Instituto Politécnico de Tomar

Escola Superior de Tecnologia de Tomar

Estimativa remota dos parâmetros de qualidade da água usando Imagens de Satélite

Projeto de Mestrado

Carlos Miguel Pinto Marques Gil

Mestrado em Engenharia Informática Internet das Coisas

Tomar, 19 de Dezembro de 2025

Escola Superior de Tecnologia de Tomar

Estimativa remota dos parâmetros de qualidade da água usando Imagens de Satélite

Dissertação de Mestrado

Carlos Miguel Pinto Marques Gil

Orientado por

Profº Doutor Manuel Barros – ESTT-IPT

Engº Hugo Magalhães – Flyrobotics

Júri

Profª Doutora Ana Lopes

Profº Doutor Pedro Correia

Dissertação apresentada ao Instituto Politécnico de Tomar para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática de Mestrado Internet das Coisas

Dedico este trabalho ao meu filho Jan, por ser a minha inspiração e a minha força em todos os dias desta caminhada. Que cada conquista aqui alcançada seja também um reflexo do amor, da coragem e da alegria que encontro em ti.

AGRADECIMENTOS

Agradeço, em primeiro lugar, à minha família, pelo apoio incondicional, compreensão e incentivo ao longo de todo este percurso. Sem a vossa presença, carinho e força nos momentos mais desafiantes, este trabalho não teria sido possível.

Um agradecimento muito especial é devido ao Professor Manuel Barros, pela sua imensa paciência, disponibilidade constante e orientação rigorosa. A sua dedicação, atenção aos detalhes e capacidade de explicar com clareza cada etapa deste processo foram fundamentais para a concretização deste trabalho, deixando uma marca profunda no meu percurso académico e pessoal.

RESUMO

A monitorização contínua da qualidade da água em corpos hídricos é essencial para a gestão sustentável de recursos aquáticos e proteção do ecossistema. Os métodos tradicionais baseados em amostragem *"in situ"* apresentam limitações significativas em termos de custos elevados, baixa resolução espacial e dificuldades de acesso a áreas remotas. Esta tese propõe o desenvolvimento e validação de modelos de aprendizagem de máquina, especificamente baseados na metodologia XGBoost (eXtreme Gradient Boosting), para estimar parâmetros de qualidade da água – nomeadamente temperatura e condutividade – utilizando dados do satélite Sentinel-2 da Agência Espacial Europeia (ESA).

A investigação integra dados de deteção remota com medições locais de sensores, empregando algoritmos avançados de aprendizagem automática para criar modelos preditivos robustos. Mediante a utilização de bandas espectrais selecionadas do Sentinel-2 e índices espectrais derivados, demonstrou-se a viabilidade de estimar parâmetros de qualidade da água com precisão aceitável. Os resultados evidenciam que o XGBoost supera métodos tradicionais em precisão e eficiência computacional, oferecendo uma plataforma escalável para monitorização ambiental em tempo quase real.

Palavras-chave: Monitorização da Qualidade da água, Deteção Remota, Sentinel-2, aprendizagem de máquina, XGBoost, Análise Espectral

ABSTRACT

Continuous monitoring of water quality in water bodies is essential for sustainable management of aquatic resources and ecosystem protection. Traditional methods based on "in situ" sampling present significant limitations in terms of high costs, low spatial resolution, and difficulties accessing remote areas. This thesis proposes the development and validation of machine learning models, specifically based on XGBoost (eXtreme Gradient Boosting) methodology, to estimate water quality parameters – namely temperature and electrical conductivity – using data from the Sentinel-2 satellite of the European Space Agency (ESA).

The research integrates remote sensing data with local sensor measurements, employing advanced machine learning algorithms to create robust predictive models. By using selected spectral bands from Sentinel-2 and derived spectral indices, the feasibility of estimating water quality parameters with acceptable accuracy has been demonstrated. Results show that XGBoost outperforms traditional methods in accuracy and computational efficiency, offering a scalable platform for environmental monitoring in near-real-time.

Keywords: Water quality, Remote Sensing, Sentinel-2, aprendizagem de máquina, XGBoost, Spectral Analysis

ÍNDICE

Estimativa remota dos parâmetros de qualidade da água usando Imagens de Satélite

Índice de figuras

Índice de tabelas

Capítulo 1- Introdução

1.1 Motivação e Enquadramento

A qualidade da água é um fator crítico e fundamental para a saúde humana, biodiversidade aquática e sustentabilidade ambiental. Com o crescimento populacional e intensificação das atividades industriais, a monitorização eficiente e contínuo da qualidade de corpos hídricos tornou-se imprescindível. Os parâmetros de qualidade da água incluem variáveis físicas como temperatura e condutividade, químicas como pH e oxigénio dissolvido, e biológicas como a concentração de clorofila-a e fitoplâncton[1-3].

Historicamente, a monitorização tem sido realizada através de amostragem "*in situ*" em pontos específicos, utilizando sondas multiparamétricas e análises laboratoriais. Esta abordagem, embora precisa, apresenta limitações substanciais: custos operacionais elevados, cobertura espacial reduzida, frequência de amostragem temporal limitada e dificuldade de acesso a áreas remotas ou de risco[1-3].

A emergência de tecnologias de deteção remota por satélite oferece uma solução complementar promissora. Os satélites de observação da Terra fornecem dados em alta resolução espacial e temporal sobre vastas áreas, possibilitando monitorização em tempo quase real. O satélite Sentinel-2, operado pela ESA no contexto do programa Copernicus, apresenta características particulares adequadas para este fim: resolução espacial de 10 a 60 metros, 13 bandas espectrais cobrindo o visível até o infravermelho de onda curta, e tempo de revisita de 5 dias[1-3].

Paralelamente, os avanços em aprendizagem de máquina e computação em nuvem criaram condições para processamento de grandes volumes de dados geoespaciais e desenvolvimento de modelos preditivos sofisticados. A fusão de dados de deteção remota com algoritmos de aprendizagem automática representa a fronteira atual em monitorização ambiental[1-3].

1.2 O Problema

Limitações dos Métodos Tradicionais:

Apesar dos avanços tecnológicos, existem ainda desafios significativos quanto à monitorização da qualidade de água. A análise dos métodos tradicionais de monitorização e aquisição de dados revela um conjunto de limitações que desafiam a implementação em larga escala e a obtenção de uma visão mais abrangente dos fenómenos em estudo.

Uma das principais barreiras reside no elevado custo operacional e de capital associado. A instalação, operação e manutenção de estações de monitorização convencionais requerem um investimento significativo não apenas em infraestrutura física (hardware robusto, instalações), mas também na alocação contínua de recursos humanos especializados e nos custos inerentes à manutenção preventiva e corretiva dos equipamentos ao longo do tempo[1-2].

Adicionalmente, estes métodos sofrem de limitações substanciais no que toca à abrangência e frequência dos dados. A natureza pontual das estações de amostragem implica uma cobertura espacial inerentemente reduzida, tornando inviável cobrir a totalidade das áreas de interesse, especialmente em regiões geográficas vastas, remotas ou de difícil cartografia. Paralelamente, a resolução temporal é frequentemente limitada, com a frequência de amostragem tipicamente restrita a ciclos de dias ou semanas. Esta restrição compromete a capacidade de detetar e analisar fenómenos de rápida evolução ou variações de curta duração[1-2].

Por fim, a acessibilidade constitui um desafio prático de relevo. As dificuldades logísticas tornam a operação inviável em áreas de difícil acesso geográfico, bem como em cenários operacionais complexos, como situações de emergência, desastres naturais ou zonas de conflito. Nestes contextos, a segurança do pessoal e a recuperação dos dados tornam-se obstáculos quase intransponíveis para as metodologias tradicionais[1-2].

Desafios da Abordagem por Satélite:

A relação entre a refletância medida por satélite e os parâmetros de qualidade da água é complexa, porque nem todos esses parâmetros são "visíveis" para os sensores: alguns não são oticamente ativos e, por isso, não se refletem diretamente na radiação que o satélite capta. Além disso, o sinal ótico que chega ao sensor é influenciado por vários fatores ambientais ao mesmo tempo, como a quantidade de sedimentos em suspensão, os pigmentos (por exemplo, clorofila) e a matéria orgânica dissolvida, o que dificulta separar o efeito de cada componente [1,2,4,5].

Para que os modelos sejam fiáveis, é essencial fazer calibração e validação com dados de referência medidos "*in situ*", ou seja, diretamente na água, de forma a ajustar e verificar as relações entre refletância e os parâmetros que se querem estimar. Há ainda o efeito atmosférico, já que a atmosfera interfere nas medições espectrais, obrigando a aplicar correções para remover a contribuição do ar e obter apenas o sinal da água. Por fim, o tempo de revisita também é uma limitação importante: a frequência com que o satélite passa sobre a mesma área, combinada com a presença de nuvens, pode fazer com que haja poucos momentos úteis de observação, dificultando o acompanhamento contínuo da qualidade da água[1,2,4,5].

Limitações da Investigação existente:

Apesar de existirem estudos de estimação de parâmetros de qualidade da água por deteção remota, são relativamente poucos os trabalhos que abordam, de forma conjunta, a estimação simultânea da temperatura e da condutividade elétrica, sobretudo em contextos de águas interiores, o que evidencia margem para investigação adicional. Também se sente a necessidade de validar modelos baseados em XGBoost em contextos específicos de monitorização da qualidade da água, para perceber se estes modelos funcionam bem em diferentes tipos de rios, lagos ou albufeiras. Além disso, a integração de forma consistente de dados de sensores locais IoT com informações obtidas por satélites, é com certeza uma combinação que pode tornar o acompanhamento da qualidade da água muito mais completo e fiável e contribuir muito para esta área[1,2,4,5].

Outro desafio, é a falta de plataformas operacionais que permitam implementar estas soluções em tempo real, desde a recolha de dados até à geração de alertas ou indicadores

úteis para a gestão. A isto acresce que ainda existe alguma escassez de bases de dados públicas, suficientemente completas e bem estruturadas, que possam ser usados de forma eficaz em projetos de investigação e desenvolvimento nesta área [1,2,4,5].

1.3 Objetivos

Objetivo Geral:

Desenvolver e validar modelos de aprendizagem de máquina baseados em XGBoost para estimar temperatura e condutividade em corpos hídricos, integrando dados do satélite Sentinel-2 com medições locais de sensores, visando criar uma plataforma escalável de monitorização ambiental.

Objetivos Específicos:

1. Integração de Dados: Estabelecer procedimentos robustos de recolha, processamento e integração de dados do Sentinel-2 com medições "*in situ*" de temperatura e condutividade.

2. Seleção de Variáveis: Identificar as bandas espectrais mais relevantes do Sentinel-2 e índices derivados para estimação de temperatura e condutividade.

3. Desenvolvimento de Modelos: Construir modelos XGBoost otimizados para cada parâmetro, com validação cruzada e ajuste de hiperparâmetros.

4. Validação e Comparação: Avaliar o desempenho dos modelos com métricas estabelecidas (RMSE, MAE, R^2) e comparar com métodos tradicionais.

5. Avaliação da relevância das variáveis: Determinar a contribuição relativa de cada variável espectral na previsão.

6. Implementação Operacional: Desenvolver um protótipo de sistema de monitorização integrando processamento em Google Earth Engine e modelos Python.

1.4 Trabalho de investigação

Perspetiva da deteção Remota

A deteção remota multiespectral permite caracterizar as propriedades espectrais da água através da medição da refletância em diferentes comprimentos de onda. Este estudo está focado na:

- Seleção e processamento de bandas Sentinel-2 para maximizar sensibilidade a temperatura e condutividade
- Derivação de índices espectrais adaptados aos parâmetros de interesse
- Correção atmosférica adequada para melhorar qualidade de dados
- Análise de correlações entre refletância e parâmetros "*in situ*"

Perspetiva da Aprendizagem Automática

As técnicas de aprendizagem de máquina permitem aprender relações não-lineares complexas entre variáveis espectrais e parâmetros de qualidade. Este estudo explora:

- As técnicas de XGBoost como algoritmo principal, explorando sua capacidade em capturar interações entre variáveis;
- A otimização de hiperparâmetros para cada contexto específico;
- Validação com técnicas de aprendizagem de máquina estabelecidas é usar métodos padronizados (como "*hold-out"* e validação cruzada) e métricas adequadas;
- A capacidade de interpretação dos modelos através de análise de importância das variáveis.

1.5 Estrutura do documento

Este documento está organizado em sete capítulos principais. O Capítulo 2 apresenta o estado da arte, fazendo uma revisão dos sistemas de monitorização local, soluções baseadas em satélite e técnicas de aprendizagem de máquina.
O Capítulo 3 estabelece fundamentos teóricos sobre deteção remota, características do Sentinel-2, propriedades óticas da água e fundamentos de temperatura e condutividade. O Capítulo 4 descreve a metodologia de investigação, a arquitetura do XGBoost e o pré-processamento de dados. O Capítulo 5 apresenta a arquitetura da plataforma IoT desenvolvida. O Capítulo 6 discute os resultados obtidos. E por fim as conclusões são feitas no Capítulo 7.

