UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS DA NATUREZA
CAMPUS LAGOA DO SINO

GABRIELLA BORGES CRISTOVAM

APRENDIZADO DE MÁQUINA SUPERVISIONADO E NÃO SUPERVISIONADO APLICADO AO MONITORAMENTO DA QUALIDADE DE ÁGUAS SUPERFICIAIS E SUBTERRÂNEAS NA BACIA DO RIO MOGI GUAÇU (UGRHI-9)

Buri - SP
2025

GABRIELLA BORGES CRISTOVAM

APRENDIZADO DE MÁQUINA SUPERVISIONADO E NÃO SUPERVISIONADO APLICADO AO MONITORAMENTO DA QUALIDADE DE ÁGUAS SUPERFICIAIS E SUBTERRÂNEAS NA BACIA DO RIO MOGI GUAÇU (UGRHI-9)

Trabalho de conclusão de curso apresentado ao Centro de Ciências da Natureza da Universidade Federal de São Carlos, para obtenção do título de Bacharel em Engenharia Ambiental.

Orientadora: Natalia de Souza Pelinson
Coorientadora: Franciane Mendonça dos Santos

Buri - SP
2025

Cristovam, Gabriella Borges

Aprendizado de máquina supervisionado e não supervisionado aplicado ao monitoramento da qualidade de águas superficiais e subterrâneas na bacia do Rio Mogi Guaçu (UGRHI-9) / Gabriella Borges Cristovam -- 2025.
38f.

TCC (Graduação) - Universidade Federal de São Carlos, campus Lagoa do Sino, Buri
Orientador (a): Natalia de Souza Pelinson
Banca Examinadora: Cláudia Marisse dos Santos Rotta, Jamil Alexandre Ayach Anache
Bibliografia

1. Monitoramento ambiental de águas. 2. Modelos estatísticos. 3. Predição . I. Cristovam, Gabriella Borges. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Lissandra Pinhatelli de Britto - CRB/8 7539

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS DA NATUREZA
CURSO DE GRADUAÇÃO EM ENGENHARIA AMBIENTAL

FOLHA DE APROVAÇÃO

Assinatura dos membros da comissão examinadora que avaliou e aprovou a Defesa de Trabalho de Conclusão de Curso da candidata Gabriella Borges Cristovam, realizada em 22/01/2025.

Documento assinado digitalmente
gov.br NATALIA DE SOUZA PELINSON
Data: 22/01/2025 23:16:38-0300
Verifique em https://validar.iti.gov.br

Dra. Natália de Souza Pelinson - Professora Orientadora
Centro de Ciências da Natureza – UFSCar – Campus Lagoa do Sino.

Documento assinado digitalmente
gov.br CLAUDIA MARISSE DOS SANTOS ROTTA
Data: 30/01/2025 16:44:54-0300
Verifique em https://validar.iti.gov.br

Dra. Cláudia Marisse dos Santos Rotta
Centro de Ciências da Natureza – UFSCar – Campus Lagoa do Sino.

Documento assinado digitalmente
gov.br JAMIL ALEXANDRE AYACH ANACHE
Data: 31/01/2025 10:01:29-0300
Verifique em https://validar.iti.gov.br

Dr. Jamil Alexandre Ayach Anache
Escola de Engenharia de São Carlos – Universidade de São Paulo

DEDICATÓRIA

Aos que fizeram de tudo para que eu chegasse até aqui, mesmo em meio a tantas dificuldades, meus pais.

AGRADECIMENTO

Agradeço primeiramente aos meus pais que abdicaram de muitos desejos de vida para que a minha saúde e educação fosse possível. Aos meus irmãos, Fernando e Heloisa, por acreditaram no meu potencial de maneira incondicional e vibrarem comigo todas as conquistas. Agradeço aos meus avós, Isabel, Angelina e José, que passaram por tantas dificuldades e abriram caminhos para que suas famílias pudessem ter uma vida mais digna e, ainda, fazem questão de demonstrarem um dos amores mais genuínos. Aos meus tios e primos, que me ajudaram em muitos momentos, de maneiras diferentes, e me fizeram sempre me sentir especial.

Agradeço àqueles que estão junto comigo desde os primeiros dias da faculdade, Giovanna Andrade, Larissa Lima e Igor Torres, que fizeram essa trajetória ser muito mais leve, engraçada e viveram comigo as melhores histórias. Além disso, agradeço também às meninas que foram como um lar pra mim e que tive o privilégio de dividir o espaço e histórias durante os últimos anos de faculdade, Milena Andrade, Gabriela Bonini, Priscila Mistro e Beatriz Oliveira. Não poderia deixar de fazer um agradecimento especial também para o meu namorado, Igor, que foi meu melhor amigo desde o início da faculdade, dividindo muitas histórias, conquistas, anseios e risadas, e que continua ocupando um dos lugares mais especiais em minha vida.

À minha orientadora, Natalia de Souza Pelinson, e coorientadora, Franciane Mendonça dos Santos, meu imenso agradecimento pela paciência, confiança, apoio e colaboração nesta etapa tão desafiadora. Vocês foram essenciais para que eu pudesse concluir essa fase tão importante em minha vida.

RESUMO

Cristovam, G. B. **Aprendizado de máquina supervisionado e não supervisionado aplicado ao monitoramento da qualidade de águas superficiais e subterrâneas na bacia do Rio Mogi Guaçu (UGRHI-9)**. Trabalho de Conclusão de Curso - Bacharelado em Engenharia Ambiental. Universidade Federal de São Carlos. Campus Lagoa do Sino. Buri, SP. 2025. 38 p.

O monitoramento ambiental de águas superficiais e subterrâneas é fundamental para que possam ser preservadas a qualidade da água e a saúde de diferentes ecossistemas. Considerando as dificuldades no processo de monitoramento, que é altamente demorado e custoso, este trabalho tem como objetivo integrar ferramentas de ciência de dados, particularmente métodos de aprendizado de máquina (*"Machine Learning"* - ML), ao monitoramento da qualidade de águas naturais com base em dados históricos da Companhia Ambiental do Estado de São Paulo (CETESB). Por meio do aprendizado supervisionado (Regressão Linear Múltipla e *Random Forest*), os modelos foram treinados para prever o IQA (Índice de Qualidade da Água), obtendo-se ótimos resultados de teste para ambos os modelos (Regressão com R^2 de 0.97 e RMSE de 2.37 e *Random Forest* com R^2 de 0.92 e RMSE de 0.06), permitindo então a predição da qualidade da água superficial futuramente. De forma análoga, o aprendizado não supervisionado (PCA, K-Means e DBSCAN) foi aplicado com o intuito de detectar padrões, analisar correlações e reduzir a dimensionalidade das variáveis, na qual poderia ter gerado melhores resultados se houvesse mais dados disponíveis. Porém, de modo geral, os algoritmos apresentaram resultados satisfatórios que podem permitir a redução de tempo e custos relacionados ao processo de monitoramento da qualidade das águas.

Palavras-chave: modelos estatísticos; monitoramento ambiental de águas; predição.

ABSTRACT

Cristovam, G. B. **Supervised and unsupervised machine learning applied to the monitoring of surface and groundwater quality in the Mogi Guaçu River Basin (UGRHI-9)**. Undergraduate Thesis - Bachelor's Degree in Environmental Engineering. Federal University of São Carlos. Lagoa do Sino Campus. Buri, SP. 2025. 38 p.

Environmental monitoring of surface and groundwater is fundamental to preserving water quality and the health of different ecosystems. Considering the difficulties in the monitoring process, which is highly time-consuming and costly, this work aims to integrate data science tools, particularly machine learning (ML) methods, into natural water quality monitoring based on historical data from the São Paulo State Environmental Company (CETESB). By means of supervised learning (Multiple Linear Regression and Random Forest), the models were trained to predict the IQA (Water Quality Index), obtaining excellent test results for both models (Regression with R^2 of 0.97 and RMSE of 2.37 and Random Forest with R^2 of 0.92 and RMSE of 0.06), thus allowing the prediction of surface water quality in the future. Similarly, unsupervised learning (PCA, K-Means and DBSCAN) was applied in order to detect patterns, analyze correlations and reduce the dimensionality of the variables, which could have generated better results if more data had been available. In general, however, the algorithms provided satisfactory results that could reduce the time and costs involved in monitoring water quality.

Keyword: statistical models; environmental water monitoring; prediction.

SUMÁRIO

1 INTRODUÇÃO

Com a intensificação das atividades antrópicas, têm sido percebidas variações nos tipos e concentrações de substâncias químicas nas águas naturais, tanto superficiais quanto subterrâneas, o que pode causar alteração, em especial, na qualidade da água e gerar impactos negativos no potencial uso deste recurso natural. Os contaminantes provenientes da falta de estrutura de Unidades de Saneamento e os lançamentos irregulares, no solo ou em corpos d'água, podem deteriorar os recursos naturais, resultando na contaminação ambiental por matéria orgânica, microrganismos patogênicos, nutrientes e outros compostos que não sejam absorvidos pelo corpo humano. Desta forma, o despejo de esgoto ainda pode figurar uma fonte de poluição uma vez que está relacionada a diversas doenças para a população, além de poder gerar problemas de desoxigenação nos corpos d'água receptores (pela adição de matéria orgânica) e desencadear a eutrofização de corpos hídricos superficiais (devido ao excesso de nutrientes) (Reis *et al.*, 2023). Neste âmbito, promover um melhor entendimento da poluição é fundamental para que intervenções possam ser sugeridas no contexto de políticas públicas e ações a partir de comitês de regulação e gerenciamento.

O monitoramento da qualidade de águas superficiais deve ser estruturado utilizando as bases de enquadramento de classes da Resolução CONAMA n° 357 (CONAMA, 2005). Simplificadamente, a rede de coleta de informações deve ser estabelecida respeitando-se as variações sazonais e, portanto, no mínimo duas coletas anuais são necessárias: uma campanha em períodos chuvosos e uma em períodos de seca. Além disso, o monitoramento da qualidade de meios físicos deve ser realizado com a qualidade necessária para se estabelecer observações espaço-temporais, com campanhas realizadas nos mesmos meses em todos os anos. Adicionalmente, dados de monitoramento pluviométricos e fluviométricos podem também ser correlacionados à qualidade de águas doces, uma vez que anomalias nos padrões de precipitação ou variações sazonais podem causar alteração nas concentrações de substâncias de interesse (Galinaro *et al*., 2022).

Uma das principais relações entre os compartimentos do Saneamento Básico está pautada na mistura das águas precipitadas com os esgotos sanitários, ou ainda as águas de drenagem urbana, com águas de corpos hídricos que poderiam, e muitas vezes são utilizados como fontes para abastecimento humano. Mesmo que relacionado ao manejo de água subterrânea, tais processos poluidores podem acarretar também em perda da qualidade de solos e da água de um corpo d’água superficial.

No Brasil, os órgãos de fiscalização da qualidade da água podem fiscalizar de forma complementar e/ou concorrente, levando-se em consideração as necessidades da Unidade de Gerenciamento de Recursos Hídricos (UGRHI). A UGRHI é a unidade básica de gerenciamento estabelecida pela Lei das Águas instituída pela Lei Federal nº 9.433/1997 (BRASIL, 1997). Desta forma, há monitoramentos com relatórios publicados regularmente tanto pela Companhia Ambiental do Estado de São Paulo (CETESB), quanto pela Agência Nacional de Águas e Saneamento (ANA). Ambos os relatórios são divulgados publicamente e o monitoramento se vale de índices para determinação da qualidade da água.

Ainda, complementarmente à aplicação de ensaios laboratoriais para construção de bases de dados sobre a qualidade da água de rios, córregos e reservatórios, ferramentas da ciência de dados têm sido exploradas para avaliação e predição de cenários em inúmeras áreas de pesquisa em ciências naturais, incluindo as ciências ambientais. Dentre estas ferramentas de estatística aplicadas, pode-se observar crescentes estudos sobre análises de qualidade de águas superficiais e subterrâneas baseados em aprendizado de máquina (ML - do inglês *"Machine Learning"*).

