ESTUDO EXPLORATÓRIO DA QUALIDADE DA ÁGUA NA BACIA DO RIO MOGI GUAÇU (UGRHI-9) POR MEIO DE ABORDAGEM DE APRENDIZADO DE MÁQUINA

Gabriella Borges Cristovam [1]; Franciane Mendonça dos Santos [2];& Natália de Souza Pelinson [3]

Abstract: The monitoring of surface and groundwater has been considered essential for preserving water quality and ecosystem health within watersheds. However, this process is marked by high costs and significant time demands. In this study, machine learning techniques were explored to assess water quality in the Mogi Guaçu River Basin (UGRHI-9). Data were obtained from historical records provided by the Environmental Company of the State of São Paulo (CETESB) for the years 2019 and 2020. Supervised models were employed to predict the Water Quality Index (WQI), with multiple linear regression and the Random Forest algorithm showing the best performance. The regression model achieved an R^2 of 0.97 and RMSE of 2.37, while Random Forest yielded an R^2 of 0.92 and RMSE of 0.06. These results demonstrated strong predictive capability. Unsupervised techniques were also applied to identify patterns and reduce variable dimensionality. PCA, K-means, and DBSCAN methods were used. Relevant patterns were identified, and distinct clusters were formed. These techniques enabled the observation of correlations among variables and spatial trends. Overall, the tested algorithms performed well. The methods applied can support water quality management and control actions, as the integration of data science tools into environmental monitoring is increasingly considered viable and promising.

Keywords: applied statistical methods; clustering; water monitoring.

Resumo: O monitoramento de águas superficiais e subterrâneas é essencial para a preservação da qualidade hídrica em bacias hidrográficas, embora envolva elevados custos e demanda de tempo. Este estudo explorou técnicas de aprendizado de máquina aplicadas à avaliação da qualidade da água na bacia do rio Mogi Guaçu (UGRHI-9), com base em dados históricos da CETESB referentes a 2019 e 2020. Foram utilizados modelos supervisionados para prever o Índice de Qualidade da Água (IQA), com destaque para a regressão linear múltipla ($R^2 = 0,97$; RMSE = 2,37) e o algoritmo Random Forest ($R^2 = 0,92$; RMSE = 0,06), ambos com bom desempenho preditivo. Métodos não supervisionados também foram aplicados para reconhecimento de padrões e redução de dimensionalidade, incluindo PCA, K-means e DBSCAN. A análise revelou agrupamentos distintos e correlações relevantes entre variáveis, evidenciando tendências espaciais nos dados. Os resultados demonstram que o uso combinado de técnicas supervisionadas e não supervisionadas pode apoiar a gestão e o controle da qualidade da água. A integração dessas abordagens ao monitoramento ambiental mostra-se viável e promissora, embora demande estudos adicionais para seu aprimoramento e aplicação em diferentes contextos.

Palavras-Chave – métodos de estatística aplicada; agrupamento estatístico; monitoramento de água.

1) Universidade Federal de São Carlos (UFSCar-CCN Campus Lagoa do Sino); Buri - SP; e-mail: gabiborgescristovam@gmail.com
2) Universidade Federal de São Carlos (UFSCar-CCN Campus Lagoa do Sino); Buri - SP; e-mail: francianems@ufscar.br
3) Universidade Federal de São Carlos (UFSCar-CCN Campus Lagoa do Sino); Buri - SP; e-mail: nataliasp@ufscar.br

INTRODUÇÃO

A intensificação das atividades humanas tem provocado alterações na composição química das águas superficiais e subterrâneas, comprometendo sua qualidade e uso. O lançamento inadequado de esgotos e a ausência de infraestrutura sanitária contribuem para a presença de matéria orgânica, patógenos e nutrientes em níveis que ameaçam os ecossistemas aquáticos e a saúde pública, podendo causar eutrofização e desoxigenação (Reis *et al*., 2023). Compreender essas dinâmicas pode ser fundamental para orientar políticas públicas e estratégias de regulação ambiental.

O monitoramento da qualidade da água superficial, no Brasil, deve considerar a Resolução CONAMA nº 357/2005 (Brasil, 2005), com coletas mínimas em períodos chuvoso e seco. A padronização temporal e espacial das campanhas permite análises mais consistentes. Dados pluviométricos e fluviométricos, quando integrados, podem indicar relações entre variações hidrológicas e oscilações na qualidade da água de rios (Galinaro *et al*., 2022). O monitoramento contínuo das águas, no país, é realizado por órgãos como a Companhia Ambiental do Estado de São Paulo (CETESB) e a Agência Nacional de Águas e Saneamento (ANA), baseando-se em índices que traduzem a condição dos corpos hídricos. Complementarmente às análises laboratoriais mais rigorosas, técnicas de ciência de dados, como o aprendizado de máquina (ML), têm ganhado espaço na análise dos parâmetros da qualidade hídrica.

Os métodos de ML podem ser divididos em abordagens supervisionadas e não supervisionadas. A primeira prevê variáveis a partir de dados rotulados, enquanto a segunda permite a identificação de padrões sem conhecimento prévio das categorias (Alloghani *et al*., 2020).