Capítulo 2 - Estado da Arte e Estudo de Tecnologia

2.1 Sistemas de Monitorização Local da Qualidade de Água

A monitorização tradicional de qualidade de água baseia-se em medições *"in situ"* utilizando instrumentação especializada, baseada em princípios clássicos de limnologia e controlo de qualidade.

Estas estações podem ser fixas (bóias, estruturas subaquáticas) ou móveis (campanhas de amostragem), fornecendo séries temporais detalhadas, mas com cobertura espacial limitada [6-9].

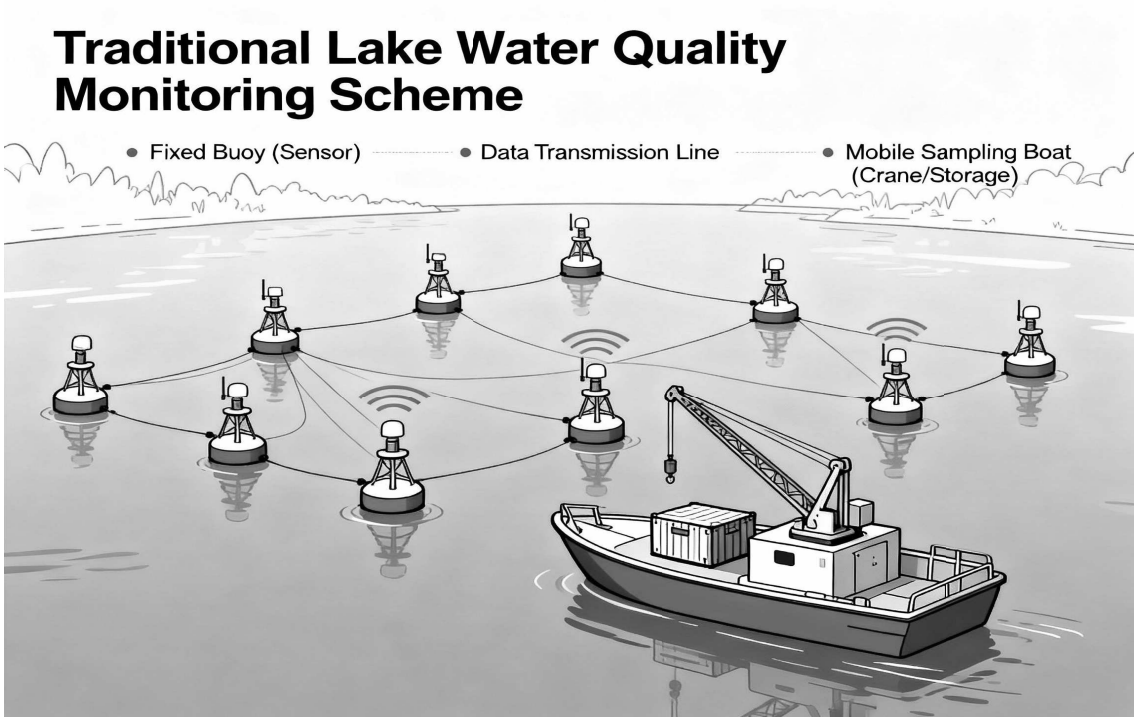

Figura 1: Forma tradicional de monitorizar a qualidade de água

Fonte: imagem gerada com apoio de GPT-5 em 06/11/2025.

Tecnologias Tradicionais:

Utilização de Sondas Multiparamétricas: Dispositivos como YSI®, Hach® ou similares medem simultaneamente múltiplos parâmetros (temperatura, pH, condutividade, oxigénio dissolvido, turbidez), oferecendo elevada precisão e repetibilidade em pontos específicos [6-9].

Análise Laboratorial: Técnicas de referência para validação e parâmetros que não podem ser medidos "*in situ*" (nutrientes, pesticidas, microrganismos). Dispendiosas e com elevado tempo de processamento.

Sistemas de Monitorização Contínua: Estações automáticas com transmissão de dados via telemetria, incluindo soluções comerciais baseadas em boias de monitorização e plataformas autónomas, como sistemas com drones aquáticos desenvolvidos por empresas especializadas em monitorização ambiental [6-9].

Um exemplo notável é o sistema implementado pela empresa Flyrobotics (https://flyrobotics.pt), que utiliza drones e plataformas autónomas para a recolha e monitorização de dados em ambientes aquáticos.

Figura 2: Exemplo de um Drone aquático da empresa Flyrobotics®

Limitações Críticas:

- Manter uma estação de monitorização no terreno implica um compromisso contínuo de recursos — entre manutenção, equipa técnica e validações laboratoriais — que, no conjunto, acaba por pesar de forma significativa no orçamento operativo [10-12,22].
- Cobertura espacial: Normalmente não se tem muitas estações por reservatório, sendo insuficiente para capturar a heterogeneidade espacial de grandes corpos hídricos.
- Resolução temporal: Medições diárias a semanais, com dificuldade em acompanhar eventos rápidos como descargas acidentais ou episódios de eutrofização súbita [10-12].
- Inacessibilidade: Impossível em áreas remotas, politicamente instáveis ou de risco, o que limita a monitorização sistemática em muitos contextos [10-12].

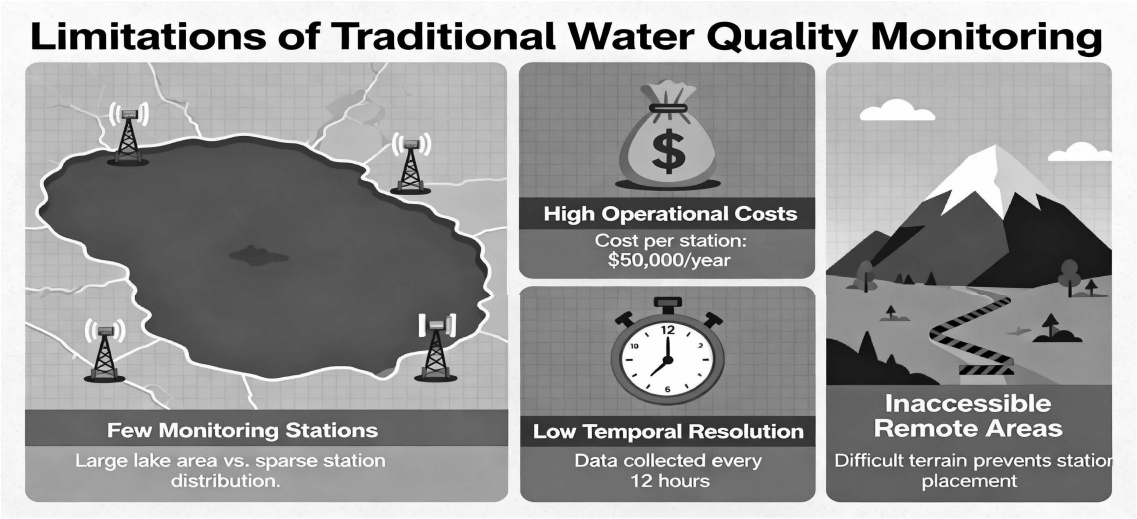

Figura 3: Figura que ilustra os desafios encontrados com os métodos tradicionais

Fonte: imagem gerada com apoio de GPT-5 em 06/11/2025.

2.2 Soluções de Monitorização baseadas em Imagens de Satélite

A deteção remota por satélite permite observação de grandes áreas com cobertura espacial alargada e com uma maior frequência temporal, constituindo um complemento às estações "*in situ*" [1-2].

Satélites relevantes para aplicação de monitorização de qualidade de água:

Caracterização do Sentinel-2 (ESA, 2015-presente):
- Resolução espacial: 10 m (bandas visível e NIR), 20 m (red-edge), 60 m (bandas de correção atmosférica) [3].
- Largura de faixa: 290 km, permitindo cobrir grandes bacias hidrográficas numa única passagem [3].
- Resolução temporal: 2-3 dias, com os três satélites Sentinel-2 atualmente em operação (S2A, S2B e S2C, com o S2A em órbita estendida, notar que esta situação é temporária e o tempo de resolução com dois satélites é de cerca de 5 dias), adequada para monitorização quase contínua em latitudes médias. De referir ainda que Sentinel-2A está em fase de

missão prolongada e tem fim de vida estimado por volta de 2026, com margem para operar até cerca de 2027 se o combustível o permitir [3].

- 13 bandas espectrais, cobrindo aproximadamente 440–2190 nm, incluindo várias bandas no red-edge com elevada sensibilidade a constituintes oticamente ativos [3].
- Acesso livre e aberto aos dados no âmbito do programa Copernicus, reduzindo barreiras de entrada para projetos de monitorização ambiental [13].
- Vantagem principal: Resolução superior à de Landsat-8 (30 m) para aplicações em lagos e reservatórios de dimensão intermédia [3,14].

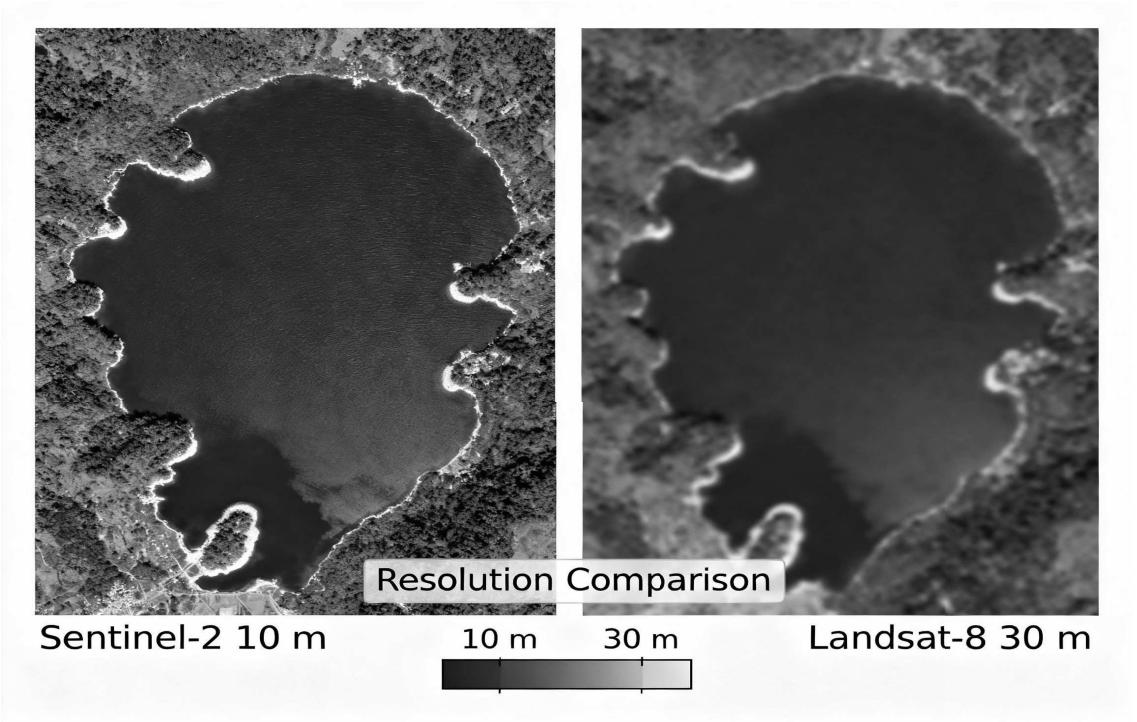

Figura 4: Imagem ilustrativa da diferença de resolução a que pode chegar o sentinel-2 por comparação e em detrimento do landsat-8

Fonte: imagem gerada com apoio de GPT-5 em 06/11/2025.

Landsat-8 (USGS, 2013-presente):

- Resolução espacial: 30 m (bandas multiespectrais) e 15 m (pancromático), com 11 bandas espectrais relevantes para aplicações em águas interiores [15-16].
- Tempo de revisita: 16 dias, com um arquivo histórico contínuo de dados de sensores ópticos (Missão Landsat) desde 1972 [15-16].
- 11 bandas espectrais

- Histórico de dados desde 1972 (sucessão de satélites)

Outros Satélites:

- MODIS: Resolução de 250–1000 m, com resolução temporal diária, adequado para monitorização de grandes lagos e sistemas costeiros em escala regional e global [34-37].
- Hiperspectrais (PRISMA, EnMAP): Elevada resolução espectral para discriminação fina de constituintes da água, mas com cobertura mais reduzida e revisita menos frequente que Sentinel-2[34-37].
- SAR (Sentinel-1): Insensível a nuvens e iluminação solar, complementar para deteção de alterações na superfície (por exemplo, derrames ou mudanças morfológicas), embora não seja sensível diretamente a parâmetros óticos de qualidade de água [34-37].
- Sentinel-3 (SLSTR): Resolução espacial de 1 km nos canais térmicos (adequada para massas de água maiores), com resolução temporal diária (usando a constelação A+B). Destaca-se pela elevada precisão (<0.3 K) devido à sua tecnologia de "dupla vista" (que corrige melhor os efeitos atmosféricos), sendo ideal para estudos de dinâmica térmica, ondas de calor e alterações climáticas em escalas regionais [18].