Os algoritmos de aprendizado de máquina podem ser supervisionados ou não supervisionados. O aprendizado supervisionado, em geral, reflete a capacidade de um algoritmo de generalizar o conhecimento dos dados disponíveis com casos rotulados, de modo que o método possa ser utilizado para prever novos casos (não rotulados) (Alloghani *et al.*, 2020). O aprendizado não-supervisionado, por outro lado, faz referência ao processo de agrupar dados em clusters usando métodos ou algoritmos automatizados em dados que não foram classificados ou categorizados anteriormente (Alloghani *et al.*, 2020). Nessa segunda situação, os algoritmos devem "aprender" os relacionamentos ou recursos subjacentes dos dados disponíveis e agrupar casos com recursos ou características semelhantes.

Nesse contexto, o presente trabalho se propõe a estabelecer uma análise da qualidade de águas naturais (superficiais e subterrâneas) utilizando os principais métodos de aprendizado de máquina supervisionados e não supervisionados, realizando um estudo de caso para a Unidade de Gerenciamento de Recursos Hídricos Mogi Guaçu (UGRHI-9). Por meio do aprendizado não supervisionado (PCA, K-Means e DBSCAN), foi detectado padrões e anomalias nos dados de qualidade da água, enquanto no aprendizado supervisionado (Regressão Linear Múltipla e *Random Forest*), os modelos foram treinados para prever o IQA (Índice de Qualidade da Água) com base em dados históricos da Companhia Ambiental do Estado de São Paulo (CETESB).

2 OBJETIVOS

2.1 Objetivo principal

O objetivo principal desta pesquisa foi avaliar os principais métodos de *Machine Learning* e sua aplicabilidade no monitoramento da qualidade das águas superficiais e subterrâneas, realizando um estudo de caso para a Unidade de Gerenciamento de Recursos Hídricos Mogi Guaçu (UGRHI-09).

2.2 Objetivos específicos

- Analisar e comparar o desempenho de diferentes métodos de aprendizado de máquinas para a construção de índices de qualidade da água e;
- Avaliar a possibilidade de seleção de parâmetros indicadores físico, químicos e biológicos para o monitoramento utilizando o banco de dados InfoAguas da Companhia Ambiental do Estado de São Paulo (CETESB).

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Monitoramento da qualidade da água superficial e subterrânea

É de conhecimento comum que a água é um recurso essencial para vida humana, possuindo importância para todo o ecossistema e diversas atividades. É entendido como águas superficiais, segundo a Resolução CONAMA n° 357/2005, águas doces, salobras, salinas, presentes em ambiente lêntico (lagos, reservatórios, entre outros) ou em ambiente lótico (como rios e córregos). As águas subterrâneas, por sua vez, são os recursos hídricos armazenados nas camadas de subsolo, responsáveis pela manutenção dos sistemas aquáticos superficiais (CETESB, 2022). O monitoramento ambiental da qualidade das águas superficiais e subterrâneas, portanto, é uma ferramenta fundamental para a geração de informações relevantes para diagnóstico, gestão, adoção de medidas preventivas e implementação de políticas públicas que visem a conservação desse valioso recurso (CETESB, 2022).

A Política Estadual de Recursos Hídricos - Lei Estadual nº 7.663, adota as bacias hidrográficas como unidade de gestão e planejamento, nas quais estas dividem-se hidrograficamente pelo estado em 22 Unidades de Gerenciamento de Recursos Hídricos. A CETESB possui programa de monitoramento das águas subterrâneas e superficiais que contemplam as 22 UGRHIs do estado de São Paulo, que são realizados por meio de um fluxo de trabalho extenso e altamente custoso, envolvendo campanhas semestrais de coleta de

amostras, transporte, análises laboratoriais, geração de resultados (destinados ao banco de dados de monitoramento da CETESB), validação dos resultados, cálculo dos índices, diagnóstico e, finalmente, elaboração e publicação dos relatórios (CETESB, 2023).

O Índice de Qualidade da Água (IQA) adotado pela CETESB é oficialmente utilizado para monitoramento da qualidade das águas interiores (doces) no Estado de São Paulo e torna possível que sejam calculados valores comparativos considerando nove parâmetros de qualidade da água, que são: pH, coliformes totais, Demanda Bioquímica de Oxigênio (DBO), nitrogênio total, fósforo total, temperatura da água, turbidez, resíduo total (sólidos) e Oxigênio Dissolvido (OD). Para cada parâmetro foi desenvolvida uma curva de variação e para o cálculo é utilizada a Equação 1.

$$IQA = \prod_{i=1}^{n} q_i^{w_i} \qquad \text{(Equação 1)}$$

Em que:

IQA = índice de qualidade da água, é uma pontuação que indica a qualidade da água, expressa como um número entre 0 e 100;

qi é a concentração ou leitura ajustada da "curva média de variação de qualidade", valores entre 0 a 100;

wi representa o peso atribuído ao parâmetro em função da sua importância para o valor final de qualidade, então a soma total do peso de cada parâmetro é 1 (ou seja, 100%);

n é o número de variáveis que entram no cálculo de IQA, ou seja, nove variáveis quantitativas.

É importante observar que o cálculo de IQA pode ser realizado, a rigor, apenas se os nove parâmetros forem quantificados naquele mesmo ponto do corpo hídrico. Pesquisas sobre análise da qualidade de águas baseada em aprendizado de máquina para realizar predições de suas tendências futuras estão cada vez mais crescentes, como por exemplo, Silva *et al.* (2023) que desenvolveram modelos preditivos e a composição de um IQA adaptado da CETESB para avaliar a qualidade da água em reservatórios no estado de São Paulo sem a necessidade de campanhas de amostragem ou análises laboratoriais caras e demoradas. Para garantir um monitoramento preciso da qualidade da água, Silva *et al.* (2023) recomendaram continuar os métodos de monitoramento, embora com uma frequência possivelmente reduzida e puderem prover um método de predição com a utilização de ferramentas de aprendizado de máquina simplificada, tais como a regressão.

3.2 Utilização de aprendizado de máquina como ferramenta para monitoramento ambiental de águas naturais

O *Machine Learning* (ML), ou Aprendizado de Máquina, é um ramo da Inteligência Artificial que permite que computadores e máquinas aprendam com as informações existentes e apliquem esse aprendizado para executar outras tarefas semelhantes, isto é, a máquina explora e aprende os padrões, tendências ou características significativas dos dados anteriores e realiza uma previsão sobre os novos dados, sem que seja necessária uma programação explícita (Hossain, 2024). Em uma abordagem antiga, Konal (1999) explorou a relação do aprendizado de máquina com a capacidade humana de raciocinar e aprender, uma vez que o modo de capacitar a máquina se assemelha a perspectiva comportamental da cognição de um ser humano, que possui capacidade de melhorar seu desempenho ao executar tarefas semelhantes.

As técnicas de *Machine Learning* são comumente divididas em aprendizado supervisionado e aprendizado não supervisionado. O aprendizado supervisionado caracteriza-se por modelar a relação entre características dos dados e os rótulos associados a esses dados, isto é, contêm tanto os dados de entrada quanto as saídas. Assim, uma vez que o modelo é determinado, ele pode ser usado para aplicar rótulos a novos dados desconhecidos, realizando previsões ou classificações com base nos dados rotulados (Vanderplas, 2023). O aprendizado não supervisionado, por sua vez, não possui nenhum rótulo associado aos conjuntos de dados e, dessa maneira, o algoritmo é capaz de identificar padrões por meio do mapeamento de características dos dados fornecidos. Ou seja, dado um conjunto de dados, o algoritmo realiza uma categorização com base nas características e, assim, fornece uma saída - na qual cada categoria é chamada de cluster (Akshay *et al.*, 2024).

No cenário ambiental, é possível observar diferentes estudos relacionados à aplicação de ML para análise de problemas complexos e identificação de tendências nos dados de monitoramento. Um estudo realizado por Pereira (2023), explorou o monitoramento da qualidade da água em uma Estação de Tratamento de Águas Residuais (ETAR ou ETE – Estação de tratamento de esgoto) e a previsão das substâncias por meio de modelos de aprendizado de máquina, obtendo-se resultados satisfatórios na identificação de padrões, tendências e relações não lineares nos dados. Além disso, Silva (2019) utilizou bases de dados da Companhia Ambiental do Estado de São Paulo (CETESB) e da *United State Geological Service* (USGS) para realizar a predição de clorofila-a em corpos hídricos, importante parâmetro indicador de processo de eutrofização, aplicando modelos de Redes Neurais Artificiais e *Random Forest,* em que obteve boa acurácia e eficiência do modelo, representando uma boa alternativa para o monitoramento e redução de custos de campo.

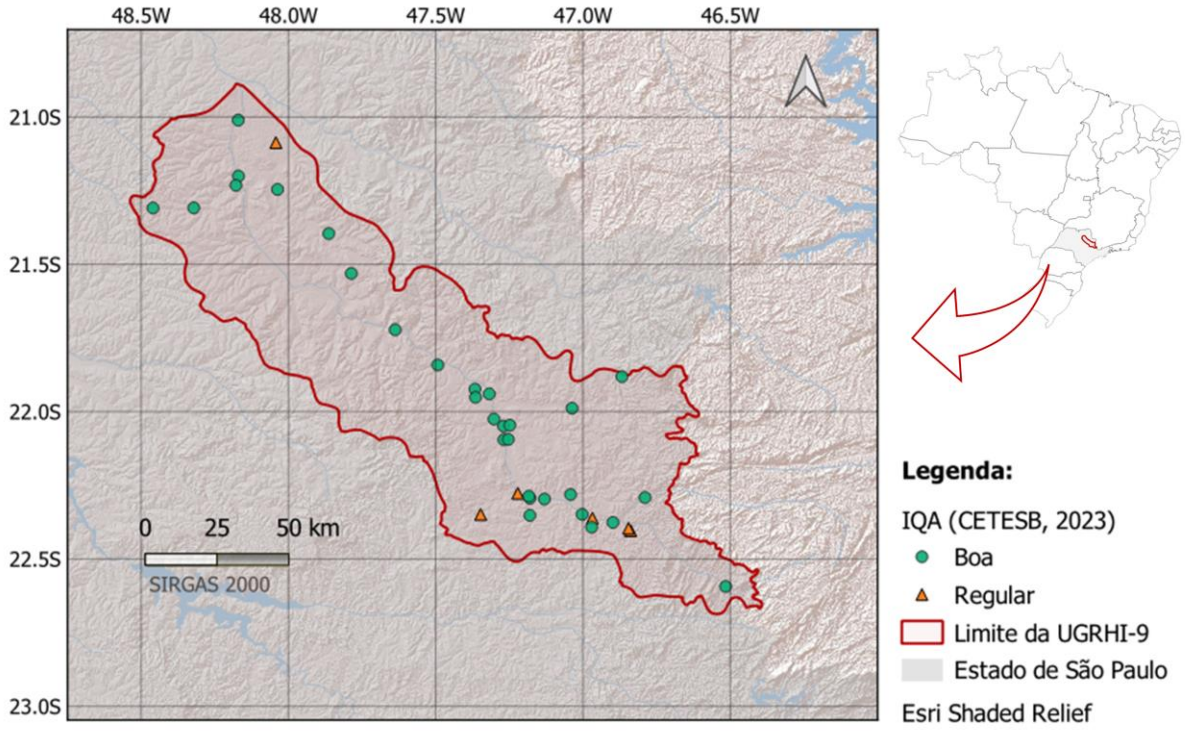
Ainda, na literatura é possível encontrar diversos estudos que utilizam o conjunto de dados *"water potability"* do repositório Kaggle, com o objetivo de classificar a potabilidade da água por meio da aplicação de diferentes bibliotecas e algoritmos de aprendizado de máquina. Poudel et. al (2022), por exemplo, utilizou os algoritmos de Regressão Logística (LR - do inglês *"Logistic Regression"*), k-NN *(k-Nearest Neighbors)*, Rede Neural Artificial (ANN - do inglês *"Artificial Neural Network"*) e *Random Forest* (RF) para prever a potabilidade da água e analisar a eficiência de cada um deles, na qual obteve como resultado o melhor desempenho advindo do algoritmo de *Random Forest* (acurácia de 70,4%). De maneira adicional, Abraham *et al.* (2022) aplicou modelos supervisionados e não supervisionados por meio de diferentes algoritmos, como KNN, LR, Árvore de Decisão e RF, obtendo como resultado acurácia de 65% para o *Random Forest* (melhor desempenho).