Diante desse cenário, este estudo tem como objetivo analisar a qualidade das águas naturais — superficiais e subterrâneas — por meio da aplicação de técnicas de aprendizado de máquina supervisionadas e não supervisionadas, com foco na Unidade de Gerenciamento de Recursos Hídricos do Mogi Guaçu, identificada como UGRHI-9. Os dados utilizados como base foram obtidos nos registros históricos da CETESB.

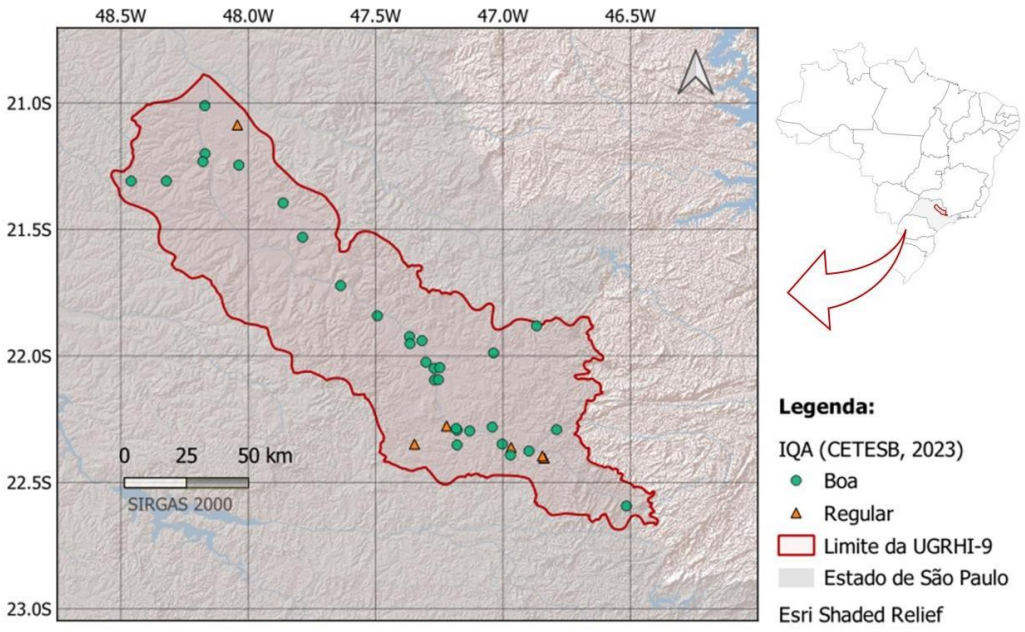
No conjunto de métodos não supervisionados, foram utilizados a Análise de Componentes Principais (ACP), o agrupamento por médias (*K-means*) e o algoritmo de agrupamento baseado na densidade com ruído (do inglês *Density-based Spatial Clustering of Applications with Noise* - DBSCAN). Esses métodos estatísticos aplicados permitiram identificar padrões recorrentes e anomalias nos dados de qualidade da água. Paralelamente, no conjunto de técnicas supervisionadas, foram aplicados a Regressão Linear Múltipla e o algoritmo de *Random Forest*, ambos empregados na construção de modelos preditivos para o Índice de Qualidade da Água (IQA).

METODOLOGIA

A área de estudo deste trabalho corresponde à Unidade de Gerenciamento de Recursos Hídricos do Rio Mogi Guaçu, oficialmente designada como UGRHI-9, situada no estado de São Paulo, Brasil (Figura 1). Esta UGHRI abrange uma extensa região composta por 43 municípios, totalizando uma população superior a um milhão de habitantes.

A bacia hidrográfica do Mogi Guaçu é caracterizada pela diversidade de usos do solo, com predomínio de atividades agrícolas e industriais, além de áreas urbanas em expansão. O rio Mogi Guaçu, que percorre dois estados brasileiros — São Paulo e Minas Gerais — possui uma extensão total de 473 km, sendo aproximadamente 378 km localizados em território paulista. Sua importância ecológica, econômica e social está relacionada tanto ao abastecimento hídrico quanto à presença de ecossistemas aquáticos vulneráveis à degradação.

Figura 1 - Área de abrangência da Unidade de Gerenciamento de Recursos Hídricos do rio Mogi Guaçu (UGRHI-9), situada no estado de São Paulo, Brasil, indicando a distribuição espacial dos pontos de monitoramento de qualidade da água superficial utilizados pela Companhia Ambiental do Estado de São Paulo (CETESB) da qualidade do Índice de Qualidade da Água (IQA).



Nos anos de 2014 e 2015, a região enfrentou eventos climáticos extremos, com significativas anomalias de precipitação que afetaram a disponibilidade de água e agravaram os problemas relacionados à sua qualidade (Galinaro *et al*., 2022). Tais características reforçam a necessidade de estudos voltados à compreensão dos padrões de qualidade hídrica na UGRHI-9, considerando suas particularidades territoriais e pressões antrópicas crescentes.

