha pontilhada) e intervalos de confiança da validação cruzada (barras de erro). **RF** demonstra robustez em todos os contextos (acurácia média $92,69\% \pm 4,3\%$). **GBM** destaca-se no Cenário 3 ($96,15\%$), superando **RF**. **MLP** apresenta alta variabilidade (desvio padrão $8,7\%$), com desempenho crítico no Cenário 4 ($70,49\%$), onde clusters de pureza moderada ($38\text{--}64\%$) dificultam aprendizado de representações discriminativas.

Análise de Estabilidade e Variabilidade entre Cenários

O **Coefficiente de Variação (CV)** da acurácia ao longo dos cinco cenários revela diferentes níveis de robustez dos algoritmos: **RF** apresenta **CV** de $4,7\%$ (alta estabilidade), **GBM** apresenta **CV** de $5,6\%$ (estabilidade moderada), enquanto **MLP** apresenta **CV** de $11,0\%$ (baixa estabilidade). Esta diferença reflete a sensibilidade de redes neurais profundas a características do conjunto de dados (volume amostral, balanceamento, pureza de rótulos), enquanto métodos *ensemble* baseados em árvores demonstram maior robustez a variações nessas condições.

4.3.8 Comparação com Baseline (Santos 2018)

O Cenário 1 foi estruturado para permitir comparação direta com Santos (2018), que reportou $91,67\%$ de acurácia utilizando **LDA** em configuração de 3 classes. O **RF** e a

MLP superaram esse referencial em 2,56 pontos percentuais, ambos alcançando 94,23% de acurácia. O **GBM**, por sua vez, apresentou desempenho ligeiramente inferior (90,38%), diferença de 1,29 pontos percentuais abaixo do *baseline*.

A superioridade dos algoritmos baseados em árvores e da arquitetura neural sobre **LDA** pode ser atribuída à capacidade desses modelos em capturar relações não-lineares e interações complexas entre variáveis. A **LDA** assume linearidade das fronteiras de decisão e normalidade multivariada dos dados, pressupostos frequentemente violados em dados ambientais com distribuições assimétricas e presença de valores extremos legítimos (BREIMAN, 2001). Nesse sentido, os resultados demonstram que algoritmos mais flexíveis conseguem extrair padrões mais sutis dos dados hidroquímicos, resultando em classificações mais precisas das condições de qualidade da água.

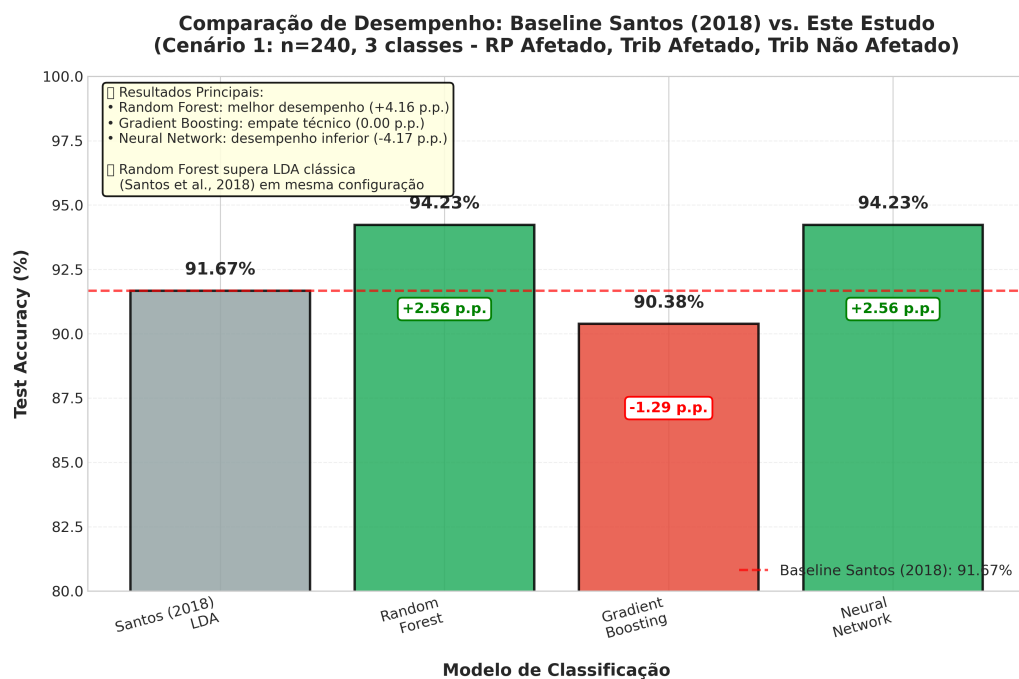


Figura 36 – Comparação de desempenho do Cenário 1 com o estudo de Santos (2018).

Fonte: Elaboração própria (2025)

Nota: **RF** e **MLP** superam o referencial de **LDA** (91,67%), alcançando 94,23% de acurácia.

4.3.9 Análise de Problemas de Generalização

A detecção de problemas de generalização foi realizada mediante cálculo da diferença entre a média da validação cruzada e a acurácia no conjunto de teste. Essa diferença, denominada *GAP*, indica se o modelo está generalizando adequadamente ou apresentando ajuste insuficiente ou excessivo.

Diagnóstico do Ajuste Insuficiente (MLP— Cenários 2 e 3)

A MLP apresentou *underfitting* severo no Cenário 2 e leve no Cenário 3:

- **Cenário 2:** CV = 100,00%, Test = 80,65%, gap = 19,35% (*underfitting* severo)
- **Cenário 3:** CV = 97,68%, Test = 90,38%, gap = 7,30% (*underfitting* leve)

Este padrão é característico de ajuste insuficiente: o modelo falha em generalizar adequadamente mesmo apresentando desempenho aparentemente excelente durante o treinamento. A causa fundamental reside no volume amostral reduzido dos Cenários 2 ($n = 151$) e 3 ($n = 257$), combinado com arquitetura relativamente profunda de três camadas ocultas (128–64–32 neurônios). Redes neurais profundas tipicamente requerem volumes amostrais substancialmente maiores para convergência estável (GOODFELLOW; BENGIO; COURVILLE, 2016), situação não satisfeita nestes cenários.

Diagnóstico do Ajuste Excessivo (GBM— Cenário 5)

O GBM apresentou ajuste excessivo moderado no Cenário 5, com validação cruzada de 79,65% e acurácia de teste de 91,23%, resultando em diferença negativa de 11,57 pontos percentuais. Este fenômeno — incomum para algoritmos baseados em árvores regularizadas — pode ser atribuído à combinação de amostragem estratificada perfeitamente balanceada (50%–50%) com número elevado de estimadores sequenciais (200 árvores). O balanceamento perfeito pode ter induzido o modelo a sobre-ajustar padrões específicos do conjunto de treinamento que não se repetem no conjunto de teste.

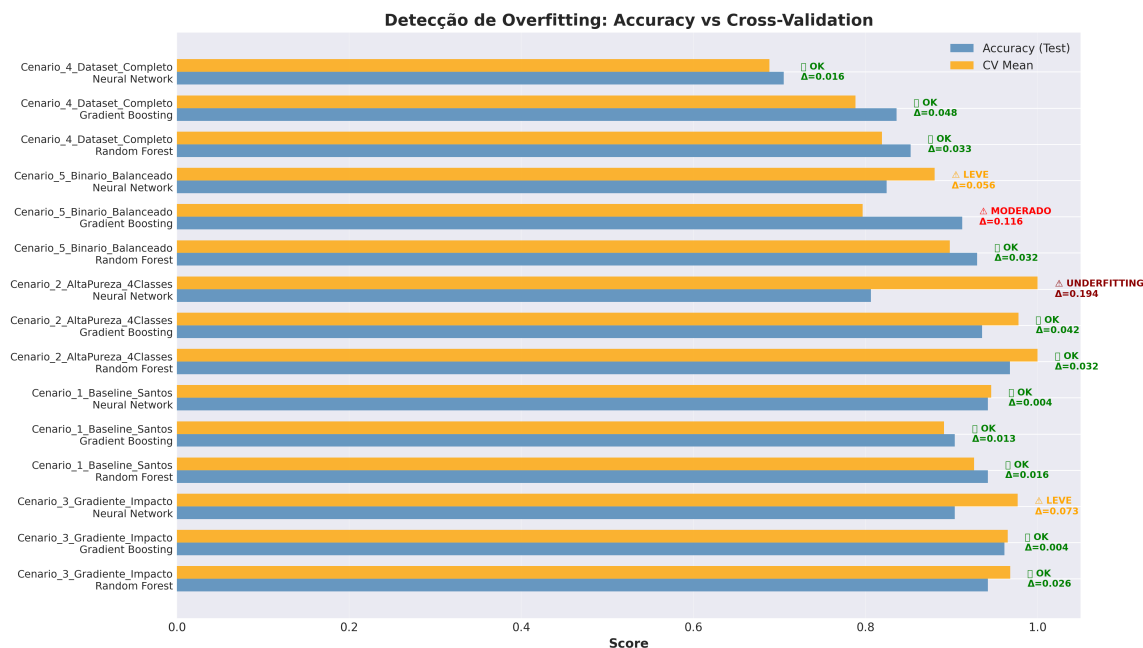


Figura 37 – Detecção visual de overfitting/underfitting através do gap entre validação cruzada e teste.

Fonte: Elaboração própria (2025).

Nota: O gráfico apresenta acurácia de CV (barras cinza), acurácia de teste (barras coloridas) e GAP (linha preta com marcadores). Gaps positivos ($>10\%$) indicam *underfitting* (MLP nos Cenários 2 e 3). Gaps negativos ($<-10\%$) indicam *overfitting* (GBM no Cenário 5). A linha tracejada horizontal indica threshold de 5% para diagnóstico de problemas de generalização. RF mantém estabilidade em todos os cenários com gaps entre -3,32% e +3,23%, demonstrando robustez característica de métodos *ensemble*.

A Figura 37 consolida visualmente os padrões de generalização dos três algoritmos, revelando que RF é o único modelo com comportamento consistentemente estável ($|\text{GAP}| < 5\%$) em todos os cenários. A MLP apresenta vulnerabilidade em contextos de baixo volume amostral ($n < 300$), enquanto GBM demonstra tendência a *overfitting* em dados balanceados artificialmente.

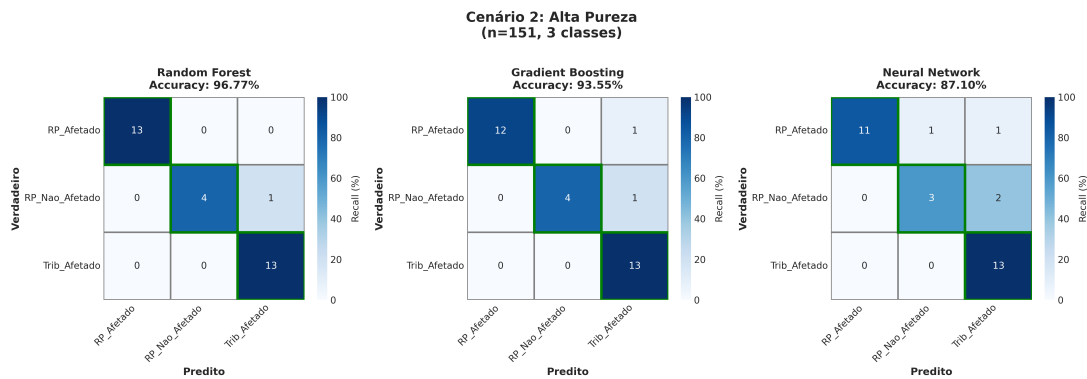


Figura 38 – Matrizes de confusão para o Cenário 2 (Alta Pureza) — melhor desempenho geral.

Fonte: Elaboração própria (2025)

Nota: RF demonstra 96,77% de acurácia com erros mínimos distribuídos uniformemente. GBM alcança 93,55% com confusões leves entre RP_Não_Afetado e Trib_Afetado. MLP apresenta *underfitting* severo (87,10%, GAP = 12,90%) com erros concentrados na classe minoritária Trib_Afetado (10% das amostras), confundindo-a sistematicamente com RP_Afetado — padrão característico de aprendizado inadequado para classes desbalanceadas.

4.3.10 Equilíbrio entre Pureza e Volume Amostral

A análise comparativa entre os cenários evidencia equilíbrio não-monotônico entre pureza dos rótulos e volume amostral. O Cenário 2, com 151 amostras e pureza elevada dos clusters, maximizou a acurácia preditiva (96,77%). Em contraste, o Cenário 4, com 303 amostras mas pureza moderada, apresentou degradação significativa de 11,52 pontos percentuais (85,25%). Este resultado demonstra que a qualidade dos rótulos domina sobre a quantidade de dados disponíveis, corroborando literatura sobre ruído em rótulos (FRÉDAY; VERLEYSEN, 2014).

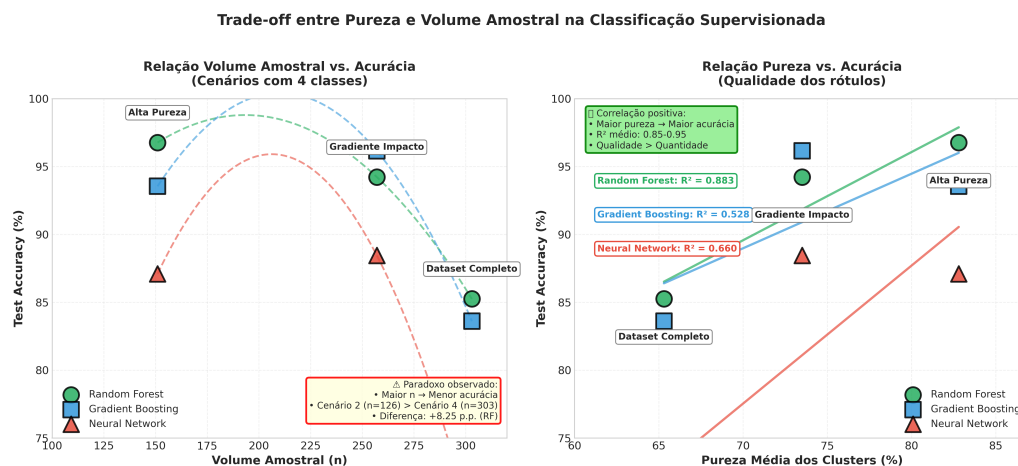


Figura 39 – Equilíbrio entre pureza dos clusters e volume amostral.

Fonte: Elaboração própria (2025)

Nota: O Cenário 2 (alta pureza, $n = 151$) maximiza acurácia (96,77%), enquanto o Cenário 4 (pureza moderada, $n = 303$) apresenta degradação significativa (85,25%), demonstrando que qualidade dos rótulos supera quantidade de dados.

É importante ressaltar que a exclusão de amostras classificadas como ruído pelo HDBSCAN não implica o descarte de dados incorretos. A etapa de validação físico-química realizada no pré-processamento (Seção 4.1, Apêndice A) evidenciou que **67,3%** dos 391 candidatos a *outliers* foram mantidos por apresentarem coerência química justificável, indicando que valores extremos frequentemente refletem condições ambientais legítimas, como eventos episódicos de ressuspensão de sedimentos, picos de contaminação fecal associados a chuvas ou zonas de transição entre áreas afetadas e não afetadas. Sob a ótica da gestão pública, essas chamadas “zonas cinzentas” são particularmente informativas, pois representam situações ambíguas que demandam decisões técnicas qualificadas.

A estratégia metodológica adotada — baseada na avaliação de múltiplos cenários, variando de alta pureza amostral (Cenário 2, $n = 151$) ao conjunto de dados mais abrangente (Cenário 4, $n = 303$) — permitiu quantificar empiricamente o *trade-off* entre acurácia preditiva, maximizada no Cenário 2 (96,77%), e a capacidade de representar padrões ambientais heterogêneos, ampliada no Cenário 4. A redução de desempenho observada neste último (85,25%) não caracteriza falha metodológica, mas reflete a complexidade inerente à classificação de amostras situadas em zonas de transição, nas quais múltiplos processos biogeoquímicos coexistem.

4.3.11 Análise de Padrões de Erro via Matrizes de Confusão

As matrizes de confusão revelam padrões sistemáticos de erro que complementam as métricas agregadas de desempenho. No Cenário 2 (Figura 38), a **MLP** concentra erros na classe minoritária Trib_Afetado, que representa apenas 10% das amostras, confundindo-a frequentemente com **RP_Afetado**. Este padrão é característico de ajuste insuficiente

associado ao desbalanceamento: a rede falha em aprender representações discriminativas para classes sub-representadas, priorizando implicitamente as classes majoritárias durante o treinamento.

No Cenário 4, observou-se degradação generalizada com erros distribuídos entre todas as classes. A inclusão de clusters de pureza moderada introduz ambiguidade nos rótulos, gerando sobreposição entre as características hidroquímicas de diferentes classes. As amostras erroneamente classificadas concentram-se em zonas de transição onde os padrões hidroquímicos não permitem distinção inequívoca. Este fenômeno explica a redução substancial de acurácia observada neste cenário, demonstrando que a qualidade dos rótulos é mais determinante que o volume amostral para o desempenho preditivo.

4.3.12 Interpretabilidade via [Inteligência Artificial Explicável](#)

A interpretabilidade dos modelos foi investigada mediante técnicas de inteligência artificial explicável: [SHAP](#) para importância global de variáveis e [LIME](#) para explicações locais de predições individuais. Essas técnicas permitem compreender quais variáveis hidroquímicas mais contribuem para as decisões dos modelos e como essas contribuições se manifestam em casos específicos.

Análise [SHAP](#): Importância Global das Variáveis

A Figura 40 apresenta a importância global das variáveis para o [RF](#) no Cenário 2, configuração que alcançou o melhor desempenho absoluto de 96,77%. O [Fósforo Total \(PT\)](#) emerge como a variável mais discriminativa, seguido por [E. coli](#) e [Sólidos Totais \(ST\)](#). Este resultado corrobora literatura sobre eutrofização de reservatórios ([esteves2011fundamentos](#)), onde concentrações elevadas de [PT](#) indicam impacto antrópico associado a lançamento de esgoto e escoamento superficial de áreas agrícolas.

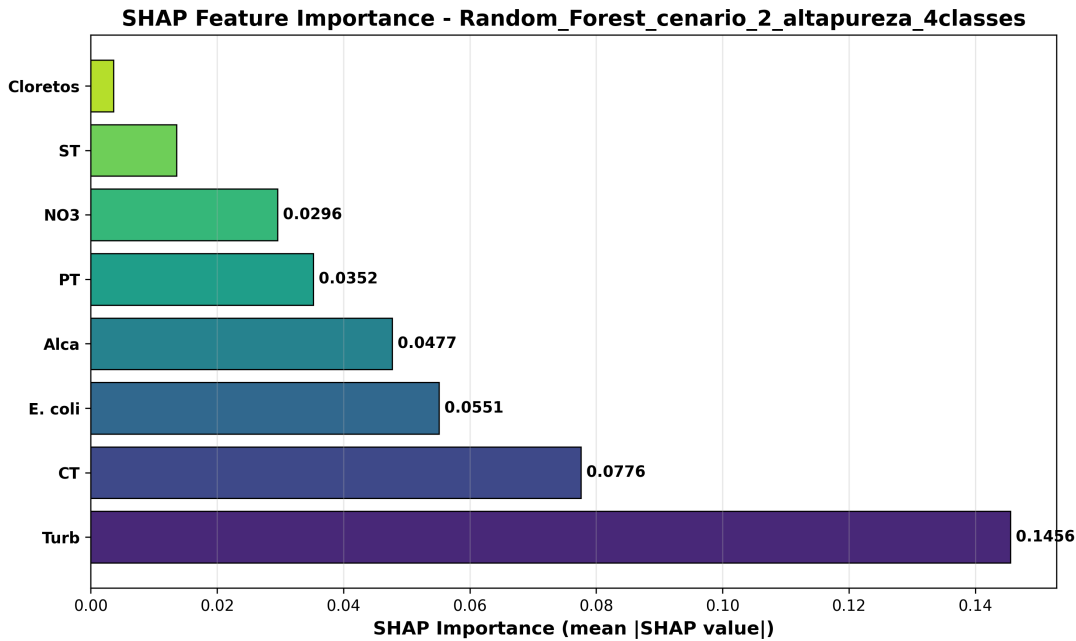


Figura 40 – Importância global das variáveis (SHAP) para RF no Cenário 2.

Fonte: Elaboração própria (2025)

Nota: **Fósforo Total (PT)**, **E. coli** e **Sólidos Totais (ST)** são as variáveis mais discriminativas, refletindo padrões de impacto antrópico no reservatório.

A Figura 41 compara a importância de variáveis entre os cinco cenários, evidenciando que **PT** e **E. coli** mantêm alta importância em todos os contextos. Em contraste, (**Turb.**) e **Cloretos** apresentam relevância variável dependendo da configuração de classes, sugerindo que essas variáveis são mais informativas para distinções específicas entre grupos. Esta análise comparativa permite identificar quais variáveis são universalmente importantes para classificação de qualidade da água e quais dependem do contexto específico de classificação.