4 MATERIAL E MÉTODOS

4.1 Área de estudo

A área de estudo desse projeto é a Unidade de Gerenciamento de Recursos Hídricos Mogi Guaçu (UGRHI-9) localizada no estado de São Paulo, Brasil (Figura 1). A UGRHI-9 é uma bacia relevante pela sua complexidade de ocupação próxima ao Rio Mogi Guaçu. O Mogi Guaçu é um rio interestadual localizado em dois estados brasileiros, totalizando 473 km de extensão (377,5 km no estado de São Paulo e 38 km no estado de Minas Gerais). A UGRHI-9 abrange 43 municípios e uma população de mais de um milhão de habitantes e o uso e a ocupação do solo podem ser categorizados em culturas agrícolas e atividades industriais. Podemos destacar, ainda, que UGRHI-09 foi afetada pelas anomalias de precipitação dos anos de 2014 e 2015, sendo um exemplo de como a disponibilidade hídrica pode afetar negativamente incrementando o risco ambiental à biota aquática devido à qualidade da água (Galinaro *et al.*, 2022).

Figura 1 - Área de estudo da Unidade de Gerenciamento do rio de Mogi Guaçu localizada no estado de São Paulo, com a localização dos pontos de monitoramento da qualidade da água superficial realizado pela CETESB para obtenção do Índice de Qualidade da Água (IQA).



Fonte: Autoria própria. Dados coletados nas Plataformas de Infraestrutura Nacional de Dados Espaciais (INDE) e DataGeo (SP).

4.2 Ferramentas utilizadas para coleta e análise de dados

A coleta de dados de águas superficiais e subterrâneas foi realizada utilizando plataformas públicas oficiais e confiáveis utilizando-se, especialmente, o banco de dados InfoAguas e as Publicações e Relatórios da Companhia Ambiental do Estado de São Paulo (CETESB) no período de 2014 a 2023.

As etapas de manipulação dos dados para aplicação dos métodos de *Machine Learning* supervisionados e não supervisionados foram realizadas utilizando-se Linguagem Python para programação na ferramenta Google Colab (*on-line*), que é um serviço Jupyter Notebook hospedado adequado para aprendizado de máquina, ciência de dados e educação. Por fim, para o desenvolvimento e treinamento dos algoritmos de aprendizado de máquina, foi necessário utilizar diversas bibliotecas e frameworks com foco nas aplicações de ciência de dados.

4.2.1 Bibliotecas Python utilizadas para análises de dados ambientais

4.2.1.1 Biblioteca Pandas

Pandas é uma biblioteca de análise e manipulação de dados de código aberto que possui como objetivo ser de alto nível para análises práticas de dados do mundo real em Python. A biblioteca permite trabalhar com dados em diversos formatos (CSV, Excel, SQL, HDF5) de maneira eficiente, na qual fornece grande variedade de funcionalidades para manipulação e transformação dos dados, como a criação de *DataFrame*, tratamento de dados ausentes, remodelagem e pivotagem do conjunto de dados, mesclagem, indexação, agrupamento, entre outros (Pandas, 2024). No presente estudo foi utilizada, principalmente, na etapa de pré-processamento de dados nos modelos de *Machine Learning* aplicados.

4.2.1.2 Biblioteca Scikit-Learn

Scikit-Learn (também conhecida como “SKLearn”) é a principal biblioteca de código aberto para ML em Python, na qual suporta o aprendizado supervisionado e não supervisionado através de diversas ferramentas que facilitam o pré-processamento de dados e o ajuste, avaliação e seleção de modelos (Scikit-learn, 2024). No presente trabalho aplicou-se a SKLearn para redução de dimensionalidade (PCA), algoritmos de *clustering*, regressão e classificação, além das ferramentas para ajuste e avaliação dos modelos.

4.2.1.3 Biblioteca Numpy

Numpy (abreviação de *Numerical Python* em inglês) é uma biblioteca para computação científica em Python que oferece ampla variedade de funções matemáticas otimizadas para lógica, álgebra linear, operações estatísticas, entre outras funções que facilitam cálculos complexos (Numpy, 2024). De maneira geral, utilizou-se a biblioteca para armazenar e manipular os conjuntos de dados numéricos dos modelos de aprendizado de máquina do presente estudo.

4.2.1.4 Biblioteca Matplotlib

Matplotlib é uma das principais bibliotecas para visualização de dados em Python, em que é possível criar e personalizar gráficos (estáticos, animados e interativos) de maneira eficaz (Matplotlib, 2024). Com isso, a biblioteca foi utilizada para visualizar e analisar graficamente os resultados de processamento dos dados de ambos os modelos de ML.

4.2.1.5 Biblioteca Plotly

Plotly é uma biblioteca de código aberto que também é utilizada para visualização de dados, principalmente para análise de séries temporais (Nielsen, 2019). Com a biblioteca é possível criar e personalizar gráficos interativos com alta qualidade de publicação online. Em especial, no presente estudo, Plotly foi utilizado identificação de padrões e tendências por meio de gráficos de dispersão, de caixa, histogramas, mapas de calor e outros gráficos disponíveis para inteligência artificial e aprendizado de máquina.

4.2.1.6 Biblioteca Scipy

Scipy (abreviação de *Scientific Python* em inglês) é uma biblioteca de código aberto que se baseia no NumPy e complementa outras bibliotecas, como o Matplotlib. O SciPy disponibiliza algoritmos e estruturas de dados para diversas operações matemáticas em Python, como otimização, integração, interpolação, problemas de autovalor, álgebra linear, estatísticas, entre outras (Scipy, 2024). Diante disso, na atual metodologia de estudo a biblioteca foi utilizada em algumas aplicações de séries temporais e estatística.

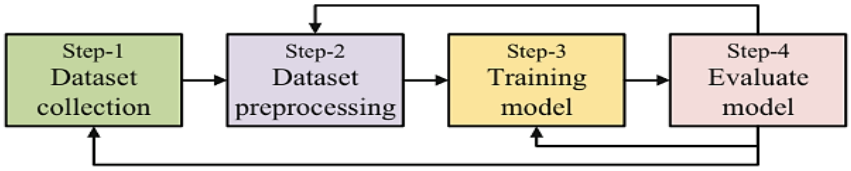
4.2.1.7 Framework Spark

O Apache Spark (comumente conhecido como “Spark”) é um *framework* para computação distribuída que dá suporte para diversas linguagens de programação (Python, Java, R, SQL e Scala) e é projetado para processar de grandes volumes de dados (Apache Spark, 2024). Para o presente trabalho foi utilizada a “Spark MLlib” (biblioteca do Spark para aplicações de Machine Learning) para desenvolver o modelo supervisionado *Random Forest.*

4.3 Metodologia de aprendizado de máquina a ser aplicado na análise dos dados de qualidade de água na UGRHI de Mogi Guaçu

A metodologia utilizada para aplicação dos algoritmos e modelos de *Machine Learning* foi baseada no fluxo de trabalho geral proposto por Hossain (2024) que contém quatro etapas principais: coleta de conjunto de dados, pré-processamento de dados, modelo de treinamento e avaliação do modelo (Figura 2). Diante disso, foi possível definir a metodologia do presente estudo objetivando a análise do aprendizado supervisionado e não supervisionado para o monitoramento da qualidade de água subterrânea e superficial na UGRHI-09 (Figura 3).

Figura 2 - Diagrama do fluxo de trabalho de Machine Learning simplificado

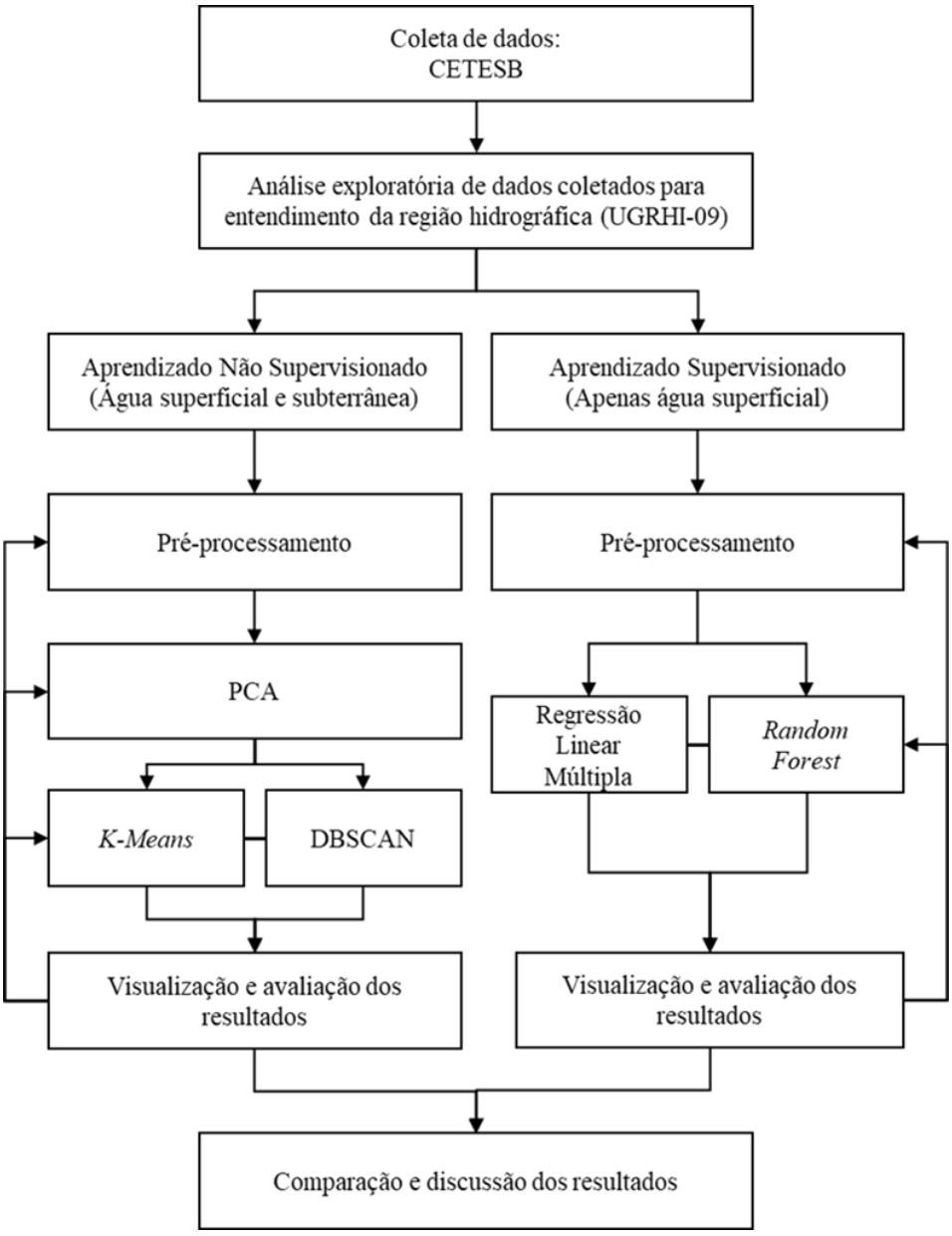


Fonte: Hossain (2024).