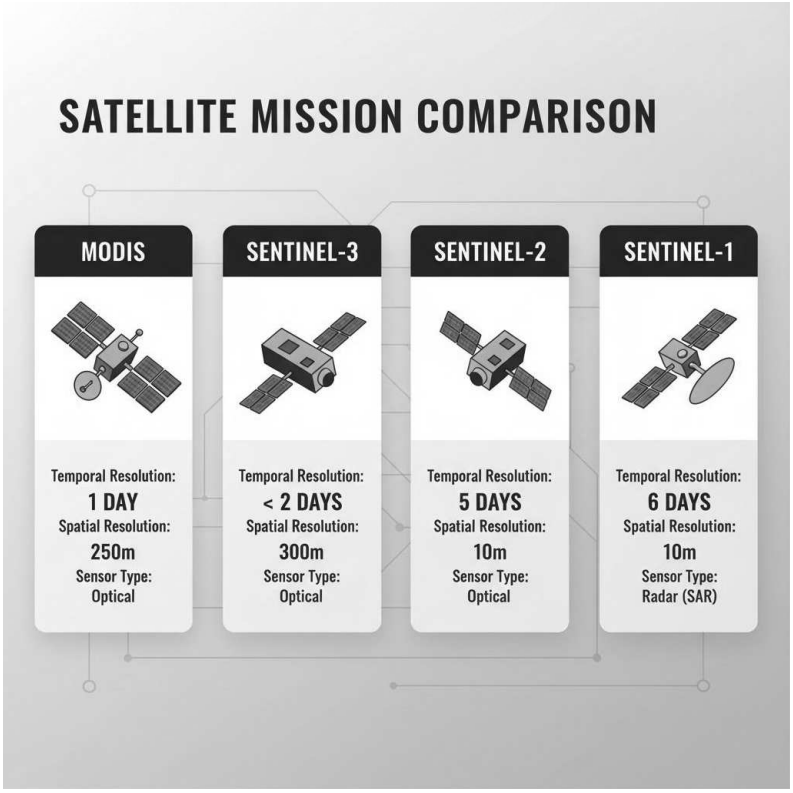

Figura 5: Comparação dos tempos de resolução temporal e revisita dos diversos satélites discutidos

Fonte: imagem gerada com apoio de GPT-5 em 03/12/2025.

Algoritmos de monitorização remota da Qualidade de Água:

Algoritmos Empíricos:
Baseiam-se em correlações diretas entre bandas espectrais (ou razões de bandas) e parâmetros de qualidade de água medidos "*in situ*", tendo sido amplamente utilizados para clorofila-a, turbidez e sólidos suspensos [38-42,1].

Exemplos incluem:

- NDCI (Normalized Difference Chlorophyll Index), que explora a diferença entre bandas no red-edge e no vermelho para estimar clorofila-a [38-42,1].
- OC3 (Ocean Color 3), originalmente desenvolvido para ambientes costeiros e oceânicos, também adaptado em alguns estudos de águas interiores [38-42,1].

Limitação geral dos algoritmos empíricos, estes modelos tendem a ser específicos de cada região e condição ótica, apresentando fraca generalização quando aplicados a outros corpos de água ou condições ambientais [38-42,1].

Modelos Semi-Analíticos, como C2RCC (Case 2 Regional Coast Colour), decompõe-na refletância em componentes (água pura, sólidos suspensos, matéria orgânica). São mais robustos, mas computacionalmente mais intensivos [38-42,1].

Vantagens de Sentinel-2:
- Resolução espacial de 20 m adequada para lagos e reservatórios de dimensão intermédia, permitindo mapear gradientes espaciais de qualidade de água [38-42,1].
- Bandas no red-edge (705, 740, 783 nm) sensíveis a pigmentos fitoplanctónicos e transições entre absorção e espalhamento, fundamentais para estimativa de clorofila-a e outros constituintes [38-42,1].
- Acesso livre aos dados, combinando-se com plataformas em nuvem como o Google Earth Engine para viabilizar projetos de monitorização operacional [38-42,1].
- Cobertura temporal adequada com os Sentinel-2A/B/C, possibilitando séries temporais de multi-ano para estudos de tendência e sazonalidade [38-42,1].

- Disponibilidade de produtos de refletância de superfície (Level-2A) já corrigidos atmosfericamente por cadeias de processamento como Sen2Cor e MAJA, simplificando a etapa de pré-processamento reduzindo a complexidade da mesma.

2.3 Tecnologias de aprendizagem de máquina para Estimação de Parâmetros da Qualidade de Água

Ao longo da última década, as técnicas de aprendizagem de máquina têm vindo a complementar, e em muitos casos, substituir progressivamente os métodos empíricos lineares na estimação de parâmetros de qualidade de água a partir de imagens de satélite [43-45].

A capacidade destes modelos em capturar relações não lineares complexas entre múltiplas bandas espectrais, índices derivados e variáveis ambientais, traduz-se em ganhos de desempenho face a modelos lineares clássicos. De seguida faz-se uma revisão de algoritmos utilizados em estudos académicos sobre a estimação de parâmetros da qualidade de água.

Máquinas de Suporte Vetorial (Support Vector Machines, SVM):

Funcionam bem com conjuntos de dados relativamente pequenos e permitem *kernels*(função que transforma os dados para um espaço onde as relações não lineares se tornam quase lineares) não lineares capazes de modelar relações complexas entre refletância e parâmetros de qualidade. No entanto, o ajuste de hiperparâmetros (C, gamma e tipo de *kernel*) pode ser complexo e computacionalmente intensivo, além de a capacidade de interpretação do modelo ser limitada. As SVM têm sido usadas em estudos com dados Landsat e MODIS para a estimação de clorofila-a, turbidez e sólidos suspensos em águas interiores e costeiras, bem como em diversas outras tarefas de classificação e regressão em aprendizado de máquina[43-45].

Redes Neuronais Artificiais (ANN):

As redes neuronais artificiais são muito flexíveis e conseguem representar relações altamente não lineares, sobretudo quando existe uma grande quantidade de dados disponíveis para treinar os modelos. Ao mesmo tempo, têm uma forte tendência para fazer "sobreajuste" (quando a rede neuronal "aprende demais" o conjunto de treino, incluindo ruído e detalhes muito específicos, em vez de aprender padrões gerais) quando os conjuntos de dados são pequenos, o que obriga a uma calibração cuidadosa da arquitetura e dos hiperparâmetros, bem como ao uso de técnicas de regularização. Além disso, são frequentemente vistas como modelos de "caixa negra", o que torna difícil perceber de forma clara como chegam às suas previsões, limitando a capacidade de interpretação.
Na área da qualidade da água, estas redes têm sido aplicadas em trabalhos recentes para classificar o estado de lagos e reservatórios a partir de imagens de satélite, nomeadamente Landsat 8 e Sentinel-2, incluindo variantes mais avançadas como as redes neuronais convolucionais (desenhado para processar imagens e dados visuais), que exploram melhor a informação espacial das imagens. Em geral, a literatura recente destaca precisamente este conjunto de características: grande capacidade de modelar padrões complexos, necessidade de muitos dados para tirar partido dessa capacidade e risco acrescido de *"*sobreajuste*"* quando os dados são insuficientes [43-45].

Random Forests (RF):

Estes modelos constituem uma abordagem de aprendizagem de máquina amplamente utilizada, destacando-se pela robustez face à presença de ruído nos dados, pela capacidade de lidar com conjuntos de preditores fortemente correlacionados e pela possibilidade de obtenção de medidas de importância das variáveis, que apoiam a interpretação dos resultados do modelo. Apesar destas vantagens, os RF tendem a ser menos eficientes do ponto de vista computacional e, em muitos casos, apresentam um desempenho ligeiramente inferior ao de métodos baseados em "*gradient boosting"* (método que constrói um modelo forte somando muitas pequenas árvores de decisão treinadas em sequência) quando estes são cuidadosamente otimizados. Ainda assim, a combinação entre robustez, bom poder preditivo e relativa simplicidade de calibração faz com que os RF sejam amplamente adotados em estudos comparativos, nos quais são frequentemente

utilizados como modelo de referência para a estimação de clorofila-a, turbidez e outros parâmetros de qualidade da água a partir de dados do Sentinel-2 e de outros sensores de observação da Terra [43-45].

Regressão Linear e Regressão Polinomial:

Os modelos de regressão linear constituem uma das abordagens estatísticas mais tradicionais na calibração espectral, sendo frequentemente adotados como primeira opção devido à sua simplicidade, elevada capacidade de interpretação e baixo custo computacional, o que facilita tanto a implementação como a análise crítica dos resultados obtidos. No entanto, estes modelos assumem tipicamente relações lineares ou polinomiais de baixa ordem entre as variáveis explicativas e a variável resposta, o que tende a ser inadequado para descrever processos complexos em qualidade da água, conduzindo muitas vezes a um desempenho preditivo limitado face a métodos não lineares mais avançados. Por essa razão, a regressão linear é amplamente utilizada como modelo de referência em estudos de monitorização remota, servindo como linha de base mínima para quantificar os ganhos obtidos com técnicas mais sofisticadas, nomeadamente na estimativa de parâmetros como clorofila-a, turbidez e outros indicadores de qualidade da água a partir de dados de sensoriamento remoto, por exemplo de missões como Sentinel-2 [43-45].

XGBoost (eXtreme Gradient Boosting):

O método baseado em gradient boosting foi selecionado como abordagem principal neste trabalho por combinar elevada capacidade preditiva com boa eficiência computacional, respondendo às exigências de modelação da relação complexa e não linear entre as variáveis espectrais do Sentinel-2 e os parâmetros de temperatura e condutividade da água. Em termos gerais, estes algoritmos constroem um modelo forte a partir de um conjunto de árvores de decisão simples ajustadas de forma sequencial, integrando mecanismos de regularização L1/L2 que limitam a complexidade do modelo e reduzem o risco de sobreajuste, o que é particularmente importante em cenários com grande número de preditores e possíveis correlações entre bandas e índices espectrais. Adicionalmente, a disponibilidade de métricas de importância de variáveis permite identificar de forma

transparente quais as bandas e combinações espectrais mais relevantes para explicar a variabilidade da temperatura e da condutividade, favorecendo uma interpretação mais fundamentada dos resultados. Neste contexto, e tendo em conta o reconhecimento de modelos de "*gradient boosting"* como estado da arte em diversas aplicações de aprendizagem de máquina, a sua adoção neste estudo procura não só maximizar o desempenho na regressão dos parâmetros de interesse, mas também estabelecer uma base metodológica sólida e atual para a monitorização remota da qualidade da água [43-45].

2.4 Estudos Comparativos: Abordagens Clássicas vs. Modernas

Estudo 1

Em estudos comparativos para estimar clorofila-a com modelos de aprendizagem automática, Random Forest, redes neuronais (ANN/MLP) e métodos da família SVM/SVR tendem a apresentar desempenhos semelhantes, e a melhor escolha depende do conjunto de dados e do desenho de validação. De forma geral, RF e ANN são opções fortes para relações não lineares, mas não existe uma hierarquia fixa face a SVM/SVR, havendo estudos em que o SVR é ligeiramente superior e outros em que a ANN se destaca [30-33].

Estudo 2

Em estudos recentes de turbidez com dados Sentinel-2, o XGBoost aparece frequentemente como uma abordagem muito competitiva face a SVR/SVM, sobretudo quando se pretende capturar relações não lineares entre bandas/índices e a turbidez. Num estudo no rio Mississippi (EUA), comparando vários algoritmos (incluindo XGBoost e SVR) com bandas e índices do Sentinel-2, reportam que os modelos baseados em árvores foram, no geral, os mais fortes, e que o XGBoost atingiu o melhor desempenho agregado ao usar todas as amostras . Em contexto de reservatório, há também evidência de que o XGBoost pode ter desempenho excelente em treino, mas que a robustez em teste pode depender do tamanho e variabilidade da amostra, tendo sido observado um caso em que

um modelo de regressão empírica superou o XGBoost na fase de teste para turbidez, enquanto um conjunto (ELR-XGBoost) melhorou a predição [28-29].

Estudo 3

Abordagens que combinam medições in situ (incluindo redes de sensores/IoT) com dados de satélite e modelos de aprendizagem automática permitem reforçar a monitorização espaço-temporal e suportar modelos de estimação de parâmetros como pH e condutividade [25-26].
De forma geral, a fusão de dados e/ou a combinação de modelos pode produzir previsões mais consistentes do que abordagens baseadas numa única fonte, reforçando o valor da integração entre observações de terreno e deteção remota [26-27].