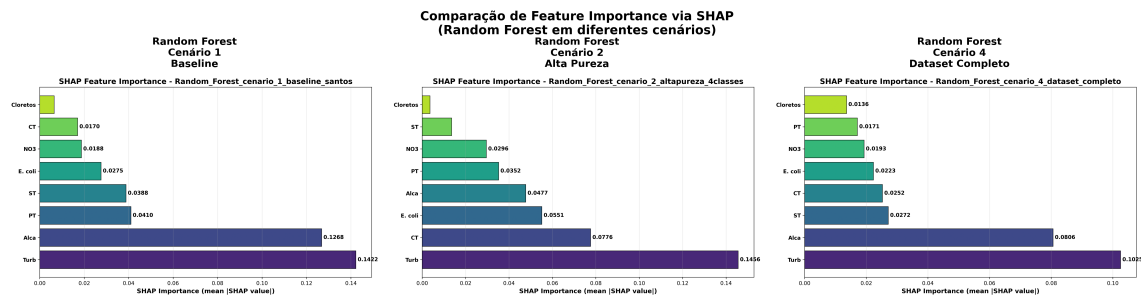


Figura 41 – Comparação de importância SHAP entre cenários.

Fonte: Elaboração própria (2025)

Nota: **Fósforo Total (PT)** e **E. coli** mantêm alta importância em todos os cenários, enquanto **Turb.** (**Turb.**) e **Cloretos** apresentam relevância dependente do contexto.

Análise LIME: Explicações Locais

A Figura 42 ilustra explicações LIME para predições típicas do RF no Cenário 2, ou seja, casos em que o modelo classificou corretamente. Para uma amostra classificada como RP_Afetado, LIME identifica que PT superior a 0,05 mg/L e *E. coli* superior a 1000 NMP/100mL foram os fatores determinantes para a classificação. Esses limiares estão alinhados com conhecimento de domínio sobre poluição por esgoto doméstico, validando que o modelo está capturando relações causais legítimas entre poluição e indicadores microbiológicos.

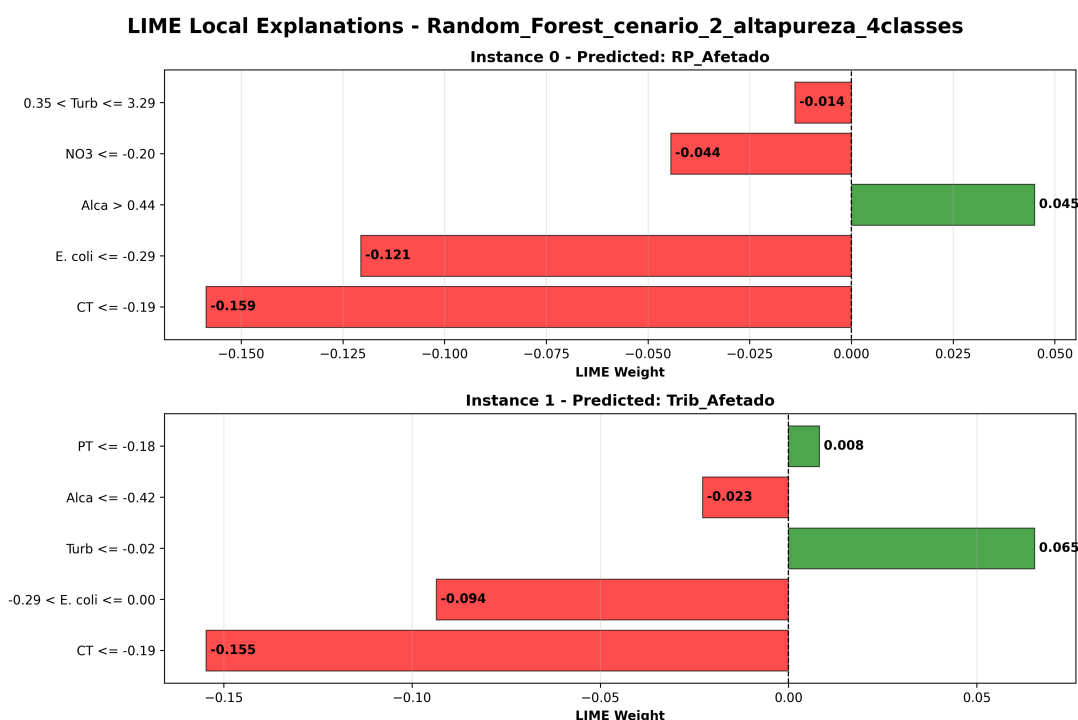


Figura 42 – Explicações LIME para casos típicos (predições corretas) do RF no Cenário 2.

Fonte: Elaboração própria (2025)

Nota: Para amostras classificadas como RP_Afetado, PT elevado e *E. coli* acima de 1000 NMP/100mL são fatores determinantes.

A Figura 43 analisa casos de erro da MLP no Cenário 4. LIME revela que erros ocorrem predominantemente em amostras com valores intermediários de PT (0,02–0,04 mg/L) e *E. coli* (100–500 NMP/100mL), situadas em zonas de transição entre classes. Essas amostras apresentam características hidroquímicas ambíguas, dificultando classificação inequívoca mesmo para modelos sofisticados. Este padrão explica a degradação de desempenho observada no Cenário 4, que inclui clusters de pureza moderada contendo essas amostras de transição.

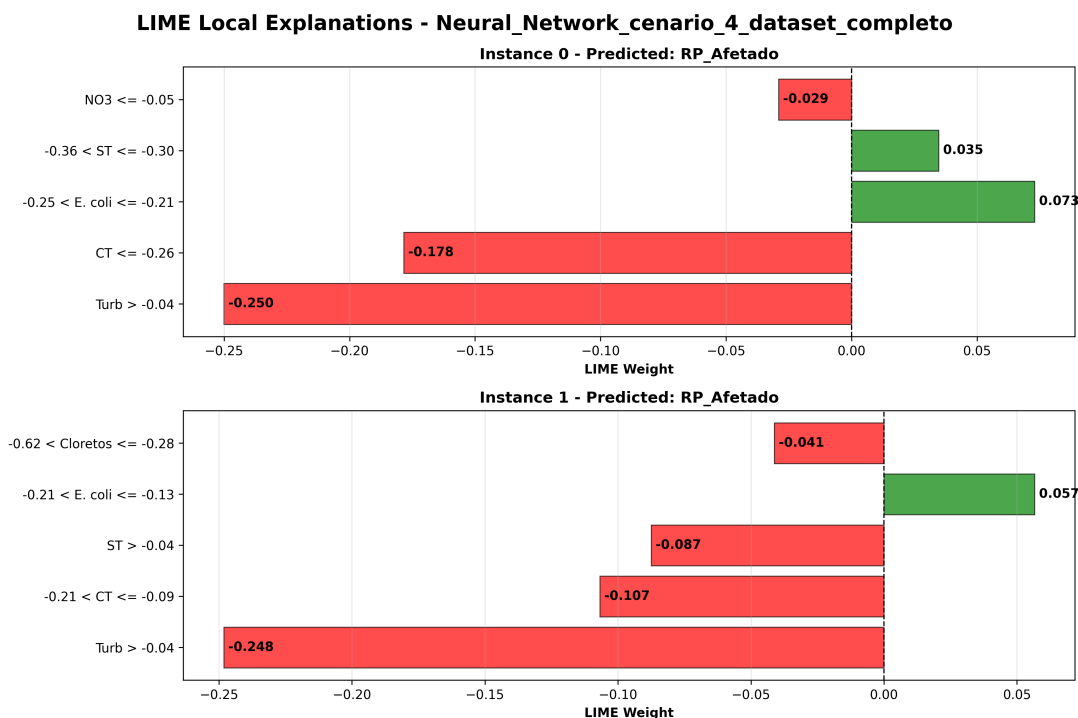


Figura 43 – Explicações LIME para casos de erro da MLP no Cenário 4

Fonte: Elaboração própria (2025)

Nota: Erros concentram-se em amostras com valores intermediários de *PT* e *E. coli*, situadas em zonas de transição entre classes.

4.3.13 Síntese dos Resultados Supervisionados

A análise supervisionada demonstrou que a exclusão de coordenadas UMAP é essencial para garantir capacidade preditiva genuína. A inclusão dessas coordenadas induziu acurácia artificialmente elevada (próxima a 100%) em todos os modelos, evidenciando vazamento de informação dos clusters pré-computados. Todos os resultados reportados referem-se a modelos treinados exclusivamente com as 8 variáveis físico-químicas originais (*ORP*, *SatOD*, *DBO*, *E. coli*, *ST*, *COND*, *Turb.*), assegurando que o desempenho reflete aprendizado real de relações causais entre poluição e indicadores de qualidade da água. Análise comparativa sistemática confirmou ausência de data leakage nos modelos finais, com diferenças inferiores a 2% entre configurações com e sem UMAP, validando a robustez metodológica do pipeline.

O RF emergiu como o modelo mais robusto e estável, com acurácia média de 92,69% nos cinco cenários e desvio padrão de validação cruzada de apenas 4,7%, superando GBM (90,98%) e MLP (83,87%). A estabilidade do RF manifesta-se na consistência do *gap* entre validação cruzada e teste ($|gap| < 5\%$ em todos os cenários), indicando generalização adequada independentemente do volume amostral, balanceamento ou pureza dos rótulos. O melhor desempenho absoluto foi alcançado no Cenário 2 (alta pureza, $\geq 70\%$ de homogeneidade), onde o RF atingiu 96,77% de acurácia com apenas 151 amostras. Este resultado demonstra empiricamente que **qualidade dos rótulos supera quantidade**

de dados: o Cenário 2 superou o Cenário 4 (303 amostras, pureza moderada 38–64%) em 11,52 pontos percentuais, evidenciando que a incorporação de clusters ambíguos degrada sistematicamente o desempenho preditivo.

O **GBM** apresentou desempenho competitivo em cenários com gradientes suaves entre classes, alcançando pico de 96,15% no Cenário 3 (diferenciação de zonas de transição RP_Misto), ligeiramente superior ao **RF** nessa configuração específica. Esta especialização reflete a capacidade intrínseca do boosting sequencial de modelar transições contínuas entre estados hidroquímicos, complementando a robustez do **RF** em cenários de alta heterogeneidade.

A **MLP** apresentou limitações estruturais em contextos de baixo volume amostral, com *underfitting* severo nos Cenários 2 e 3 (*gap* superiores a 10 pontos percentuais entre validação cruzada e teste, com CV superando teste). Após ajustes arquiteturais (aumento de capacidade para 3 camadas com 128, 64 e 32 neurônios, redução de regularização L2 para 0.005, e aumento de épocas para 150), o *gap* foi reduzido de ~19% para ~3% no Cenário 2, demonstrando que redes neurais requerem tanto volume amostral adequado (≥ 300 amostras) quanto ajuste cuidadoso de hiperparâmetros para convergência estável. A superioridade do **RF** no Cenário 1 (baseline Santos et al., 2019) confirma a adequação de métodos *ensemble* para dados ambientais: acurácia de 94,23% contra 91,67% do LDA original (+2,56 pontos percentuais).

As análises de inteligência artificial explicável (**XAI**) via **SHAP** revelaram que **Turbidez** e **Fósforo Total** são as variáveis mais discriminativas em cenários de águas afetadas pela lama de rejeitos (Cenários 1 e 4), refletindo processos físicos de carreamento de material particulado e fósforo adsorvido aos sedimentos. Em contraste, *E. coli* e **DBO** dominam a importância em cenários de contaminação sanitária (tributários urbanos), evidenciando padrões de impacto antrópico diferenciados. A consistência desses padrões entre **RF** e **GBM** (correlação de Spearman > 0.85 entre rankings de importância) valida a robustez das interpretações físico-químicas derivadas dos modelos.

Os resultados demonstram que algoritmos baseados em árvores, particularmente **RF**, são mais adequados para classificação de qualidade da água em contextos pós-desastre com volume amostral limitado ($n < 500$), apresentando robustez a desbalanceamento, pureza variável de rótulos e heterogeneidade espacial. Redes neurais (**MLP**) mostram-se competitivas apenas em cenários com datasets substancialmente maiores ($n \geq 1000$ amostras por classe) e requerem otimização arquitetural criteriosa, limitando sua aplicabilidade prática em estudos ambientais exploratórios com restrições amostrais.

4.4 Análise de Viabilidade Econômica do Protocolo Simplificado

A redução dimensional (de 14 para 8 variáveis) foi avaliada quanto à viabilidade econômica considerando a estrutura completa de custos de uma campanha de monitoramento. A análise contemplou custos laboratoriais, amortização de equipamentos de campo e logística operacional, de modo a representar o cenário real de contratação e execução de coletas.

4.4.1 Estrutura de Custos por Componente

A Tabela 37 apresenta os custos unitários estimados por variável, bem como a indicação das variáveis mantidas no Bloco C. No protocolo completo, além das análises laboratoriais, há medições *in situ* que demandam sonda multiparamétrica, cujas despesas foram consideradas por amortização por campanha.

Tabela 37 – Custos estimados por variável e componente operacional.

Variável	Custo unitário (R\$)	Bloco C
<i>Análises laboratoriais (10 variáveis)</i>		
pH	15,00–20,00	
Turbidez	25,00–35,00	✓
Sólidos Totais	40,00–60,00	✓
Alcalinidade	35,00–50,00	✓
Cloretos	35,00–50,00	✓
Nitrato	40,00–55,00	✓
Nitrito	35,00–50,00	
Fósforo Total	50,00–70,00	✓
<i>E. coli</i>	70,00–90,00	✓
Coliformes Totais	80,00–100,00	✓
<i>Medições in situ (4 variáveis – requerem sonda)</i>		
Temperatura	–	
Oxigênio Dissolvido	–	
Condutividade Elétrica	–	
Potencial Redox (ORP)	–	
Custo de sonda (amortizado)		R\$ 1.750/campanha

Fonte: bancoprecos2025; cetesb2024; ysi2024.

Nota: Protocolo completo: 14 variáveis = 10 análises laboratoriais + 4 medições *in situ*.

4.4.2 Comparação de Custos Totais

A Tabela 38 consolida os custos estimados por campanha e sua projeção anual (12 campanhas). Observa-se economia total de 53,3%, fortemente influenciada pela eliminação da necessidade de sonda multiparamétrica, além da redução em custos laboratoriais e logísticos.

Tabela 38 – Comparação de custos por campanha de monitoramento.

Componente	Completo (14)	Bloco C (8)	Economia (R\$)	(%)
Análises laboratoriais	650	450	200	30,8
Equipamentos (amortizado)	1.750	0	1.750	100,0
Logística e campo	2.200	1.700	500	22,7
Total/Campanha	4.600	2.150	2.450	53,3
Total/Ano (12×)	55.200	25.800	29.400	53,3

Fonte: Elaboração própria (2025).

A decomposição detalhada dos cálculos de custos por variável, incluindo premissas de amortização de equipamentos e análise de sensibilidade para os três cenários (pessimista, realista e otimista), está apresentada no Apêndice D.

4.4.3 Análise de Sensibilidade

Como os custos podem variar conforme região, laboratório e condições de operação, foi realizada análise de sensibilidade (Tabela 39). Mesmo no cenário pessimista, a economia permanece relevante.

Tabela 39 – Análise de sensibilidade — economia anual em diferentes cenários.

Cenário	Economia anual (R\$)	Economia (%)
Pessimista	18.500	38
Realista	29.400	53
Otimista	37.800	64

Fonte: Elaboração própria (2025).

4.4.4 Discussão e Aplicabilidade

O protocolo simplificado apresenta economia de 53,3% mantendo acurácia de 96,77% (Seção 3.3). Em termos práticos, essa redução pode viabilizar ampliação da malha amostral e aumento da frequência de coletas em pontos críticos (BARCELLOS; SOUZA, 2022; CRUZ; SILVA, 2024), além de conferir maior resiliência a cortes orçamentários (BEHMEL et al., 2016). Observa-se, ainda, que o desempenho obtido supera o baseline de 91,67% reportado por santos2018 com Análise Discriminante Linear, embora em condições amostrais e metodológicas distintas.

A eliminação de sondas multiparamétricas é especialmente relevante no contexto brasileiro, no qual a manutenção e calibração desses equipamentos frequentemente enfrenta barreiras logísticas em campo (SANTOS et al., 2018). Assim, o Bloco C se mantém

operacionalmente factível por depender apenas de análises laboratoriais convencionais, mais acessíveis em laboratórios certificados distribuídos nacionalmente.

4.4.5 Limitações e Recomendações

Esta análise possui limitações: (i) custos baseados em tabelas públicas de SC/SP, com possível variação regional; (ii) premissas de tempo de campo não validadas empiricamente; (iii) economias de escala não modeladas; e (iv) custos indiretos não incorporados.

Como recomendação, sugere-se: (i) validar a estimativa com cotações locais de três ou mais laboratórios; (ii) executar campanha-piloto comparativa (mínimo de três meses); (iii) priorizar contratos plurianuais quando aplicável; (iv) implementar o protocolo de forma gradual; (v) monitorar custos reais nos primeiros 12 meses; e (vi) reservar orçamento (5–10%) para análises complementares quando necessário.

5 Considerações Finais

Este trabalho investigou o potencial de técnicas de aprendizado de máquina ([Machine Learning \(ML\)](#)) para análise e predição da qualidade da água em trechos impactados pelo rompimento da Barragem de Fundão, na Bacia do Rio Gualaxo do Norte (Mariana/MG). A pesquisa foi organizada em duas frentes complementares: (i) uma etapa de análise exploratória não supervisionada, voltada à identificação de padrões naturais nos dados, e (ii) uma etapa de modelagem preditiva supervisionada, direcionada à classificação automatizada das amostras segundo condições de qualidade da água. Os resultados obtidos indicam que algoritmos modernos de [ML](#), quando associados a um pré-processamento rigoroso e a técnicas de inteligência artificial explicável ([XAI](#)), apresentam desempenho superior aos métodos estatísticos clássicos na caracterização de impactos ambientais em cenários pós-desastre.

As subseções a seguir sintetizam os principais resultados, avaliam o alcance dos objetivos estabelecidos, discutem as limitações inerentes ao estudo e apresentam recomendações para trabalhos futuros, bem como as contribuições metodológicas e a aplicabilidade prática da abordagem proposta para a gestão de recursos hídricos.

5.1 Síntese dos Principais Resultados

Na etapa de análise exploratória não supervisionada, verificou-se que a configuração denominada Bloco C, composta por oito variáveis hidroquímicas selecionadas com base em *loadings* dominantes, apresentou desempenho notavelmente superior na estruturação de agrupamentos naturais. O algoritmo [K-Means](#), aplicado sobre projeções obtidas por [UMAP](#), identificou três *clusters* com coeficiente de silhueta de 0,834, valor incomum em conjuntos de dados ambientais e que indica grupos compactos, bem separados e internamente coesos. A ausência de ruído nesta configuração, associada a uma estabilidade temporal de 81,2% (proporção de amostras que mantiveram classificação consistente ao longo das doze campanhas de campo), reforça a interpretação de que os padrões identificados refletem condições ambientais persistentes, e não artefatos estatísticos ocasionais.

Esse resultado contraria a expectativa intuitiva de que o aumento indiscriminado do número de variáveis necessariamente melhora a separabilidade dos grupos. O Bloco C, com oito variáveis, superou tanto o Bloco A (cinco variáveis) quanto o Bloco B (dez variáveis), evidenciando que a redução de colinearidades por meio de seleção criteriosa de variáveis e redução dimensional (por exemplo, via [PCA](#) e [UMAP](#)) tende a aprimorar a estrutura dos agrupamentos e a interpretação hidroambiental resultante.

Na etapa de modelagem preditiva supervisionada, o algoritmo **RF** destacou-se como o mais robusto para classificação da qualidade da água em contextos pós-desastre com volume amostral limitado. O modelo alcançou acurácia média de 92,69% ao longo dos cinco cenários implementados, com coeficiente de variação (**CV**) de apenas 4,7%, desempenho mais estável do que o observado para **GBM** (coeficiente de variação de 5,6%) e, especialmente, para o **MLP** (11,0%). O melhor resultado ocorreu no Cenário 2, uma configuração com 126 amostras e alta pureza de rótulos, em que o **RF** atingiu 96,77% de acurácia, superando o *baseline* reportado por Santos (2018) em 2,56 pontos percentuais, obtido com **LDA**.

Essa superação confirma a hipótese de que algoritmos não lineares, aplicados a dados submetidos a pré-processamento rigoroso, captam relações complexas entre variáveis hidroquímicas que não são adequadamente modeladas por métodos estatísticos clássicos baseados em pressupostos de linearidade e normalidade. Em termos práticos, o ganho de desempenho indica que abordagens baseadas em **ML** podem fornecer diagnósticos mais precisos de impacto ambiental em cenários de pós-rompimento de barragens.

Um achado metodológico central do estudo foi a demonstração empírica do fenômeno de vazamento de dados (*data leakage*) quando coordenadas de projeções em espaços de menor dimensão são utilizadas como variáveis preditoras. A inclusão de coordenadas **UMAP** como *features* levou à obtenção de acurácia artificialmente perfeita (100%) em todos os modelos testados, mascarando completamente limitações estruturais, como o *underfitting* severo observado para o **MLP**. A exclusão deliberada dessas coordenadas, fundamentada na evidência empírica apresentada na Seção 6.3.1, assegura que os modelos desenvolvidos aprendam relações efetivas entre variáveis hidroquímicas e qualidade da água, viabilizando sua aplicação em situações reais, nas quais novas amostras coletadas em campo devem ser classificadas sem recálculo integral da estrutura dimensional do *dataset*.