A coleta de dados de águas superficiais e subterrâneas foi realizada diretamente na Plataforma InfoAguas e nas Publicações e Relatórios da Companhia Ambiental do Estado de São Paulo (CETESB) no período de 2014 a 2023. Entretanto, dificuldades foram encontradas na obtenção de dados de água subterrânea no sistema do InfoAguas, na qual foi possível coletar apenas os dados referentes ao ano de 2022 e 2023. Os parâmetros considerados para água subterrânea foram: Cloreto (Cl^-); Cobalto Total (Co); Cobre Total (Cu); Condutividade (Campo) (µS/cm); Crômio Total (Cr); Cromo Hexavalente (Cr^{6+}); Dureza Total (mg/L $CaCO_3$); Alcalinidade Bicarbonato ($CaCO_3$); Alcalinidade Carbonato ($CaCO_3$); Alcalinidade Hidróxido ($CaCO_3$); Alumínio Total (Al); Antimônio Total (Sb); Arsênio Total (As); Bactérias Heterotróficas (UFC/mL); Bário Total (Ba); Boro Total (B); Cádmio Total (Cd); Cálcio Total (Ca); Carbono Orgânico Dissolvido (DOC) (mg/L); Chumbo Total (Pb); Cloreto.1 (Cl^-); Estanho Total (Sn); Estrôncio Total (Sr); Ferro Total (Fe); Fluoreto (F^-); Fósforo Total (P); Lítio Total (Li); Magnésio Total (Mg); Manganês Total (Mn); Molibdênio Total (Mo); Níquel Total (Ni); Nitrogênio Nitrato (NO_3^--N); Nitrogênio Nitrito (NO_2^--N); Nitrogênio Total (N); pH (Campo); Potássio Total (K); Prata Total (Ag); Selênio Total (Se); Sódio Total (Na); Sólidos Dissolvidos Totais (mg/L); Sulfato (SO_4^{2-}); Temperatura da Água (Campo) (°C); Titânio Total (Ti); Urânio Total (U); Vanádio Total (V); Zinco Total (Zn).

Em relação aos dados de água superficial, os parâmetros considerados foram as nove variáveis consideradas relevantes para a avaliação da qualidade das águas (incorporadas no Índice de Qualidade das Águas – IQA), que são: Oxigênio Dissolvido (OD), Demanda Bioquímica de Oxigênio ($DBO_{5,20}$), pH, *Escherichia coli* (em substituição aos coliformes termotolerantes), Temperatura da água, Nitrogênio total, Fósforo total, Turbidez e Sólido total (equivalente ao resíduo total).

Figura 3 - Diagrama da metodologia específica aplicada no presente estudo, incluindo coleta de dados, processamento dos dados, análises estatísticas e visualização para interpretação e avaliação dos resultados.



Fonte: Autoria própria.

Além disso, a média mensal de precipitação na UGRHI-09 no período de 2014 a 2023 também foi coletada com o intuito de compor as análises do estudo, tanto para o método supervisionado, como para o não supervisionado. Ao final da coleta, a base de dados de água superficial totalizou 1783 linhas de dados, comparativamente a 65 referente à água subterrânea (decorrente da dificuldade encontrada na obtenção dos dados no sistema InfoAguas).

A disponibilização dos algoritmos, descritos a seguir, encontra-se hospedada no GitHub (Link: https://github.com/gabriellabc/monitoramento_agua_superficial_subterranea).

4.3.1 Métodos de aprendizado não supervisionado para análise dos dados ambientais

O aprendizado não supervisionado foi aplicado tanto para água subterrânea, como para água superficial com o objetivo de detectar padrões, analisar as correlações das variáveis e para a redução de dimensionalidade.

4.3.1.1 Pré-processamento de dados coletados acerca da qualidade da água superficial

A importação do conjunto de dados no Google Colab foi realizada utilizando-se arquivos em formato CSV de maneira separada para água subterrânea e superficial. Após a coleta de todos os conjuntos de dados necessários, as técnicas de pré-processamento iniciaram-se pela etapa de formatação dos conjuntos de dados (exclusão de colunas desnecessárias para análise e inclusão da coluna de média mensal de precipitação) e, em seguida, a verificação e limpeza de dados duplicados e inconsistentes, além da conversão na representação dos dados: "*object"* para *"float"* nos valores de cada parâmetro. A Análise Exploratória dos Dados (AED) foi aplicada por meio de análise estatística descritivas dos valores dos parâmetros, criação de histogramas (para visualizar a distribuição dos dados), *boxplots* (para visualização dos outliers), média móvel (para visualizar a tendência dos dados) e matrizes de correlação de Pearson e Spearman (para medir a associação/correlação entre as variáveis).

Posteriormente, realizou-se a verificação e tratamento de dados desconhecidos/nulos utilizando o método de imputação pela mediana e o tratamento de dados discrepantes (outliers), na qual para água subterrânea foi aplicado a técnica estatística de Winsorização, uma vez que o conjunto de dados é pequeno e as técnicas de remoção de outliers eliminaria toda a variabilidade dos dados do conjunto e, para água superficial, a remoção desses valores foi feita através do método de Intervalo Interquartil - IQR, definindo limites superiores e inferiores para os valores discrepantes. Por fim, a padronização dos dados foi realizada utilizando-se a técnica "*StandardScaler"* da biblioteca *scikit-learn*, para que cada variável tivesse média igual a 0 e desvio padrão igual a 1, resultando em uma importância relativa comparável (Porto, 2024).

4.3.1.2 Análise de Componentes Principais para redução da dimensionalidade dos dados

A Análise de Componentes Principais (PCA - *Principal Component Analysis* em inglês) foi utilizada para reduzir a dimensionalidade dos conjuntos de dados, visando a seleção de variáveis que concentrem a maior parte da variabilidade dos dados, além da melhoria de desempenho dos algoritmos de aprendizado não supervisionado (uma vez que a PCA auxilia no posterior agrupamento e detecção de padrões).

Assim, utilizando principalmente as bibliotecas SKLearn, Numpy e Matplotlib, foi realizada a verificação da variância explicada dos componentes, matriz de componentes (loadings) e a correlação das variáveis originais e dos componentes principais.

4.3.1.3 Agrupamento K-Means (K-Means Clustering) aplicado a dados de córregos e rios

O K-Means foi utilizado como algoritmo de agrupamento para o aprendizado não supervisionado. Os dados utilizados no algoritmo foram os dados transformados pelo PCA. De maneira inicial, foi utilizado o método do cotovelo para determinar o número ideal de clusters (K), com ajuda da ferramenta *"KElbowVisualize"* para visualizar o número ótimo de clusters. No entanto, o algoritmo foi ajustado até encontrar um número de clusters que satisfizesse melhor as métricas *Silhouette Score, Calinski-Harabasz Index e Davies-Bouldin Index.*

4.3.1.4 Clusterização Espacial Baseada em Densidades de Aplicações de Ruídos - DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

O DBSCAN também foi utilizado como algoritmo de agrupamento para o aprendizado não supervisionado. Os dados utilizados no algoritmo foram os dados transformados pelo PCA. Foi necessário a realização de diversos ajustes dos parâmetros do algoritmo *(eps e min_sample)* para encontrar a configuração mais ideal aos dados. Para avaliação da qualidade do algoritmo, foram utilizadas as métricas *Silhouette Score, Calinski-Harabasz Index e Davies-Bouldin Index.*

4.3.2 Métodos de aprendizado supervisionado para análise de dados

O aprendizado supervisionado foi aplicado apenas para água superficial, uma vez que o objetivo foi desenvolver um modelo capaz de prever o IQA (Índice de Qualidade da Água).

4.3.2.1 Pré-processamento dos dados de água superficial

A importação do conjunto de dados no Google Colab foi realizada por meio de um arquivo em formato CSV. Após a coleta de todos os conjuntos de dados necessários, as técnicas de pré-processamento iniciaram-se pela formatação dos conjuntos de dados (exclusão de colunas desnecessárias para análise, pivotagem do DataFrame e inclusão da coluna de média mensal de precipitação) e, em seguida, a verificação e limpeza de dados duplicados e inconsistentes, além da conversão na representação dos dados: "*object"* para *"float"* nos valores de cada parâmetro. De maneira adicional, o cálculo do IQA foi realizado, com base na fórmula disponibilizada pela CETESB, acrescentando-se a coluna com os resultados do cálculo

para cada ponto de monitoramento. A Análise Exploratória dos Dados (AED) foi aplicada por meio da correlação de Pearson (para medir a associação/correlação entre as variáveis), criação de histogramas (para visualizar a distribuição da frequência da variável dependente e assimetria do gráfico), *boxplots* (para visualização o comportamento da variável dependente) e *pairplot* (para analisar dispersão das variáveis independentes). A transformação dos dados foi realizada apenas para a Regressão Linear Múltipla, uma vez que não é necessário o escalonamento dos dados para o *Random Forest* (Hossain, 2024). Assim, o método Log-Log foi utilizado nas variáveis dependente e independentes para melhorar as relações lineares e linearizar as relações não lineares para o modelo de Regressão Linear Múltipla. O pré-processamento foi finalizado dividindo os dados entre conjunto de treinamento e teste (utilizando proporção de 70:30 - 70% para treino e 30% para teste).

4.3.2.2 Regressão Linear Múltipla

Após as etapas de pré-processamento e preparação dos dados, foram realizados testes para escolha de variáveis adequadas para o modelo (a partir de aplicação dos testes F e t) e, assim, realizou-se a remoção de variáveis não estatisticamente significativas. Diante disso, foi possível seguir para a criação, ajuste do modelo e obtenção dos coeficientes de regressão linear. Para avaliação da qualidade do modelo foi utilizado o teste do R^2 (Coeficiente de Determinação) tanto para dados de treino, como para os dados de teste. Por fim, foi criada e analisada uma nova base de dados - com novos valores aleatórios dos parâmetros - para testar a previsão do modelo.

4.3.2.3 Algoritmo de análise 'Random Forest' aplicada aos dados de qualidade da água

O *Random Forest Regressor* foi implementado por meio do framework Spark e foi utilizado como segundo modelo supervisionado para prever o IQA. De maneira inicial, foi realizada a vetorização dos dados e a avaliação das variáveis estatisticamente significativas. Para a criação do modelo, considerou-se a profundidade da árvore *"maxDepth"* de 7 e o número de árvores de decisão *"numTrees"* de 10. Após os ajustes e previsões dos modelos, foi calculada as métricas de avaliação do modelo RMSE (Raiz do Erro Quadrático Médio) e R^2 (Coeficiente de Determinação). Por fim, foi criado e analisado um novo conjunto de dados - com novos valores aleatórios dos parâmetros - para testar a previsão do modelo.

5 RESULTADOS E DISCUSSÃO

5.1 Aprendizado não supervisionado para detecção de padrões na qualidade das águas

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).

A Análise de Componentes Principais (PCA) é uma técnica para redução de dimensionalidade dos dados, podendo ser considerada como uma ferramenta de compressão, na qual é esperado que um menor número de componentes principais consiga manter a maior variabilidade possível do conjunto de dados originais (Zimmermann *et al.*, 2008).

O algoritmo K-Means, adicionalmente, é definido como um algoritmo de agrupamento *(clustering)*, na qual busca dividir um conjunto de dados em K grupos distintos *(clusters)* e cada grupo é representado por um centroide (Nielsen, 2019). O algoritmo DBSCAN, por sua vez, é um algoritmo de agrupamento baseado em densidade, isto é, agrupa pontos que estão próximos uns dos outros através de medida de densidade, permitindo a identificação de clusters de forma arbitrária e a detecção de ruídos (Vanderplas, 2023).