O monitoramento da qualidade das águas superficiais foi realizado com base nos dados da Companhia Ambiental do Estado de São Paulo (CETESB) entre 2014 e 2023. A metodologia aplicada seguiu o fluxo proposto simplificadamente por Hossain (2024): coleta de dados, pré-processamento, treinamento de modelos e avaliação. O de talhamento completo dos códigos e processos metodológicos da análise estatística da qualidade das águas está disponível no repositório GitHub: https://github.com/gabriellabc/monitoramento_agua_superficial_subterranea.

Os dados foram analisados por meio de técnicas de aprendizado de máquina supervisionado e não supervisionado, com uso da linguagem Python e execução na plataforma Google Colab. Foram empregadas bibliotecas e *frameworks* especializados: Pandas (manipulação de dados), Scikit-learn (modelos de ML), NumPy (operações matemáticas), Matplotlib e Plotly (visualização gráfica), SciPy (análises estatísticas), e Spark MLlib (treinamento do modelo *Random Forest* em larga escala).

Coleta e análise de dados de águas superficiais

Foram coletados dados de qualidade da água em rios da Unidade de Gerenciamento de Recursos Hídricos do Mogi Guaçu (UGRHI-9), com base em registros da Companhia Ambiental do Estado de São Paulo (CETESB) e na plataforma InfoÁguas, abrangendo o período de 2014 a 2023. As variáveis analisadas correspondem aos nove parâmetros que compõem o Índice de Qualidade da Água (IQA), como oxigênio dissolvido, pH, turbidez e Escherichia coli, entre outros. A base de dados totalizou 1.783 registros. Após o pré-processamento (remoção de outliers, tratamento de nulos, padronização),

aplicaram-se algoritmos de aprendizado não supervisionado (Análise de Componentes Principais – PCA; K-Means; DBSCAN) para identificar padrões e reduzir a dimensionalidade. Para previsão do IQA, foram utilizados modelos supervisionados: Regressão Linear Múltipla e Random Forest, com validação por R^2 e RMSE. As análises também integraram dados mensais de precipitação para avaliar correlações climáticas.

Coleta e análise de dados de águas subterrâneas

Os dados de aquíferos foram obtidos para os anos de 2022 e 2023 na plataforma InfoÁguas, devido à limitada disponibilidade histórica. A base resultante somou 65 registros. Foram considerados mais de 40 parâmetros físico-químicos e microbiológicos, como metais (arsênio, cádmio, chumbo), nutrientes (nitrato, nitrito), indicadores microbiológicos e características físico-químicas (pH, condutividade, dureza, etc.). Após o pré-processamento (incluindo Winsorização para outliers, devido ao pequeno volume de dados), aplicaram-se algoritmos não supervisionados para análise exploratória e agrupamento: PCA para redução de variáveis, e os métodos K-Means e DBSCAN para identificação de padrões e possíveis anomalias. O aprendizado supervisionado não foi aplicado a este conjunto, devido ao volume limitado e ausência de variável-resposta padronizada como o IQA.

RESULTADOS E DISCUSSÃO

O aprendizado não supervisionado foi utilizado para o reconhecimento de padrões e correlações, para águas subterrâneas e superficiais, uma vez que neste método nenhuma informação de resposta é fornecida para os algoritmos (Costa, 2024).

Algoritmos aplicados para a análise água subterrânea

Foram analisados 40 parâmetros físico-químicos e microbiológicos em amostras de água subterrânea, incluindo metais como arsênio, cádmio e chumbo; nutrientes como nitrato e nitrito; indicadores microbiológicos; além de variáveis físico-químicas, como pH, condutividade elétrica, dureza total e dados de precipitação acumulada (pluviometria).

A aplicação da Análise de Componentes Principais (PCA) permitiu a redução da dimensionalidade do conjunto original, resultando em oito componentes principais que explicaram, conjuntamente, 71,2% da variância total. Esse valor indica uma boa capacidade de representação das informações relevantes, favorecendo a interpretação de padrões latentes nos dados (Zimmermann *et al*., 2008). Entre os componentes, a variabilidade foi: PC1 explicou 16,27%, seguido de PC2 (14,14%), PC3 (11,76%) e PC4 (9,11%), não representando bem a redução dos dados analisados.