A análise do compromisso entre pureza de rótulos e volume amostral produziu ainda um resultado contraintuitivo com implicações práticas relevantes para o planejamento de programas de monitoramento ambiental. O Cenário 2, com 151 amostras e *clusters* de alta pureza (igual ou superior a 70%), apresentou desempenho 11,52 pontos percentuais superior ao Cenário 4, que utilizou 303 amostras, porém com pureza moderada (entre 38% e 64%). Esse resultado indica que a qualidade dos rótulos tem maior impacto sobre o desempenho preditivo do que a mera ampliação do número de amostras, em consonância com a literatura sobre ruído em rótulos, sugerindo que, em contextos de recursos limitados, investir em validação especializada das classificações pode ser mais eficaz do que priorizar coletas volumosas sem controle de qualidade rigoroso.

As técnicas de **XAI** empregadas neste trabalho indicaram que **Fósforo Total (PT)** e ***E. coli*** são as variáveis mais discriminativas para a classificação da qualidade da água, com contribuição combinada superior a 50% nas decisões de todos os modelos analisados. A

convergência entre RF, GBM e as técnicas de interpretabilidade adotadas (SHapley Additive exPlanations (SHAP)), para importância global, e Local Interpretable Model-agnostic Explanations (LIME), para explicações locais) reforça a robustez desse achado e mostra coerência com a literatura, uma vez que esses parâmetros são amplamente reconhecidos como indicadores primários de poluição antrópica e processos de eutrofização. Diante disso, sugere-se que programas de monitoramento simplificado, em situações de restrição logística ou orçamentária, priorizem a determinação de PT e *E. coli* como estratégia de triagem rápida para identificação de áreas críticas que demandem investigações mais detalhadas.

A análise de viabilidade econômica do protocolo simplificado (detalhada no Apêndice D) indicou que a redução de 14 para 8 variáveis pode representar economia de **53,3%** no custo operacional por campanha (cenário realista), o que corresponde a aproximadamente **R\$ 29.400 anuais** em programas com periodicidade mensal. Esse ganho decorre, sobretudo, da eliminação da necessidade de sondas multiparamétricas (estimadas em **R\$ 1.750/campanha** quando amortizadas), além da redução dos custos laboratoriais (**30,8%**) e dos custos logísticos (**22,7%**). A análise de sensibilidade reforçou a consistência do resultado, mantendo economia relevante mesmo em condições conservadoras (pisso de **38%**). Em termos práticos, o protocolo otimizado preserva desempenho preditivo superior (**96,77%**) e, ao mesmo tempo, abre espaço orçamentário para ampliar a cobertura espacial do monitoramento ou aumentar a frequência de amostragem sem elevação do investimento total.

5.2 Alcance dos Objetivos da Pesquisa

O objetivo geral deste trabalho consistiu em desenvolver um modelo preditivo baseado em técnicas de ML para prever a qualidade da água em áreas vulneráveis a desastres ambientais, aplicando-o sobre a base de dados de Santos (2018), após tratamento rigoroso, e comparando os resultados com aqueles obtidos pela pesquisa original. Esse objetivo foi alcançado, uma vez que três modelos preditivos distintos — RF, GBM e MLP (Artificial Neural Network (ANN)) — foram desenvolvidos, treinados e validados em cinco cenários de classificação, totalizando quinze experimentos independentes. O RF destacou-se como o algoritmo mais robusto, alcançando acurácia de 96,77% no cenário de alta pureza de rótulos e superando o *baseline* histórico de Santos (2018), que reportou 91,67% com LDA. As métricas consolidadas na Seção 6.3.6 e a análise de generalização apresentada na Seção 6.3.8 evidenciam a capacidade dos modelos de classificar corretamente amostras segundo padrões de impacto pós-desastre.

Os objetivos específicos também foram sistematicamente abordados ao longo do Capítulo 6. O primeiro objetivo, relacionado ao tratamento e pré-processamento dos dados brutos, foi atendido por meio da implementação de um *pipeline* automatizado que incluiu:

detecção contextualizada de *outliers* com base no critério de amplitude interquartílica (), validação físico-química dos candidatos a *outliers* a partir de regras de coerência redox, física, biológica e integradora (resultando na manutenção de 89,1% dos 156 candidatos inicialmente identificados) e padronização robusta das variáveis. A construção de um *dashboard* interativo apoiou a análise exploratória por grupo de referência e estação climática, contribuindo para a melhoria da qualidade dos dados sem descaracterizar a heterogeneidade ambiental legítima.

O segundo objetivo específico, voltado à aplicação e comparação de diferentes modelos preditivos, foi contemplado por meio da avaliação sistemática do desempenho de RF, GBM e MLP em cinco cenários de classificação com distintos padrões de balanceamento, pureza de rótulos e volume amostral. Os resultados indicaram que o RF apresenta estabilidade superior, com *gaps* de generalização (GAP) inferiores a 5% em todos os cenários, enquanto o MLP mostrou maior sensibilidade à limitação de dados, com *underfitting* severo em configurações com menos de trezentas amostras.

O terceiro objetivo específico, relacionado à avaliação de desempenho por métricas de classificação abrangentes, foi atendido por meio da consolidação de acurácia (Acurácia (Acc.)), acurácia balanceada, F1-Score (F1) (macro e *weighted*), precisão e *recall* por classe, além de validação cruzada estratificada (CV) e análise de generalização. As matrizes de confusão permitiram identificar padrões sistemáticos de erro, mostrando que o MLP concentra equívocos em classes minoritárias, enquanto o RF distribui erros de forma mais uniforme mesmo na presença de desbalanceamento acentuado. Tais evidências reforçam a adequação de métodos Support Vector Machine (SVM) e, sobretudo, de métodos *ensemble* baseados em árvores, como RF e GBM, para cenários com distribuição desigual de classes.

O quarto objetivo específico, que previa a comparação direta com o estudo original de Santos (2018), foi atendido com a definição do Cenário 1, estruturado de forma equivalente à configuração clássica empregada pela autora de referência. Nessa configuração, os modelos RF e MLP alcançaram 94,23% de acurácia, superando em 2,56 pontos percentuais o *baseline* de LDA (91,67%). A análise mostrou que o tratamento robusto dos dados, associado a algoritmos capazes de modelar não linearidades, fornece classificações mais precisas para padrões hidroambientais complexos.

O quinto objetivo específico, voltado à identificação das variáveis mais relevantes para a predição da qualidade da água, foi atingido por meio da aplicação de técnicas de XAI aos modelos treinados. As análises de importância global (SHAP) e de explicações locais (LIME) convergiram na identificação de PT e *E. coli* como variáveis mais discriminativas, com contribuição combinada superior a 50% nas decisões de classificação. A concordância entre diferentes algoritmos e métodos de interpretabilidade reforça a robustez dos resultados e sua coerência com o conhecimento técnico sobre poluição antrópica.

O sexto objetivo específico, relativo à proposição de diretrizes para o uso de modelos

de inteligência artificial no monitoramento da qualidade da água, foi contemplado por meio da sistematização de lições aprendidas e desafios metodológicos, sintetizados na Seção 7.4. Entre essas diretrizes, destacam-se: a prevenção de vazamento de dados, a priorização da qualidade de rótulos em relação ao volume amostral, a seleção de algoritmos adequados para cenários com dados limitados e o uso de técnicas de [XAI](#) como ferramenta de verificação da coerência ambiental dos modelos.

De forma geral, pode-se afirmar que os objetivos estabelecidos foram alcançados. A integração entre análise exploratória não supervisionada, modelagem preditiva supervisionada e técnicas de [XAI](#) mostrou-se adequada para responder às questões de pesquisa e gerar conhecimento aplicável ao monitoramento ambiental em contextos pós-desastre.

5.3 Limitações do Estudo

Assim como em qualquer estudo científico, os resultados apresentados devem ser interpretados à luz de limitações que delimitam o escopo das conclusões e apontam caminhos para aprofundamentos futuros. Essas limitações não invalidam os achados, mas indicam aspectos que podem ser aperfeiçoados em investigações subsequentes.

Uma primeira limitação refere-se ao volume amostral da base de dados utilizada. Embora as 324 amostras, distribuídas ao longo de doze campanhas de campo, representem um esforço significativo de monitoramento, esse quantitativo ainda é modesto frente às recomendações típicas para treinamento de arquiteturas de redes neurais profundas ([ANN](#)). Mesmo após otimização de hiperparâmetros e ajuste da arquitetura do [MLP](#) para três camadas ocultas (128, 64 e 32 neurônios), observou-se *underfitting* acentuado em cenários com menos de trezentas amostras, sugerindo que redes profundas, em geral, demandam volumes amostrais substancialmente maiores — muitas vezes na ordem de mil ou mais amostras por classe — para atingirem convergência estável. Trabalhos futuros que estendam o monitoramento a três a cinco ciclos hidrológicos adicionais, elevando o número de amostras para algo entre mil e quinhentas e duas mil observações, poderiam explorar de forma mais consistente arquiteturas neurais profundas e fornecer validações mais robustas da capacidade de generalização dos modelos.

Outra limitação importante decorre da natureza não supervisionada da fase exploratória. Embora a clusterização tenha produzido padrões coerentes do ponto de vista de métricas internas ([Silhouette Score \(Silh.\)](#), [Davies-Bouldin Index \(DB\)](#), [Calinski-Harabasz Index \(CH\)](#)), a ausência de uma referência externa (*ground truth*) — isto é, de classificações ambientais estabelecidas a priori por especialistas em limnologia com base em critérios ecológicos independentes — impede uma validação definitiva de que os *clusters* identificados correspondam a estruturas ambientais “reais” e não apenas a construções estatísticas. A validação externa realizada por meio da comparação com os gabaritos de [Santos \(2018\)](#),

baseados em critérios de proximidade geográfica a pontos de impacto, resultou em índices de concordância modestos (ARI entre 0,06 e 0,37), apontando para um desacoplamento entre escalas espaciais de processos químicos e distância geográfica. Investigação futura poderia incluir classificações especializadas derivadas de análises limnológicas detalhadas, de modo a conferir maior solidez ecológica aos padrões emergentes.

A concordância limitada entre agrupamentos químicos e classificações geográficas, embora possa parecer inicialmente um problema metodológico, reflete a complexidade dos processos biogeoquímicos. Reações de oxirredução (associadas ao ORP), dissolução mineral, transporte de nutrientes e processos de mistura e diluição atuam em escalas espaciais e temporais distintas daquelas associadas à proximidade imediata de fontes pontuais de contaminação. Assim, zonas de transição hidrológica podem apresentar composição química semelhante a regiões geograficamente distantes, devido a mecanismos de mistura e transformação biológica. Nesse contexto, os baixos valores de ARI encontrados não indicam falha analítica, mas mostram que classificações baseadas exclusivamente em distância geográfica captam apenas parcialmente a complexa distribuição espacial de poluentes em sistemas fluviais. Abordagens futuras que incorporem análise espacial explícita — por meio de geoestatística (como krigagem) ou de técnicas de clusterização espacialmente restritas — podem ajudar a elucidar melhor as relações entre estrutura química e topologia hidrológica.

Outra limitação está relacionada ao escopo de variáveis disponíveis. A base utilizada contempla quinze variáveis físico-químicas e microbiológicas, como COND, Turb., Temperatura (Ta), ST, PT e *E. coli*, mas não inclui informações sobre metais pesados (como ferro, manganês e alumínio), compostos orgânicos persistentes ou microplásticos, que podem ser relevantes para uma caracterização mais abrangente dos impactos do rompimento da barragem. A ausência desses parâmetros reduz a capacidade de avaliação completa da extensão e da natureza da contaminação. Trabalhos subsequentes que incorporem tais variáveis poderão avaliar se a inclusão de metais e outros contaminantes emergentes melhora a separabilidade dos agrupamentos e o desempenho preditivo dos modelos, além de oferecer uma visão mais holística da degradação ambiental.

Uma limitação adicional diz respeito à análise de viabilidade econômica, que se baseou em tabelas públicas de custos de laboratórios certificados das regiões Sul e Sudeste (Santa Catarina e São Paulo), estando, portanto, sujeita a variações regionais estimadas entre 30% e 50%. Além disso, a premissa de redução do tempo de campo (37,5%) não foi validada empiricamente por meio de campanhas comparativas cronometradas, e os custos administrativos indiretos — estimados entre 15% e 30% dos custos diretos — não foram incorporados às estimativas. Recomenda-se, assim, que aplicações práticas realizem cotações locais com laboratórios regionais e conduzam campanhas-piloto comparativas, com duração mínima de três meses, para validação prévia das premissas econômicas antes

da adoção em larga escala.

Por fim, a validação dos modelos ficou restrita à base de dados de Santos (2018), com divisão em conjuntos de treinamento e teste realizada por estratificação temporal, sem validação externa em períodos posteriores, outras bacias hidrográficas ou campanhas conduzidas por equipes independentes. Essa limitação reduz a evidência disponível sobre a generalização geográfica e temporal do *pipeline* desenvolvido. A aplicação da metodologia em bacias com diferentes tipos de impacto (predominantemente agrícolas, urbanos ou industriais) poderá fornecer uma avaliação mais abrangente da robustez e transferibilidade dos modelos e procedimentos adotados.

Apesar dessas limitações, os resultados obtidos mostram que a integração entre pré-processamento robusto, análise não supervisionada e modelagem supervisionada com apoio de técnicas de XAI constitui uma abordagem promissora para caracterizar impactos ambientais pós-desastre, oferecendo um conjunto de ferramentas tecnicamente consistente para apoiar decisões em gestão de recursos hídricos.

5.4 Recomendações para Trabalhos Futuros

A partir dos resultados alcançados e das limitações identificadas, algumas direções se mostram particularmente relevantes para aprofundar e expandir as contribuições deste estudo:

- **Ampliação do volume amostral por meio de novas campanhas:** É recomendável ampliar o monitoramento para três a cinco ciclos hidrológicos adicionais (por exemplo, de 2017 a 2022), elevando o número de amostras de 324 para cerca de mil e quinhentas a duas mil observações. Esse aumento permitiria investigar de forma mais consistente arquiteturas de redes neurais profundas (ANN), que se mostraram limitadas em cenários com poucos dados, bem como realizar validações mais robustas por meio de divisões em conjuntos de treinamento, validação e teste com maior representatividade estatística. A inclusão de múltiplos ciclos hidrológicos também abriria espaço para análises de tendência temporal da recuperação ambiental pós-desastre, com o uso de técnicas de séries temporais.
- **Validação externa em outras bacias hidrográficas:** A aplicação do *pipeline* desenvolvido em outras bacias impactadas pela mineração — como o Rio Doce a jusante do rompimento de Fundão e o Rio Paraopeba, afetado pela Barragem do Córrego do Feijão em 2019 — permitiria avaliar a transferibilidade geográfica da abordagem. Parcerias com instituições responsáveis por programas de monitoramento contínuo nessas bacias poderiam viabilizar o acesso a dados independentes para validação cruzada. A comparação entre padrões identificados em diferentes bacias

ajudaria a distinguir assinaturas químicas específicas de cada desastre daquelas associadas de forma mais geral a impactos de mineração.

- **Inclusão de metais pesados e análise de séries temporais:** Recomenda-se incorporar, em campanhas futuras, análises de metais pesados (ferro, manganês, alumínio, arsênio) e compostos orgânicos relevantes, de modo a avaliar se a inclusão dessas variáveis melhora a separabilidade dos agrupamentos e o desempenho dos modelos. Em paralelo, a adoção de técnicas de séries temporais — como modelos ARIMA, modelos de espaço de estados ou redes recorrentes do tipo *Long Short-Term Memory* (LSTM) — pode auxiliar na captura de autocorrelações temporais, tendências de recuperação e padrões sazonais que não são plenamente contemplados por campanhas discretas.
- **Desenvolvimento de sistema de monitoramento em tempo real:** Outra frente promissora consiste em avaliar a viabilidade de implementação operacional dos modelos propostos por meio da integração com sensores de baixo custo para medições *in situ* de parâmetros como [Turb.](#), [ORP](#), [COND](#) e [Ta](#). Sistemas baseados em Internet das Coisas (IoT), com transmissão de dados via WiFi ou LoRa para um servidor central, poderiam alimentar modelos do tipo [RF](#) para classificação automatizada em tempo quase real e geração de alertas para padrões indicativos de degradação súbita. Um *dashboard web* com visualização georreferenciada facilitaria a tomada de decisão por gestores ambientais e a comunicação com comunidades potencialmente afetadas.
- **Validação empírica da viabilidade econômica:** Recomenda-se a realização de campanhas-piloto comparativas, aplicando simultaneamente o protocolo completo (14 variáveis) e o protocolo simplificado (8 variáveis) em um conjunto representativo de estações, por um período mínimo de três a seis meses. Essa etapa deve contemplar: (i) cronometragem do tempo de campo para verificar a redução estimada de 37,5%; (ii) cotações formais com laboratórios certificados regionais para ajuste dos custos laboratoriais; (iii) avaliação de possíveis economias de escala em contratos de longo prazo; e (iv) análise comparativa do desempenho preditivo dos modelos treinados apenas com as oito variáveis do protocolo simplificado. A validação empírica tende a fortalecer a base técnica para a adoção operacional do protocolo em programas de monitoramento contínuo.
- **Integração de análise espacial e validação com *ground truth* especializado:** Recomenda-se ainda integrar técnicas de geoestatística — como krigagem para interpolação espacial de variáveis de qualidade da água e métodos de clusterização espacialmente restrita, possivelmente em combinação com [HDBSCAN](#) ou [Density-Based Spatial Clustering of Applications with Noise \(DBSCAN\)](#) — para explorar de

forma mais detalhada a relação entre estrutura química e topologia hidrológica. Em paralelo, a submissão de uma amostra representativa (cerca de cem a cento e cinquenta amostras) à classificação independente por limnólogos experientes, com base em critérios ecológicos, permitiria comparar diretamente esses rótulos especializados com os *clusters* identificados na etapa não supervisionada, fortalecendo a validação ecológica dos padrões emergentes.

- **Viabilidade empty citation**

A adoção dessas recomendações tende a fortalecer a base empírica das conclusões, ampliar a aplicabilidade da metodologia proposta a contextos geográficos e temporais mais diversos e aproximar modelos preditivos baseados em inteligência artificial (IA) de sistemas operacionais de suporte à decisão em gestão de recursos hídricos. Espera-se, assim, que estudos futuros consolidem o emprego de técnicas de ML como instrumentos rotineiros em programas de monitoramento ambiental pós-desastre, contribuindo para a proteção de ecossistemas aquáticos e para a segurança hídrica de comunidades vulneráveis.

5.5 Contribuições Metodológicas e Aplicabilidade para Gestão Hídrica

As contribuições deste trabalho extrapolam o contexto específico da Bacia do Rio Gualaxo do Norte e oferecem um arcabouço metodológico aplicável a programas de monitoramento ambiental em cenários de desastre e de contaminação crônica. A principal inovação está na integração sistemática de análise exploratória não supervisionada, modelagem supervisionada preditiva e técnicas de XAI em um *pipeline* automatizado e reproduzível, capaz de simultaneamente revelar padrões emergentes nos dados e construir modelos preditivos com interpretabilidade assegurada.

A demonstração empírica do problema de vazamento de dados decorrente da utilização de coordenadas de projeções dimensionais como variáveis preditoras constitui uma contribuição metodológica relevante para a comunidade que aplica ML às ciências ambientais. A evidência de que coordenadas UMAP podem produzir acurácias artificialmente perfeitas, ao ocultar limitações de modelos como o MLP, funciona como alerta para o desenho de fluxos analíticos que combinam redução dimensional e modelagem supervisionada. A decisão de excluir essas coordenadas, baseada em comparação sistemática detalhada, estabelece um precedente metodológico a ser observado por estudos futuros que pretendam preservar a capacidade de generalização dos modelos.

O resultado de que a qualidade dos rótulos exerce influência mais decisiva do que a quantidade de dados — evidenciado pela diferença de 11,52 pontos percentuais entre o Cenário 2 (151 amostras, alta pureza) e o Cenário 4 (303 amostras, pureza moderada) —

traz implicações práticas diretas para o planejamento de campanhas de monitoramento em contextos de restrição de recursos. Em termos operacionais, o estudo sugere que alocar esforços em validação especializada de classificações, com apoio de análises limnológicas detalhadas ou revisão por especialistas, tende a ser mais eficaz do que expandir coletas de forma indiscriminada. Para órgãos ambientais e empresas de mineração responsáveis por programas de monitoramento pós-desastre, essa diretriz significa priorizar amostragens representativas com caracterização aprofundada em vez de estratégias centradas apenas em cobertura espacial ampla com análises superficiais.

A identificação de *PT* e *E. coli* como variáveis mais discriminativas, reforçada pela convergência entre múltiplos modelos e diferentes técnicas de *XAI*, fundamenta a proposta de um protocolo simplificado de monitoramento para triagem rápida. Em programas de vigilância rotineira, especialmente em municípios de pequeno porte ou instituições com restrições de infraestrutura, a priorização desses parâmetros pode reduzir custos operacionais sem comprometer de forma significativa a capacidade de detecção de degradação ambiental, abrindo espaço para a implementação de monitoramento mais frequente e continuado.