5.1.1 Algoritmos aplicados para água subterrânea

As variáveis utilizadas foram descritas anteriormente (ver 4.3), com o acréscimo da variável de precipitação. De maneira exploratória inicial, visualizou-se as correlações utilizando-se o Coeficientes de *Pearson* (para avaliação da linearidade) e o Coeficientes de *Spearman* (avaliação da não-parametrização), conforme pode ser observado nas Figuras 4 e 5. Os valores mais próximos de 1 (azul) apresentam maior correlação positiva e valores mais próximos de -1 (vermelho) correlações mais negativas. Ainda, nota-se que as correlações foram maiores quando selecionado o coeficiente de *Spearman*, uma vez que ele mede a relação entre variáveis que podem não ser linearmente relacionadas (ao contrário de *Pearson* que assume apenas correlações lineares).

Com a aplicação da PCA, obteve-se oito componentes principais (PCs) que explicam 71,2% da variabilidade total dos dados originais, isto é, foi possível reduzir a dimensionalidade do conjunto de dados originais mantendo uma boa representatividade das informações contidas, podendo ser utilizado para detecção de padrões (Zimmermann *et al.*, 2008). Ainda, também foi identificada a variância dos dados explicada por cada componente principal, na qual os quatro primeiros componentes explicam a maior variabilidade dos dados, em que o PC1 explicou 16.27%, seguido de 14.14% (PC2), 11.76% (PC3) e 9.11%.

Figura 4 - Correlação de Pearson dos parâmetros físico-químicos de água subterrânea. Correlações positivas ficam evidenciadas em tons de azul, enquanto as correlações negativas são marcadas pelos tons de vermelho.

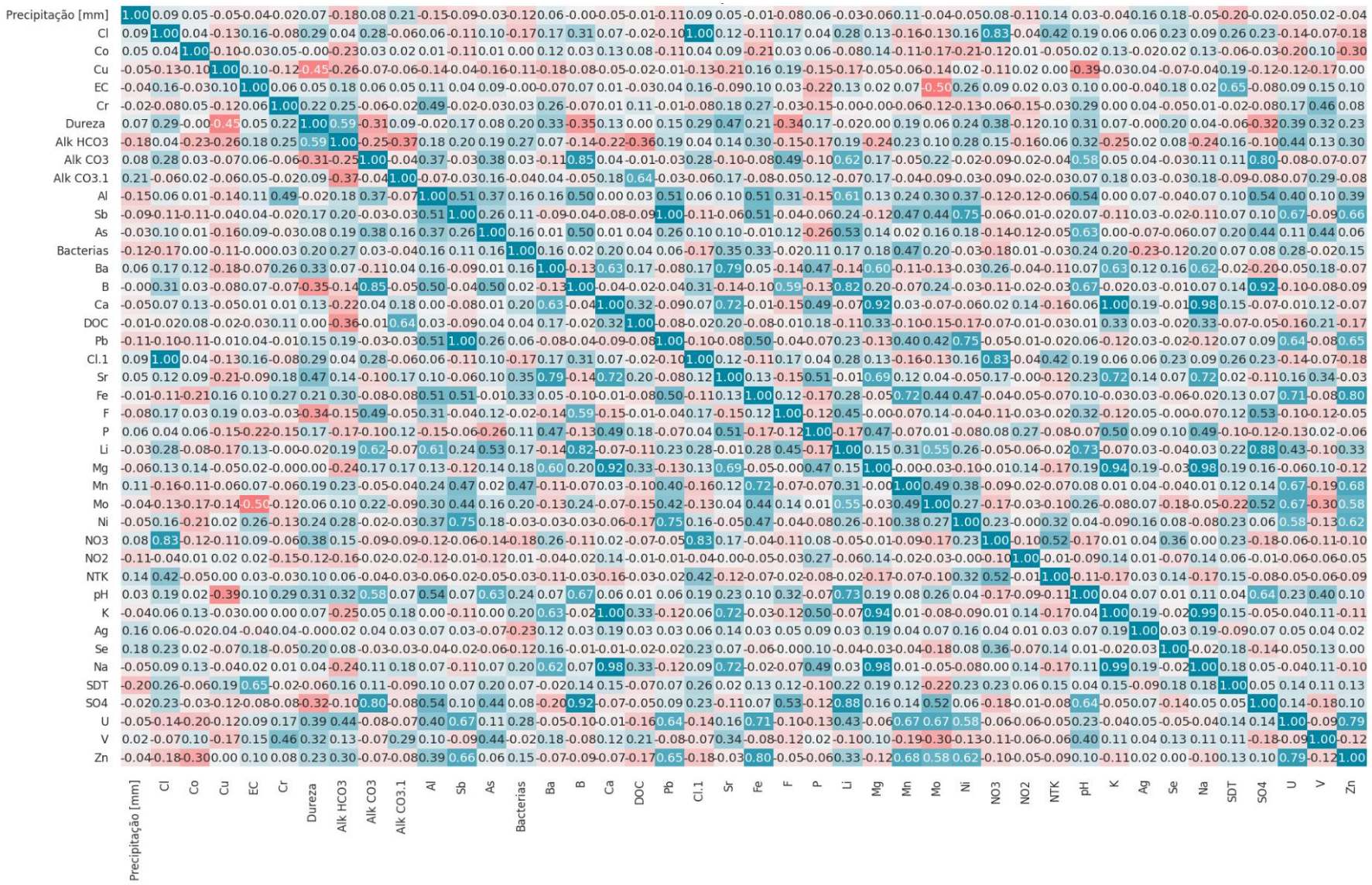
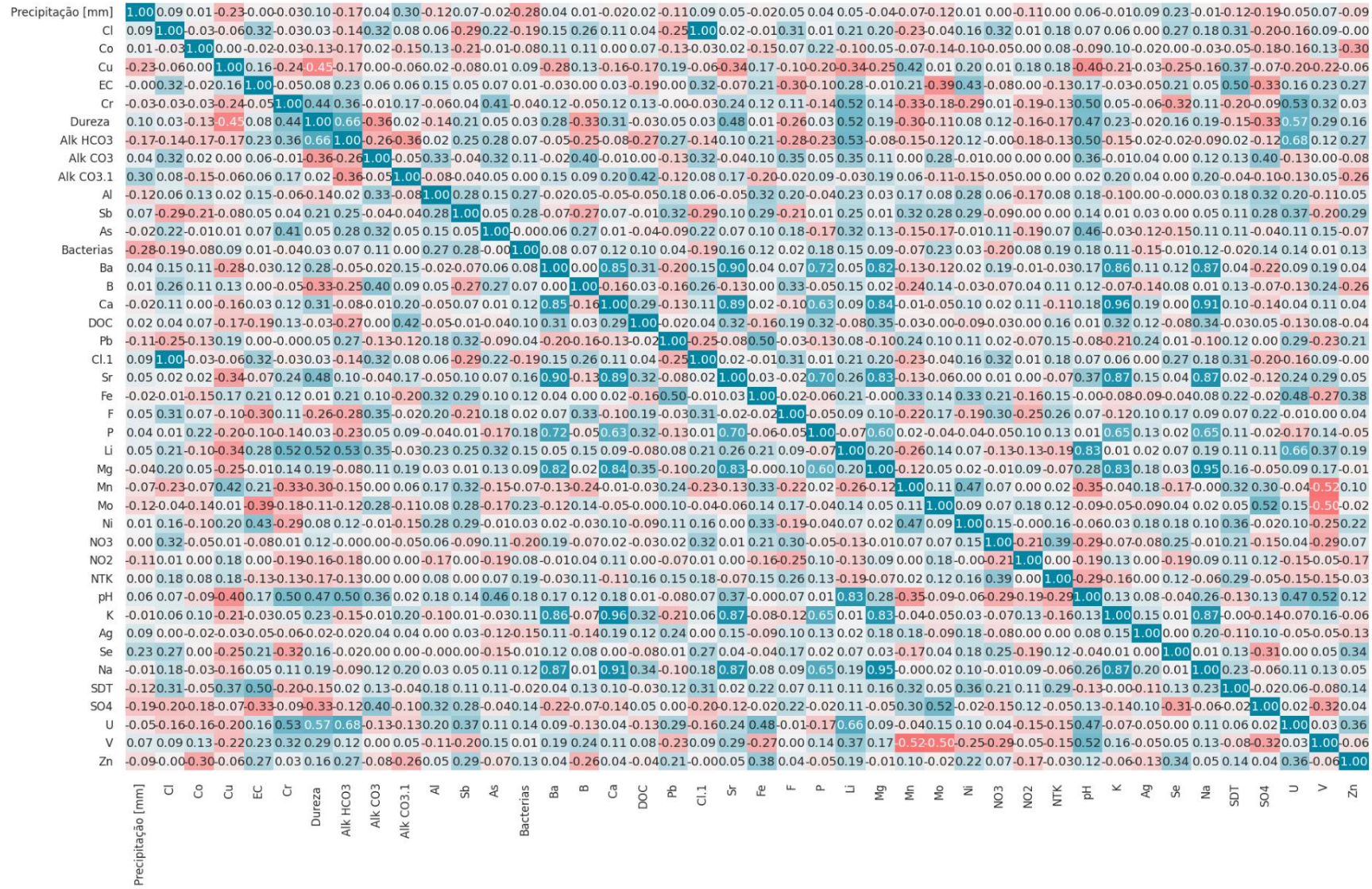


Figura 5 - Correlação de *Spearman* dos parâmetros físico-químicos de água subterrânea. Correlações positivas ficam evidenciadas em tons de azul, enquanto as correlações negativas são marcadas pelos tons de vermelho.

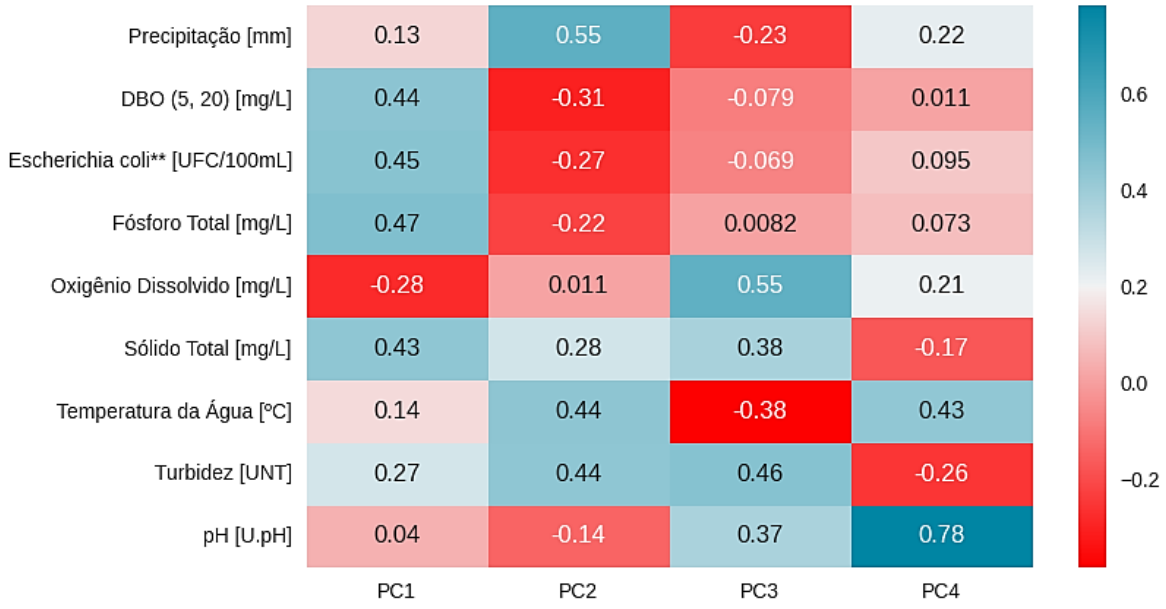


Os resultados de água subterrânea para o algoritmo *K-means* não foram satisfatórios, uma vez que a quantidade de dados disponíveis foi insuficiente e os dados se apresentaram pouco representativos para uma análise mais assertiva. As métricas de avaliação encontradas de *Silhouette Score*, *Calinski-Harabasz Index* e *Davies-Bouldin Index* foram, respectivamente: 0.30, 16.88 e 0.90, demonstrando que os clusters não foram bem definidos e houve presença de sobreposição. De modo similar, os resultados do algoritmo DBSCAN também não foram satisfatórios, provavelmente pelo mesmo motivo relacionado aos dados. Os ajustes no algoritmo e a divisão entre os clusters apresentou as métricas de avaliação de *Silhouette Score*, *Calinski-Harabasz Index* e *Davies-Bouldin Index*, respectivamente: 0.28, 10.11 e 0.89, resultando novamente em clusters não bem definidos e sobreposição.