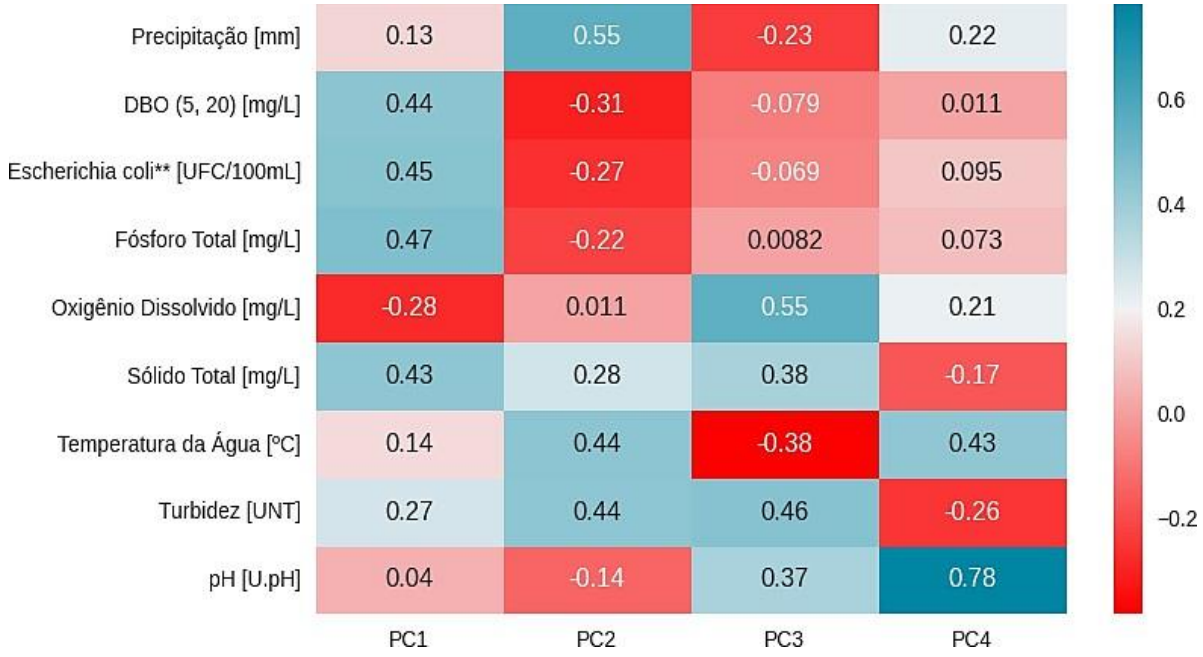
Desta forma, a análise por meio do algoritmo de agrupamento K-means não apresentou resultados satisfatórios, possivelmente devido à baixa representatividade e à limitação amostral dos dados. As métricas obtidas – Silhouette Score de 0,30, Calinski-Harabasz Index de 16,88 e Davies-Bouldin Index de 0,90 – indicaram sobreposição entre os grupos e pouca separabilidade entre clusters. Resultado semelhante foi observado com o uso do algoritmo DBSCAN. Mesmo após ajustes de parâmetros, as métricas obtidas (Silhouette Score = 0,28; Calinski-Harabasz = 10,11; Davies-Bouldin = 0,89) reforçam a limitação, sugerindo que os dados disponíveis não foram suficientes para uma segmentação robusta.

Algoritmos aplicados para análise da água superficial na UGRHI-9

A aplicação da Análise de Componentes Principais (PCA) ao conjunto de dados de qualidade da água superficial permitiu a identificação de quatro componentes principais (PCs), que juntos explicaram 80,74% da variabilidade dos dados originais. Isso demonstra a eficácia da redução de dimensionalidade sem perda significativa de informação. Os componentes PC1, PC2, PC3 e PC4 explicaram, respectivamente, 31,72%, 20,59%, 16,63% e 11,80% da variância.

Na Figura 2, estão apresentados os *loadings* das variáveis originais, indicando a contribuição relativa de cada uma na composição dos componentes, com valores positivos ou negativos próximos de 1 (ou -1) refletindo forte influência, e valores próximos de 0 indicando influência reduzida.

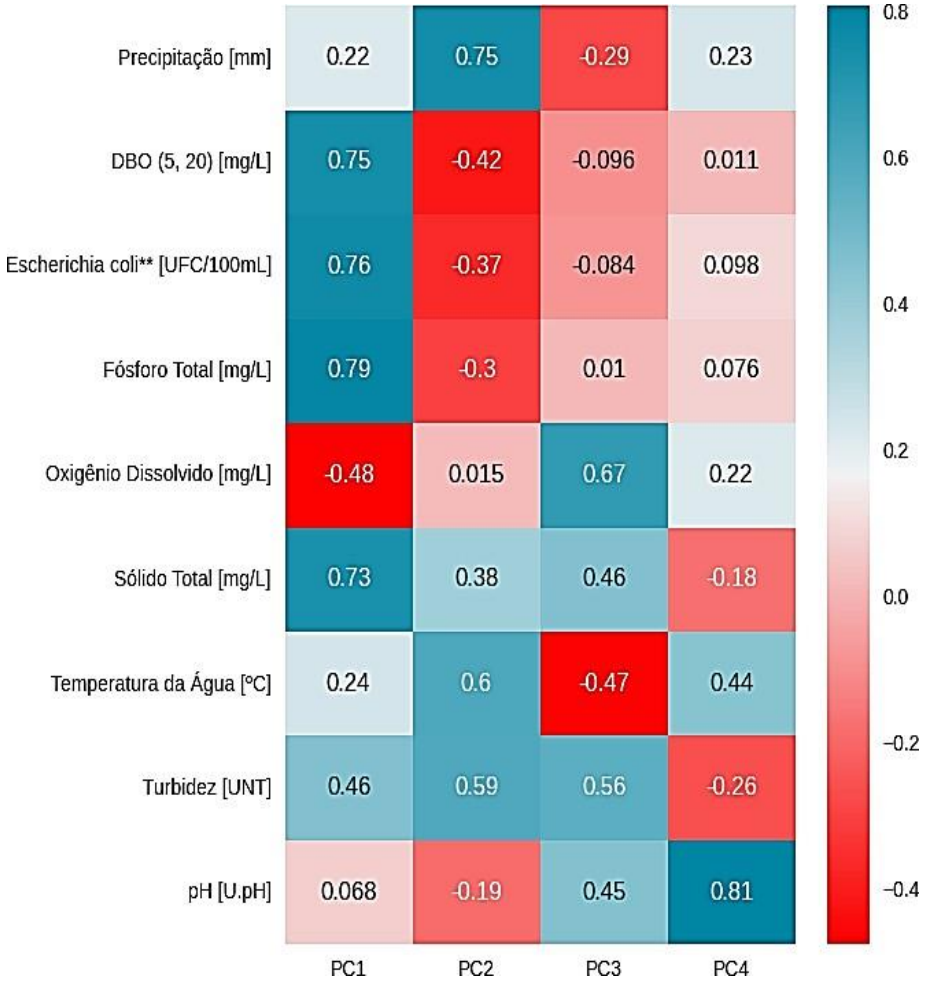
Figura 2 - Matriz de componentes (*loadings*) do PCA em relação aos parâmetros de água superficial. Influências positivas são representadas em azul e negativas em vermelho.