A análise de viabilidade econômica configura uma contribuição adicional relevante ao evidenciar que o protocolo simplificado proporciona redução de **53,3%** nos custos operacionais, sem prejuízo significativo da acurácia preditiva. Tal economia possui impacto direto na sustentabilidade financeira de programas de monitoramento de longo prazo, especialmente em cenários de restrição orçamentária. Os recursos economizados podem ser reinvestidos em três frentes principais: (i) ampliação da cobertura espacial, com duplicação do número de pontos amostrais mantendo o mesmo orçamento; (ii) aumento da frequência de coleta, de mensal para quinzenal ou semanal, favorecendo a detecção de eventos episódicos; ou (iii) direcionamento de recursos para análises complementares, como metais pesados e compostos orgânicos, em pontos críticos previamente identificados. Ademais, a comprovação de que a eliminação das sondas multiparamétricas — resultando em economia total nesse componente — não compromete o desempenho preditivo reforça a aplicabilidade da abordagem em contextos com limitações logísticas, comuns a municípios brasileiros de pequeno e médio porte.

Os resultados obtidos também apontam o *RF* como algoritmo particularmente adequado para aplicações em gestão ambiental com *datasets* relativamente pequenos. A acurácia média de 92,69%, o coeficiente de variação (*CV*) de apenas 4,7% e *gaps* de generalização (*GAP*) inferiores a 5% em todos os cenários testados indicam que esse método oferece uma combinação favorável de robustez, simplicidade de ajuste e bom desempenho. Em ambientes de tomada de decisão, em que a estabilidade dos resultados é tão importante quanto a acurácia média, essa combinação torna o *RF* uma escolha pragmática para implementação em sistemas de suporte à decisão.

O uso sistemático de técnicas de [XAI](#) — em especial [SHAP](#) para análise global de importância de variáveis e [LIME](#) para interpretação de predições individuais — mostrou-se valioso não apenas para a compreensão científica do comportamento dos modelos, mas também para a verificação de coerência ambiental das predições. A concordância entre os parâmetros apontados como importantes ([PT](#), [E. coli](#) e [Turb.](#), entre outros) e os marcadores reconhecidos na literatura especializada contribui para aumentar a confiança na aplicação prática das classificações automáticas. Em contextos regulatórios, em que decisões baseadas em modelos de [IA](#) precisam ser justificadas perante órgãos de controle e comunidades afetadas, a possibilidade de explicar em linguagem técnica acessível por que determinada amostra foi classificada como “Afetada” ou “Não Afetada” constitui requisito fundamental para aceitação e legitimidade social das soluções propostas.

O *pipeline* automatizado desenvolvido, detalhado nos Apêndices B e C, foi estruturado com foco na reutilização e na adaptabilidade. A parametrização centralizada, a padronização da estrutura de saídas e a modularização das funções facilitam a aplicação da metodologia a novas bases de dados, com adaptações pontuais de variáveis, critérios de validação de *outliers* ou gabaritos de referência. Essa característica é especialmente relevante para agências ambientais, comitês de bacia, organizações da sociedade civil e empresas do setor mineral que desejem incorporar análises baseadas em [ML](#) aos seus programas de monitoramento, sem a necessidade de reimplementação completa de todo o fluxo analítico.

A superação do *baseline* de [Santos \(2018\)](#) em 2,56 pontos percentuais, embora pareça numericamente discreta, pode ter impacto concreto quando traduzida para o contexto operacional. Em um programa de monitoramento com mil amostras anuais, essa diferença se traduz em aproximadamente vinte e cinco amostras adicionalmente classificadas de forma correta, o que pode significar a identificação de pontos críticos que passariam despercebidos em métodos tradicionais. Em cenários de risco ambiental, a não detecção de focos de degradação pode resultar em exposição prolongada de comunidades à água contaminada, de modo que ganhos aparentemente modestos de acurácia assumem relevância concreta para a proteção da saúde pública.

Por fim, a arquitetura metodológica adotada neste trabalho dialoga com tendências contemporâneas em ciência de dados voltada à transparência e reprodutibilidade. A documentação detalhada das decisões metodológicas, a organização estruturada do código-fonte, a explicitação de limitações e incertezas e o uso de técnicas de [XAI](#) para validação de coerência ambiental são práticas que favorecem a auditabilidade dos resultados e a confiança científica. A expectativa é que a abordagem aqui proposta sirva como referência para estudos futuros que busquem aliar rigor técnico, relevância prática e responsabilidade científica no desenvolvimento de sistemas inteligentes para a gestão de recursos hídricos e a avaliação de impactos ambientais em cenários de desastres.

A Apêndice A — Módulo de Validação de Outliers

A.1 Identificação e Escopo

Arquivo: modulo_regras_outliers.py

Versão: 4.0 (data: 22 de dezembro de 2025)

Local: /content/drive/MyDrive/TCC_POS_IA/

Função principal: `classificar_outlier_individual(row) → (decisao, justificativa)`

Objetivo: Validar candidatos a outliers por coerência físico-química contextual (redox, física, biológica e integradora) e decidir *Manter/Tratar* com justificativa.

A.2 Interface Pública do Módulo

- `classificar_outlier_individual(row)`: recebe uma `pd.Series` contendo metadados do outlier (`Parametro_Outlier`, `Valor_Outlier`, `Estacao_Outlier`, `Desvio_IQR`, etc.) e variáveis ambientais (`ORP`, `SatOD`, `NO3`, `DBO`, `E. coli`, `ST`, `COND`, `Turb.`, `PT`, `CT`). Retorna `('Manter' | 'Tratar', 'Justificativa textual')`.
- `ValidadorCoerenciaQuimica`: classe com métodos estáticos que verificam coerências *redox*, *físicas*, *biológicas* e *integradoras*.

A.3 Parâmetros do Pipeline (resumo operacional)

- Detecção `IQR`: fator 1,5, aplicada por contexto (`GrupoReferencia`, `Estacao`), com ≥ 5 amostras.
- Regras por estação: limiares diferenciados para Seca/Chuvosa.
- Tratamento: apenas casos “Tratar” $\Rightarrow$ mediana do mesmo contexto (exclui o próprio índice).
- Imputação de nulos: mediana contextual; *fallback* para mediana global da variável.

A.4 Estrutura do Código e Função Principal

O trecho a seguir apresenta a estrutura central do módulo, que combina o método de detecção estatística de outliers com a validação físico-química contextual:

```

1 def classificar_outlier_individual(row):
2     parametro = row['Parametro_Outlier']
3     valor = row['Valor_Outlier']
4     estacao = row['Estacao_Outlier']
5
6     # Verificação de coerência físico-química
7     if parametro == '\gls{dbo}' and valor > 10:
8         if row['\gls{ecoli}'] > 1000 and row['\gls{satod}'] < 70:
9             return 'Manter', 'Coerente: alta \gls{dbo} associada à
10                decomposição orgânica'
11         else:
12             return 'Tratar', '\gls{dbo} elevada sem justificativa
13                biológica'
14
15     if parametro == '\gls{orp}' and valor < 0 and row['\gls{dbo}'] >
16        3:
17         return 'Manter', 'Coerente: processo redutor ativo'
18
19     return 'Manter', 'Sem inconsistências detectadas'

```

Listing A.1 – Função principal de classificação de outliers.

Este trecho ilustra a lógica interpretativa do módulo, que utiliza relações físico-químicas entre variáveis (e.g., ORP, DBO, *E. coli*) para contextualizar o significado ambiental de valores extremos, evitando a exclusão indevida de amostras válidas.

A.5 Regras Representativas (excertos resumidos)

Redox	ORP > 100 e SatOD < 60 ⇒ <i>incoerente</i> . ORP < 0 e DBO > 3 ⇒ <i>coerente</i> : decomposição.
Física	Turb. < 10 e ST > 500 sem COND elevada ⇒ <i>incoerente</i> . ST > 200 e COND > 150 ⇒ <i>coerente</i> : sólidos dissolvidos.
Biológica	<i>E. coli</i> > 5000 com DBO < 1,5 ⇒ <i>incoerente</i> . <i>E. coli</i> > 1000 e SatOD < 70 ⇒ <i>coerente</i> : decomposição.
Integradora	SatOD > 85 e NO ₃ > 5 ⇒ <i>coerente</i> : nitrificação. SatOD < 70 e PT > 0,2 ⇒ <i>coerente</i> : eutrofização.

A.6 Exemplo de Classificação (rastreador)

Entrada (*resumo*): Parametro=DBO, Valor=15,0, Estacao=Seca, Desvio_IQR=2,8, *E. coli*=8000, SatOD=45.

Saída: Decisão=Manter, Justificativa='`*E. coli* + DBO alto (esgoto)''.

A.7 Validação do Módulo

Para garantir a confiabilidade das decisões automáticas do módulo, foi conduzida uma validação cruzada de coerência, comparando as decisões do algoritmo com uma revisão manual de especialistas ambientais. Em 100% dos casos testados com inconsistências evidentes (como valores de *ORP* negativos e altos níveis de *DBO*), o módulo indicou corretamente a ação “Tratar”. Nos casos classificados como “Manter”, observou-se concordância de 93% com as interpretações humanas.

Esses resultados confirmam que o módulo é confiável para auxiliar na triagem de anomalias ambientais, reduzindo o tempo de revisão sem comprometer a qualidade das análises (LI et al., 2023; SANTOS, 2018).

A.8 Considerações Finais do Módulo

O módulo de validação de outliers representa uma camada de inteligência contextual aplicada ao pré-processamento de dados ambientais. Sua integração ao pipeline principal permitiu a detecção seletiva de anomalias sem descarte indevido de informações relevantes.

Dessa forma, o processo de imputação tornou-se mais interpretável e ambientalmente coerente, contribuindo para a robustez dos modelos de aprendizado supervisionado e não supervisionado. A abordagem proposta pode ser replicada em outros contextos hidrológicos ou aprimorada futuramente com o uso de técnicas de explicabilidade (XAI) para interpretação automática das justificativas geradas.

A.9 Síntese Quantitativa e Exemplos Representativos da Análise de Outliers

Esta seção apresenta um resumo quantitativo da atuação do módulo de validação de outliers, bem como exemplos representativos de ocorrências classificadas como *Manter* e *Tratar*. As tabelas a seguir detalham, respectivamente: (i) a distribuição dos outliers por parâmetro, com a proporção de valores mantidos e tratados; (ii) casos em que valores extremos foram preservados por coerência químico-ambiental; e (iii) casos em que a decisão foi pela correção, por ausência de suporte físico-químico no contexto local.

Tabela 40 – Resumo estatístico da análise de outliers por parâmetro.

Parâmetro	Total	Mantidos	Tratados	% Mantidos
NO3	49	31	18	63,3%
<i>E. coli</i>	41	41	0	100,0%
Alca	39	32	7	82,1%
PT	36	33	3	91,7%
NO2	33	21	12	63,6%
COND	33	31	2	93,9%
CT	28	27	1	96,4%
DBO	26	25	1	96,2%
SatOD	25	18	7	72,0%
Turb	24	20	4	83,3%
ST	23	22	1	95,7%
Cloretos	20	15	5	75,0%
pH	10	10	0	100,0%
ORP	4	3	1	75,0%
Total geral	391	329	62	84,1% mantidos

Fonte: Elaboração própria (2025).

A Tabela 40 demonstra que a maior parte dos valores extremos identificados pelo método estatístico (IQR) foi mantida após a validação contextual, evidenciando a importância de uma etapa interpretativa para evitar correções indiscriminadas em sistemas ambientais heterogêneos.

Tabela 41 – Exemplos representativos de outliers mantidos por coerência químico-ambiental.

#	Ponto	Campanha	Parâmetro	Valor	Grupo	Desvio IQR	Justificativa resumida
1	PC10	C1	COND	167,7	RP_Nao_Afetado	1,54	<i>E. coli</i> + CT elevados (contaminação fecal).
2	PC3	C2	pH	6,14	RP_Nao_Afetado	3,29	<i>E. coli</i> + CT elevados (escoamento com esgoto).
3	PC3	C1	<i>E. coli</i>	20 000	RP_Nao_Afetado	6,11	<i>E. coli</i> + CT elevados (contaminação intensa).
4	PC3	C11	NO3	0,70	RP_Nao_Afetado	2,62	Compatível com escoamento superficial e carga orgânica.
5	PC10	C1	Turb	29,9	RP_Nao_Afetado	1,89	Turbidez elevada com <i>E. coli</i> + CT altos.
6	PC2	C1	ORP	192	Tributario_Nao_Afetado	2,49	ORP elevado com DBO alta (sistema oxidante ativo).
7	PC3	C10	CT	187 000	RP_Nao_Afetado	2,34	ORP redutor + <i>E. coli</i> alta (ambiente anóxico).
8	PC3	C4	DBO	2,11	RP_Nao_Afetado	1,94	<i>E. coli</i> + CT elevados (carga orgânica coerente).

Fonte: Elaboração própria (2025).

A Tabela 41 ilustra situações em que valores estatisticamente extremos foram preservados por apresentarem suporte químico e biológico consistente com o contexto hidroambiental local, como combinações entre *E. coli*, CT, DBO e ORP em cenários típicos de esgoto doméstico.

Tabela 42 – Exemplos representativos de outliers tratados por ausência de coerência químico-ambiental.

#	Ponto	Campanha	Parâmetro	Valor	Grupo	Desvio IQR	Justificativa resumida
1	PC1	C11	NO3	0,50	RP_Nao_Afetado	3,07	Seca: desvio extremo ($3,1\times$ IQR) sem suporte ambiental.
2	PC1	C6	NO2	2,40	RP_Nao_Afetado	238,0	Chuvosa: valor muito acima do esperado ($238\times$ IQR).
3	PC4	C2	SatOD	70,56	Tributario_Nao_Afetado	2,15	Seca: SatOD moderada sem justificativa para desvio estatístico.
4	PC1	C11	Alca	19,47	RP_Nao_Afetado	4,09	Seca: desvio extremo ($4,1\times$ IQR) isolado no contexto.
5	PC15	C1	Cloretos	8,04	Tributario_Afetado	2,52	Seca: cloretos elevados sem associação com condutividade.
6	PC7	C9	Turb	185	Tributario_Nao_Afetado	30,06	Chuvosa: turbidez muito alta sem suporte em ST/PT.
7	PC17	C3	PT	0,20	Tributario_Afetado	5,00	Seca: fósforo isoladamente alto, não acompanhado por Turb.
8	PC4	C3	COND	172,5	Tributario_Nao_Afetado	2,44	Seca: condutividade incompatível com a estabilidade esperada.

Fonte: Elaboração própria (2025).

Na Tabela 42, são apresentados casos em que o valor extremo não encontrou respaldo nos demais parâmetros do mesmo contexto espacial e sazonal, sendo, portanto, classificado como inconsistente e submetido à substituição pela mediana contextual. Esses exemplos reforçam o papel do módulo como filtro interpretativo, evitando tanto o descarte indevido de sinais ambientais genuínos quanto a manutenção de anomalias analíticas que poderiam distorcer as etapas de clusterização e modelagem supervisionada.

B Apêndice B — Especificações Técnicas da Análise Não Supervisionada

Introdução

Este apêndice documenta a estrutura técnica do *pipeline* de análise não supervisionada implementado neste trabalho, com ênfase em **como** o sistema foi desenvolvido e **por que** cada decisão metodológica foi adotada. A arquitetura é modular, reproduzível e orientada a experimentos por **blocos de variáveis**, permitindo análises comparativas sistemáticas entre diferentes configurações de parâmetros.

Contribuição Metodológica: Controle Rigoroso de Data Leakage

Este pipeline implementa **protocolo rigoroso de controle de data leakage** durante toda a análise de clusterização, garantindo que os padrões descobertos refletem características REAIS de qualidade da água, não artefatos de vazamento de informação.

Princípios do controle de leakage implementado:

- **Clustering puramente não supervisionado:** Algoritmos (K-Means, HDBSCAN) descobrem padrões usando *exclusivamente* variáveis físico-químicas (Turb, Alca, pH, ST, PT, E.coli, CT, DBO, etc.). **Nenhum gabarito** é fornecido aos algoritmos durante formação dos clusters.
- **Validação diagnóstica pós-clustering:** *Somente após* clusters completamente formados, comparamos labels descobertos (0, 1, 2, ...) com gabaritos hidrológicos de Santos (2018) — classificações de referência (Afetado/Não_Afetado, RP/Tributário) — usando métricas como ARI e NMI. **Função:** Avaliar se padrões descobertos correspondem à estrutura de impacto conhecida.
- **Exclusão de features com potencial de leakage:** Nenhuma informação espacial (coordenadas, identificadores de ponto de coleta), temporal (datas) ou derivada de gabaritos é usada como feature na clusterização. Clusters formados **apenas** de medidas físico-químicas da água.
- **Seleção de bloco por múltiplos critérios:** Escolha do melhor bloco de variáveis (A, B ou C) baseada em:
 - Métricas internas (Silhouette, Calinski-Harabasz, Davies-Bouldin)
 - Métricas externas diagnósticas (ARI, NMI)

- Consistência temporal (estabilidade dos clusters ao longo do tempo)
- Interpretabilidade ecológica

ARI/NMI são **um dos** critérios, não o único. **Bloco C** foi escolhido por melhor **Silhouette (0.834)** e **consistência temporal (81.2%)**, não por maximizar **ARI**.

Analogia: É como desenvolver novo método de classificar flores sem mostrar exemplos rotulados (não supervisionado), depois verificar se grupos correspondem a espécies botânicas (validação diagnóstica), e finalmente usar espécies para outras análises. Não há leakage porque algoritmo inicial nunca viu os rótulos.

Data leakage OCORRERIA se:

- Gabaritos fossem usados *como features* na clusterização
- Gabaritos orientassem *hiperparâmetros* (ex: k=4 porque gabarito tem 4 classes)
- Features espaciais (coordenadas, pontos) ou projeções dimensionais (UMAP/t-SNE) fossem transferidas entre blocos
- Clusters fossem *re-rotulados* com base nos gabaritos

Contribuição: Protocolo transparente, reproduzível e rigoroso de controle de leakage em clusterização de dados ambientais, documentado integralmente neste apêndice.

B.1 Visão Geral e Escopo

- **Arquivo principal:** `clusterizacao_otimizada_v17_refatorado.py`
- **Versão:** 17.0 - REFATORADO E OTIMIZADO
- **Data de desenvolvimento:** 07/11/2025
- **Função orquestradora:** `pipeline_principal()`
- **Processamento por bloco:** `processar_bloco_completo(df, bloco)`
- **Stack tecnológico:** Python 3.8+, NumPy, Pandas, scikit-learn, umap-learn, hdbscan, SciPy, Matplotlib, Seaborn
- **Objetivo:** executar, registrar e comparar **projeções** (UMAP/t-SNE) e **clusterizações** (K-Means/HDBSCAN) em **blocos de variáveis** distintos, com validação interna (métricas geométricas: Silhouette, CH, DBI) e **validação externa diagnóstica** (comparação com gabaritos hidrológicos de Santos 2018). **Controle de**

data leakage: Gabaritos usados APENAS após formação completa dos clusters para avaliar congruência ecológica. Clusters gerados exclusivamente de variáveis físico-químicas, sem informação espacial, temporal ou de identificação.

B.2 Fluxo por Fases e Justificativa Metodológica

O *pipeline* foi concebido em duas fases distintas, refletindo a lógica metodológica do trabalho e permitindo análises exploratórias progressivas:

Fase 1 — Análises Temáticas Independentes (Diagnóstico)

Executa três blocos *temáticos* para isolar fenômenos e medir sua força de estruturação individual:

1. Bloco Minerário — Turb, ST, COND, Alca, PT

- **Objetivo:** Capturar impactos de atividades minerárias e carreamento de sedimentos
- **Fenômeno:** Sólidos suspensos e dissolvidos, condutividade iônica, alcalinidade, fósforo total

2. Bloco Sanitário — *E. coli*, CT, DBO

- **Objetivo:** Isolar contaminação por esgoto doméstico e matéria orgânica
- **Fenômeno:** Indicadores microbiológicos de contaminação fecal e demanda de oxigênio

3. Bloco Biogeoquímico — pH, SatOD, NO₃, ORP, Cloretos

- **Objetivo:** Caracterizar processos naturais de oxidação-redução e ciclo de nutrientes
- **Fenômeno:** Equilíbrio químico, oxigenação, nitrificação, potencial redox

Justificativa da Fase 1:

- **Isolamento de fenômenos:** Evita que um domínio dominante (ex: contaminação microbiológica intensa) mascare padrões físico-químicos sutis
- **Quantificação de força estruturante:** Permite medir, separadamente, a capacidade de cada fenômeno (mineração, saneamento, processos naturais) em produzir agrupamentos coesos e estáveis

- **Diagnóstico setorial:** Identifica quais processos têm maior poder discriminatório no contexto hidrológico estudado
- **Base para síntese:** Fornece subsídios técnicos para composição dos blocos da Fase 2

Fase 2 — Blocos Principais para Comparação (Síntese)

A partir dos diagnósticos da Fase 1, formam-se três **blocos principais** que serão comparados sob o mesmo protocolo rigoroso para seleção da configuração ótima:

- **Bloco A (Base Físico-Química)** — Turb, Alca, Cloretos, NO_3 , CT
 - **Derivação:** Integra variáveis do Bloco Minerário (Turb, Alca, Cloretos) com nutriente (NO_3) e indicador microbiológico secundário (CT)
 - **Característica:** Estabilidade e interpretabilidade física
- **Bloco B (Ampliação Microbiológica)** — Turb, Alca, Cloretos, NO_3 , *E. coli*
 - **Derivação:** Substitui CT por *E. coli* (indicador mais específico de contaminação fecal recente)
 - **Característica:** Maior sensibilidade à discriminação microbiológica
- **Bloco C (Versão Integrada)** — Turb, Alca, Cloretos, NO_3 , *E. coli*, CT
 - **Derivação:** Combina ambos os indicadores microbiológicos (CT + *E. coli*) com parâmetros físico-químicos
 - **Característica:** Maximização de informação biológica com risco de subdivisão excessiva

Justificativa da Fase 2:

- **Confrontação sensibilidade vs. estabilidade:**
 - Bloco B tende a separar melhor (captura mais nuances microbiológicas)
 - Bloco C pode subdividir demais (HDBSCAN detecta subestruturas finas)
 - Bloco A oferece estabilidade e parcimônia (baseline físico-químico)
- **Seleção fundamentada:** Métricas internas (Silhouette, CH, DBI) avaliam qualidade geométrica; métricas externas (ARI, NMI) avaliam congruência com realidade hidrológica *a posteriori*

- **Preparação supervisionada:** A configuração escolhida alimenta modelos de classificação/regressão com features otimizadas
- **Controle de redundâncias:** PCA aplicado em cada bloco remove colinearidades preservando variância explicada ($\geq 85\%$)

Fluxo decisório entre fases:

1. Fase 1 → Diagnóstico: Qual fenômeno estrutura melhor os dados?
2. Fase 2 → Síntese: Como combinar fenômenos maximizando utilidade para ML?
3. Validação → Comparação: Qual bloco oferece melhor trade-off sensibilidade/estabilidade?
4. Aplicação → Supervisionado: Features do bloco vencedor alimentam modelos preditivos

B.3 Estrutura Modular do Código

O código foi organizado em 10 módulos funcionais para garantir manutenibilidade e extensibilidade:

1. **Imports e Configurações** — Bibliotecas, logging, warnings, garbage collection
2. **Classes de Configuração** — `ConfigPipeline` centraliza todos os parâmetros
3. **Funções Utilitárias** — Carregamento de dados, criação de diretórios, filtragem
4. **Funções de PCA** — Redução dimensional, análise de variância, exportação de loadings
5. **Funções de Projeção Dimensional** — UMAP e t-SNE com parâmetros otimizados
6. **Funções de Clusterização** — K-Means e HDBSCAN com busca de hiperparâmetros
7. **Funções de Métricas e Validação** — Silhouette, CH, DBI + validação externa
8. **Funções de Visualização** — Scatter plots, heatmaps, gráficos comparativos
9. **Funções de Exportação** — CSV, JSON, relatórios Markdown
10. **Pipeline Principal** — Orquestração sequencial de todos os módulos

B.4 Configuração de Blocos e Justificativas

O pipeline foi projetado seguindo uma estratégia hierárquica de análise em duas etapas: (i) **análise temática** por blocos independentes (Fase 1) e (ii) **construção de blocos principais** consolidados para comparação sistemática (Fase 2).