5.1.2 Algoritmos aplicados para água superficial

A variável de nitrogênio total foi excluída das análises por apresentar muitos valores nulos. Com a aplicação da PCA, obteve-se quatro componentes principais (PCs) que explicam 80,74% da variabilidade total dos dados originais, isto é, foi possível reduzir a dimensionalidade do conjunto de dados originais mantendo boa representatividade das informações Figura 6).

Figura 6 - Matriz de componentes (*loadings*) do PCA em relação aos parâmetros de água superficial. Influências positivas ficam evidenciadas em tons de azul, enquanto as influências negativas são marcadas pelos tons de vermelho.



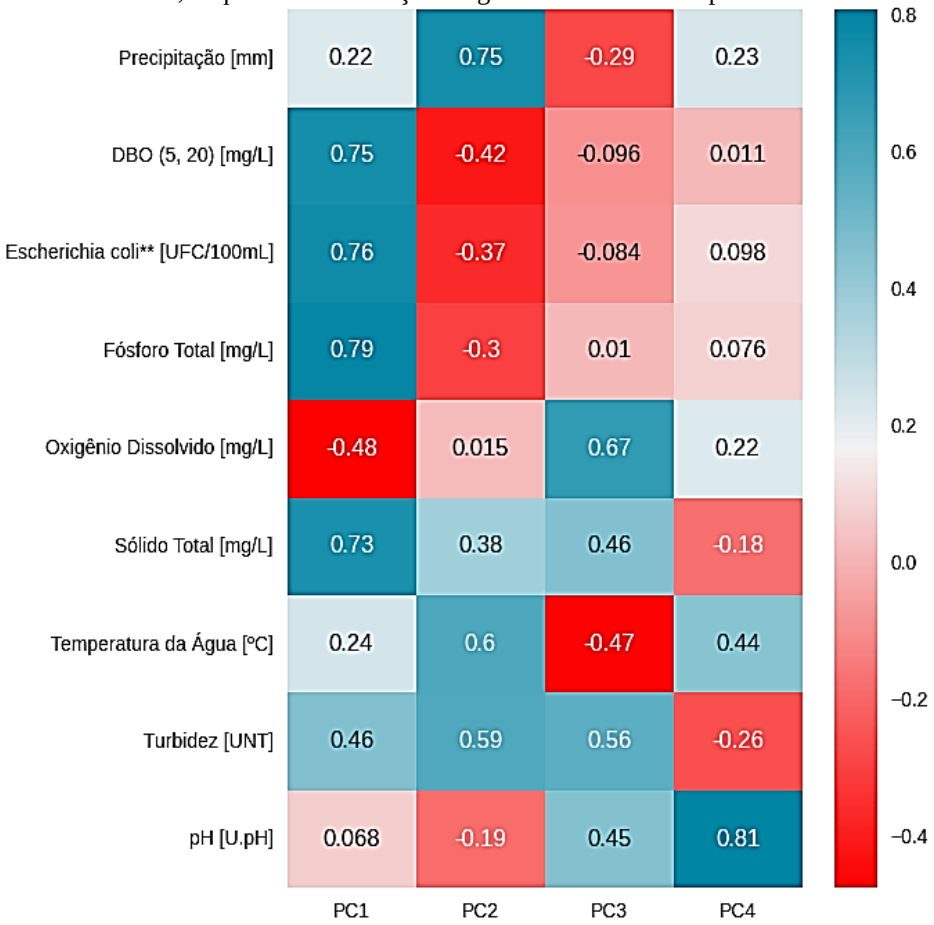
Fonte: Autoria própria.

Ainda, também foi identificada a variância dos dados explicada por cada componente principal, na qual o PC1 explicou 31,72% da variabilidade total dos dados, seguido de 20,59% (PC2), 16,63% (PC3) e 11,80% (PC4). Na Figura 6, está indicada a contribuição *(loading)* de cada variável original para a construção de cada componente principal, ou seja, indica quais variáveis são mais importantes para definir cada um deles (Hossain, 2024). Valores positivos ou negativos (próximos de 1 ou -1) representam forte influência no componente principal, enquanto valores próximos de 0 indicam baixa influência.

Por fim, visualizou-se a correlação (força e direção da relação linear) entre as variáveis originais e os scores dos componentes principais, como observado na Figura 7, sugerindo algumas interpretações. Importante ressaltar que valores positivos ou negativos (próximos de 1 ou -1) representam forte correlação com o componente principal, enquanto valores próximos de 0 indicam baixa correlação (Hossain, 2024).

O PC1 apresentou forte correlação positiva com as variáveis de fósforo Total, *Escherichia coli*, DBO e sólido Total, enquanto a maior correlação negativa foi com o Oxigênio Dissolvido, isto é, o PC1 conseguiu captar a variabilidade de dados relacionados à poluição da água, podendo estar associada aos impactos por esgoto doméstico e efluentes industriais, por exemplo (CETESB, 2023). Já o PC2 apresentou forte correlação positiva principalmente com a variável de precipitação, temperatura da água e turbidez, enquanto as correlações negativas foram com o DBO, *Escherichia coli*, fósforo total e pH, sugerindo a relação dinâmica entre os fatores climáticos e a qualidade da água, supondo por exemplo, que em eventos de alta precipitação a água tende a ter maior turbidez e menor concentração de poluentes. O PC3, por sua vez, apresentou forte correlação positiva principalmente com OD, turbidez e sólido total, enquanto a negativa foi correlacionada principalmente com a temperatura da água, sugerindo relações mais complexas (por exemplo processos de oxidação e suspensão de partículas) com a temperatura da água. O fator PC4, por fim, apresenta forte correlação positiva com o pH.

Figura 7 - Correlação entre as variáveis originais e os componentes principais (PCs). Correlações positivas ficam evidenciadas em tons de azul, enquanto as correlações negativas são marcadas pelos tons de vermelho.

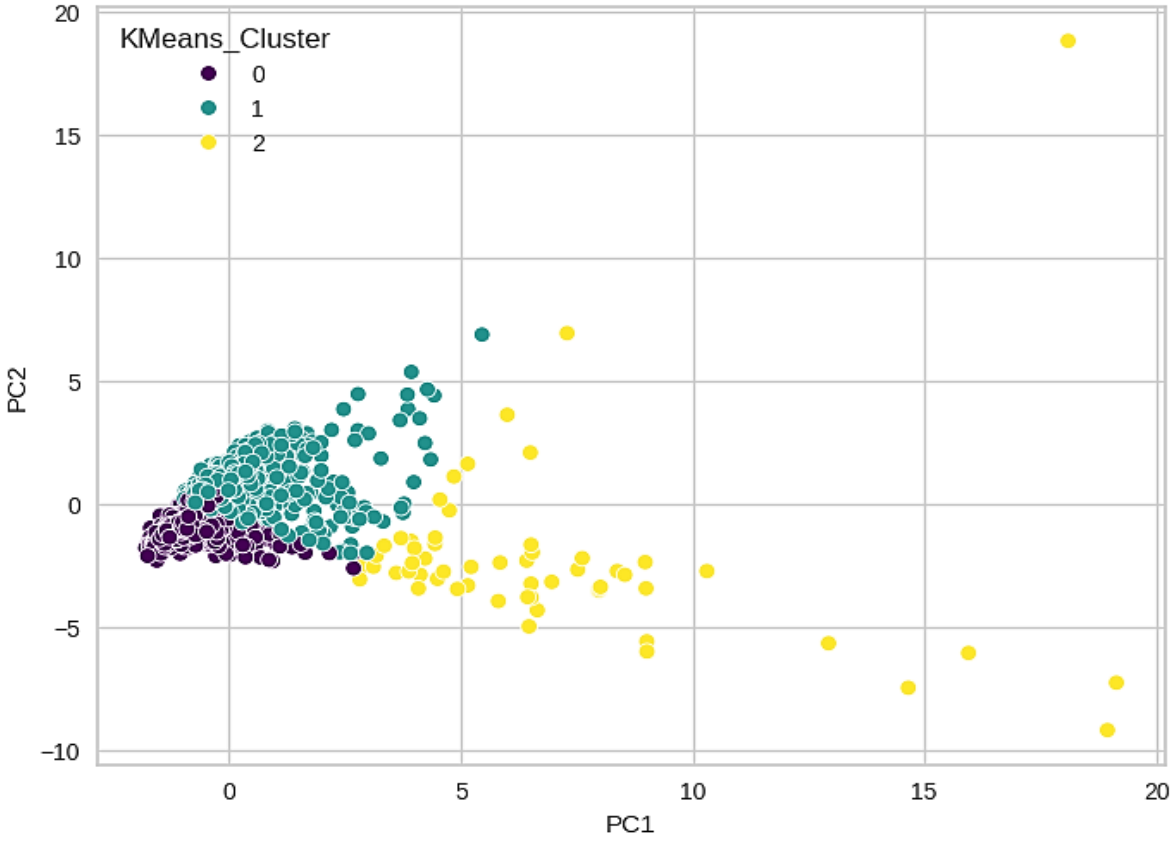


Fonte: Autoria própria.

Na Figura 8, há representação visual da análise dos dados dos agrupamentos pelo método K-Means nos dois primeiros componentes principais, PC1 e PC2, que são os componentes que capturam a maior variabilidade do conjunto de dados. A divisão entre 3 clusters apresentou as melhores métricas de avaliação encontrada de *Silhouette Score*, *Calinski-Harabasz Index* e *Davies-Bouldin Index* que mediram, respectivamente: 0.44, 855.51 e 0.86. No entanto, o resultado das métricas indicaram agrupamento razoável (relativamente eficaz) dos *clusters* e com sobreposições dos dados.

Complementarmente, os resultados obtidos com o algoritmo DBSCAN apresentaram diferenças aos identificados com o algoritmo K-Means. Os ajustes no algoritmo e a divisão entre 3 clusters apresentou as métricas de avaliação de *Silhouette Score*, *Calinski-Harabasz Index* e *Davies-Bouldin Index* que mediram, respectivamente: 0.68, 242.01 e 1.55. Além disso, o DBSCAN encontrou 28 ruídos nos dados, que podem ser interpretados como amostras com características atípicas.

Figura 8- Clusters K-Means entre o PC1 e PC2 (Água superficial)



Fonte: Autoria própria.

5.2 Aprendizado supervisionado para predição do índice de qualidade de água

O aprendizado supervisionado foi realizado por meio dos métodos de modelagem preditiva de Regressão Linear Múltipla e *Random Forest.* A regressão é capaz de modelar a relação entre uma variável dependente (ou resposta) e duas ou mais variáveis independentes (ou preditoras), assumindo relação linear entre as variáveis (Hossain, 2024). O algoritmo de *Random Forest*, no entanto, baseia-se em árvores de decisão, criando um conjunto de árvores de decisão (floresta) e, posteriormente, realiza a previsão a partir da média das previsões de todas as árvores (Vanderplas, 2023). Ainda, observa-se o algoritmo *Random Forest* sendo muito utilizado para avaliação e para propor alternativas de gestão para os recursos hídricos (Silva, 2019).

As variáveis independentes (preditoras) escolhidas para os modelos foram: Oxigênio Dissolvido (OD), Demanda Bioquímica de Oxigênio ($DBO_{5,20}$), pH, *Escherichia coli*, Temperatura da água, Nitrogênio total, Fósforo total, Turbidez, Sólido total e Precipitação, enquanto a variável dependente (resposta) foi o IQA. Inicialmente, para a Regressão Linear Múltipla as variáveis foram linearizadas mediante transformação logarítmica (Brusa, 2004) e, assim, foi realizada a avaliação estatística do modelo, na qual foi possível verificar quais eram as variáveis mais significativas estatisticamente (boas preditoras). Diante disso, removeu-se duas variáveis que não apresentavam significância estatística, temperatura da água e nitrogênio total (Tabela 1).