Na Figura 3, há a apresentação da correlação entre os componentes principais e as variáveis. O PC1 se correlacionou positivamente com fósforo total, *Escherichia coli*, DBO e sólidos totais, e negativamente com oxigênio dissolvido, indicando possível relação com contaminação por esgoto ou efluentes industriais. O PC2 mostrou correlação positiva com precipitação, temperatura da água e turbidez, e negativa com DBO, *E. coli*, fósforo total e pH, sugerindo influências de eventos de chuva sobre a qualidade da água. O PC3 apresentou correlação positiva com OD, turbidez e sólidos totais, e negativa com temperatura da água, enquanto o PC4 apresentou correlação positiva com o pH.

O algoritmo DBSCAN, por sua vez, apresentou resultados distintos: Silhouette Score = 0,68, Calinski-Harabasz = 242,01 e Davies-Bouldin = 1,55, e detectou 28 amostras como ruídos (*outliers*), evidenciando características atípicas no conjunto.

Figura 3 - Correlação entre variáveis originais e componentes principais (PCs). Azul indica correlação positiva e vermelho, por outro lado, representa a correlação negativa.



A modelagem preditiva foi realizada com dois algoritmos supervisionados: Regressão Linear Múltipla e Random Forest. As variáveis preditoras utilizadas foram OD, DBO, pH, *E. coli*, temperatura, nitrogênio total, fósforo total, turbidez, sólido total e precipitação; a variável dependente foi o Índice de Qualidade da Água (IQA). Após o treinamento, o modelo de regressão obteve R^2 = 0,94 (treino) e 0,92 (teste), com RMSE = 0,06 para ambos, indicando excelente ajuste.

O resultado da aplicação do algoritmo K-Means considerando os dois primeiros componentes (PC1 e PC2) pode ser observado na Figura 4. A divisão em três agrupamentos apresentou valores de *Silhouette Score* = 0,44, Calinski-Harabasz = 855,51 e Davies-Bouldin = 0,86, indicando agrupamentos razoáveis, embora com sobreposição entre grupos.

Enquanto, na Figura 5, pode ser observado o gráfico de dispersão entre valores previstos e observados de IQA, na Figura 6, está ilustrada a influência das variáveis sobre o IQA, destacando que OD, precipitação e fósforo total contribuem positivamente para o índice, enquanto pH, turbidez, sólido total, DBO e *E. coli* apresentaram relações negativas.