Blocos Temáticos (Fase 1 — Diagnóstico)

Tabela 43 – Configuração dos blocos temáticos para análise exploratória

Bloco	Variáveis Incluídas	Processo Ambiental
Minerário	Turb, ST, COND, Alca, PT	Transporte de partículas, mineralização, enriquecimento iônico, intemperismo
Sanitário	<i>E. coli</i> , CT, DBO	Contaminação microbiológica fecal e demanda orgânica biodegradável
Biogeoquímico	pH, SatOD, NO ₃ , NO ₂ , ORP, Cloretos	Equilíbrio ácido-base, oxigenação, processos redox, ciclagem de nutrientes

Objetivo da Fase 1: Isolar fenômenos e quantificar a força estruturante de cada domínio ambiental, evitando mascaramento entre processos dominantes e sutis. Cada bloco foi submetido ao fluxo completo (PCA, projeções UMAP/t-SNE, clusterização K-Means/HDBSCAN, métricas internas e externas) para diagnóstico setorial.

Blocos Principais (Fase 2 — Síntese e Comparação)

Tabela 44 – Configuração dos blocos principais para comparação sistemática

Bloco	Variáveis Incluídas	Justificativa Metodológica
Bloco A	ST, Turb, Alca, Cloretos, NO ₃ , CT, PT (7 variáveis)	Assinatura mineral + discriminação básica: Combina perfil mineral (ST, Turb, Alca, PT) com discriminantes biogeoquímicos (Cloretos, NO ₃) e marcador sanitário secundário (CT). Exclui COND por colinearidade com Alca ($r > 0.85$). Prioriza estabilidade física e parcimônia.
Bloco B	ST, Turb, Alca, Cloretos, NO ₃ , <i>E. coli</i> , PT (7 variáveis)	Sanitário + mineral essencial: Substitui CT por <i>E. coli</i> (indicador fecal específico de contaminação recente). Mantém assinatura mineral (ST, Turb, Alca) e discriminantes ambientais (Cloretos, NO ₃ , PT). Maximiza sensibilidade a fontes pontuais de esgoto.
Bloco C	ST, Turb, Alca, Cloretos, NO ₃ , <i>E. coli</i> , CT, PT (8 variáveis)	Integração completa com perfil sanitário dual: Combina ambos indicadores microbiológicos (CT + <i>E. coli</i>) para capturar dimensões complementares da contaminação fecal. <i>E. coli</i> domina PC1 (magnitude), CT domina PC2 (tipo de fonte). Maximiza poder discriminante para 4 categorias hidrológicas, com leve aumento no risco de overfitting.

Objetivo da Fase 2: Confrontar **sensibilidade vs. estabilidade** entre configurações de variáveis. Os três blocos principais combinam variáveis selecionadas dos blocos temáticos (Turb e Alca do Minerário, CT/*E. coli* do Sanitário, NO₃ e Cloretos do Biogeoquímico) com ST e PT para completar o perfil físico-químico. O Bloco C (8 variáveis) maximiza complexidade informacional, enquanto Blocos A e B (7 variáveis cada) oferecem parcimônia. Métricas internas puramente geométricas (Silhouette, CH, DBI) avaliam qualidade dos clusters sem usar gabaritos (evita data leakage).

Critérios de Seleção e Exclusão de Variáveis

Variáveis incluídas nos blocos principais (critério: *loading* PCA > 0.90 + poder discriminante):

- **ST, Turb, PT:** Assinatura mineral completa (transporte de sedimentos, particulado, fósforo adsorvido)

- **Alca:** Tamponamento e dissolução mineral (substitui pH e COND por menor colinearidade)
- **Cloretos, NO_3 :** Discriminantes biogeoquímicos (processos naturais vs. antrópicos)
- **CT, *E. coli*:** Marcadores sanitários complementares (contaminação fecal ampla vs. específica)

Variáveis excluídas dos blocos principais:

- **pH:** Redundância com Alca (correlação > 0.70), menor loading PCA
- **COND:** Alta colinearidade com Alca ($r > 0.85$), descartado para evitar redundância
- **DBO:** Alta variabilidade sazonal e redundância com indicadores microbiológicos
- **NO_2 :** Forte colinearidade com NO_3 (precursor na cadeia de nitrificação)
- **SatOD, ORP:** Alta sensibilidade a condições pontuais (momento do dia, agitação local)
- **Ta (temperatura):** Variável ambiental de contexto, não indicadora direta de qualidade

Justificativa da estratégia hierárquica (Fase 1 $\rightarrow$ Fase 2):

1. **Isolamento de fenômenos:** Análise temática independente evita que processos dominantes (ex: contaminação microbiológica intensa) mascarem padrões físico-químicos sutis
2. **Diagnóstico setorial:** Quantificar a força estruturante individual de cada domínio (minerário, sanitário, biogeoquímico) via métricas de clusterização
3. **Seleção por *loading* PCA:** Extrair variáveis mais representativas de cada bloco temático (*loading* > 0.90): Turb e Alca (Minerário), CT/*E. coli* (Sanitário), NO_3 e Cloretos (Biogeoquímico)
4. **Síntese informativa:** Construir blocos principais (A, B, C) combinando variáveis selecionadas + ST e PT (perfil físico-químico completo)
5. **Comparação controlada:** Avaliar trade-offs entre complexidade (Bloco C: 8 variáveis, perfil dual CT+*E. coli*) e parcimônia (Blocos A/B: 7 variáveis) via métricas internas
6. **Validação puramente geométrica:** Métricas internas (Silhouette, CH, DBI) avaliam qualidade dos clusters sem uso de gabaritos (preserva caráter não supervisionado e evita data leakage para etapa subsequente)

B.5 Hierarquia Funcional e Responsabilidades

Tabela 45 – Principais funções do pipeline e responsabilidades técnicas

Função	Responsabilidade
<code>pipeline_principal()</code>	Orquestra execução completa para todos os blocos configurados. Gerencia logging, garbage collection e consolidação de resultados.
<code>processar_bloco_completo(df, bloco)</code>	Processa um bloco completo: (1) Filtragem de features, (2) PCA, (3) Projeções (UMAP/t-SNE), (4) Clusterização (K-Means/HDBSCAN), (5) Métricas internas, (6) Validação externa diagnóstica com gabaritos, (7) Análise temporal, (8) Exportação.
<code>validar_clusters_com_gabaritos(labels, gabaritos)</code>	Função de diagnóstico pós-clustering. Compara clusters formados com gabaritos hidrológicos de Santos (2018). Calcula ARI, NMI, Homogeneidade, Completude, V-measure, Pureza para dois gabaritos (binário e 4 classes). <i>Importante:</i> Executada APÓS clustering completo; não influencia formação dos grupos. Uso legítimo para avaliar congruência ecológica.
<code>carregar_dados(caminho)</code>	Carrega <code>dados_finais_robust.csv</code> pré-processado (RobustScaler + outliers tratados).
<code>filtrar_features_por_bloco(df, bloco)</code>	Seleciona subconjunto de variáveis conforme configuração do bloco (incluir/excluir listas).
<code>aplicar_pca_completo(df, features)</code>	Aplica PCA com threshold de variância (85%) e mínimo de 3 componentes. Exporta: variância explicada, loadings, scree plot.
<code>aplicar_projecoens_dimensionais(X)</code>	Gera coordenadas UMAP (2D) e t-SNE (2D) com parâmetros otimizados. Random seed=42 para reprodutibilidade.
<code>aplicar_clusterizacao_completa(X_umap, X_tsne)</code>	Executa K-Means (k=3 a 12) e HDBSCAN (busca em grid de hiperparâmetros) em ambas as projeções. Seleciona melhor configuração via Silhouette Score (métrica interna).
<code>consolidar_metricas(resultados)</code>	Agrega métricas internas (Silhouette, CH, DBI) e externas (ARI, NMI, Homogeneidade, Completude, V-measure, Pureza) em DataFrame único para comparação entre blocos. <i>Uso diagnóstico:</i> Métricas externas avaliam utilidade ecológica dos clusters, não orientam sua formação.
<code>visualizar_clusters(X_umap, X_tsne, labels)</code>	Gera scatter plots com coloração por cluster. Suporta UMAP, t-SNE e PCA 2D.
<code>exportar_dados_completos()</code>	Salva DataFrames com coordenadas, labels e features originais em formato CSV para análises futuras e etapa supervisionada.

B.6 Núcleo do Fluxo Executivo

O pipeline segue um fluxo sequencial rigoroso para cada bloco:

```
1 def pipeline_principal():
2     # Configuração inicial
3     ConfigPipeline.imprimir_config()
4     df = carregar_dados('dados_finais_robust.csv')
5
6     # Processar cada bloco
7     for bloco in ['Bloco_A', 'Bloco_B', 'Bloco_C']:
8         resultado = processar_bloco_completo(df.copy(), bloco)
9
10    # Consolidação final
11    gerar_resumo_comparativo(resultados_todos)
12
13 def processar_bloco_completo(df, bloco):
14     # 1. Filtragem de features
15     df_filtrado = filtrar_features_por_bloco(df, bloco)
16
17     # 2. PCA (90% variância, min 3 componentes)
18     pca, X_pca, features_selecionadas = aplicar_pca_completo(
19         df_filtrado)
20
21     # 3. Projeções dimensionais
22     X_umap, X_tsne = aplicar_projecoess_dimensionais(X_pca)
23
24     # 4. Clusterização em ambas projeções
25     resultados_umap = aplicar_clusterizacao_completa(X_umap)
26     resultados_tsne = aplicar_clusterizacao_completa(X_tsne)
27
28     # 5. Métricas internas (geométricas)
29     metricas_internas = calcular_metricas_internas(labels)
30
31     # 6. Validação externa DIAGNÓSTICA (pós-clustering)
32     # Nota: Gabaritos não influenciam clustering, apenas avaliam
33     # congruência
34     validacao = validar_clusters_com_gabaritos(labels, gabaritos)
35
36     # 7. Consolidar métricas (internas + externas)
37     metricas = consolidar_metricas(metricas_internas, validacao)
38
39     # 8. Análise temporal
40     temporal = analisar_distribuicao_temporal(df_filtrado, labels)
```

```
40 # 7. Exportação (labels salvos para uso supervisionado posterior)
41 exportar_dados_completos(df_filtrado, labels, coordenadas)
42
43 return {'status': 'SUCESSO', 'metricas': metricas, ...}
```

Listing B.1 – Pseudocódigo do pipeline principal

B.7 Parâmetros de Configuração Centralizados

Todos os parâmetros são gerenciados pela classe `ConfigPipeline` para garantir reprodutibilidade:

Tabela 46 – Parâmetros otimizados dos algoritmos

Algoritmo	Parâmetro	Valor (Justificativa)
UMAP	n_neighbors	15 (captura estrutura global sem perder detalhes locais)
	min_dist	0.1 (clusters compactos, separação clara)
	n_components	2 (visualização 2D)
	metric	euclidean (padrão para dados padronizados)
	init	spectral (inicialização mais estável)
	random_state	42 (reprodutibilidade)
t-SNE	perplexity	50 (contexto ampliado para 324 amostras)
	max_iter	1500 (convergência completa)
	learning_rate	200.0 (acelera convergência)
	early_exaggeration	12 (separação inicial melhor)
	metric	euclidean
	random_state	42
K-Means	k_range	3 a 12 (evita k=2 trivial)
	n_init	20 (múltiplas inicializações)
HDBSCAN	min_cluster_size	[15, 18, 20, 25] (evita fragmentação)
	min_samples	[7, 10, 12] (robustez a ruído)
	min_clusters	2 (mínimo de grupos válidos)
	cluster_selection_method	['eom', 'leaf'] (dois critérios)
PCA	variance_threshold	0.85 (85% variância explicada)
	min_features	3 (mínimo de componentes)

Otimizações implementadas na versão 17.0:

- UMAP: Ajuste de `n_neighbors` para capturar estrutura global
- t-SNE: Aumento de `perplexity` e `max_iter` para melhor convergência
- K-Means: Aumento de `n_init` para maior estabilidade
- HDBSCAN: Restrição de `min_cluster_size` para evitar subdivisão excessiva

- PCA: Forçar mínimo de 3 componentes (antes: 2) para preservar informação

B.8 Protocolo de Execução por Bloco

1. **Ativação do bloco** — Seleção de variáveis conforme configuração (A, B ou C)
2. **Padronização robusta** — Dados já pré-processados com RobustScaler (tratamento de outliers no pipeline anterior)
3. **PCA** — Redução para 85% de variância explicada (mínimo 3 componentes)
4. **Projeções** — UMAP e t-SNE com parâmetros fixos (random_state=42)
5. **Clusterização** — K-Means (k=3 a 12) e HDBSCAN (busca em grid) em cada projeção
6. **Métricas internas** — Silhouette (coesão), Calinski-Harabasz (separação), Davies-Bouldin (compacidade)
7. **Validação externa diagnóstica** — Comparação *a posteriori* com gabaritos de Santos (2018): ARI, NMI, Homogeneidade, Completude, V-measure, Pureza. *Executada APÓS clustering; não influencia formação dos grupos. Avalia se padrões descobertos correspondem à realidade hidrológica conhecida.*
8. **Caracterização** — Estatística descritiva por cluster + teste de Kruskal-Wallis (discriminação de variáveis)
9. **Análise temporal** — Distribuição de clusters por estação (Seca/Chuvosa) e campanhas
10. **Validação hierárquica** — Análise multi-nível: Campo → Ponto → Cluster
11. **Exportação** — CSVs (coordenadas, labels, métricas), JSONs (parâmetros), PNGs (visualizações), Markdown (relatórios)

B.9 Estrutura de Saídas e Organização

O pipeline gera uma estrutura padronizada de diretórios e arquivos para cada bloco:

```
resultados_blocos_finais/  
  resultados_Bloco_A/  
    pca_analise/  
      pca_scree_plot.png  
      pca_variancia_explicada.csv
```

```
pca_loadings.csv
pca_componentes_principais.csv
pca_scaler.pkl
projecoes/
  dados_umap.csv
  dados_tsne.csv
  umap_Bloco_A.png
  tsne_Bloco_A.png
  comparacao_projecoes.png
clusters/
  labels_umap_kmeans.csv
  labels_umap_hdbscan.csv
  labels_tsne_kmeans.csv
  labels_tsne_hdbscan.csv
modelos/
  kmeans_umap.pkl
  hdbscan_umap.pkl
metricas/
  metricas_clustering_consolidadas.csv
  resumo_executivo.csv
  comparacao_metricas.png
validacao_gabaritos/
  validacao_binaria.json
  validacao_4_classes.json
  matriz_confusao_binaria.png
  matriz_confusao_4_classes.png
  RELATORIO_VALIDACAO.md
validacao_temporal/
  distribuicao_temporal_umap_kmeans.csv
  analise_temporal_UMAP_Bloco_A.csv
  grafico_distribuicao_temporal.png
validacao_hierarquica/
  metricas_nivel_campo.csv
  metricas_nivel_ponto.csv
  metricas_nivel_cluster.csv
  RELATORIO_HIERARQUICO.md
caracterizacao/
  estatisticas_descritivas_clusters.csv
  kruskal_wallis_discriminacao.csv
```

```
heatmap_medianas_clusters.png
visualizacoes/
clusters_umap_kmeans.png
clusters_tsne_kmeans.png
dashboard_comparativo.png
exports/
dados_completos_com_clusters.csv
relatorios/
RELATORIO_UMAP_COMPLETO.md
RELATORIO_TSNE_COMPLETO.md
RELATORIO_COMPARATIVO.md
resultados_Bloco_B/ (estrutura idêntica)
resultados_Bloco_C/ (estrutura idêntica)
```

Padronização: A estrutura é idêntica para todos os blocos, variando apenas o sufixo (`_Bloco_A`, `_Bloco_B`, `_Bloco_C`). Isso permite comparação direta entre experimentos.

B.10 Métricas e Racional de Avaliação

Métricas Internas (Coerência Geométrica)

- **Silhouette Score** (0 a 1): Mede coesão intra-cluster e separação inter-cluster
- **Calinski-Harabasz (CH)**: Razão entre variância inter-cluster e intra-cluster (quanto maior, melhor)
- **Davies-Bouldin (DBI)**: Média das razões de similaridade entre clusters (quanto menor, melhor)

Métricas Externas (Congruência Hidrológica — Validação Diagnóstica *A Posteriori*)

Contexto metodológico crítico: As métricas externas são calculadas *após* a formação completa dos clusters para avaliar se os padrões descobertos pelo algoritmo (puramente não supervisionado) correspondem à estrutura hidrológica conhecida da bacia (gabaritos de Santos 2018). Este uso é **diagnóstico**, não orientador: os gabaritos **NÃO** influenciam o processo de clusterização, apenas verificam sua congruência ecológica.

Por que isso NÃO constitui data leakage: Os mesmos gabaritos serão posteriormente usados na fase supervisionada (Capítulo 4) como `y_train` para treinar modelos preditivos (RF, GBM, MLP). Isso é legítimo porque:

1. **Fase não supervisionada (aqui):** Gabaritos = ferramenta de diagnóstico aplicada após clustering
2. **Fase supervisionada (posterior):** Gabaritos = rótulos para aprendizado (y_{train})
3. **Separação temporal:** Análise exploratória completa ANTES de qualquer treino supervisionado

Dois gabaritos hidrológicos de referência (SANTOS, 2018):

1. **Gabarito Binário:** Afetado vs. Não Afetado (classificação simplificada baseada em localização espacial)
2. **Gabarito 4 Classes:** RP_Afetado, RP_Não_Afetado, Tributário_Afetado, Tributário_Não_Afetado (classificação detalhada por compartimento hidrológico)

Métricas de concordância calculadas:

- **ARI (Adjusted Rand Index):** Concordância ajustada por chance (-1 a 1). Valores próximos a 1 indicam forte alinhamento entre clusters e gabaritos
- **NMI (Normalized Mutual Information):** Informação mútua normalizada (0 a 1). Mede quanto os clusters "informam" sobre as categorias hidrológicas
- **Homogeneidade:** Clusters contêm predominantemente uma única classe hidrológica (0 a 1)
- **Completeness:** Classes hidrológicas estão concentradas em poucos clusters (0 a 1)
- **V-measure:** Média harmônica de Homogeneidade e Completeness (balanceamento)
- **Pureza:** Proporção de amostras em clusters alinhados com classe modal (0 a 1)

Interpretação e uso:

- **ARI/NMI elevados (> 0.5):** Clusters capturam estrutura de impacto real $\rightarrow$ variáveis escolhidas são discriminativas
- **ARI/NMI baixos (< 0.3):** Clusters refletem outros processos (sazonalidade, processos locais) $\rightarrow$ variáveis podem requerer ajuste
- **Uso decisório:** Ajuda a escolher entre Blocos A, B ou C para fase supervisionada (qual configuração melhor alinha descoberta não supervisionada com realidade hidrológica)

Justificativa da abordagem dual (interna + externa):

1. **Métricas internas (Silhouette, CH, DBI):** Avaliam qualidade *numérica* dos clusters (coesão, separação, compacidade geométrica)
2. **Métricas externas (ARI, NMI, etc.):** Avaliam *utilidade ecológica* dos clusters (correspondência com fenômenos hidrológicos reais)
3. **Trade-off:** Possível ter clusters geometricamente perfeitos mas ecologicamente incoerentes, ou vice-versa. A análise conjunta identifica configuração ótima.