2.5 Conclusão dos estudos comparativos - Limitações Identificadas

- São poucos os estudos que focam especificamente a estimação simultânea de temperatura e condutividade elétrica da água a partir de Sentinel-2 em reservatórios de água doce [1,5,43,46,48].
- Existe a necessidade de uma validação sistemática de modelos XGBoost e outros métodos de "Gradient boosting" em contextos europeus com características particulares de climatologia e uso do solo [1,5,43,46,48].
- A integração entre sensores IoT, plataformas de processamento em nuvem, como o Google Earth Engine, e *"pipelines"* de aprendizagem automática permanece ainda em fase incipiente, observando-se uma escassez de *"frameworks operacionais"* bem estruturados e efetivamente replicáveis na prática, isto é, que possam ser facilmente reproduzidos por diferentes utilizadores e transferidos para distintos contextos espaciais. [1,5,43,46,48].
- A escassez de implementações operacionais em tempo quase real que liguem diretamente dados Sentinel-2, sensores locais e as plataformas de apoio à decisão para entidades gestoras de recursos hídricos [1,5,43,46,48].

Este capítulo contextualizou a evolução das técnicas de monitorização de qualidade de água e identificou as limitações científicas que motivam esta investigação. Para superar estas limitações, é necessário estabelecer os alicerces teóricos sobre os quais se constrói a metodologia proposta. O Capítulo 3 apresenta, portanto, os princípios físicos da deteção remota, as características específicas do satélite Sentinel-2, e os processos ópticos que governam a qualidade da água em reservatórios de água doce, estabelecendo assim a base conceptual para a abordagem metodológica que se segue.

Capítulo 3 – Fundamentação Teórica

3.1 Princípios de Deteção Remota Multiespectral

A deteção remota mede a radiação eletromagnética refletida ou emitida por objetos na superfície terrestre, permitindo inferir propriedades físicas e bioquímicas sem contacto direto com o alvo [41,43,49-52].
No contexto da qualidade da água, trabalha-se sobretudo com radiação refletida no intervalo aproximadamente entre 0.4 e 2.2-2.3 μm, onde se encontram as bandas visíveis, infravermelho próximo (NIR) e infravermelho de ondas curtas (SWIR) utilizadas por sensores como o Sentinel-2 e o Landsat-8 [41,43,49-52].

O processo físico pode ser resumido em várias etapas: a radiação solar incide na superfície da água, uma parte é refletida, a outra penetra na coluna de água e sofre absorção e espalhamento, regressando parcialmente à superfície como radiação ascendente que é posteriormente registada pelo sensor a bordo do satélite [41,43,49-52].
Ao longo deste percurso, a radiação interage com a atmosfera, que contribui com uma fração significativa do sinal medido, sendo por isso necessária a aplicação de procedimentos de correção atmosférica para recuperar a refletância de superfície [41,43,49-52].

A refletância espectral é definida como a fração de radiação incidente que é refletida em cada comprimento de onda, constituindo uma assinatura característica que pode ser associada à composição da água, aos seus constituintes oticamente ativos e ao estado da superfície [41,43,49-52].
Esta assinatura permite distinguir entre diferentes tipos de massas de água, níveis de turbidez, concentrações de fitoplâncton e presença de matéria orgânica dissolvida, desde que existam modelos adequados que relacionem refletância com parâmetros de qualidade de água [41,43,49-52].

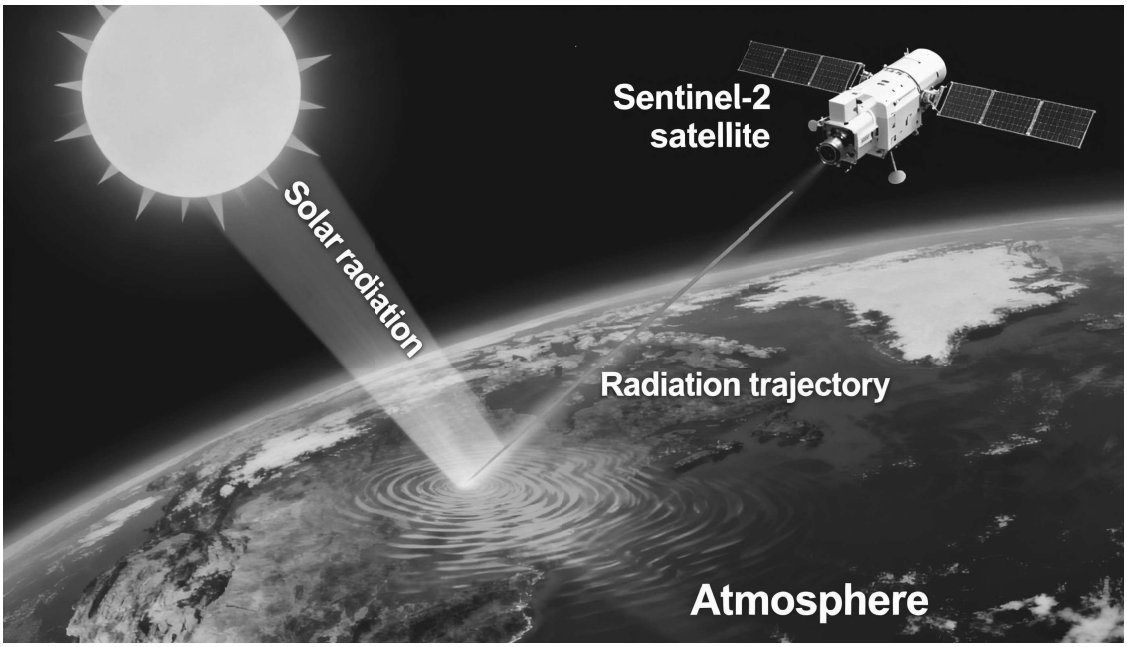

Figura 6: Esquema simplificado do processo de deteção remota da qualidade da água por satélite, desde a incidência da radiação solar até ao registo pelo sensor.

Fonte: imagem gerada com apoio de GPT-5 em 21/11/2025.

Entre as principais vantagens da deteção remota para monitorização de recursos hídricos destacam-se a natureza não invasiva das medições, a elevada cobertura espacial, a repetição frequente e a possibilidade de construir séries temporais longas para análise de tendências [41,43,49-52].

Por outro lado, existem desafios relevantes, como o contributo dominante da atmosfera no sinal medido, o efeito do fundo em águas rasas, a mistura de vários constituintes com respostas espectrais distintas e a necessidade de dados in situ de qualidade para calibração e validação dos modelos [41,43,49-52].

3.2 Características do Satélite Sentinel-2

O Sentinel-2 (ver Figura 7) é um par de satélites óticos multi-espectrais do programa Copernicus (Fig. 7). Atualmente, a constelação do Sentinel-2 é composta por três satélites em operação (Sentinel-2A, Sentinel-2B e Sentinel-2C) desenhado para observar a superfície terrestre com elevada resolução espacial e temporal, incluindo zonas costeiras e águas interiores [53-55].

Cada satélite opera numa órbita heliossíncrona a cerca de 786 km de altitude, com hora de passagem local normalmente até ao meio-dia, garantindo condições de iluminação relativamente constantes entre aquisições [53-55].

Figura 7: Um dos satélites da missão Sentinel-2 https://dataspace.copernicus.eu/explore-data/data-collections/sentinel-data/sentinel-2 (acesso em 06-12-2025)

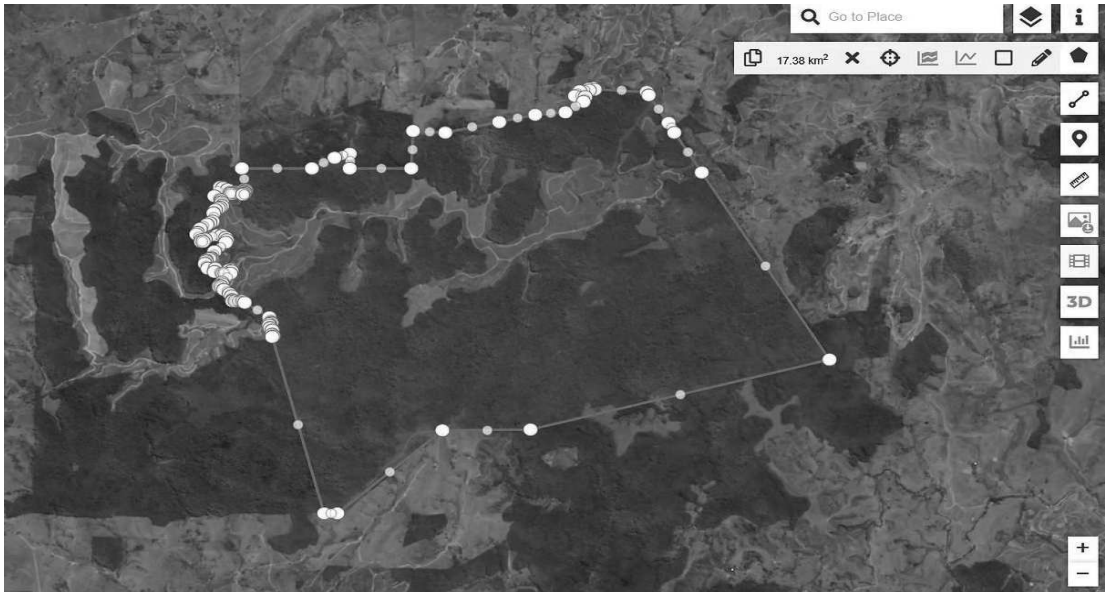

Figura 8: Utilização do Copernicus Browser para visualizar e recolher imagens Sentinel. Fonte: Copernicus Data Space Ecosystem (Copernicus Browser).https://browser.dataspace.copernicus.eu (acesso em 06-12-2025)

A missão Sentinel-2 foi inicialmente concebida como uma constelação de dois satélites idênticos, Sentinel-2A e Sentinel-2B, em órbita heliossíncrona, separados por 180°. Com o lançamento do Sentinel-2C em 2024 e a decisão de manter o Sentinel-2A em operação, a constelação passou, de forma excecional, a funcionar com três satélites em simultâneo. Esta configuração aumentou a frequência de tempo de revisita e a disponibilidade de dados de observação da Terra.

As especificações relevantes do Sentinel-2 incluem resolução espacial de 10 m nas bandas azul, verde, vermelho e NIR (B2, B3, B4, B8), 20 m nas bandas red-edge e SWIR (B5, B6, B7, B11, B12) e 60 m nas bandas dedicadas à correção atmosférica (B1, B9, B10) [3,4,9]. A largura de faixa (*swath*) de aproximadamente 290 km e o ciclo de tempo de revisita combinado de 5 dias com as duas plataformas permitem uma cobertura frequente de áreas extensas, o que é particularmente útil para monitorização quase contínua de grandes bacias hidrográficas e reservatórios [53-55].

A tabela seguinte, resume as bandas espectrais mais relevantes para aplicações em qualidade de água com Sentinel-2, bem como o seu propósito principal [41,55-60].

Banda	**Comprimento de onda aproximado**	**Propósito principal**
B1	443 nm (aerossol)	Correção atmosférica, deteção de bruma
B2	490 nm (azul)	Penetração na água, sensível a clorofila e sedimentos finos
B3	560 nm (verde)	Sensível a turbidez e sólidos suspensos
B4	665 nm (vermelho)	Forte absorção por clorofila-a
B5	705 nm (red-edge)	Contorno da banda de absorção da clorofila-a
B6	740 nm (red-edge)	Máxima sensibilidade a pigmentos fitoplanctónicos
B7	783 nm (red-edge)	Resposta a matéria orgânica e colóides
B8	842 nm (NIR)	Elevada refletância em águas turvas e com muitos sólidos
B11	1610 nm (SWIR)	Forte absorção por água pura, útil para índices de água
B12	2190 nm (SWIR)	Sensível à composição mineral de sedimentos

Tabela 1: Bandas espectrais principais do Sentinel-2 e respetiva função em aplicações de qualidade de água.

Para estudos de qualidade de água, as bandas visíveis (B2 – azul, B3 – verde, B4 – vermelho), NIR e red-edge (sobretudo B5, B6, B7 e B8) são particularmente importantes, pois capturam variações na concentração de pigmentos, sólidos suspensos e matéria orgânica que se relacionam com parâmetros como turbidez e clorofila-a [41,55-60].
Em muitos trabalhos sobre lagos e reservatórios combinam-se especialmente B2, B3, B4 e B8, por terem 10 m de resolução, e acrescentam-se bandas red edge (B5–B7) ou mesmo

B11/B12 (SWIR) para melhorar a estimação de parâmetros de qualidade da água [41,55-60].

O Sentinel-2 disponibiliza produtos de refletância de superfície (nível L2A) que já incluem correção atmosférica, podendo ser obtidos gratuitamente através do Copernicus Open Access Hub (https://colhub.copernicus.eu/) ou de plataformas em nuvem como o Google Earth Engine (https://earthengine.google.com), o que reduz barreiras de entrada e facilita a implementação de *pipelines* operacionais de monitorização [9,12].