Figura 4 - Clusters K-Means entre PC1 e PC2 para água superficial da UGRHI-6

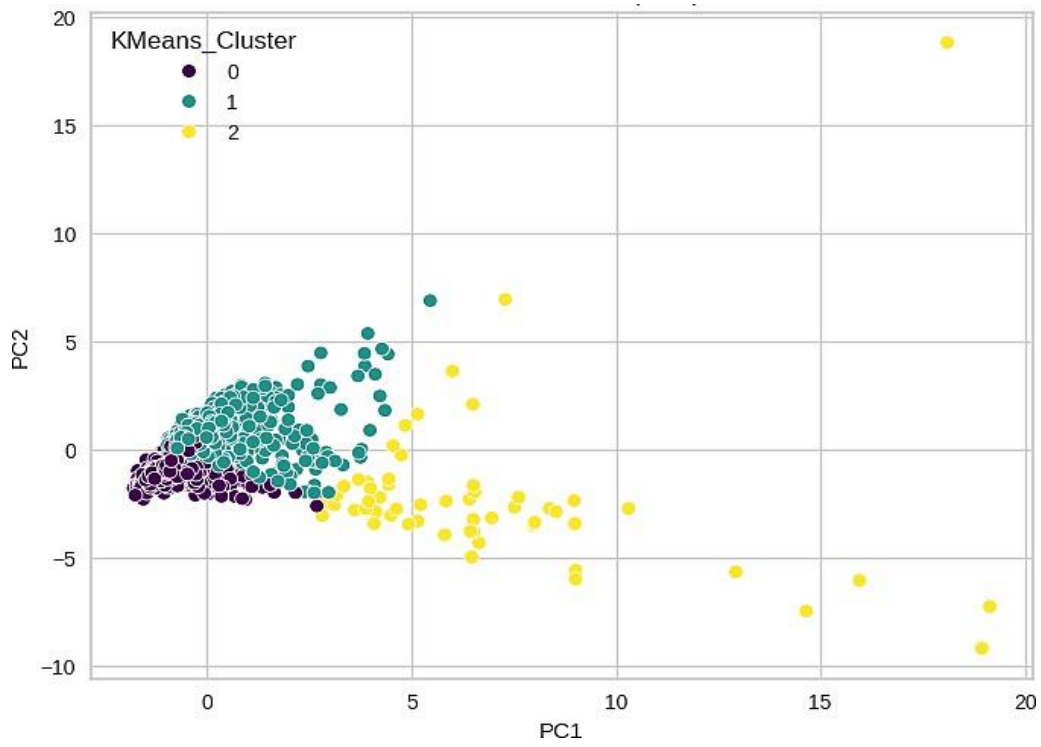


Figura 5 - Gráfico de dispersão entre IQA previsto e observado (Regressão Linear)

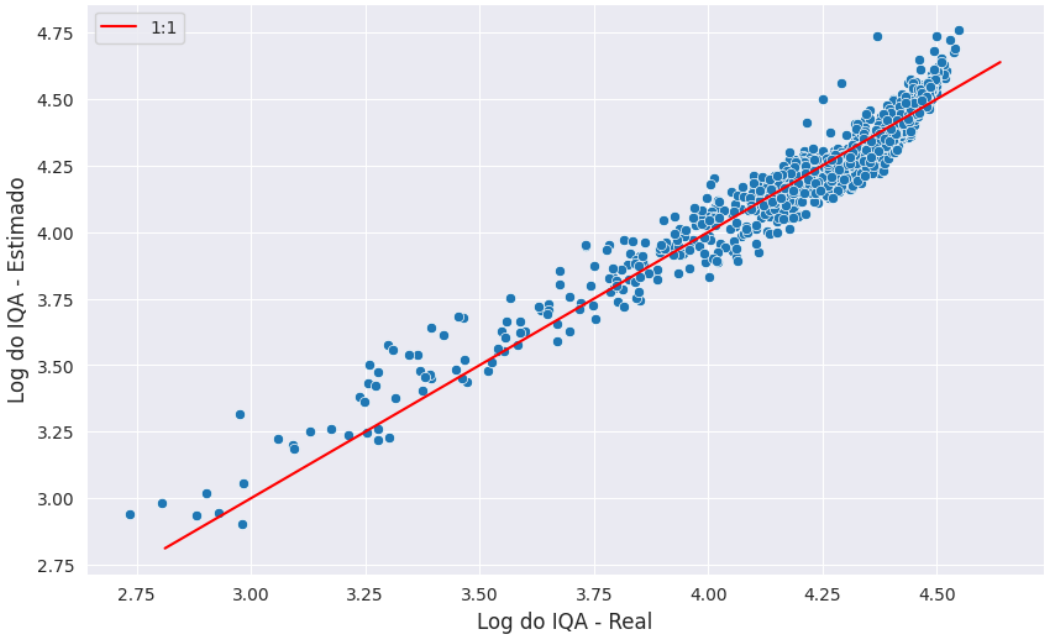
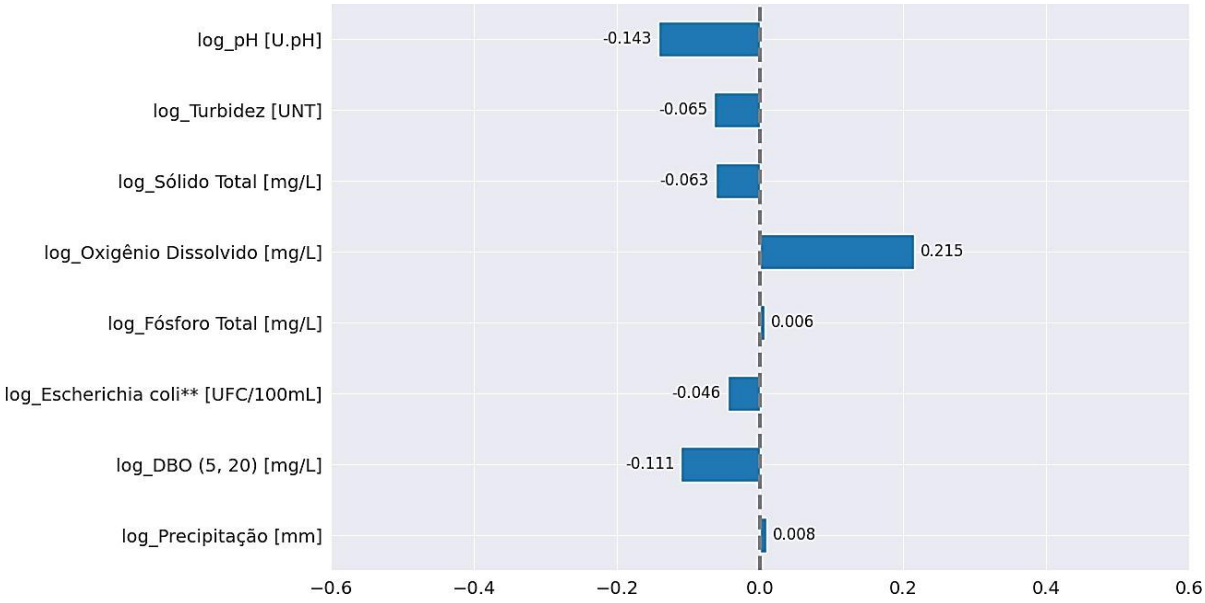