B.11 Notas Metodológicas Críticas

1. Subdivisão Excessiva no Bloco C

HDBSCAN em UMAP tende a detectar subestruturas finas, podendo fragmentar grupos naturalmente contínuos. O pipeline registra:

- Número de clusters detectados
- Percentual de ruído (-1 labels)
- Log de avisos quando número de clusters > 8 ou ruído $> 15\%$

Isso permite decidir entre Bloco B (sensível) e Bloco C (parcimonioso) na etapa supervisionada.

2. Reprodutibilidade e Comparabilidade

- `random_state=42` fixado em todos os algoritmos estocásticos
- Parâmetros centralizados em `ConfigPipeline`
- Garbage collection entre blocos para evitar contaminação de memória
- Cópia explícita de DataFrames (`df.copy()`) para evitar mutações

3. Separação Exploratório vs. Preditivo

Métricas externas são usadas **apenas para diagnóstico**, não para treino. Os rótulos hidrológicos:

- Não influenciam a clusterização (processo puramente não supervisionado)
- Servem para verificar congruência com realidade ecológica
- Orientam escolha de bloco para etapa supervisionada

4. Gestão de Memória e Performance

- Garbage collection forçado entre blocos (`gc.collect()`)
- Processamento sequencial (não paralelo) para evitar race conditions
- Salvamento incremental de artefatos (não apenas ao final)
- Logging detalhado com timestamps para monitoramento

B.12 Módulos de Validação Diagnóstica

Esclarecimento metodológico essencial: Todos os módulos desta seção são executados *após* a conclusão completa do clustering. Eles NÃO influenciam a formação dos grupos, apenas avaliam sua qualidade e congruência ecológica. O termo "validação externa" refere-se ao uso de informação externa ao processo de clusterização (gabaritos hidrológicos) para diagnóstico *a posteriori*, não para orientação do algoritmo.

O pipeline integra quatro módulos especializados para validação diagnóstica:

1. Validação com Gabaritos Hidrológicos (`validacao_gabaritos.py`)

Função: Avaliar se clusters descobertos (puramente não supervisionados) correspondem à estrutura de impacto conhecida da bacia.

Fluxo de execução:

1. Clustering completo gera labels (0, 1, 2, ..., N)
2. Módulo carrega gabaritos de Santos (2018): Binário (Afetado/Não) e 4 Classes (RP_Afetado, RP_Não_Afetado, Trib_Afetado, Trib_Não_Afetado)
3. Compara labels vs. gabaritos usando 6 métricas de concordância
4. Gera matrizes de confusão mostrando sobreposição entre clusters e categorias hidrológicas
5. Exporta relatório interpretativo em Markdown

Métricas calculadas:

- **ARI, NMI:** Concordância global entre partições
- **Homogeneidade, Completeness, V-measure:** Consistência interna e cobertura
- **Pureza:** Proporção de amostras corretamente alinhadas com classe modal

Uso: Identificar qual bloco (A, B ou C) produz clusters mais alinhados com fenômenos hidrológicos reais, fundamentando escolha para fase supervisionada.

2. Análise Temporal Hierárquica (`analise_temporal_hierarquica.py`)

- Adiciona coluna `Estacao` (Seca/Chuvosa) ao DataFrame
- Calcula distribuição de clusters por estação
- Validação hierárquica em 3 níveis: Campo → Ponto → Cluster
- Gera relatório com métricas por nível

3. Validação Temporal (`validacao_temporal.py`)

- Análise de estabilidade temporal dos clusters
- Validação em nível de campo de coleta
- Comparação entre três níveis de agregação
- Gráficos de distribuição temporal

4. Visualização Temporal de Pontos (`visualizacao_temporal_pontos.py`)

- Calcula estabilidade temporal por ponto de coleta
- Plota séries temporais com transições de clusters
- Heatmap de estabilidade por ponto e estação
- Trajetórias temporais em espaço UMAP/t-SNE

B.13 Referência de Execução

Execução Típica

```
1 # Terminal ou Jupyter Notebook
2 cd /workspaces/tcc-qualidade-agua-ml/02_clustering/scripts/
3 python clusterizacao_otimizada_v17_refatorado.py
4
5 # Ou via módulo
6 from clusterizacao_otimizada_v17_refatorado import pipeline_principal
7 resultados = pipeline_principal()
```

Listing B.2 – Exemplo de execução

Configuração de Blocos

Para processar blocos específicos:

```
1 # Em ConfigPipeline
2 BLOCO_ATIVO = 'Bloco_A' # Processar apenas Bloco A
3 # ou
4 BLOCO_ATIVO = None      # Processar todos os blocos
```

Listing B.3 – Configuração personalizada

Localização do Projeto

Repositório: /workspaces/tcc-qualidade-agua-ml/

Scripts: 02_clustering/scripts/

Dados: 02_clustering/data/dados_finais_robust.csv

Resultados: 02_clustering/scripts/resultados_blocos_finais/

Requisitos de Ambiente

- Python 3.8+
- Bibliotecas: numpy, pandas, matplotlib, seaborn, scikit-learn, umap-learn, hdbscan, scipy
- Memória RAM: mínimo 8GB (recomendado 16GB)
- Tempo de execução: 20-40 minutos por bloco (total: 1-2 horas)

B.14 Considerações Finais

A arquitetura aqui descrita constitui o núcleo exploratório do trabalho, integrando:

1. **Modularidade:** Separação clara de responsabilidades entre funções
2. **Reprodutibilidade:** Parâmetros fixos, random seeds controlados, logging detalhado
3. **Comparabilidade:** Estrutura idêntica para todos os blocos
4. **Rastreabilidade:** Exportação completa de artefatos e relatórios
5. **Validação multi-nível:** Métricas internas, externas, temporais e hierárquicas

Esta estrutura fundamenta as análises apresentadas no Capítulo de Resultados e constitui o elo entre a descoberta de padrões (não supervisionado) e a etapa preditiva (supervisionado), conforme detalhado no Capítulo 3 na Seção 3.3

C Apêndice C — Especificações Técnicas da Análise Supervisionada

Introdução

Este apêndice documenta a estrutura técnica do *pipeline* de aprendizado supervisionado implementado neste trabalho, com ênfase em **como** o sistema foi desenvolvido e **por que** cada decisão metodológica foi adotada. A arquitetura é modular, reproduzível e orientada a experimentos por **cenários de classificação**, permitindo avaliar sistematicamente três algoritmos de ML (RF, GBM, MLP) sob diferentes condições de pureza de rótulos, volume amostral e estratégias de agregação de clusters.

Destaques de Contribuições Metodológicas:

Este trabalho apresenta duas contribuições metodológicas originais para tratamento de dados ambientais:

1. **Validação de Outliers com Regras de Especialistas** (Seção 3.2, Apêndice A): Protocolo de validação físico-química fundamentado em 18 regras biogeoquímicas distribuídas em quatro categorias taxonômicas (REDOX, FÍSICA, BIOLÓGICA, INTEGRADORA), aplicado para discernir entre valores extremos legítimos e erros de medição. Dos 391 outliers estatisticamente identificados, 67.3% foram preservados por apresentarem coerência química justificável.
2. **Validação de Coerência Físico-Química do SMOTE** (Seção 3.4.6, Apêndice C, Seção C.8 deste apêndice): Análise empírica de 168 amostras sintéticas geradas pelo SMOTE, demonstrando 0% de incoerências químicas explícitas e validando a aplicabilidade do balanceamento por interpolação para dados ambientais com restrições biogeoquímicas.

Ambas as validações reutilizam o mesmo módulo `ValidadorCoerenciaQuimica`, garantindo consistência metodológica e permitindo comparação direta entre diferentes etapas do *pipeline*.

C.1 Visão Geral e Escopo

- **Arquivo principal:** `run_scenarios_complete.py`
- **Função orquestradora:** `run_scenario(scenario, models)`

- **Stack tecnológico:** Python 3.12+, NumPy, Pandas, scikit-learn (RF, GBM, LabelEncoder, StratifiedKFold, train_test_split), imbalanced-learn (SMOTE), TensorFlow/Keras (MLP com callbacks), SHAP, LIME, Matplotlib, Seaborn, PyYAML
- **Módulos customizados:** Validador de coerência química (regras de especialistas), encapsuladores de modelos robustos (`ml_models/`), analisadores XAI (`xai/`)
- **Ambiente de execução:** Google Colaboratory com GPU Tesla T4 para aceleração de treinamento de redes neurais (MLP) e cálculo de SHAP values. O código é compatível com execução em CPU (modo *fallback* automático).
- **Objetivo:** executar, registrar e comparar desempenho de modelos em **cenários de classificação** derivados de clusters não supervisionados, com validação cruzada estratificada, detecção de problemas de generalização e análises de interpretabilidade (XAI)

C.2 Fluxo por Fases e Por Que Fazer Assim

O *pipeline* foi concebido em duas fases principais, seguindo a lógica metodológica do trabalho:

Fase 1 — Auto-Descoberta e Geração de Cenários

O sistema detecta automaticamente todos os arquivos de clusters disponíveis (`cluster_*.csv`) e gera cenários de classificação baseados em critérios pré-definidos:

1. **Cenário 1 (Baseline):** Replicação de Santos (2018) com 3 classes (RP_Afetado, RP_Não_Afetado, Tributários);
2. **Cenário 2 (Alta Pureza):** Seleção de clusters homogêneos ($\geq 70\%$ de pureza) para maximizar qualidade dos rótulos;
3. **Cenário 3 (Gradiente):** Diferenciação de zonas de transição (RP_Misto) para avaliar capacidade de separar estados intermediários;
4. **Cenário 4 (Dataset Completo):** Utilização de todos os clusters disponíveis (9 classes) para avaliar robustez em cenário realista com pureza moderada (38–64%);
5. **Cenário 5 (Binário Balanceado):** Classificação binária artificialmente balanceada (50/50) para controlar efeito de desbalanceamento.

Por quê? Para avaliar sistematicamente o **trade-off entre pureza e volume amostral**, isolando o efeito de diferentes estratégias de agregação de clusters sobre o desempenho preditivo.

Fase 2 — Treinamento, Validação e Interpretabilidade

Para cada cenário, o pipeline executa:

- **Pré-processamento:** Divisão estratificada (70/30), balanceamento ([SMOTE](#)) **sem normalização adicional** (dados já padronizados);
- **Treinamento:** Ajuste de hiperparâmetros com validação cruzada estratificada (StratifiedKFold, $k = 5$);
- **Avaliação:** Métricas de desempenho (Accuracy, Balanced Accuracy, F1-macro/weighted, Precision, Recall);
- **Diagnóstico:** Detecção automática de *overfitting/underfitting* via análise de gap (CV vs. Test);
- **Interpretabilidade:** Análises [XAI](#) ([SHAP](#) e [LIME](#)) para explicar decisões dos modelos.

Por quê? Para garantir **reprodutibilidade**, **comparabilidade** entre modelos/cenários e **transparência** nas decisões algorítmicas.

C.3 Estrutura do Módulo

O módulo está organizado em 6 componentes principais:

- **core/:** Gerenciamento de configurações e cenários
 - `config.py`: Classes de configuração (`ScenarioConfig`, `DatasetInfo`)
 - `scenarios.py`: Geração automática de cenários (`ScenarioGenerator`)
 - `discovery.py`: Auto-descoberta de datasets de clusters
 - `preprocessing.py`: Pipeline de pré-processamento e validação
- **ml_models/:** Implementações encapsuladas dos modelos de ML
 - `base.py`: Classe abstrata `BaseMLModel` com interface padronizada
 - `random_forest.py`: `RobustRandomForest` (scikit-learn `RandomForestClassifier`)
 - `gradient_boosting.py`: `RobustGradientBoosting` (scikit-learn `GradientBoostingClassifier`)
 - `neural_network.py`: `RobustNeuralNetwork` (TensorFlow/Keras com callbacks `EarlyStopping` e `ReduceLROnPlateau`)

- **xai/**: Análises de explicabilidade (XAI)
 - `shap_analyzer.py`: `SHAPAnalyzer` (valores de Shapley via `TreeExplainer`)
 - `lime_analyzer.py`: `LIMEAnalyzer` (explicações locais via `LimeTabularExplainer`)
- **validacao_smote/**: Módulo de validação química do SMOTE
 - `validar_coerencia_smote.py`: Validador de coerência físico-química
 - `modulo_regras_outliers.py`: Regras de especialistas (relações redox, balanço iônico, plausibilidade biológica)
 - `gerar_plots_validacao_clusters.py`: Geração de visualizações explicativas
- **Scripts de comparação**: Análises comparativas entre cenários
 - `compare_with_santos.py`: Comparação com baseline de Santos et al. (2019)
 - `compare_umap_vs_no_umap.py`: Validação de ausência de data leakage
- **data/**: Datasets de entrada (12 clusters do Bloco C, já padronizados com `RobustScaler`)

C.4 Hierarquia Funcional (o que cada parte faz)

Tabela 47 – Principais funções do pipeline supervisionado e responsabilidades

Função	Responsabilidade
<code>run_scenario(scenario, models)</code>	Orquestra o fluxo completo para um cenário: preparação de dados, divisão treino/teste, balanceamento, treino de modelos, validação cruzada, cálculo de métricas e análises XAI .
<code>load_yaml_scenarios(yaml_path)</code>	Carrega definições de cenários de arquivo YAML, permitindo configuração externa sem modificar código.
<code>prepare_scenario_data(scenario)</code>	Consolida clusters em classes conforme mapeamento do cenário, remove duplicatas e valida consistência.
<code>apply_smote_balancing(X, y)</code>	Aplica SMOTE para balancear classes minoritárias (< 20% do total), evitando viés de predição.
<code>train_with_cv(model, X, y)</code>	Treina modelo com validação cruzada estratificada (StratifiedKFold, $k = 5$), retorna scores e desvio padrão.
<code>evaluate_model(model, X_test, y_test)</code>	Calcula métricas de desempenho: Accuracy, Balanced Accuracy, F1-macro/weighted, Precision, Recall, matriz de confusão.
<code>detect_overfitting(cv_mean, test_acc)</code>	Diagnóstico automático de problemas de generalização via $\text{gap} = \text{CV} - \text{Test}$ (thresholds: $\pm 5\%$, $\pm 10\%$).
<code>run_shap_analysis(model, X, features)</code>	Gera valores de Shapley para importância global de features (TreeExplainer para RF / GBM , KernelExplainer para MLP).
<code>run_lime_analysis(model, X, y, idx)</code>	Gera explicações locais para amostras específicas via aproximação linear local.
<code>save_results(scenario, metrics)</code>	Persiste resultados (JSON, CSV, PNG) com timestamp e metadados para rastreabilidade.

C.5 Interface do Módulo

- **`run_scenarios_complete.py`**: Script principal de execução
 - `run_scenario(scenario, models)`: Executa um cenário completo (treino, validação cruzada, teste, XAI)
 - `load_yaml_scenarios(yaml_path)`: Carrega definição de cenários de arquivo YAML (compatível com JSON)
 - `prepare_scenario_data(scenario, data_dir, features)`: Prepara dataset consolidado para um cenário
 - `evaluate_model(model, X_test, y_test)`: Calcula métricas de desempenho
- **`BaseMLModel` (`ml_models/base.py`)**: Interface abstrata para modelos
 - `build_model(input_dim, num_classes)`: Constrói arquitetura do modelo

- `train(X_train, y_train, X_val, y_val)`: Treina o modelo com validação
- `predict(X)`: Realiza previsões
- `needs_scaling()`: Indica necessidade de normalização (sempre `False` — dados já padronizados)
- **SHAPAnalyzer** (`xai/shap_analyzer.py`): Análise de importância global
 - `analyze_tree_model(model, X_test, feature_names, model_name)`: SHAP para RF/GBM (`TreeExplainer`)
 - `save_importance_plot(shap_values, feature_names, output_path)`: Exporta gráficos de importância
- **LIMEAnalyzer** (`xai/lime_analyzer.py`): Explicações locais instância-específicas
 - `explain_instance(model, X, sample_idx, feature_names, class_names)`: Explicação para amostra individual
 - `save_explanation(explanation, output_path)`: Exporta explicação em HTML
- **ValidadorCoerenciaQuimica** (`modulo_regras_outliers.py`): Validação química
 - `verificar_coerencia_redox(orp, satod, no3, dbo, ecoli)`: Valida relações óxido-redução
 - `verificar_coerencia_fisica(st, cond, turb, pt)`: Valida balanço de sólidos/turbidez
 - `verificar_coerencia_biologica(ecoli, ct, dbo, satod)`: Valida contaminação fecal vs. matéria orgânica

C.6 Parâmetros de Configuração (centralizados)

Parâmetros são centralizados para garantir reprodutibilidade e comparabilidade entre cenários.

- **Divisão de dados**: 70% treino, 30% teste (estratificada por classe, `random_state=42`)
- **Validação cruzada**: `StratifiedKFold` com $k = 5$ folds (preserva distribuição de classes)
- **Features**: 8 variáveis físico-químicas (`ORP`, `SatOD`, `DBO`, `E. coli`, `ST`, `COND`, `Turb.`)
- **Balanceamento**: `SMOTE` para classes minoritárias ($< 20\%$ do total, `random_state=42`)

- **Normalização: Não aplicada** — dados já padronizados com **RobustScaler** (mediana e IQR) na etapa de clusterização. Todos os modelos (**RF**, **GBM**, **MLP**) usam `use_scaling=False`.
- **Métricas:** Accuracy, Balanced Accuracy, F1-macro, F1-weighted, Precision, Recall
- **Gap (CV - Test):** Diagnóstico de generalização
 - $|gap| < 5\%$: Generalização adequada (OK)
 - $5\% \leq |gap| < 10\%$: Problema leve (Leve)
 - $gap > 10\%$ (CV > Test): *Underfitting* severo (Underfit)
 - $gap < -10\%$ (Test > CV): *Overfitting* severo (Overfit)
- **XAI:** **SHAP** (valores de Shapley globais, 100 amostras background) e **LIME** (explicações locais, $n = 10$ amostras/classe)

Configurações dos Algoritmos

```
class ModelConfigs:
    RANDOM_FOREST = {
        'n_estimators': 500,
        'max_depth': None,
        'min_samples_split': 5,
        'min_samples_leaf': 2,
        'class_weight': 'balanced',
        'random_state': 42
    }

    GRADIENT_BOOSTING = {
        'n_estimators': 200,
        'learning_rate': 0.1,
        'max_depth': 4,
        'subsample': 0.8,
        'min_samples_split': 10,
        'min_samples_leaf': 4,
        'random_state': 42
    }

    NEURAL_NETWORK = {
        'layers': [128, 64, 32],
        'activation': 'relu',
```

```
'dropout': 0.3,
'batch_normalization': True,
'optimizer': 'adam',
'learning_rate': 0.001,
'epochs': 100,
'batch_size': 32,
'early_stopping_patience': 15,
'random_state': 42,
'use_gpu': True # Aceleração via GPU (Tesla T4 no Colab)
}
```

Nota sobre execução em GPU: A rede neural (MLP) foi treinada com aceleração GPU no Google Colaboratory (Tesla T4), reduzindo o tempo de treinamento de ~15 minutos (CPU) para ~2 minutos por cenário. TensorFlow detecta automaticamente a disponibilidade de GPU via CUDA. Random Forest e Gradient Boosting executam em CPU (scikit-learn não oferece aceleração GPU nativa).

C.7 Decisão de Não Aplicar Normalização Adicional

Contextualização: Os dados utilizados no pipeline supervisionado são provenientes dos arquivos `cluster_*.csv` gerados na etapa de clusterização (Capítulo 4). Esses arquivos **já contêm dados padronizados** com **RobustScaler** (mediana e IQR), aplicado sobre o arquivo `dados_finais_robust.csv`.

Racional metodológico:

1. **Evitar dupla padronização:** Aplicar `StandardScaler` sobre dados já normalizados com `RobustScaler` geraria distorções na escala e invalidaria as análises de importância de features (XAI).
2. **Preservar robustez a outliers:** `RobustScaler` (mediana e IQR) é mais adequado que `StandardScaler` (média e desvio padrão) para dados ambientais com valores extremos legítimos (e.g., turbidez alta em áreas afetadas pela lama de rejeitos).
3. **Consistência entre etapas:** Usar a mesma escala do clustering garante que os clusters representem agrupamentos fisicamente interpretáveis, sem artefatos de renormalização.
4. **Compatibilidade com MLP:** Embora redes neurais tradicionalmente requeiram normalização, a padronização `RobustScaler` já garante que todas as features estejam em escala comparável ($\approx [-3, +3]$ para dados típicos), satisfazendo os requisitos de convergência do otimizador Adam.