Tabela 1- Resultados do teste t nas constantes. Valores acima de 0.05 indicam as variáveis que não contribuem significativamente para o modelo.

Constante (variável)	Teste t (significância)
Precipitação	0.035
DBO (5, 20)	0.000
*Escherichia coli***	0.000
Fósforo Total	0.015
Nitrogênio Total	**0.245**
Oxigênio Dissolvido	0.000
Sólido Total	0.000
Temperatura da água	**0.186**
Turbidez	0.000
pH	0.000

Fonte: Autoria própria.

Posteriormente, realizou-se o treinamento do modelo e obteve como avaliação R^2 dos dados de treino e teste, respectivamente, 0.94 e 0.92, e ambos RMSE de 0.06, valores que indicam um ótimo ajuste do modelo aos dados de treinamento e boa generalização do modelo à dados novos, demonstrando ainda a capacidade do modelo de prever mesmo com a redução de variáveis. De maneira adicional, o bom ajuste do modelo de regressão também pôde ser observado através do gráfico de dispersão entre os valores previstos e reais, em que se observa boa correlação e concentração entre os pontos (Figura 9).

Figura 9- Gráfico de dispersão entre IQA previsto e real do modelo de Regressão Linear com transformação logarítmica.

Fonte: Autoria própria.

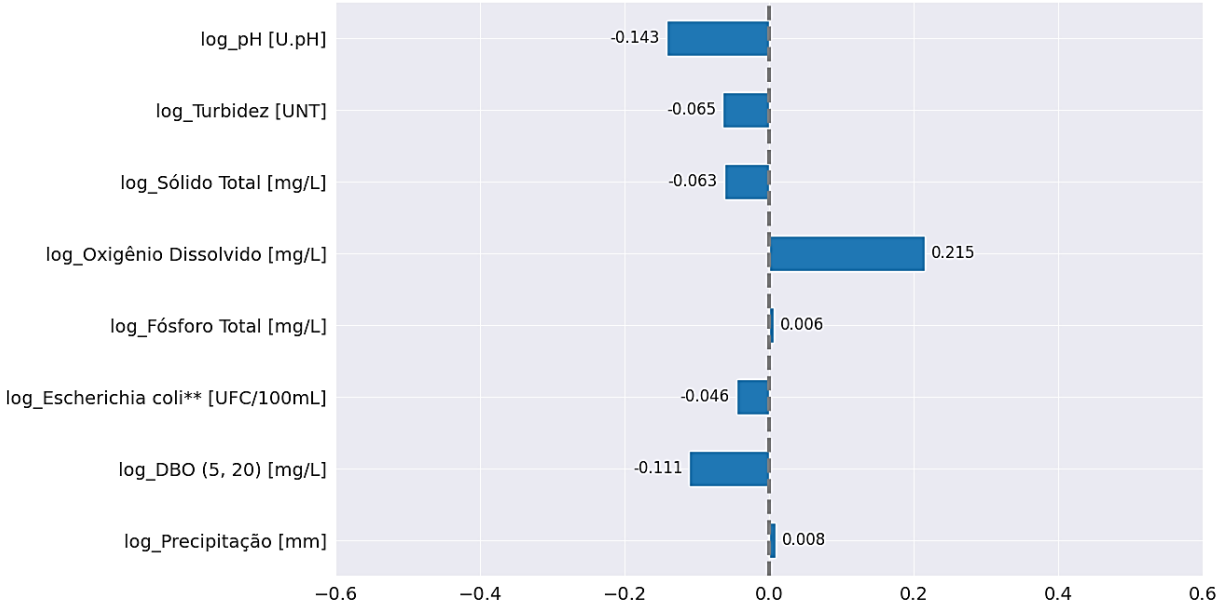
Além disso, foi analisada a influência das variáveis preditoras para a variável resposta (Figura 10), considerando que a determinação das variáveis de maior peso permite identificar os elementos que determinam a maior ou menor alteração da variável resposta (Arraes *et al.*, 2009). As relações positivas entre a variável preditora e a variável resposta é indicada pela barra à direita do eixo vertical, demonstrando que se houver um aumento na variável preditora, também haverá uma tendência de aumento na variável resposta (IQA). Analogamente, as relações negativas entre a variável preditora e a variável resposta é indicada pela barra à esquerda do eixo vertical, demonstrando que se houver um aumento na variável preditora, haverá uma tendência de diminuição na variável resposta (IQA).

Dessa forma, as relações que apresentaram relação positiva para o índice da qualidade da água foram observadas nos parâmetros de OD, precipitação e fósforo total, enquanto as negativas foram observadas nos parâmetros de pH, turbidez, sólido total, DBO e *Escherichia coli*. Importante ressaltar que a relação de influência é interpretada mantendo-se os valores de todas as outras variáveis explicativas iguais a zero (constantes), podendo ser interpretada, por exemplo, que um acréscimo de 1% (1 unidade no logaritmo) no Oxigênio Dissolvido gera, em média, um acréscimo de 0.215% (0.0215 unidades no logaritmo) no valor do IQA.

Do mesmo modo, um acréscimo das variáveis de fósforo total e precipitação geram, respectivamente, acréscimo de 0.006% e 0.008% no valor previsto de IQA e, de maneira oposta,

um acréscimo isolado de 1% de pH, turbidez, sólido total, DBO e *Escherichia coli* geram, respectivamente, uma diminuição de 0.143%, 0.065%, 0.063%, 0.111% e 0.046% no valor de IQA.

Figura 10- Influência das variáveis independentes na previsão do IQA do modelo de Regressão Linear com transformação logarítmica.

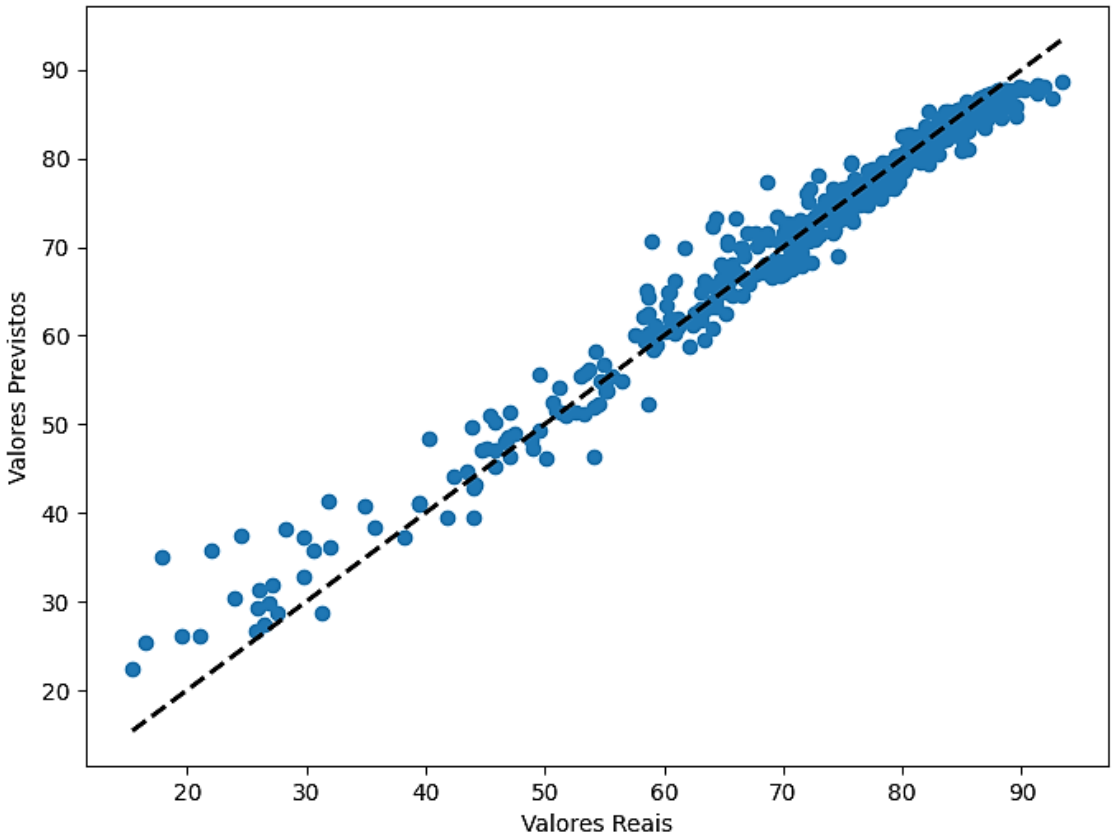


Fonte: Autoria própria.

As variáveis independentes (preditoras) e de resposta selecionadas para o *Random Forest* foram as mesmas do modelo de Regressão Linear Múltipla, na qual todas foram mantidas até o final do modelo.

Após os ajustes necessários e a divisão dos dados entre 70% para treino e 30% para teste, obteve-se como avaliação R^2 dos dados de treino e teste, respectivamente, 0.97 e 0.96, e RMSE de 2.02 e 2.81, demonstrando excelente ajuste aos dados e boa generalização do modelo aos dados desconhecidos. Também pôde ser observado, no gráfico de dispersão entre os valores previstos e reais, em que há boa correlação e concentração entre os pontos (Figura 11).

Figura 11- Gráfico de dispersão entre valores previstos e reais do modelo *Random Forest*

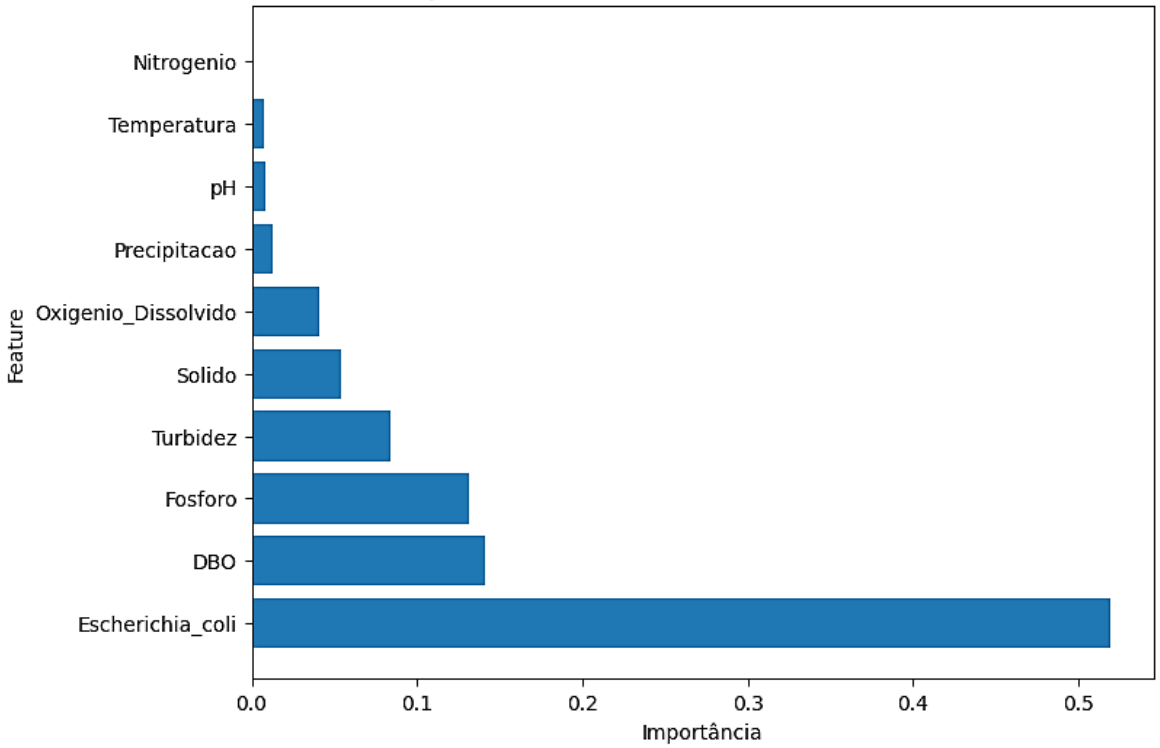


Fonte: Autoria própria.