Estas características tornam o Sentinel-2 particularmente adequado para este trabalho, que requer resolução espacial suficiente para detalhar o Reservatório de Castelo do Bode, bem como cobertura temporal regular para alimentar modelos de previsão de temperatura e condutividade [41,55-60].

3.3 Propriedades Óticas da Água

A resposta espectral da água resulta da combinação de vários constituintes com comportamentos de absorção e espalhamento distintos, incluindo a própria água pura, sólidos suspensos, matéria orgânica dissolvida e fitoplâncton, entre outros [5,9].

Compreender estas propriedades óticas é fundamental para interpretar os sinais medidos por sensores de satélite e para desenvolver modelos que relacionem refletância com parâmetros físico-químicos de interesse [2,38,61-63].

A água pura tem absorção relativamente baixa na zona do azul (com um mínimo no azul) e essa absorção aumenta à medida que avançamos para comprimentos de onda maiores — do verde para o vermelho e, sobretudo, para o infravermelho próximo (NIR) e o infravermelho de ondas curtas (SWIR).

Por isso, em águas profundas e limpas, a refletância tende a ser mais elevada no azul e a diminuir acentuadamente a partir do vermelho, sendo geralmente muito reduzida no NIR/SWIR. [2,38,61-63].

Em contrapartida, sólidos suspensos (minerais e matéria suspensa orgânica) aumentam a refletância em todo o espectro, com impacto particular nas bandas verde e NIR, estando fortemente associados à turbidez e à carga sedimentar [2,38,61-63].

A matéria orgânica dissolvida colorida (CDOM ou MOD) absorve fortemente na região do azul e do ultravioleta, com absorção que decresce com o aumento do comprimento de onda, influenciando principalmente as bandas mais curtas e modulando a cor aparente da água [2,38,61-63].

O fitoplâncton, por sua vez, possui pigmentos (sobretudo clorofila-a) com bandas de absorção características no azul (cerca de 440–460 nm) e no vermelho (cerca de 665 nm) e um pico de refletância na região red-edge (700–750 nm), o que permite estimar a sua abundância através de índices específicos [2,38,61-63].

Em águas interiores, estes componentes costumam variar ao mesmo tempo, o que gera respostas espectrais complexas e difíceis de separar. Diferentes combinações de turbidez, matéria orgânica e fitoplâncton podem originar sinais muito parecidos no satélite, o que complica a estimação direta e fiável dos parâmetros de qualidade da água [2,38,61-63].

Modelos empíricos e de aprendizagem automática são, por isso, frequentemente utilizados para capturar estas relações não lineares, desde que exista um conjunto representativo de dados "*in situ*" para calibração e validação [2,38,61-63].

No contexto desta tese, embora a temperatura e a condutividade não sejam parâmetros opticamente ativos no sentido estrito, admite-se que possam ser inferidas de forma indireta através de correlações estatísticas com variáveis óticas (por exemplo, turbidez/TSS, pigmentos e CDOM) e com a dinâmica do sistema.

Ainda assim, a existência e a estabilidade dessas relações podem variar com o tipo de massa de água, a sazonalidade e os processos dominantes, pelo que é essencial calibrar localmente e validar com dados *'in situ'*. [2,38,61-63].

Assim, a exploração conjunta de bandas espectrais, índices e modelos de aprendizagem de máquina permite capturar estas relações indiretas e inferir parâmetros fisicamente relevantes a partir de assinaturas espectrais complexas [2,38,61-63].

3.4 Temperatura e Condutividade (Fundamentos Físicos)

A temperatura da água é um parâmetro central em limnologia (ciência que estuda as águas continentais) e ecologia aquática, influenciando praticamente todos os processos físicos, químicos e biológicos que ocorrem em lagos e reservatórios [17,43,45,64-66].
Controla a taxa de reações bioquímicas, afeta a solubilidade de gases como o oxigénio e o dióxido de carbono e determina a densidade da água, condicionando padrões de estratificação térmica e mistura vertical [17,43,45,64-66].

Em sistemas temperados como o reservatório de Castelo do Bode, a temperatura superficial da água pode variar de valores baixos no inverno para valores elevados no verão, gerando camadas térmicas distintas (epilímnio, metalímnio e hipolímnio) que têm implicações diretas na distribuição de nutrientes, de oxigénio dissolvido e da biota. Epilímnio, metalímnio e hipolímnio são as três camadas em que se costuma dividir a coluna de água de um lago ou albufeira estratificada. O epilímnio é a camada superficial, mais quente e geralmente mais bem oxigenada; o metalímnio é a camada intermédia de transição, onde a temperatura muda rapidamente e o hipolímnio é a camada profunda, mais fria, mais densa e muitas vezes com menos oxigénio [17,43,45,64-66].

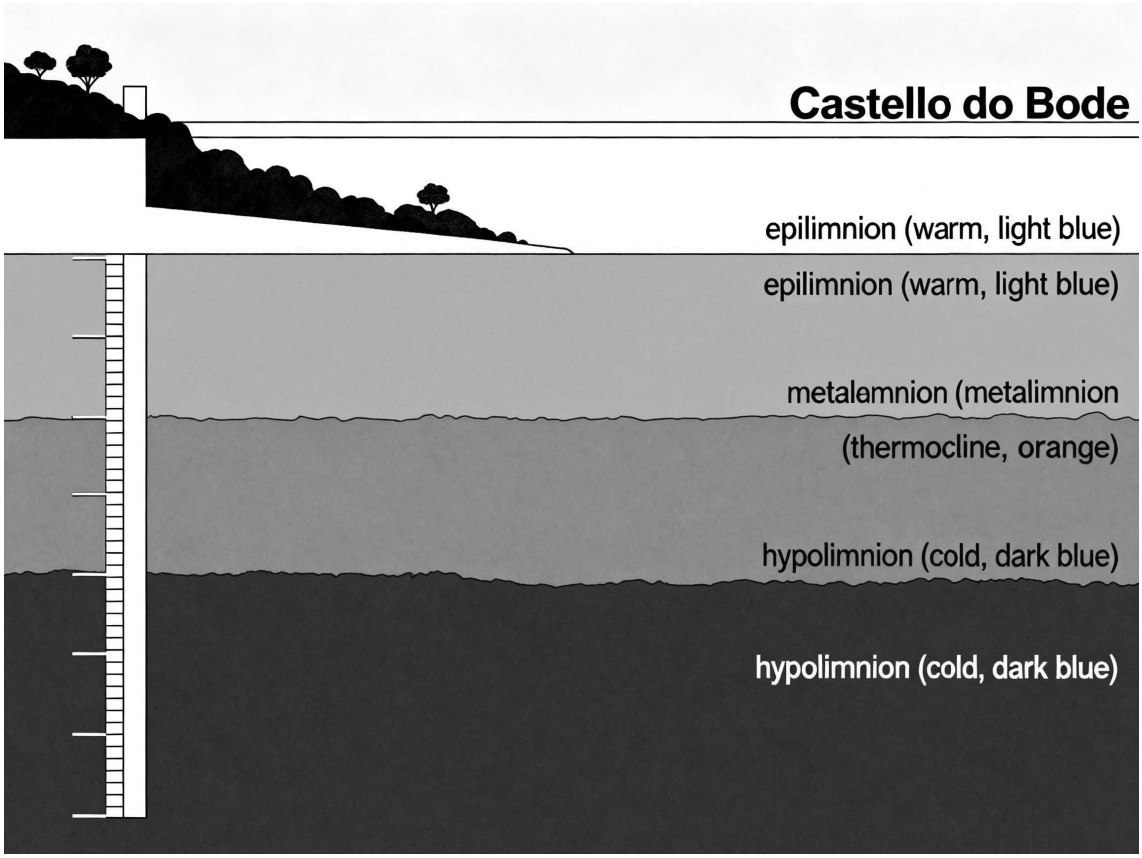

Figura 9: Perfil esquemático da estratificação térmica na albufeira do Castelo do Bode, com epilímnio, metalímnio e hipolímnio.

Fonte: imagem gerada com apoio de GPT-5 em 22/11/2025.

Esta estratificação sazonal pode favorecer o desenvolvimento de "*blooms"* de fitoplâncton em determinadas épocas, alterando a qualidade da água e afetando usos como abastecimento para consumo humano e recreação.

Ao longo do ano, a camada superficial da água da albufeira do Castelo do Bode tem variações sazonais: A temperatura, mistura vertical, concentração de clorofila-a, matéria orgânica, sedimentos em suspensão e profundidade penetração da luz variam de inverno para verão. Isso faz com que a relação entre refletância e parâmetros de qualidade da água também mude com a estação, podendo levar a modelos calibrados no verão a funcionar pior no inverno, e vice-versa [17,43,45,64-66].

A condutividade elétrica (CE) representa a capacidade de a água conduzir corrente elétrica e é aproximadamente proporcional à concentração total de iões dissolvidos, sendo

por isso um indicador útil do grau de mineralização e, em muitos casos, dos sólidos dissolvidos totais (TDS).
Valores típicos de condutividade em águas doces variam entre cerca de 50 e 500 µS/cm, dependendo da geologia da bacia, da influência de efluentes e de processos de mistura e evaporação, situando-se os reservatórios utilizados para abastecimento geralmente numa faixa intermédia desta gama [17,43,45,64-66].

Do ponto de vista ecológico, a condutividade afeta a forma como os organismos aquáticos regulam o equilíbrio de sais e água no seu corpo. Além disso, pode sinalizar mudanças na composição iónica da água, quer por processos naturais, como o intemperismo das rochas ou a intrusão de água salgada, quer por atividades humanas, como descargas industriais ou a lixiviação de fertilizantes e outros químicos dos solos agrícolas.
Em sistemas de abastecimento, variações significativas de condutividade podem sinalizar mudanças na mistura de fontes de água ou eventos anómalos, pelo que a sua monitorização contínua é recomendada em planos de segurança da água [17,43,45,64-66].

A condutividade é fortemente dependente da temperatura, aumenta tipicamente da ordem de ~2%/°C (dependendo da composição), sendo por isso comum normalizar para 25 °C, para permitir comparações entre campanhas e locais distintos.
Esta correção é também relevante para a calibração de modelos preditivos, uma vez que separa o efeito puramente térmico do contributo associado a alterações na concentração iónica real da água [17,43,45,64-66].

Embora temperatura e condutividade não tenham assinaturas espectrais diretas no visível e NIR comparáveis às de constituintes ópticos, a sua relação com processos como estratificação térmica, transporte de sedimentos, balanço de evaporação-precipitação e entradas de afluentes faz com que se correlacionem, em maior ou menor grau, com padrões espectrais observáveis em bandas e índices específicos [1,2,5].
A abordagem adotada nesta tese tira partido destas correlações indiretas, combinando dados "*in situ"* de temperatura e condutividade com refletância multiespectral Sentinel-2 e técnicas de XGBoost para estimar estes parâmetros com precisão operacionalmente útil [17,43,45,64-66].

3.5 Visão Geral de Algoritmos de Machine Learning

Aprendizagem de máquina, é um campo da inteligência artificial que permite aos modelos aprenderem padrões a partir de dados, em vez de serem explicitamente programados com regras determinísticas, o que é particularmente útil quando as relações entre refletância espectral e parâmetros de qualidade da água são complexas e não lineares.
No contexto desta tese, o foco recai sobre algoritmos de aprendizagem supervisionada de regressão, em que se pretende mapear vetores de variáveis de entrada (bandas e índices do Sentinel-2) para variáveis alvo contínuas (temperatura e condutividade), utilizando conjuntos de treino em que os valores de referência são obtidos por medições "*in situ"* [30,67-68].

De forma geral, podem distinguir-se três grandes paradigmas de aprendizagem de máquina: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço, ainda que apenas o primeiro seja diretamente relevante para este trabalho.
Na aprendizagem supervisionada, os modelos são treinados com pares entrada-saída, como no caso desta tese em que cada registo contém refletância multiespectral e as medições de temperatura ou condutividade associadas, permitindo construir funções de previsão que generalizam para novas observações [30,67-68].

A aprendizagem não supervisionada trabalha apenas com variáveis de entrada, sem rótulos, sendo tipicamente usada para agrupamento ou redução de dimensionalidade, podendo ter interesse complementar em análises exploratórias, mas não sendo utilizada como núcleo da modelação aqui apresentada.
Por sua vez, a aprendizagem por reforço envolve um agente que interage com um ambiente e aprende políticas ótimas com base em recompensas e penalizações, não se enquadrando diretamente no problema de previsão de parâmetros de qualidade de água abordado nesta investigação [30,67-68].