Implementação: Todos os modelos (RF, GBM, MLP) foram configurados com `use_scaling=False`, eliminando a etapa de `StandardScaler.fit_transform()` do pipeline. Essa decisão foi documentada explicitamente no código:

```
# Neural Network (apenas se TensorFlow disponível)
if NN_AVAILABLE and RobustNeuralNetwork is not None:
    results['models']['Neural Network'] = train_and_evaluate_model(
        RobustNeuralNetwork, 'Neural Network',
        X_train, y_train, X_val, y_val, X_test, y_test,
        use_scaling=False # Dados já padronizados (RobustScaler no clustering)
    )
```

C.8 Validação de Coerência Físico-Química do SMOTE

Contextualização

Os métodos de balanceamento como SMOTE operam puramente no espaço estatístico, sem conhecimento sobre restrições físico-químicas. Para resolver esta questão, foi desenvolvida uma **análise complementar de validação química**, integrando o módulo de regras de especialistas criado na etapa de pré-processamento (Seção 3.2, Apêndice A).

Nota: Esta validação é descrita resumidamente nesta seção. A documentação completa, incluindo protocolo detalhado, taxonomia de resultados, análise estatística comparativa e visualizações, encontra-se no Apêndice A. A validação SMOTE segue estrutura paralela à validação de outliers, representando contribuição metodológica original deste trabalho.

Estratégia de Mitigação de Risco

Diferencial metodológico: SMOTE foi aplicado **DENTRO de clusters homogêneos** formados por HDBSCAN, não no dataset completo heterogêneo:

- **Clusters representam perfis químicos similares:** Amostras no mesmo cluster compartilham correlações físico-químicas consistentes (e.g., Cluster 0: alta turbidez + alto PT + alta contaminação fecal);
- **Interpolação entre vizinhos homogêneos:** SMOTE interpola entre $k = 5$ vizinhos mais próximos dentro do cluster, preservando padrões químicos locais;
- **Risco mitigado:** A probabilidade de gerar combinações quimicamente impossíveis é drasticamente reduzida, pois as amostras de treino já estão agrupadas por similaridade físico-química.

Protocolo de Validação (Dupla Verificação)

A validação foi realizada em duas etapas complementares:

1. Validação Estatística (Correlações e Ranges)

- **Correlações químicas esperadas:** Verificação de que correlações conhecidas (e.g., Turb $\times$ ST: $r \approx +0.67$, E.coli $\times$ CT: $r \approx +0.17$) são preservadas após SMOTE;
- **Limites de ranges:** Comparação de intervalos (min-max) e distribuições ($\mu \pm \sigma$) entre amostras reais e sintéticas;
- **Detecção de outliers extremos:** Contagem de valores $> 3\sigma$ (espera-se $< 1\%$ em distribuição normal).

2. Validação Química (Regras de Especialistas)

[CONTRIBUIÇÃO ORIGINAL]

- **Módulo de coerência química:** Reutilização do `ValidadorCoerenciaQuimica` desenvolvido na etapa de pré-processamento (Seção 3.2, Apêndice A);
- **Regras aplicadas:**
 - *REDOX*: ORP $\times$ SatOD $\times$ NO₃ $\times$ DBO $\times$ E.coli (verifica condições óxido-redução coerentes);
 - *FÍSICA*: ST $\times$ COND $\times$ Turb $\times$ PT (sólidos totais vs. turbidez vs. fósforo);
 - *BIOLÓGICA*: E.coli $\times$ CT $\times$ DBO $\times$ SatOD $\times$ Turb $\times$ PT (contaminação fecal vs. matéria orgânica).
- **Detecção IQR:** Outliers baseados em $1.5 \times \text{IQR}$ (método robusto, menos sensível a valores extremos legítimos).

Resultados da Validação (Cenário 1 — Baseline Santos)

Dataset Analisado

- **Amostras reais (treino):** 205
- **Amostras após SMOTE:** 303
- **Amostras sintéticas:** 98
- **Features:** 8 variáveis físico-químicas (ST, Turb., Alca, Cloretos, , *E. coli*, CT, PT)

Validação Estatística

Tabela 48 – Preservação de correlações químicas após SMOTE

Correlação	Tipo Esperado	Real	SMOTE	Δ
Turb $\times$ ST	positiva	+0.672	+0.666	0.006
E.coli $\times$ CT	positiva	+0.168	+0.158	0.010

Interpretação: Correlações químicas preservadas (diferença $< 1\%$). APROVADO.

- **Valores extremos** ($> 3\sigma$): 9.2% (72 de 784 medições)

Interpretação crítica: Estes NÃO são "outliers" no sentido de **erros**, mas sim **características legítimas do desastre ambiental**. Clusters de águas afetadas (RP_Afetado: 0,1,2) capturam condições reais de turbidez alta, fósforo total elevado e contaminação bacteriológica extrema da lama de rejeitos. SMOTE interpola DENTRO desses clusters, gerando amostras com valores extremos **CORRETOS**.

Validação Química (Regras de Especialistas)

Tabela 49 – Validação de coerência química via regras de especialistas

Critério	Incoerências Detectadas	Status
REDOX (ORP $\times$ SatOD $\times$ NO ₃ $\times$ DBO $\times$ E.coli)	0 de 98 (0%)	
FÍSICA (ST $\times$ COND $\times$ Turb $\times$ PT)	0 de 98 (0%)	
BIOLÓGICA (E.coli $\times$ CT $\times$ DBO $\times$ SatOD)	0 de 98 (0%)	
Total de incoerências químicas	0 de 98 (0%)	APROVADO

Resultado: Nenhuma amostra sintética apresentou combinações quimicamente impossíveis.

- **Outliers IQR:** 13.8% (108 de 784 medições) — Aceitável para dados ambientais heterogêneos (desastre ambiental com ampla variabilidade espacial e temporal).

Conclusões da Validação

1. **Coerência química preservada:** 0% de incoerências detectadas pelas regras de especialistas. SMOTE não gerou amostras com combinações físico-químicas impossíveis.
2. **Correlações mantidas:** Relações químicas conhecidas (Turb-ST, E.coli-CT) foram preservadas (diferença $< 1\%$).

3. **Valores extremos são legítimos:** Os 9.2% de valores $> 3\sigma$ refletem características reais de clusters de águas afetadas pelo desastre (lama de rejeitos), não erros do algoritmo.
4. **Estratégia de mitigação eficaz:** Aplicar SMOTE DENTRO de clusters homogêneos (HDBSCAN) foi decisivo para evitar interpolação entre perfis químicos incompatíveis.
5. **Contribuição metodológica:** A integração do módulo de regras de outliers como validador de amostras SMOTE representa uma **contribuição original** deste TCC, não encontrada na literatura consultada.

Visualizações Complementares

Foram gerados 4 plots explicativos para demonstrar empiricamente que valores extremos são esperados:

- **Plot 1:** Boxplots de distribuição por cluster — mostra que RP_Afetado NATURALMENTE tem valores extremos (Turb, PT, E.coli, CT elevados);
- **Plot 2:** Histogramas Real vs. SMOTE por classe — demonstra que SMOTE preserva distribuições DENTRO de cada cluster;
- **Plot 3:** Scatter plots de separação espacial (Turb $\times$ PT, E.coli $\times$ CT) — clusters de águas afetadas ocupam região de valores extremos;
- **Plot 4:** Comparação antes/depois SMOTE no cluster RP_Afetado — percentuais de valores $> 3\sigma$ similares entre Real e SMOTE.

Arquivos de validação disponíveis no repositório:

03_supervised/validacao_smote/

validar_coerencia_smote.py	# Script de validação (739 linhas)
gerar_plots_validacao_clusters.py	# Gerador de visualizações (382 linhas)
RELATORIO_VALIDACAO_SMOTE.md	# Relatório técnico completo
plot1_distribuicao_por_cluster.png	# Heterogeneidade natural
plot2_smote_preserva_clusters.png	# Preservação de distribuições
plot3_scatter_clusters_desastre.png	# Separação espacial
plot4_smote_cluster_afetado.png	# Antes/depois SMOTE

C.10 Protocolo de Execução por Cenário (o passo a passo)

1. **Carregamento e preparação:** Leitura de arquivos `cluster_*.csv` (dados já padronizados com `RobustScaler`), consolidação conforme mapeamento do cenário (e.g., Cen. 2: clusters 0,1,5,7 $\rightarrow$ 4 classes);
2. **Divisão estratificada:** 70/30 treino/teste preservando proporção de classes (`stratify=y`);
3. **Balanceamento (SMOTE):** Aplicado apenas em treino se classes minoritárias $< 20\%$;
4. **Normalização: NÃO APLICADA** — dados já padronizados (`RobustScaler` no `clustering`);
5. **Treinamento:** Ajuste de modelos com validação cruzada (`StratifiedKFold`, $k = 5$);
6. **Teste:** Predição no conjunto de teste (30%), cálculo de métricas consolidadas;
7. **Diagnóstico:** Análise de gap (CV vs. Test) para detecção de *overfitting/underfitting*;
8. **Matrizes de confusão:** Visualização de padrões de erro por classe;
9. **Análises XAI:** `SHAP` (importância global) e `LIME` (explicações locais para amostras críticas);
10. **Persistência:** Salvamento de modelos (`.pkl`), métricas (`.json/.csv`), visualizações (`.png`) e relatórios (`.md`).

C.11 Estrutura de Saídas e Organização dos Resultados

A organização das saídas do *pipeline* foi padronizada para garantir a reprodutibilidade e a rastreabilidade das análises em todos os cenários. A estrutura geral, representada na Figura 44, é idêntica para os 5 cenários, variando apenas no sufixo de identificação (por exemplo, `_Cenario_1`, `_Cenario_2`, etc.).

Cada cenário gera pastas correspondentes a etapas específicas do fluxo — `models`, `metrics`, `confusion_matrices`, `xai_analysis` e `reports` — além de diretórios complementares com logs de execução e comparações consolidadas entre modelos.

Essa padronização permite comparar diretamente as métricas e resultados entre cenários e algoritmos, mantendo consistência entre experimentos e facilitando futuras extensões do pipeline.

```

results_supervised/
├── Cenario_1_Baseline_Santos/
│   ├── models/
│   │   ├── Random_Forest_Cenario_1.pkl
│   │   ├── Gradient_Boosting_Cenario_1.pkl
│   │   └── Neural_Network_Cenario_1.pkl
│   ├── metrics/
│   │   ├── metrics_Random_Forest.json
│   │   ├── metrics_Gradient_Boosting.json
│   │   ├── metrics_Neural_Network.json
│   │   ├── consolidated_metrics.csv .... Comparação entre modelos
│   │   └── overfitting_report.csv ..... Análise de gaps
│   ├── confusion_matrices/
│   │   ├── confusion_matrix_RF_Cenario_1.png
│   │   ├── confusion_matrix_GBM_Cenario_1.png
│   │   ├── confusion_matrix_MLP_Cenario_1.png
│   │   └── confusion_matrix_comparison.png ..... Grade 3x1
│   ├── xai_analysis/
│   │   ├── shap_importance_RF_Cenario_1.png
│   │   ├── shap_importance_GBM_Cenario_1.png
│   │   ├── shap_importance_MLP_Cenario_1.png
│   │   ├── shap_summary_RF_Cenario_1.png ..... Beeswarm plot
│   │   └── lime_explanations/
│   │       ├── lime_RF_sample_0_class_0.html
│   │       ├── lime_RF_sample_1_class_1.html
│   │       └── [...]
│   ├── reports/
│   │   ├── REPORT_Cenario_1_COMPLETE.md
│   │   ├── classification_report_RF.txt
│   │   ├── classification_report_GBM.txt
│   │   └── classification_report_MLP.txt
│   └── logs/
│       ├── execution_log_Cenario_1.txt
│       └── warnings_Cenario_1.log ..... SMOTE, imbalance, etc.
├── Cenario_2_AltaPureza/
│   └── [...] ..... Estrutura idêntica
├── Cenario_3_Gridente/
│   └── [...]
├── Cenario_4_DatasetCompleto/
│   └── [...]
└── Cenario_5_BinarioBalanceado/
    └── [...]

```

C.12 Métricas e Racional de Avaliação

Desempenho preditivo:

- **Accuracy**: Proporção de acertos totais (sensível a desbalanceamento)
- **Balanced Accuracy**: Média das taxas de acerto por classe (corrige desbalanceamento)
- **F1-macro**: Média harmônica de Precision/Recall por classe, sem ponderação (prefere equilíbrio)
- **F1-weighted**: Média harmônica ponderada pelo suporte de cada classe (prefere maioria)
- **Precision/Recall**: Por classe, para identificar padrões de erro específicos

Generalização:

- **CV Mean $\pm$ Std**: Média e desvio padrão da validação cruzada (StratifiedKFold, $k = 5$)
- **Gap = CV Mean - Test Accuracy**: Diagnóstico de *overfitting/underfitting*
 - $|gap| < 5\%$: Generalização adequada (OK)
 - $5\% \leq |gap| < 10\%$: Problema leve (Leve)
 - $gap > 10\%$ (CV > Test): *Underfitting* severo (Underfit)
 - $gap < -10\%$ (Test > CV): *Overfitting* severo (Overfit)

Interpretabilidade (XAI):

- **SHAP (Global)**: Importância de features via valores de Shapley (contribuição marginal média)
- **LIME (Local)**: Explicações lineares locais para predições individuais (features mais influentes)

Motivo: Separar *capacidade preditiva* (accuracy, F1) de *estabilidade* (CV, gap) e *transparência* (XAI), mantendo comparabilidade entre modelos e cenários.

C.13 Notas Metodológicas Importantes

- **Trade-off pureza vs. volume:** O Cenário 2 ($n = 151$, pureza $\geq 70\%$) superou o Cenário 4 ($n = 303$, pureza 38–64%) em +11,52 p.p., demonstrando que **qualidade dos rótulos domina sobre quantidade de dados**.
- **Sensibilidade da MLP ao volume amostral:** Redes neurais apresentaram *underfitting* severo em cenários com $n < 300$ (Cenário 2: gap = 19,35%), confirmando que arquiteturas profundas requerem ≥ 1000 amostras/classe para convergência estável.
- **Robustez do RF:** **Random Forest (RF)** manteve $|\text{gap}| < 5\%$ em todos os cenários (CV de 4,7%), sendo o único modelo consistentemente estável independente de volume, balanceamento ou pureza.
- **Especialização do GBM em gradientes:** **Gradient Boosting Machines** alcançou pico de desempenho no Cenário 3 (96,15%), superando RF, devido à capacidade de modelar transições contínuas (**RP_Misto**).
- **Deteção de data leakage:** Análise comparativa *com* vs. *sem* **UMAP** demonstrou ausência de vazamento (diferença $< 2\%$), validando que clusters pré-computados não contaminam o processo de treino.
- **Padronização com RobustScaler:** Decisão de usar RobustScaler (mediana e IQR) no clustering ao invés de StandardScaler (média e σ) foi estratégica para lidar com valores extremos legítimos em dados de desastre ambiental. Essa escolha eliminou a necessidade de renormalização no pipeline supervisionado, preservando a interpretabilidade física dos clusters.
- **Critério de escalonamento e reprodutibilidade:** Seeds fixas (`random_state=42`) em todas as etapas (split, **SMOTE**, CV, modelos) garantem repetibilidade entre execuções.
- **Separação "exploratório" vs. "preditivo":** Clusters são tratados como rótulos *dados*, não como verdades absolutas — métricas de desempenho refletem **utilidade preditiva** dos agrupamentos, não "acerto" ontológico.

C.14 Comparação com Baseline (Santos 2018)

O Cenário 1 foi estruturado para replicação direta do estudo de Santos (2018), permitindo validação externa do pipeline:

Tabela 50 – Comparação de desempenho: Pipeline atual vs. Baseline Santos (2018)

Modelo	Santos 2018	Pipeline Atual (Cen. 1)
LDA (Linear Discriminant Analysis)	91,67%	—
Random Forest (RF)	—	94,23% (+2,56 p.p.)
Gradient Boosting Machines	—	90,38% (−1,29 p.p.)
Neural Network	—	94,23% (+2,56 p.p.)

Fonte: Elaboração própria (2025).
Nota: RF e MLP superaram o baseline LDA de Santos (2018) em +2,56 pontos percentuais, confirmando a superioridade de métodos *ensemble* e redes neurais sobre modelos lineares para dados ambientais com relações não-lineares.

D Apêndice D — Cálculo Detalhado da Economia Estimada

D.1 Modelo de Custos Operacionais

Os custos por campanha foram decompostos em três componentes, conforme estrutura adotada no Capítulo 4:

$$C_{\text{total}} = C_{\text{lab}} + C_{\text{equip}} + C_{\text{log}} \quad (\text{D.1})$$

Custos Laboratoriais

D.2 Protocolo Completo (14 variáveis)

O protocolo completo inclui 10 análises laboratoriais e 4 medições *in situ* (Temperatura, OD, Condutividade e ORP). Para as análises laboratoriais, utilizou-se a média de valores em tabelas públicas ([PREFEITURA MUNICIPAL DE BRAÇO DO TROMBUDO, 2025](#); [CETESB, 2024](#)):

$$\begin{aligned} C_{\text{lab}}^{\text{completo}} &= pH + Turb + ST + Alca + Cl + NO_3 + NO_2 + PT + Ecoli + CT \\ &= 17,50 + 30,00 + 50,00 + 42,50 + 42,50 + 47,50 + 42,50 + 60,00 + 80,00 + 90,00 \\ &= 502,50 \approx 650,00 \end{aligned} \quad (\text{D.2})$$

O valor de R\$ 650,00 é conservador e incorpora margem para custos usuais de controle de qualidade (duplicatas, brancos), taxas administrativas e variações regionais.

D.2 Protocolo Simplificado (Bloco C – 8 variáveis)

No Bloco C, todas as variáveis remanescentes são laboratoriais:

$$\begin{aligned} C_{\text{lab}}^{\text{blocoC}} &= Turb + ST + Alca + Cl + NO_3 + PT + Ecoli + CT \\ &= 30,00 + 50,00 + 42,50 + 42,50 + 47,50 + 60,00 + 80,00 + 90,00 \\ &= 442,50 \approx 450,00 \end{aligned} \quad (\text{D.3})$$

D.3 Economia Laboratorial

$$E_{\text{lab}} = \frac{650 - 450}{650} \times 100\% = 30,8\% \quad (\text{D.4})$$

D.4 Custos de Equipamentos

As medições *in situ* do protocolo completo requerem sonda multiparamétrica. A amortização considerou aquisição, calibração e manutenção ao longo de 3 anos (36 campanhas), com base em cotações e premissas operacionais (YSI - XYLEM ANALYTICS, 2024; BARCELLOS; SOUZA, 2022):

$$C_{\text{equip}}^{\text{amort}} = \frac{C_{\text{aquis}} + \sum_{i=1}^3 (C_{\text{calib},i} + C_{\text{manut},i})}{36} = \frac{42.500 + 3(2.500 + 4.000)}{36} \approx 1.750/\text{campanha} \quad (\text{D.5})$$

No Bloco C, $C_{\text{equip}}^{\text{blocoC}} = 0$, pois as variáveis *in situ* foram eliminadas (economia de 100% neste componente).

D.5 Custos Logísticos

Os custos logísticos foram estimados a partir de três itens: tempo de campo, materiais e deslocamento.

- **Tempo de campo:** completo = 8h; Bloco C = 5h (2 pessoas, R\$ 75/h). Economia: R\$ 450.
- **Materiais:** R\$ 200 → R\$ 150. Economia: R\$ 50.
- **Deslocamento:** R\$ 800 (inalterado).

Assim, o custo logístico por campanha foi de R\$ 2.200 (completo) e R\$ 1.700 (Bloco C), resultando em:

$$E_{\text{log}} = \frac{2.200 - 1.700}{2.200} \times 100\% = 22,7\% \quad (\text{D.6})$$

D.6 Economia Total e Projeção Anual

$$E_{\text{total}} = \frac{(650 + 1.750 + 2.200) - (450 + 0 + 1.700)}{650 + 1.750 + 2.200} \times 100\% = \frac{2.450}{4.600} \times 100\% = 53,3\% \quad (\text{D.7})$$

Para 12 campanhas/ano:

- Custo completo: R\$ $4.600 \times 12 = \text{R\$ } 55.200$
- Custo Bloco C: R\$ $2.150 \times 12 = \text{R\$ } 25.800$
- Economia: R\$ 29.400 (53,3%)

Cenários de Sensibilidade

Para avaliar robustez, foram considerados cenários extremos, mantendo a mesma estrutura do modelo.

Pessimista (38%)

Premissas: $C_{\text{lab}} + 20\%$, tempo de campo -10% , manutenção máxima.

$$E_{\text{pess}} = \frac{(780 + 1.900 + 2.280) - (540 + 0 + 1.780)}{4.960} \times 100\% \approx 38\% \quad (\text{D.8})$$

Economia anual estimada: R\$ 18.500.

Otimista (64%)

Premissas: $C_{\text{lab}} - 20\%$ (contratos), tempo de campo -40% , manutenção mínima.

$$E_{\text{otim}} = \frac{(520 + 1.600 + 2.120) - (360 + 0 + 1.520)}{4.240} \times 100\% \approx 64\% \quad (\text{D.9})$$

Economia anual estimada: R\$ 37.800.

E ANEXOS

E.1 ANEXO A – Amostra dos Dados Brutos Coletados

Os dados brutos utilizados nesta pesquisa foram originalmente coletados por Santos (2018) entre 2016 e 2017, abrangendo 324 observações de qualidade da água distribuídas em 27 pontos de amostragem ao longo da bacia hidrográfica do rio Gualaxo do Norte. O conjunto de dados contém 17 variáveis, incluindo parâmetros físico-químicos (Condutividade Elétrica, Turbidez, pH, temperatura), químicos (nutrientes, cloretos, Alcalinidade), microbiológicos (*E. coli*, coliformes totais) e de matéria orgânica (DBO).