Figura 6 - Influência das variáveis na previsão do IQA (Regressão Linear)



No modelo Random Forest, as mesmas variáveis foram utilizadas. Os resultados obtidos foram R^2 = 0,97 (treino) e 0,96 (teste), com RMSE = 2,02 e 2,81, respectivamente. Na Figura 7, o gráfico de dispersão entre valores previstos e observados, evidenciando boa correlação, pode ser visualizado, enquanto, a Figura 12, está representada a importância relativa das variáveis, com destaque para *E. coli*, DBO, fósforo total e turbidez.

Figura 7 - Gráfico de dispersão entre valores previstos e reais utilizando *Random Forest*

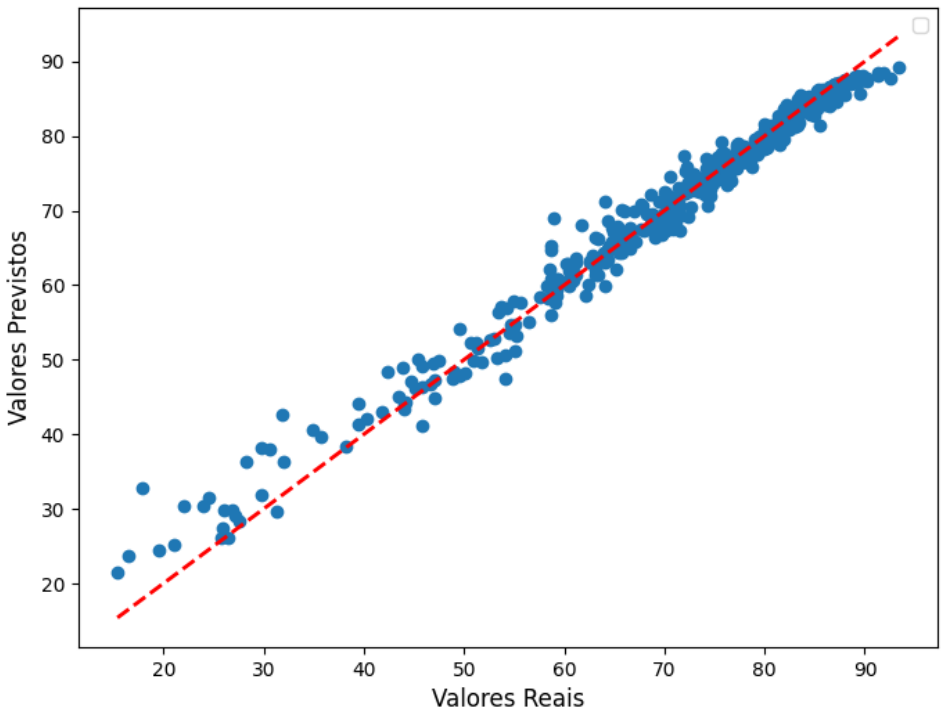
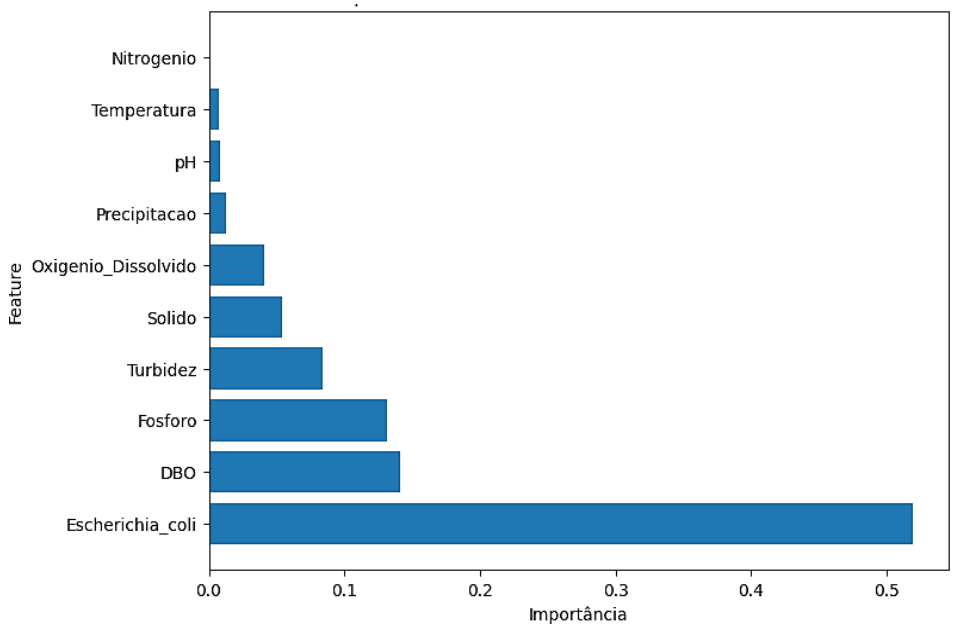


Figura 8 - Importância das variáveis preditoras (Random Forest)



Na Tabela 1, está apresentada a comparação entre os dois modelos. O Random Forest obteve maior R^2, enquanto a regressão linear apresentou menor RMSE, indicando menor erro de previsão.