Entre os algoritmos supervisionados mais usados em estudos anteriores de qualidade de água destacam-se Support Vector Machines (SVM), Redes Neuronais Artificiais (ANN),

Random Forests (RF) e métodos de g*radient boosting* como o XGBoost, frequentemente comparados em revisões e estudos de caso com imagens Landsat e Sentinel-2 [5][6][7][8]. Em vários estudos recentes, métodos de conjuntos — como o Random Forest e, sobretudo, técnicas de *gradient boosting* (por exemplo, o XGBoost) — têm apresentado resultados melhores do que abordagens mais simples (incluindo regressões lineares) e, nalguns casos, do que modelos do tipo SVR/SVM, especialmente quando a relação entre as variáveis é claramente não linear e os dados contêm algum ruído [30,67-68].

Depois de estabelecidos os fundamentos teóricos da deteção remota e de aprendizagem de máquina, o Capítulo 4 apresenta de forma detalhada a metodologia operacional adotada nesta investigação. Em particular, é explicado o funcionamento do algoritmo XGBoost, justificando a razão pela qual foi escolhido em detrimento de outras alternativas, e são apresentadas as métricas e os procedimentos de validação que asseguram a robustez científica dos resultados obtidos.

Capítulo 4 - Metodologia de Investigação e implementação

4.1 XGBoost: Arquitetura e Funcionamento

XGBoost (eXtreme Gradient Boosting) foi desenvolvido por Tianqi Chen a partir de 2014 e tornou-se amplamente adotado em competições de a*prendizagem de máquina* e indústria. Baseia-se na metodologia de "*gradient boosting"* mas com otimizações significativas. É uma implementação otimizada da família de métodos de "*gradient boosting"*, na qual o modelo final é construído como uma soma de múltiplas árvores de decisão simples, cada uma treinada para ajustar-se aos pseudo-resíduos — os gradientes negativos da função de perda — calculados a partir do modelo anterior.
Esta abordagem aditiva permite capturar relações altamente não lineares entre as variáveis espectrais e os parâmetros de qualidade da água, preservando ao mesmo tempo uma estrutura baseada em árvores que facilita a obtenção de medidas de importância de variáveis [45].

Conceitos Fundamentais:

Gradient Boosting:
- Combina múltiplas árvores de decisão fracas (*weak learners*)
- Cada nova árvore aprende dos erros das árvores anteriores,
- Iterativo: árvore n+1 melhora predições da árvore n,
- Redução gradual de erro (daí *gradient*).

Matematicamente:

$F_k(x) = F_{k-1}(x)+\eta hk(x)$

Onde:
- Fk: Modelo com k árvores
- ηhk: Nova árvore, com taxa de aprendizagem η

O objetivo é minimizar uma função de perda L (loss function):

$L = \sum_i l(y_i, F_{k(xi)}) + \sum_j \Omega(hj)$

Onde:
- l: Perda individual para cada predição
- Ω: Termo de regularização

No caso deste trabalho, adotam-se funções de perda típicas de regressão, como o erro quadrático médio para temperatura e condutividade, complementadas com termos de regularização que penalizam modelos excessivamente complexos, reduzindo assim o risco de *"sobreajuste"* [45].

Vantagens do XGBoost sobre *Gradient Boosting* simples:

1. Regularização Incorporada:
   - Penalizações L1 (Lasso) e L2 (Ridge)
   - Reduz sobreajuste (sobreajuste)
   - Permite usar mais árvores sem risco excessivo

2. Tratamento de Dados Ausentes:
   - XGBoost aprende automaticamente para que direção enviar valores ausentes
   - Não requer imputação manual de dados

3. Escalabilidade:
   - Permite uma paralelização eficiente
   - Suporte para GPU

- Permite o processamento de bancos de dados (*conjunto de dadoss)* de grandes dimensões.

4. Importância de Variáveis:
   - Fornece as medidas da contribuição de cada *feature*(característica/atributo dos dados que usas para fazer a previsão)
   - Um dos melhores algoritmos em termos de capacidade de interpretação

5. Flexibilidade de Perda:
   - Múltiplas funções de perda disponíveis
   - Personalização possível

Arquitetura XGBoost:

Entrada é constituída por: conjunto de dados (x, y) e por Hiperparâmetros

Inicializar: F0(x) = média(y)

Para cada iteração *k=1* até *K*:
1. Calcular resíduos: $r_i,k = y_i - F_{k-1}(x_i)$
2. Treinar árvore hk para prever resíduos
3. Encontrar melhor taxa de aprendizagem α
4. Atualizar: $F_k(x) = F_{k-1}(x) + \alpha \cdot hk(x)$
5. Atualizar função de perda

Saída: Modelo Final = F0 + h1 + h2 + ... + hK

Hiperparâmetros Críticos:

Hiperparâmetro	Descrição	Intervalo típico
n_estimators	Número de árvores	50–1000
max_depth	Profundidade máxima de cada árvore	3–10
learning_rate (eta)	Taxa de aprendizagem	0.01–0.3
subsample	Fração de amostras por iteração	0.5–1.0
colsample_bytree	Fração de variáveis por árvore	0.5–1.0
lambda (L2)	Regularização L2	0–10
alpha (L1)	Regularização L1	0–10
min_child_weight	Peso mínimo em folha	1–10

Tabela 2: Hiperparâmetros principais do modelo XGBoost e intervalos típicos de valores considerados na afinação dos modelos de temperatura e condutividade[45,69]

Aplicação a Regressão (Qualidade de Água):

Para regressão, a função de perda típica é:

$L(y, \hat{y}) = (y - \hat{y})^2$

Ou, utilizando uma equação mais robusta, baseada em erro absoluto:

$L(y, \hat{y}) = |y - \hat{y}|$

XGBoost minimiza esta perda iterativamente, construindo modelos aditivos.

A aplicação do XGBoost a problemas de qualidade de água baseados na monitorização remota tem sido reportada em estudos recentes, onde modelos de g*radient boosting* apresentam vantagens claras face a SVM e regressão linear na previsão de parâmetros como turbidez e clorofila-a, em particular quando se combinam múltiplas bandas e índices espectrais.

Os resultados obtidos nesta tese alinham-se com essa evidência, mostrando que o XGBoost atinge erros mais baixos e coeficientes de determinação mais elevados para temperatura e condutividade quando comparado com regressão linear, *Random Forest* e SVM, reforçando a adequação desta escolha metodológica [70].

4.2 Validação e Métricas de Desempenho

Para perceber se um modelo funciona bem fora dos dados com que foi treinado, é essencial separar os dados em partes diferentes: treino, validação e teste. Neste trabalho, os "*matchups"* (pares de dados Sentinel-2 e medições in situ no mesmo local e em datas próximas) foram divididos em cerca de 70% para treino, 15% para validação e 15% para teste.

O conjunto de teste deve ficar "intocado" até ao fim, para dar uma medida mais honesta do desempenho final. Tudo o que aprende parâmetros a partir dos dados (por exemplo, normalização/padronização) deve ser feito apenas com o conjunto de treino e depois aplicado da mesma forma à validação e ao teste, para evitar fuga de informação ("*data leakage*"). O mesmo se aplica à escolha de hiperparâmetros: o ajuste deve ser feito com os dados de treino (e validação cruzada), sem usar o teste para tomar decisões.

As métricas usadas foram RMSE, MAE e R^2. O RMSE penaliza mais os erros grandes, o MAE dá um erro médio fácil de interpretar nas mesmas unidades do parâmetro e o R^2 mostra quanta variabilidade o modelo consegue explicar [69,71].

Processamento/Divisão de Dados:

conjunto de dados dividido em:

- Treino (70%): Usado para aprender parâmetros do modelo
- Validação (15%): Usado para ajuste de hiperparâmetros
- Teste (15%): Usado para avaliação final

Validação Cruzada (Cross-Validation)[69,71]:

Técnica para melhorar estimativa de desempenho:

1. Dividir dados em *k* partes (*folds*) (típ. k=5 ou 10);
2. Treinar *k* modelos, cada um usando os k-1 subconjuntos (treino);
3. Avaliar com os dados dado sub-conjunto restante (teste);
4. Para o desempenho final do modelo, calcular a média de métricas nos *k* testes.

Esta técnica reduz a variância em estimativas e usa melhor os dados limitados.

Métricas para um problema de Regressão:

RMSE (Root Mean Squared Error):

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

- Penaliza erros grandes (quadrático)
- Unidade: mesma da variável alvo
- Interpretação intuitiva

MAE (Mean Absolute Error):

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

- Mais robusto a "*outliers"* que RMSE
- Interpretação direta: erro médio em unidades originais

R^2 (Coeficiente de Determinação):

$R^2 = 1 - (SS\_res / SS\_tot)$

Onde:

- $SS\_res = \sum(y_i - \hat{y}_i)^2$
- $SS\_tot = \sum(y_i - \bar{y})^2$

- Varia de 0 a 1 (valores negativos indicam modelo pior que média)
- Interpretação: Fração de variância explicada pelo modelo

4.3 Tratamento de Dados e Pré-processamento

Antes de treinar os modelos, os dados foram limpos e preparados para reduzir erros e incoerências. Este passo é importante para evitar que o modelo aprenda padrões errados por causa de valores em falta, "*outliers"* ou problemas na imagem/sensor [69,71].

4.3.1 Qualidade das medições "*in situ*"

As medições de temperatura e condutividade foram analisadas para detetar valores ausentes e valores muito fora do normal. Quando há leituras repetidas ou medições muito próximas, é importante confirmar se são consistentes, porque leituras erradas podem aumentar muito o erro do modelo. Também deve ser indicado se a condutividade é apresentada "tal como medida" ou já compensada para 25°C (porque isso muda a interpretação e pode criar dependência com a temperatura), o que não aconteceu no nosso caso, uma vez que a condutividade da estação estava compensada e a medida pelo drone também o estava [69,71].

4.3.2 Critérios de "*matchup*" (tempo e espaço)

Os "*matchups"* foram filtrados com base em: (1) uma janela temporal (±3 dias entre imagem e medição), (2) a posição do ponto de medição dentro de um pixel de água e (3) qualidade da imagem (limite de nuvens). A janela de ±3 dias é um compromisso, dá mais

"*matchups"*, mas aumenta o risco de a água já ter mudado entre a imagem e a medição. Para reforçar a robustez, recomenda-se acrescentar uma análise simples que mostre o efeito de janelas menores no número de amostras e no desempenho [69,71].

4.3.3 Variáveis do Sentinel-2 e escala

Foram usadas bandas do Sentinel-2 (B2–B8, B11 e B12) e índices/razões como NDVI, NDCI, NDMI, NDBI, B3/B2 e B4/B8. Estas variáveis ajudam a resumir informação espectral relacionada com fitoplâncton, sedimentos e outros constituintes que podem estar associados, de forma indireta, à temperatura e à condutividade.
Quando as bandas vêm em formato inteiro (por exemplo, 0–10000), deve-se converter para refletância (0–1) e validar a conversão com estatísticas e histogramas. Mesmo que o XGBoost (árvores) não precise de normalização para funcionar bem, se for usada padronização por consistência do pipeline, ela deve ser ajustada apenas no conjunto de treino para evitar fuga de informação [69,71].

4.3.4 Colinearidade e interpretação

Bandas e índices podem estar muito correlacionados entre si. Isso não impede necessariamente boas previsões, mas pode atrapalhar a interpretação da importância das variáveis (o modelo pode "escolher" uma variável e ignorar outra muito parecida). Por isso, é útil discutir importâncias por grupos (visível, red-edge, NIR, SWIR, índices) e, se possível, complementar com métodos como SHAP ou testes de remoção de grupos de variáveis [69,71].

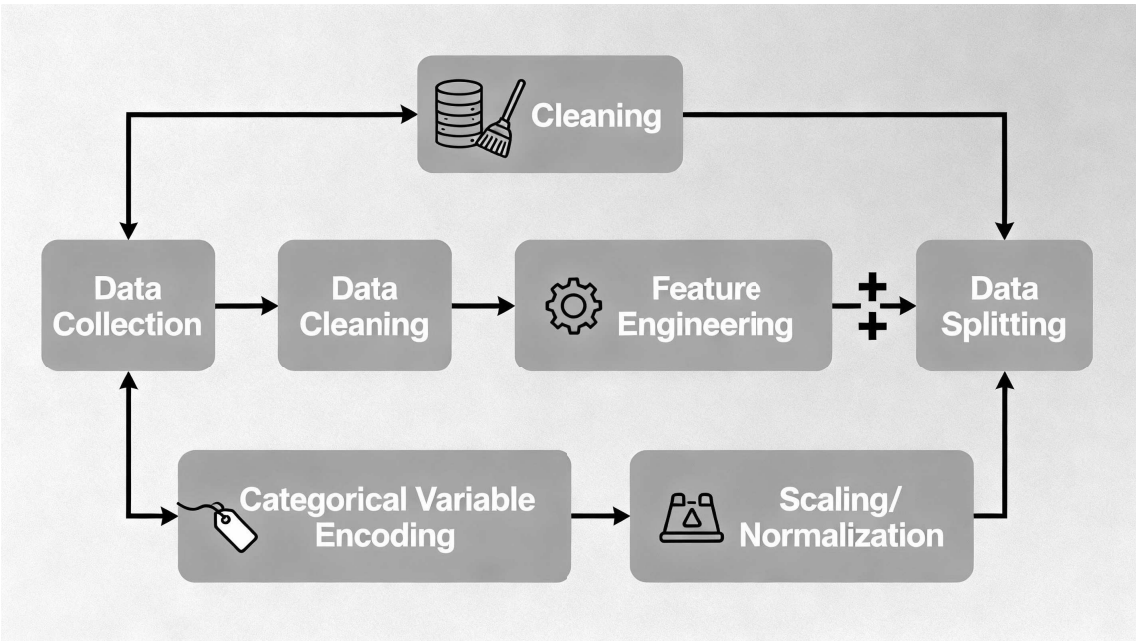

Figura 10: Fluxograma do pré-processamento para XGBoost

Fonte: imagem gerada com apoio de GPT-5 em 22/11/2025.