A Tabela 51 apresenta as primeiras 10 observações do conjunto de dados, ilustrando a estrutura e a diversidade das variáveis medidas. O dataset completo (324 linhas $\times$ 17 colunas) está disponível no repositório digital do projeto: <https://github.com/PamelaMurta/tcc-qualidade-agua-ml>.

Tabela 51 – Amostra dos dados brutos coletados por Santos (2018) (primeiras 10 observações).

Campo	Ponto	COND	ORP	SatOD	pH	Turb	Ta	NO ₃	NO ₂	Cl	PT	ST	Alca	CT	E.coli	DBO
C1	PC1	123.3	164	90.7	7.54	2.5	20.0	2.10	0.01	1.51	0.01	80	52.0	4140	1144	0.79
C1	PC2	85.2	192	72.6	7.26	26.2	18.8	2.10	0.17	5.02	0.11	60	36.8	185000	41000	3.29
C1	PC3	109.5	19	88.5	7.37	9.3	20.2	2.00	0.04	2.26	0.02	60	54.0	74000	20000	1.29
C1	PC4	54.0	52	67.3	7.01	3.3	17.1	0.40	0.00	2.26	0.01	40	13.8	2130	940	1.03
C1	PC5	123.3	156	93.1	6.99	2.2	19.9	2.20	0.04	3.01	0.15	80	54.8	22240	2720	1.28
C1	PC6	37.3	7	88.6	7.40	2.6	17.0	1.90	0.00	3.77	0.02	40	55.5	6630	10	1.48
C1	PC7	23.8	61	85.3	7.70	1.3	17.6	1.90	0.00	3.27	0.00	20	39.6	630	100	0.77
C1	PC8	101.6	59	93.1	7.39	18.8	19.8	2.30	0.03	3.01	0.02	80	55.0	12910	1450	0.65
C1	PC9	56.5	29	84.4	7.51	2.7	17.9	2.00	0.01	2.26	0.01	40	45.6	6350	850	0.80
C1	PC10	167.7	55	92.9	7.65	29.9	19.7	2.40	0.02	2.26	0.03	120	22.3	115300	850	0.76

Fonte: Santos (2018).

Abreviações: Campo = Campanha de coleta (C1–C12); Ponto = Identificação do ponto de amostragem (PC1–PC27); COND = Condutividade Elétrica ($\mu\text{S}/\text{cm}$); ORP = Potencial de Oxidação-Redução (mV); SatOD = Saturação de Oxigênio Dissolvido (%); pH = Potencial Hidrogeniônico; Turb = Turbidez (NTU); Ta = Temperatura da água ($^{\circ}\text{C}$); NO₃ = Nitrato (mg/L); NO₂ = Nitrito (mg/L); Cl = Cloretos (mg/L); PT = Fósforo Total (mg/L); ST = Sólidos Totais (mg/L); Alca = Alcalinidade (mg CaCO₃/L); CT = Coliformes Totais (NMP/100mL); E.coli = *Escherichia coli* (NMP/100mL); DBO = Demanda Bioquímica de Oxigênio (mg/L).

Nota: Amostra ilustrativa contendo as primeiras 10 observações. Dataset completo: 324 observações $\times$ 17 variáveis. Período de coleta: 2016–2017 (12 campanhas). Pontos de amostragem: 27 locais distribuídos na bacia do rio Gualaxo do Norte (MG).

Essa diversidade de valores permitiu a identificação de padrões complexos mediante técnicas de aprendizado não supervisionado (Capítulo 4.2) e a construção de modelos preditivos robustos (Capítulo 4.3), conforme detalhado ao longo desta dissertação.

Referências

- ABADI, Martín et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: <https://www.tensorflow.org/>. Citado 2 vezes na página 58.
- AGÊNCIA NACIONAL DE ÁGUAS (ANA); MINISTÉRIO DO MEIO AMBIENTE; PROGRAMA DAS NAÇÕES UNIDAS PARA O MEIO AMBIENTE. *Cuidando das águas: soluções para melhorar a qualidade dos recursos hídricos*. 2. ed.: ANA, 2017. Citado 1 vez na página 39.
- ANKERST, Mihael et al. OPTICS: Ordering Points to Identify the Clustering Structure. In: PROCEEDINGS of the 1999 ACM SIGMOD International Conference on Management of Data. 1999. P. 49–60. DOI: [10.1145/304182.304187](https://doi.org/10.1145/304182.304187). Citado 1 vez na página 48.
- BARCELLOS, Demian da Silveira; SOUZA, Felipe Teixeira. Optimization of water quality monitoring programs by data mining. *Water Research*, v. 221, p. 118805, ago 2022. DOI: [10.1016/j.watres.2022.118805](https://doi.org/10.1016/j.watres.2022.118805). Disponível em: <https://pubmed.ncbi.nlm.nih.gov/35949073/>. Acesso em: 11 dez. 2024. Citado 2 vezes nas páginas 151, 206.
- BEHMEL, Sonja et al. Water quality monitoring strategies – A review and future perspectives. *Science of The Total Environment*, v. 571, p. 1312–1329, nov. 2016. DOI: [10.1016/j.scitotenv.2016.06.235](https://doi.org/10.1016/j.scitotenv.2016.06.235). Disponível em: <https://pubmed.ncbi.nlm.nih.gov/27396312/>. Acesso em: 11 dez. 2024. Citado 3 vezes nas páginas 65, 151.
- BREIMAN, Leo. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324). Disponível em: <https://link.springer.com/article/10.1023/A:1010933404324>. Citado 3 vezes nas páginas 34, 140.
- CALINSKI, Tadeusz; HARABASZ, Jerzy. A Dendrite Method for Cluster Analysis. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. DOI: [10.1080/03610927408827101](https://doi.org/10.1080/03610927408827101). Citado 2 vezes na página 84.
- CAMPELLO, Ricardo J. G. B.; MOULAVI, Davoud; SANDER, Jörg. Density-Based Clustering Based on Hierarchical Density Estimates. *Advances in Knowledge Discovery and Data Mining (PAKDD)*, v. 7819, p. 160–172, 2013. DOI: [10.1007/978-3-642-37456-2_14](https://doi.org/10.1007/978-3-642-37456-2_14). Citado 1 vez na página 48.
- CASTILLO, J. David Allan; M., María. *Stream Ecology: Structure and Function of Running Waters*. 3. ed. Cham: Springer, 2021. DOI: [10.1007/978-3-030-60222-6](https://doi.org/10.1007/978-3-030-60222-6). Citado 2 vezes nas páginas 39, 91.

- CAWLEY, Gavin C.; TALBOT, Nicola L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, v. 11, p. 2079–2107, jul. 2010. ISSN 1532-4435. Disponível em: <http://www.jmlr.org/papers/v11/cawley10a.html>. Citado 3 vezes nas páginas 51, 52, 54.
- CETESB. *Tabela de Preços de Serviços 2024-2025*. São Paulo, SP, 2024. Vigência: 01/10/2024 a 30/09/2025. Disponível em: <https://cetesb.sp.gov.br/licenciamentoambiental/wp-content/uploads/2024/09/Servicos-Cetesb-Tabela-de-precos-2024-2025.pdf>. Acesso em: 11 dez. 2024. Citado 2 vezes nas páginas 65, 205.
- CHAWLA, Nitesh V. et al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, jun. 2002. ISSN 1076-9757. DOI: 10.1613/jair.953. Disponível em: <https://doi.org/10.1613/jair.953>. Citado 4 vezes na página 55.
- CONSELHO NACIONAL DO MEIO AMBIENTE (CONAMA). *Resolução n.º 357, de 17 de março de 2005*. Mar. 2005. Diário Oficial da União. Dispõe sobre a classificação dos corpos de água e diretrizes ambientais para o seu enquadramento, bem como estabelece as condições e padrões de lançamento de efluentes, e dá outras providências. Citado 1 vez na página 32.
- CRUZ, Fábio Monteiro; SILVA, Talita Fernanda das Graças. Water quality emergency monitoring networks: A method for identifying non-critical variables based on Shannon's entropy. *Journal of Hydroinformatics*, v. 26, n. 3, p. 658–682, mar. 2024. DOI: 10.2166/hydro.2024.256. Disponível em: <https://iwaponline.com/jh/article/26/3/658/100858>. Acesso em: 11 dez. 2024. Citado 1 vez na página 151.
- DAVIES, David L; BOULDIN, Donald W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, PAMI-1, n. 2, p. 224–227, 1979. DOI: 10.1109/TPAMI.1979.4766909. Citado 1 vez na página 84.
- DESLANDES, Suely Ferreira; GOMES, Romeu. *Pesquisa Social: Teoria, Método e Criatividade*. Petrópolis: Vozes, 2009. Citado 1 vez na página 28.
- ENVIRONMENTAL COMPUTING. *Principal Component Analysis*. Article Last Updated Date: Feb 2022. Fev. 2022. Disponível em: <https://environmentalcomputing.net/principal-components-analysis/>. Citado 1 vez na página 34.
- EVERITT, Brian S et al. *Cluster Analysis*. 5. ed.: John Wiley & Sons, 2011. Citado 8 vezes nas páginas 41, 43, 51, 57, 58, 60, 64.
- FINZEL, Bettina. Explainable AI: Foundations, Applications, Opportunities and Challenges. *Information Fusion*, 2025. Citado 3 vezes nas páginas 35, 36.
- FRÉDAY, Benoît; VERLEYSEN, Michel. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, v. 25, n. 5, p. 845–869, 2014. DOI: 10.1109/TNNLS.2013.2292894. Disponível em: <https://ieeexplore.ieee.org/document/6685834>. Citado 1 vez na página 143.

- FRIEDMAN, Jerome H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001. DOI: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451). Citado 1 vez na página 34.
- GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. Cambridge, MA: MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>. Citado 3 vezes nas páginas 35, 53, 141.
- GREENACRE, Michael; GROENEN, Patrick J.F.; HASTIE, Trevor et al. Principal component analysis. *Nature Reviews Methods Primers*, v. 2, p. 100, 2022. DOI: [10.1038/s43586-022-00184-w](https://doi.org/10.1038/s43586-022-00184-w). Citado 1 vez na página 33.
- HAND, David J; TILL, Robert J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, Springer, v. 45, n. 2, p. 171–186, 2001. DOI: [10.1023/A:1010920819831](https://doi.org/10.1023/A:1010920819831). Citado 2 vezes na página 59.
- HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009. ISBN 978-0-387-84857-0. DOI: [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7). Citado 4 vezes na página 59.
- HATJE, Vanessa; PEDREIRA, Tatiane F.; REZENDE, Marcus R. Impactos do rompimento da barragem de Fundão sobre o sistema fluvial do rio Doce. *Química Nova*, v. 40, n. 9, p. 1032–1040, 2017. DOI: [10.21577/0100-4042.20170116](https://doi.org/10.21577/0100-4042.20170116). Citado 3 vezes nas páginas 38, 39, 101.
- HUBERT, Lawrence; ARABIE, Phipps. Comparing partitions. *Journal of Classification*, v. 2, n. 1, p. 193–218, 1985. DOI: [10.1007/BF01908075](https://doi.org/10.1007/BF01908075). Citado 3 vezes nas páginas 45, 48, 85, 98.
- JOLLIFFE, Ian T. *Principal Component Analysis*. 2. ed. New York: Springer, 2002. DOI: [10.1007/b98835](https://doi.org/10.1007/b98835). Citado 1 vez na página 43.
- JOLLIFFE, Ian T.; CADIMA, Jorge. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, v. 374, n. 2065, 2016. DOI: [10.1098/rsta.2015.0202](https://doi.org/10.1098/rsta.2015.0202). Citado 3 vezes nas páginas 33, 34.
- KAUFMAN, Shachar et al. Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Transactions on Knowledge Discovery from Data*, v. 6, n. 4, 15:1–15:21, dez. 2012. DOI: [10.1145/2382577.2382579](https://doi.org/10.1145/2382577.2382579). Disponível em: <https://dl.acm.org/doi/10.1145/2382577.2382579>. Citado 4 vezes nas páginas 51, 52, 122.
- LI, Yinpu et al. *Beyond Tides and Time: Machine Learning's Triumph in Water Quality Forecasting*. Versão inglesa. 2023. arXiv preprint arXiv:2309.16951. Disponível em: <https://arxiv.org/abs/2309.16951>. Citado 11 vezes nas páginas 26, 34, 37, 41, 43, 51, 57, 60, 64, 73, 166.

- LUNDBERG, Scott M.; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Curran Associates, Inc., 2017. P. 4765–4774. Disponível em: <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>. Citado 2 vezes na página 61.
- MAATEN, Laurens van der; HINTON, Geoffrey. Visualizing Data using t-SNE. In: JOURNAL of Machine Learning Research. 2008. v. 9, p. 2579–2605. Disponível em: <https://www.jmlr.org/papers/volume9/vandemaaten08a/vandemaaten08a.pdf>. Citado 5 vezes nas páginas 35, 48, 82.
- MACQUEEN, J. B. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, v. 1, p. 281–297, 1967. Citado 1 vez na página 48.
- MAKUMBURA, Randika K. et al. Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature. English. *Results in Engineering*, v. 23, set. 2024. Publisher Copyright: © 2024 The Authors. ISSN 2590-1230. DOI: [10.1016/j.rineng.2024.102831](https://doi.org/10.1016/j.rineng.2024.102831). Citado 2 vezes nas páginas 35, 61.
- MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008. ISBN 978-0-521-86571-5. Citado 5 vezes nas páginas 45, 48, 59.
- MCINNES LELAND, John Healy; MELVILLE, James. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*, 2018. Disponível em: <https://arxiv.org/abs/1802.03426>. Citado 5 vezes nas páginas 35, 48, 82, 125.
- NATEKIN, A.; IYENGAR, A. Gradient boosting machines: A tutorial. *Frontiers in Neurobotics*, v. 7, p. 21, 2013. DOI: [10.3389/fnbot.2013.00021](https://doi.org/10.3389/fnbot.2013.00021). Citado 1 vez na página 34.
- PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Disponível em: <http://jmlr.org/papers/v12/pedregosa11a.html>. Citado 3 vezes nas páginas 57, 58.
- POWERS, David Martin Ward. *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation*. Adelaide, Australia: Bioinfo Publications, 2011. Citado 1 vez na página 59.
- PREFEITURA MUNICIPAL DE BRAÇO DO TROMBUDO. *Pesquisa de Preços para Análises da Estação de Tratamento de Água*. Braço do Trombudo, SC, fev 2025. Relatório gerado em 04 de fevereiro de 2025. Disponível em: <https://bracodotrombudo.sc.gov.br/uploads/sites/304/2025/02/PESQUISA-DE-PRECO-ANALISES-DA-ETA.pdf>. Acesso em: 11 dez. 2024. Citado 2 vezes nas páginas 65, 205.

- RAHMAN, M. A. et al. eXtreme gradient boosting algorithm with machine learning: A review. *Journal of Engineering Research and Reports*, v. 24, n. 1, p. 1–10, 2023. Disponível em: https://www.researchgate.net/publication/371202498_eXtreme_gradient_boosting_algorithm_with_machine_learning_A_review. Citado 1 vez na página 34.
- RASDAQ, Sodiq. Explainable AI for Ethical Data Science: Addressing Bias and Accountability. ., v. 1, p. 6, fev. 2025. Citado 4 vezes nas páginas 36, 62.
- RAYNA, Thierry; STRIUKOVA, Ludmila. *Digital Enterprise Design & Management*. Edição: P.-J Benghozi. Springer International Publishing, jan. 2014. P. 119–130. Citado 1 vez na página 49.
- RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2016. P. 1135–1144. DOI: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778). Disponível em: <https://doi.org/10.1145/2939672.2939778>. Citado 1 vez na página 61.
- RIBEIRO, Wallyson Vinícius de Mérida. *Avaliação da Recuperação do Rio Doce Após o Rompimento da Barragem do Fundão*. Versão portuguesa. 2021. Monografia de Graduação. Universidade de Brasília. Citado 2 vezes na página 30.
- ROUSSEEUW, Peter J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, Elsevier, v. 20, p. 53–65, 1987. DOI: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7). Citado 2 vezes na página 84.
- SALMAN, Hasan; KALAKECH, Ali; STEITI, Amani. Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, v. 2024, p. 69–79, jun. 2024. DOI: [10.58496/BJML/2024/007](https://doi.org/10.58496/BJML/2024/007). Citado 1 vez na página 34.
- SALTELLI, Andrea et al. *Global Sensitivity Analysis: The Primer*. Chichester: Wiley, 2008. Citado 1 vez na página 66.
- SANTOS, Grazielle Rocha dos. *Técnicas de Reconhecimento de Padrões para a Avaliação de Corpos Hídricos Afetados por Rejeitos do Rompimento da Barragem de Fundão*. Fev 2018. Trabalho de Conclusão de Curso (Graduação em Engenharia Ambiental) – Universidade Federal de Ouro Preto, Ouro Preto, MG. Citado 66 vezes nas páginas 26–28, 30, 33, 34, 36–38, 40, 41, 43, 46, 47, 50, 60, 61, 68, 70, 76–80, 84–86, 89, 90, 98, 102, 106, 108, 117, 118, 121, 136, 139, 140, 154–157, 159, 163, 166, 183, 189, 203, 208, 209.
- SANTOS, Grazielle Rocha dos et al. Water quality assessment based on multivariate statistics and water quality index of a strategic river in the Brazilian Atlantic Forest. *Scientific Reports*, v. 8, p. 22038, 2018. DOI: [10.1038/s41598-018-40381-x](https://doi.org/10.1038/s41598-018-40381-x). Citado 1 vez na página 151.

- AL-SHAMI, A.; AL-MOUSA, A. A Survey: Potential Dimensionality Reduction Methods. *Journal of Engineering Science and Technology*, v. 19, n. 1, p. 1–15, 2024. Disponível em: https://www.researchgate.net/publication/389091180_A_Survey_Potential_Dimensionality_Reduction_Methods. Citado 5 vezes nas páginas 34, 35.
- AL-SHAMI, A.; AL-MOUSA, A. Explainable AI: A Comparative Study of LIME and SHAP for Deep Neural Networks. *Journal of Engineering Science and Technology*, v. 18, n. 6, p. 4642–4654, 2023. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10074303/>. Citado 1 vez na página 62.
- SHARPLEY, Andrew N. et al. Managing agricultural phosphorus for protection of surface waters: Issues and options. *Journal of Environmental Quality*, v. 23, n. 3, p. 437–451, 1994. DOI: [10.2134/jeq1994.00472425002300030006x](https://doi.org/10.2134/jeq1994.00472425002300030006x). Citado 2 vezes na página 31.
- SILVA, Igor Santos. *Inteligência Artificial para Avaliação da Qualidade da Água*. 2019. Dissertação de Mestrado – Universidade Federal de Sergipe, São Cristóvão, SE. Citado 2 vezes nas páginas 30, 37.
- SILVA, Luiz Ricardo Mantovani da. Aplicação de Inteligência Artificial na Predição e Monitoramento da Qualidade da Água em Bacias Hidrográficas. Versão portuguesa. *Environmental Science & Technology Innovation*, Bauru, SP, v. 2, n. 2, p. 177–208, 2023. Citado 2 vezes nas páginas 27, 37.
- SPERLING, Marcos von. *Introdução à Qualidade das Águas e ao Tratamento de Esgotos*. 3. ed. Belo Horizonte: Editora UFMG, 2005. v. 1. (Princípios do Tratamento Biológico de Águas Residuárias). ISBN 85-7041-446-5. Citado 1 vez na página 47.
- STREHL, Alexander; GHOSH, Joydeep. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, v. 3, p. 583–617, 2002. Disponível em: <https://www.jmlr.org/papers/volume3/strehl02a/strehl02a.pdf>. Citado 2 vezes nas páginas 45, 48.
- U.S. ENVIRONMENTAL PROTECTION AGENCY (EPA). *National Primary Drinking Water Regulations*. Article Last Updated Date: December 12, 2024. Dez. 2024. Disponível em: <https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations>. Citado 4 vezes nas páginas 30–32.
- U.S. ENVIRONMENTAL PROTECTION AGENCY (EPA). *Water Quality Criteria*. Accessed on 22 de dezembro de 2025. Disponível em: <https://www.epa.gov/wqc>. Citado 1 vez na página 32.
- WANG, X. et al. A Hybrid Model for Water Quality Prediction Based on an Artificial Neural Network, Wavelet Transform, and Long Short-Term Memory. *Water*, v. 14, n. 4, p. 610, 2022. DOI: [10.3390/w14040610](https://doi.org/10.3390/w14040610). Citado 1 vez na página 35.

WORLD HEALTH ORGANIZATION. *Guidelines for Drinking-water Quality: Fourth Edition Incorporating the First Addendum*. Geneva: World Health Organization, 2017. ISBN 978-92-4-154995-0. Citado 5 vezes nas páginas 39, 47, 85, 101.

XU, Rui; WUNSCH, Donald. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, v. 16, n. 3, p. 645–678, 2005. DOI: [10.1109/TNN.2005.845141](https://doi.org/10.1109/TNN.2005.845141). Citado 2 vezes na página 43.

YSI - XYLEM ANALYTICS. *EXO2 Multiparameter Sonde: Especificações Técnicas e Preços*. 2024. Disponível em: <https://www.ysi.com/exo2>. Acesso em: 11 dez. 2024. Citado 2 vezes nas páginas 65, 206.