Tabela 1 - Métricas de desempenho dos modelos supervisionados aplicados à água superficial na Bacia Mogi Guaçu (valores de R^2 próximos de 1 e RMSE próximos de 0 indicam melhor desempenho).

Métricas		**Regressão linear**	***Random Forest***
Treino	R^2	0.94	0.97
	RMSE	0.07	2.39
Teste	R^2	0.92	0.98
	RMSE	0.07	2.38

Dessa forma, a regressão linear se destacou como o modelo mais adequado para previsão do IQA nesta pesquisa, considerando a simplicidade, capacidade de generalização e o bom desempenho mesmo com a redução de variáveis.

CONSIDERAÇÕES FINAIS

Os resultados do estudo evidenciam o potencial do uso de aprendizado de máquina na análise da qualidade da água. Para águas superficiais, as técnicas aplicadas resultaram em análises exploratórias consistentes e previsões eficazes do Índice de Qualidade da Água (IQA), mesmo com a redução do número de variáveis, sugerindo caminhos para monitoramentos mais eficientes e econômicos. A inclusão da precipitação como variável demonstrou sua relevância na dinâmica da qualidade hídrica.

A Análise de Componentes Principais (PCA) possibilitou a identificação de padrões e relações entre os parâmetros físico-químicos, microbiológicos e climáticos, enquanto os modelos supervisionados, especialmente a Regressão Linear Múltipla, apresentaram alto desempenho preditivo.

Por outro lado, as análises para água subterrânea foram limitadas pela baixa disponibilidade de dados, o que comprometeu os resultados dos algoritmos de agrupamento. As métricas obtidas confirmam a necessidade de conjuntos de dados mais robustos para a aplicação de técnicas não supervisionadas nesse contexto. Tais limitações reforçam a urgência de aprimorar o acesso e a consistência dos dados nos sistemas oficiais de monitoramento, como o InfoÁguas.

De forma geral, o estudo demonstra que a aplicação de técnicas de inteligência artificial pode ampliar as possibilidades de interpretação e gestão da qualidade das águas, desde que apoiadas por bases de dados adequadas.

REFERÊNCIAS

ALLOGHANI, MOHAMED; AL-JUMEILY, DHIYA; MUSTAFINA, JAMILA, ABIR HUSSAIN; ALJAAF, AHMED J. *Chapter 1 A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*. In: BERRY, M. B.; MOHAMED, A.; YAP, B. W. Unsupervised and Semi-Supervised Learning. ISBN 978-3- 030-22474-5 ISBN 978-3-030-22475-2 (eBook), Springer Nature: Cham, Switzerland, 2020. https://doi.org/10.1007/978-3-030-22475-2

COSTA, C. C. R. *Desempenho de uma língua eletrônica impedimétrica com algoritmo de aprendizado de máquina na análise de águas residuais*. 2024. 93 f., il. Dissertação (Mestrado em Química) — Universidade de Brasília, Brasília, 2024. http://repositorio.unb.br/handle/10482/51010.

GALINARO, C. A.; SPADOTO, M.; AQUINO, F. W. B.; PELINSON, N. S.; VIEIRA, E. M. Environmental risk assessment of parabens in surface water from a Brazilian river: the case of Mogi Guaçu Basin, São Paulo State, under precipitation anomalies. *Environmental Science and Pollution Research*, v. 29, p. 8816-8830, 2022.

HOSSAIN, E. *Machine Learning Crash Course for Engineers*. Springer. 2024. 465p. https://doi.org/10.1007/978-3-031-46990-9

POUDEL, D; SHRESTHA, D; BHATTARAI, S; GHIMIRE, A. *Comparison of Machine Learning Algorithms in Statistically Imputed Water Potability Dataset*. 2022. https://www.researchgate.net/publication/362154236_Comparison_of_Machine_Learning_Algorithms_in_Statistically_Imputed_Water_Potability_Dataset

VANDERPLAS, Jake. *Python Data Science Handbook*: Essential Tools for Working with Data. 2023. O'Reilly Media, Inc., Gravenstein Highway North, Sebastopol, CA 95472. 591p.

ZIMMERMANN, C; GUIMARÃES, O; PERALTA-ZAMORA, P. *Avaliação da qualidade do corpo hídrico do rio Tibagi na região de Ponta Grossa utilizando análise de componentes principais (PCA)*. 2008. https://doi.org/10.1590/S0100-40422008000700025.