4.4 Área de Estudo

A área de estudo está focado no Reservatório de Castelo do Bode, localizado no centro de Portugal.

Localização Geográfica:
- Coordenadas: 39°32'N, 8°19'W
- Bacia: Rio Zêzere (afluente do Rio Tejo)
- Área: ~35 km²
- Volume: ~902.5 hm³

O Castelo do Bode é um dos maiores reservatórios em Portugal e tem um papel estratégico na gestão da água e da energia no país. Localiza-se no rio Zêzere, entre os concelhos de Tomar, Abrantes e Ferreira do Zêzere, e está ligado a uma vasta área de abastecimento e de usos múltiplos da água.

Em termos de produção hidroelétrica, o reservatório alimenta uma central importante, contribuindo de forma significativa para a geração de eletricidade a partir de uma fonte renovável e ajudando a estabilizar a rede elétrica em períodos de maior consumo. No abastecimento de água potável, a albufeira é uma das principais origens de água para populações da região de Lisboa e centro do país, pelo que a qualidade da água bruta e a sua monitorização são prioritárias.
Castelo do Bode também é muito usado para recreação, com praias fluviais, desportos náuticos (como vela, canoagem e mota de água) e turismo de natureza, o que o torna um importante polo de lazer e de economia local. Além disso, a albufeira suporta atividades de aquacultura e pesca, formais ou recreativas, contribuindo para a produção de peixe e para a manutenção de comunidades piscícolas que dependem da qualidade ecológica do ecossistema aquático [72].

Apresenta características que o tornam ideal para este estudo:
- Cobertura Sentinel-2 adequada
- Variabilidade sazonal de temperatura e condutividade
- Dados históricos disponíveis
- Fácil acesso para coleta "*in situ*" devido a proximidade do local

Variabilidade Esperada:

- Temperatura: 11°C (inverno) a 26°C (verão) (aproximadamente pelos dados da estação)
- Condutividade: 60-350 μS/cm (intervalo que varia com a mineralogia da bacia, localização e sólidos totais na água na altura da medição)
- Estratificação: Verificada em períodos quentes

4.5 Recolha e Integração de Dados

Dados Sentinel-2:

- Fonte: Copernicus Open Access Hub, Google Earth Engine
- Períodos: Dados históricos 2017-2024, para máxima cobertura temporal
- Processamento: refletância de superfície (L2A), correção atmosférica incluída
- Frequência: Passagens sem cobertura de nuvens (filtro: <30% nuvens)
- Extração: Bandas B2-B8, B11-B12; cálculo de índices espectrais

Dados "*In Situ"*:

- Fonte: Monitorização local com sondas multiparamétricas e de recolha da estação pt16h12c da rede SNIRH(Sistema Nacional de Informação de Recursos Hídricos) e do drone asv1 da Flyrobotics [73-74].
- Localização:
    pt16h12c – latitude: 39,546252 longitude: -8,303988
    asv1 - latitude: 39.551228 longitude: -8.292163
- Frequência: quando possível e determinado por campanha
- Parâmetros: Temperatura (°C), Condutividade (µS/cm)
- Período: 2017-2024

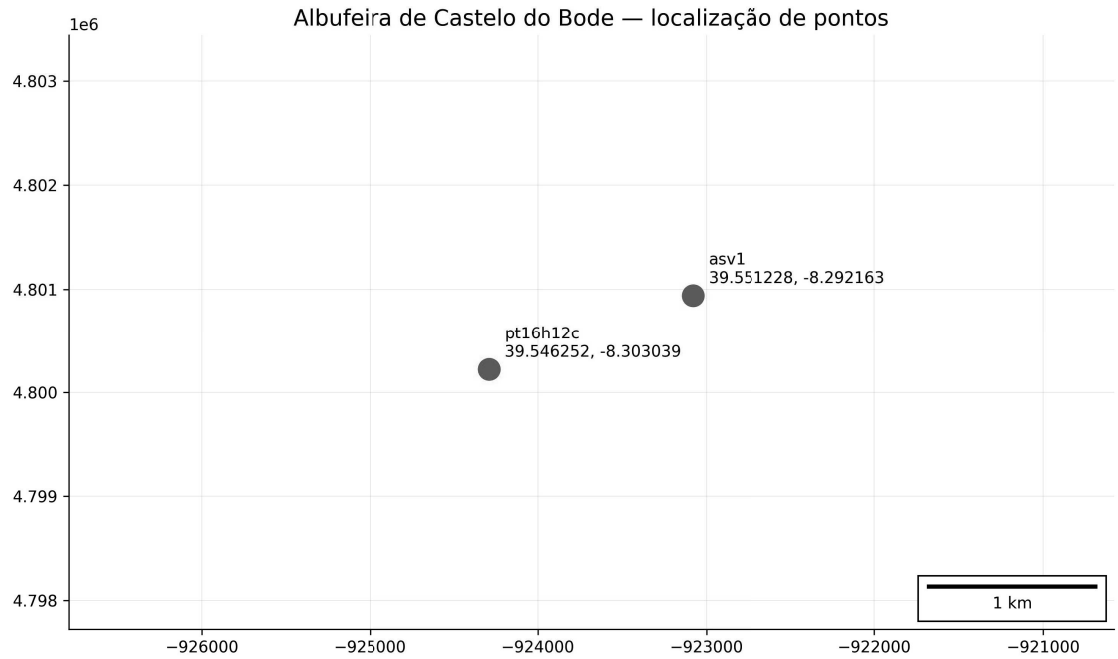

Figura 11: Visualização espacial da distância entre a estação e o drone

Fonte: imagem gerada com apoio de GPT-5 em 27/11/2025.

Figura 12: Drone asv1 da Flyrobotics® perto da ilha do vale do manso na albufeira do Castelo do Bode [74]

Integração espaço-temporal:

- "*Matchup"*: Associar ou fazer a correspondência entre os dados Sentinel-2 com os dados de observações in situ (colocar na primeira aparição)
- Critério temporal: Imagem Sentinel-2 no máximo 3 dias antes/depois da medição
- Critério espacial: Pixel Sentinel-2 (10m/20m) contém ponto de medição
- Validação: Verificar concordância entre múltiplas leituras próximas

Banco de Dados Resultante:

Parâmetro	Valor
Número de *matchups*	98
Período temporal	2017–2024
Variáveis espectrais	12 (bandas B2–B8, B11, B12)
Índices derivados	8 (NDVI, NDCI, NDBI, etc.)
Variáveis alvo	2 (Temperatura, Condutividade)
Cobertura espacial	2 estações

Tabela 3: Características do banco de dados resultante de "matchups" satélite–terrestre

4.6 Planeamento de amostragens em função da trajetória do Sentinel-2

A sincronização entre as campanhas de amostragem "*in situ"* e as passagens do Sentinel-2 é crítica para reduzir a incerteza temporal e maximizar a qualidade dos "*matchups"* satélite-terrestre. Neste trabalho, o planeamento das campanhas de campo teve em conta o ciclo de tempo de revisita de 5 dias do Sentinel-2, bem como as efemérides das órbitas sobre o Reservatório de Castelo do Bode, de forma a concentrar as medições em janelas temporais de ±3 dias em torno das imagens sem cobertura significativa de nuvens.

A estratégia planeada consistiu em gerar, para horizontes de 15 a 30 dias, um calendário de passagens potencialmente úteis do Sentinel-2, combinando resolução temporal, previsões de cobertura de nuvens e restrições operacionais de acesso ao reservatório.

As campanhas foram agendadas para coincidir com o dia da passagem do satélite ou o primeiro dia disponível na janela de tolerância temporal (±3 dias), garantindo representatividade espacial do píxel Sentinel-2 e repetibilidade nas estações A (pt16h12c) e B (drone asv1).

Este planeamento visou aumentar o número de *matchups* válidos e minimizar variabilidade não explicada por mudanças rápidas nas condições da água, reforçando a robustez estatística dos modelos XGBoost.

4.7 Processamento de Imagens Sentinel-2

Pipeline de Processamento:

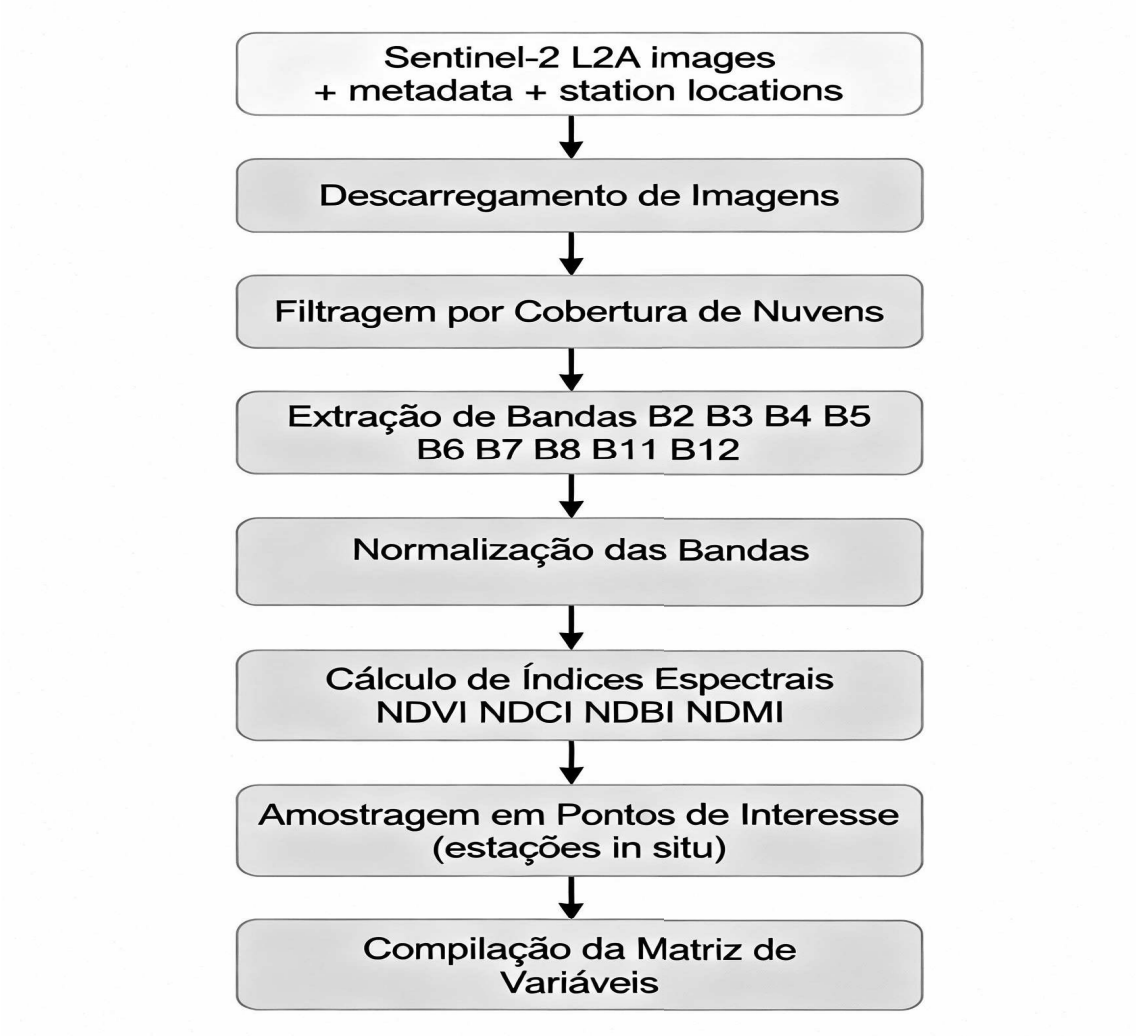

Figura 13: Pipeline de processamento de imagens Sentinel-2 para cálculo de índices espectrais e extração de variáveis

Fonte: imagem gerada com apoio de GPT-5 em 27/11/2025.