Ademais, foi analisada a influência das variáveis preditoras para a previsão do modelo *Random Forest* (Figura 12). A barra representa a magnitude da influência da variável independente, em que quanto maior a barra, maior a importância dela para a previsão do modelo. Desta maneira, conclui-se que a variável de maior influência é a *Escherichia coli*, seguido do DBO, fósforo, turbidez, sólido, OD, precipitação, temperatura, pH e, por fim, nitrogênio.

Ainda, para fins de comparação entre os modelos de aprendizado supervisionado, analisou-se as diferenças das métricas de desempenho (Tabela 2), na qual o de *Random Forest* obteve o melhor ajuste dos dados (observado por meio dos resultados do coeficiente de determinação - R^2). Por outro lado, os ajustes de Regressão Linear Múltipla tiveram melhores resultados em relação aos erros entre os valores reais e os valores previstos (observado por meio do RMSE). Nota-se que os coeficientes de determinação foram maiores que 0.60, indicando um bom ajuste entre os dados observados e previstos (Silva *et al.*, 2023).

Figura 12- Importância das variáveis preditoras na previsão do IQA do modelo *Random Forest.*



Fonte: Autoria própria.

Tabela 2 - Comparação entre as métricas de desempenho dos modelos de Regressão Linear Múltipla e Random Forest. Valores de R^2 próximo de 1 indicam melhor ajuste dos dados, enquanto os valores de RMSE próximos de 0 indicam maior capacidade de predição dos modelos.

Métricas		**Regressão Linear Múltipla**	***Random Forest***
Treino	**R^2**	0.9405	0.9734
	RMSE	0.0668	2.3900
Teste	**R^2**	0.9219	0.9761
	RMSE	0.0651	2.3790

Fonte: Autoria própria.

Ainda, valores de RMSE próximos de zero indicam maior a capacidade de predição do modelo (Silva, 2019). Portanto, considerando que o modelo de regressão também obteve uma métrica boa de ajuste dos dados e, ainda, foi capaz de reduzir variáveis preditoras e obter previsões com a menor margem de erro possível, foi considerado como o melhor modelo para previsão de índice de qualidade de água para a presente pesquisa.

5.3 Importância dos algoritmos e modelos para o monitoramento da qualidade da água

Pelas análises realizadas, foi possível extrair diversas informações relevantes acerca das características da água da UGRHI de Mogi Guaçu. Os diferentes resultados de água superficial e subterrânea - e de modelos supervisionados e não supervisionados - demonstram como o aprendizado de máquina pode ser uma ferramenta relevante para diferentes aplicações no monitoramento ambiental.

De maneira geral, o objetivo da pesquisa foi satisfeito através da aplicação dos modelos. Para águas subterrâneas, no entanto, ficou explícito como a falta de dados disponíveis interferiu negativamente na análise, resultando em constatações pouco eficientes. Adversamente, com os dados completos de água superficial, as análises realizadas foram altamente satisfatórias e relevantes para o estudo. Através do aprendizado não supervisionado constatou-se importantes correlações e padrões entre as variáveis estudadas e, com o aprendizado supervisionado, foi possível prever adequadamente o Índice de Qualidade de Água (IQA).

Nota-se ainda, que no modelo de regressão linear, variáveis preditoras puderam ser removidas sem impacto significativo na previsão do índice, demonstrando grande oportunidade como alternativa à redução de custos de monitoramento, uma vez que a fórmula oficial do IQA necessita obrigatoriamente da coleta e análise de todas as variáveis (o que pode ser altamente demorado e custoso).

Por fim, a inclusão da variável de precipitação (que não é originalmente utilizada para calcular o IQA) e os resultados positivos obtidos, indicou não só como o aprendizado de máquina possui alto potencial exploratório e é capaz de capturar relações complexas dos dados, mas também a possibilidade de um índice de qualidade da água com parâmetros mais diversificados.

5.4 Disponibilidade de dados em plataformas públicas

A questão da disponibilidade de dados demonstrou-se como um grande empecilho para o desenvolvimento mais completo e aprofundado do presente estudo. Os dados de água subterrânea, disponibilizados no sistema InfoAguas (pertencente à CETESB), foram impossibilitados de serem coletados devido a falhas de funcionamento do sistema.

6 CONSIDERAÇÕES FINAIS

É notável os avanços tecnológicos atuais, principalmente relacionados à utilização de Inteligência Artificial. Os resultados advindos da pesquisa permitiram explorar e satisfazer o objetivo do estudo por meio da implementação de algoritmos e modelos de *Machine Learning* no contexto ambiental, especificamente da qualidade de águas superficiais e subterrâneas.

Os modelos analisados no estudo apresentaram resultados eficazes, com destaque à água superficial, que apresentou análises assertivas e predições altamente satisfatórias para ambos modelos supervisionados. O modelo de Regressão Linear Múltipla foi selecionado como o melhor modelo para previsão de índice de qualidade de água para a presente pesquisa, indicando o grande potencial de contribuição para a área de monitoramento de parâmetros físico-químicos e biológicos, podendo estar relacionada a implementação de monitoramentos mais viáveis técnica e economicamente.

Para o aprendizado não supervisionado para água superficial, foi possível confirmar a possibilidade de seleção e análise de relações de parâmetros indicadores físico, químicos e biológicos para o monitoramento utilizando o banco de dados InfoAguas da Companhia Ambiental do Estado de São Paulo (CETESB). No entanto, em relação à água subterrânea, o estudo também demonstrou por meio das análises e métricas obtidas como é necessária a obtenção de um conjunto de dados mais completo para que seja possível realizar análises mais significativas e, assim, ter uma maior precisão para os algoritmos não supervisionados.

7 SUGESTÃO DE TRABALHOS FUTUROS

Os resultados obtidos com essa monografia constituem um ponto de partida de ideias para trabalhos futuros. A incorporação de outras variáveis preditoras nos modelos, como dados hidrológicos (vazão, nível do rio, etc.) e de uso e ocupação do solo, podem ser promissoras para o futuro da pesquisa. Além disso, a fórmula do IQA é limitada, uma vez que não considera diversos parâmetros relevantes como substâncias tóxicas, protozoários patogênicos e substâncias que interferem nas propriedades organolépticas da água, sendo interessante explorar variáveis preditoras com esses novos parâmetros.

Por fim, também seria proveitoso encontrar um conjunto de dados mais robusto para água subterrânea, a fim de alcançar um melhor desempenho na aplicação do aprendizado não supervisionado e, até mesmo, para incluir a predição de parâmetros por meio do aprendizado supervisionado.

8 REFERÊNCIAS

ABRAHAM, A., LIVINGSTON, D., GUERRA, I.; YANG, J. (2022, September). **Exploring the Application of Machine Learning Algorithms to Water Quality Analysis**. IEEE Explore. https://doi.org/10.1109/BCD54882.2022.9900636

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

APACHE SPARK: PySpark Overview. [S. l.], 2024. Disponível em: https://spark.apache.org/docs/latest/api/python/index.html. Acesso em: 02 dez. 2024.

ARRAES, F. D. D.; ANDRADE, E. M.; PALÁCIO, H. A. Q.; FROTA, J. I. J.; SANTOS, J. C. N. **Identificação dos íons determinantes da condutividade elétrica nas águas superficiais da bacia do Curu, Ceará.** Revista Ciência Agronômica, Fortaleza, v. 40, p. 346-3, 2009.

BRASIL (1997) Lei Federal Nº 9.433, 1997. **Política Nacional de Recursos Hídricos e criação do Sistema Nacional de Gerenciamento de Recursos Hídricos**. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9433.htm.

BRUSA, Luis. **Aprimoramento estatístico da regionalização de vazões máximas e médias. Aplicação a bacias hidrográficas do Rio Grande do Sul e Santa Catarina.** Dissertação (Doutorado). Programa de Pós-Graduação em Recursos Hídricos e Saneamento Ambiental - Universidade Federal do Rio Grande do Sul, 2004.

CETESB. **Relatório de Qualidade das Águas Interiores no Estado de São Paulo 2022.** Disponível em: https://cetesb.sp.gov.br/aguas-interiores/wp-content/uploads/sites/12/2023/09/Relatorio-de-Qualidade-das-Aguas-Interiores-no-Estado-de-Sao-Paulo-2022.pdf

CETESB. **Relatório de Qualidade das Águas Interiores no Estado de São Paulo 2023.** Disponível em: https://cetesb.sp.gov.br/aguas-interiores/wp-content/uploads/sites/12/2024/11/RAI-2023-Relatorio-de-Qualidade-de-Aguas-Interiores-2023.pdf

CETESB. **Relatório de Qualidade das Águas Subterrâneas no Estado de São Paulo 2022.** Disponível em: https://cetesb.sp.gov.br/aguas-subterraneas/wp-content/uploads/sites/13/2023/10/Qualidade-das-Aguas-Subterraneas-no-Estado-de-Sao-Paulo-2022.pdf

COSTA, Camila Calandriny Rocha da. **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, Eklas. **Machine Learning Crash Course for Engineers. Springer.** 2024. 465p. https://doi.org/10.1007/978-3-031-46990-9

MATPLOTLIB: Visualization with Python. [S. l.], 2024. Disponível em: https://matplotlib.org/. Acesso em: 07 nov. 2024.

NIELSEN, A. **Practical time series analysis:** Prediction with statistics and machine learning. 1. ed. Sebastopol, CA: O'Reilly Media, 2019. 500p.

NUMPY: The fundamental package for scientific computing with Python. [S. l.], 2024. Disponível em: https://numpy.org/. Acesso em: 15 out. 2024.

SÃO PAULO (1991) Lei Estadual nº 7.663, 1991. **Política Estadual de Recursos Hídricos.** Disponível em: https://www.al.sp.gov.br/repositorio/legislacao/lei/1991/lei-7663-30.12.1991.html

SCIKIT-LEARN: Machine Learning in Python. [S. l.], 2024. Disponível em: https://scikitlearn.org/. Acesso em: 07 nov. 2024.

SCIPY: The fundamental package for scientific computing with Python. [S. l.], 2024. Disponível em: https://scipy.org/. Acesso em: 18 nov. 2024.

SILVA, S. S. **Inteligência artificial para avaliação da qualidade da água.** 2019. Dissertação de Mestrado em Recursos Hídricos. Universidade Federal de Sergipe.

SILVA, P. L. C.; BORGES, A. C.; LOPES, L. S.; ROSA, A. P. **Developing a Modified Online Water Quality Index: A Case Study for Brazilian Reservoirs. Hydrology**, 2023, 10, 115. https://doi.org/10.3390/hydrology10060115

PANDAS: Python Data Analysis Library. [S. l.], 2024. Disponível em: https://pandas.pydata.org/index.html. Acesso em: 15 out. 2024.

PEREIRA, J. P. R. Monitorização da qualidade da água numa Estação de Tratamento de Águas Residuais: Uma abordagem baseada em Machine Learning. 2023. Dissertação de Mestrado (Mestrado em Engenharia Informática). Universidade do Minho, 2023.

PLOTLY: Interactive graphing library for Python. [S. l.], 2024. Disponível em: https:/plotly.com/. Acesso em: 15 out. 2024.

PORTO, Wellington. **Uso de modelos de aprendizado supervisionado para classificação da potabilidade da água.** 2024. 82p. Dissertação (mestrado) – Universidade Católica de Pelotas, Programa de Pós-Graduação, Mestrado em Engenharia Eletrônica e Computação. - Pelotas, 2024.

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.