Índices Espectrais Calculados:

Para captar diferentes aspectos da qualidade da água, calculou-se:

Índice	Fórmula	Significado
NDVI	(B8-B4)/(B8+B4)	Vegetação aquática, fitoplâncton
NDCI	(B6-B4)/(B6+B4)	Clorofila-a
NDBI	(B11-B8)/(B11+B8)	Sólidos suspensos, água turbulenta
NDMI	(B8-B11)/(B8+B11)	Conteúdo de água, humidade
Razão B3/B2	B3/B2	Verde/Azul: sensibilidade a sedimentos
Razão B4/B8	B4/B8	Red/NIR: fitoplâncton

Tabela 4: Índices espectrais calculados a partir de bandas Sentinel-2 e respetiva interpretação onde se incluem indicadores indiretos como descritores espectrais.

4.7.1 Pré-processamento e Extração de Índices Sentinel-2 com o auxílio Google Earth Engine (GEE)

GEE fornece acesso a repositório de imagens Sentinel-2 e capacidade de processamento distribuído [21].

No Apêndice B, encontra-se o código define primeiro um ponto e um buffer de 5 km à volta para ser a área de estudo, depois carrega a coleção Sentinel-2 de refletância de superfície entre 2017 e 2024, apenas dentro dessa região e com menos de 30% de nuvens. Em seguida, para cada imagem, calcula três índices (NDVI, NDCI e NDBI) a partir das bandas apropriadas e adiciona esses índices como novas bandas. Depois amostra, em cada imagem, os valores de todas as bandas e índices dentro da região com resolução de 20 m,

junta todas essas amostras numa única tabela e, por fim, exporta essa tabela condensada em formato CSV com o nome S2_data_Castelo.

4.8 Construção de Modelos XGBoost

Foram treinados dois modelos separados: um para prever temperatura e outro para prever condutividade. O processo foi o mesmo para ambos e segue uma ordem simples: preparar dados, dividir em treino/validação/teste, escolher hiperparâmetros, treinar com "*early stopping"* e avaliar no teste.
Primeiro, escolhem-se as colunas de entrada (bandas e índices) e define-se a variável alvo (temperatura ou condutividade). A divisão treino/validação/teste deve acontecer antes de qualquer transformação que "aprenda" valores, para evitar fuga de informação.
Depois, é feita a procura de hiperparâmetros usando apenas o treino, normalmente com validação cruzada interna. A grelha inclui parâmetros que controlam a complexidade do modelo (número de árvores e profundidade), a taxa de aprendizagem e regularização, para reduzir sobreajuste.
Por fim, o modelo é treinado com "*early stopping"* usando o conjunto de validação, parando quando deixa de melhorar. No final, o desempenho que conta para reportar é o do conjunto de teste, porque é o que simula melhor "dados novos".

No Apêndice C, indica-se o desenvolvimento para treinar e avaliar dois modelos de regressão com XGBoost a partir de um ficheiro CSV: um modelo para estimar a temperatura da água e outro para estimar a condutividade, usando como variáveis de entrada bandas e índices derivados (por exemplo NDVI, NDCI, NDMI, etc.). Primeiro, o código lê o CSV e converte a coluna do tempo para formato de data/hora, e depois procura no conjunto de dados uma coluna que indique a partição já definida para separar automaticamente os registos de "treino" e de "teste".
A seguir, dentro do conjunto de treino, é criada uma partição adicional para validação: como a ideia é ter um esquema 70/15/15 (treino/validação/teste), e o "treino" já corresponde a 70% do total, o script retira do treino uma fração equivalente a 15% do total (isto é, 15/70 do treino) para formar o conjunto de validação. Antes de treinar, remove

linhas com valores em falta no alvo, garantindo que o modelo não falha por causa de NaNs.

Para escolher os melhores hiperparâmetros, o código faz uma escolha de hiperparâmetros apenas com o subconjunto de treino (sem tocar na validação nem no teste), usando validação cruzada a 5 folds e RMSE como critério de seleção. Esta pesquisa é feita através de um processo que inclui um transformador de escala seguido de um regressor XGB, o que é importante porque garante que o escalonamento é ajustado dentro de cada fold da validação cruzada e assim evita "data leakage" (informação do conjunto de validação a influenciar o pré-processamento do conjunto de treino).

Depois de encontrados os melhores hiperparâmetros, o script faz um treino final "limpo": ajusta o transformador de escala no treino, aplica a transformação à validação e ao teste, converte os dados para DMatrix e treina com a API nativa do XGBoost usando *early stopping*. O *early stopping* pára o treino quando o RMSE na validação deixa de melhorar durante um certo número de iterações, o que ajuda a reduzir sobreajuste e a escolher automaticamente um número eficaz de rondas de boosting.

Por fim, para cada alvo (temperatura e condutividade), o código guarda em disco o scaler e o modelo treinado (.joblib), calcula métricas de desempenho (RMSE, MAE e R^2) em treino, validação e teste, imprime um resumo no ecrã e grava um CSV final (training_results.csv) com os resultados e os melhores parâmetros encontrados.

4.9 Validação e Avaliação de Modelos

Como descrito no apêndice D, existem três etapas finais da avaliação do modelo de temperatura: calcula previsões e métricas de desempenho em cada subconjunto (treino, validação, teste), analisa os resíduos (erros) e estima a importância de cada variável no modelo XGBoost.

Primeiro, o modelo gera as previsões de temperatura para o conjunto de treino, para o conjunto de validação e para o conjunto de teste. Em seguida, a função das métricas recebe os valores reais, os valores previstos e o nome do conjunto, e devolve três métricas: RMSE (raiz do erro quadrático médio), MAE (erro absoluto médio) e R^2 (coeficiente de

determinação), calculadas com as funções padrão do scikit-learn. Essas métricas são calculadas para treino, validação e teste, guardadas num ficheiro temporário, convertidas em conjunto e impressas numa tabela, o que permite comparar rapidamente se o modelo está a sobreajustar (erro muito baixo em treino e muito grande em teste) ou a generalizar bem.

Depois, o código analisa os erros no conjunto de teste, definidos como resíduos (diferença entre entre os dados previstos e os do conjunto de teste), ou seja, valor real menos valor previsto. Com estes resíduos, produz um histograma (para ver a distribuição dos erros) e um gráfico de dispersão resíduos vs. valores previstos (para verificar se há padrões sistemáticos, como tendência de erro maior para temperaturas altas ou baixas); idealmente, os resíduos devem estar centrados em zero e sem padrão evidente em função dos valores previstos. Em seguida, aplica o teste de shapiro–wilk (shapiro(resíduos)), obtém o valor de p e interpreta, se o valor de p for maior que 0.05, considera que os resíduos são aproximadamente normais. Já se o valor de p for menor que 0.05, rejeita a normalidade. Este teste é uma forma clássica de verificar a hipótese de normalidade dos resíduos em modelos de regressão.

Por fim, o código calcula a importância global de cada variável explicativa no modelo XGBoost. Esta função contém, para cada variável de entrada (cada coluna de variáveis), uma medida de contribuição para as divisões das árvores do modelo; quanto maior o valor, mais aquela variável ajudou a reduzir o erro durante o treino. A lista resultante ordena os índices das variáveis por importância decrescente, e o gráfico de barras horizontal mostra visualmente quais são as variáveis mais influentes (no eixo vertical, os nomes das variáveis; no horizontal, o valor de importância). Em seguida, a tabela de importância organiza estes resultados, listando para cada variável o valor bruto de importância e a respetiva percentagem em relação ao total, o que facilita a leitura e a descrição no texto: por exemplo, pode-se dizer que “a variável NDVI contribui com x % da importância total do modelo”.

4.9.1 Análise exploratória de dados (EDA)

Antes da fase de modelação, foi realizada uma análise exploratória de dados (Exploratory Data Analysis, EDA) com o objetivo de caracterizar as distribuições de temperatura e condutividade, bem como a estrutura de correlações entre as variáveis espectrais do Sentinel-2 e os parâmetros de qualidade da água.

Esta análise incluiu histogramas, boxplots e estatísticas descritivas (média, desvio padrão, assimetria) para as variáveis alvo, permitindo identificar outliers e padrões sazonais relevantes, em linha com as recomendações clássicas de EDA.

Adicionalmente, foi calculada uma matriz de correlação de Pearson entre bandas espectrais, índices derivados (NDVI, NDCI, NDMI, razões B3/B2 e B4/B8) e as medições "*in situ"*, com representação gráfica através de um mapa de calor para facilitar a interpretação.

A matriz de correlação evidenciou forte colinearidade entre bandas (incluindo NIR/SWIR) e relações coerentes entre índices (NDVI/NDMI vs. NDBI); a relevância preditiva de bandas/índices para os alvos foi avaliada posteriormente através das importâncias do XGBoost.

A EDA constituiu também um passo essencial para definir critérios de remoção ou tratamento de outliers, assegurando que valores manifestamente inconsistentes (por exemplo, leituras isoladas fora de intervalos fisicamente plausíveis) não enviesassem a estimação dos modelos nem a avaliação de desempenho, conforme preconizado na literatura de EDA e de estatística aplicada à ciência de dados, a descrição do código está no apêndice G [75].

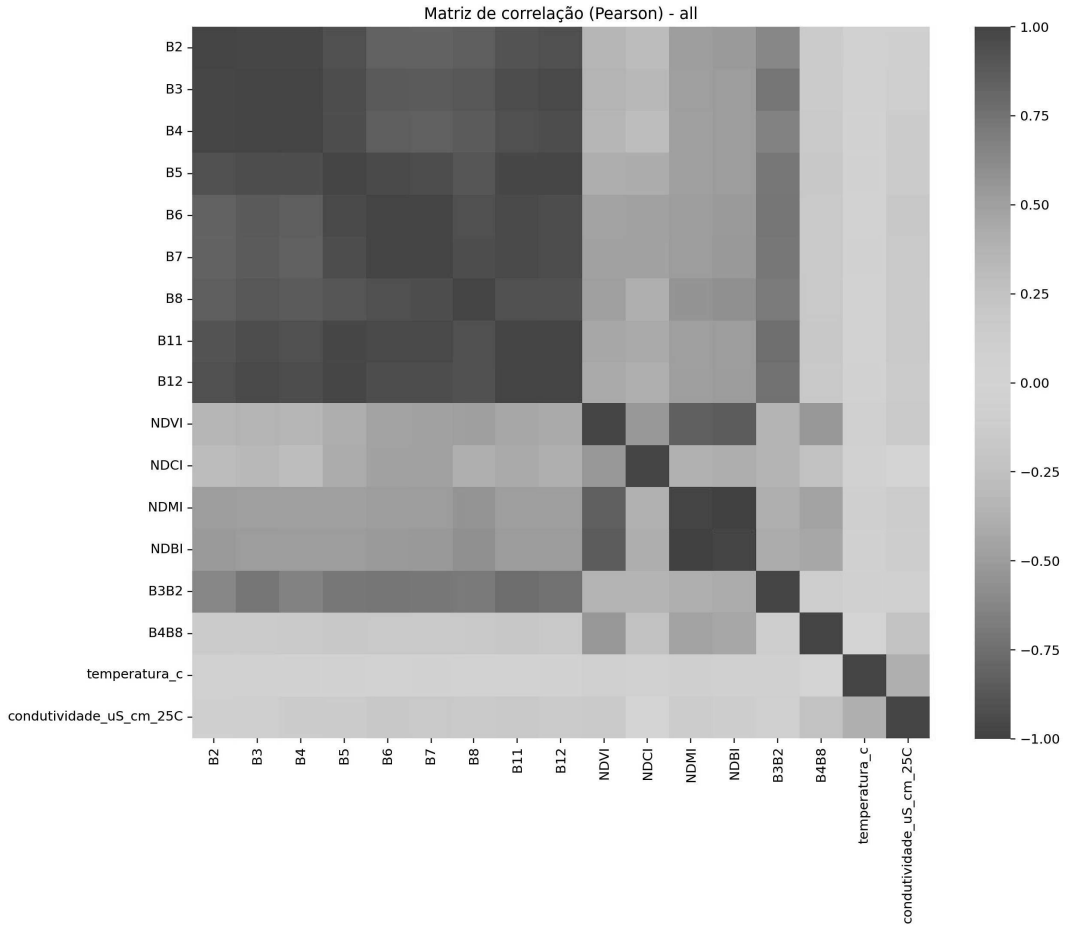

Figura 14: Matriz usado a correlação de Pearson para o dataset onde se usa os dados dodrone com treino e os da estação como teste

A matriz mostra muita colinearidade entre as bandas espectrais (B2–B8, B11, B12): quase todas estão fortemente e positivamente correlacionadas entre si (blocos vermelhos), o que indica variáveis muito redundantes.

Os índices derivados comportam-se como esperado: NDVI/NDCI/NDMI tendem a correlacionar-se positivamente entre si e com bandas do NIR/red-edge, enquanto o NDBI aparece em oposição (correlações negativas, tons azuis) sobretudo face a NDVI/NDMI, sugerindo que capta um "sinal" diferente.

Em relação às variáveis-alvo, temperatura e condutividade, têm correlações lineares fracas a moderadas com a maioria das bandas/índices (cores perto do branco), e entre si parecem ter uma correlação positiva moderada (tom alaranjado).

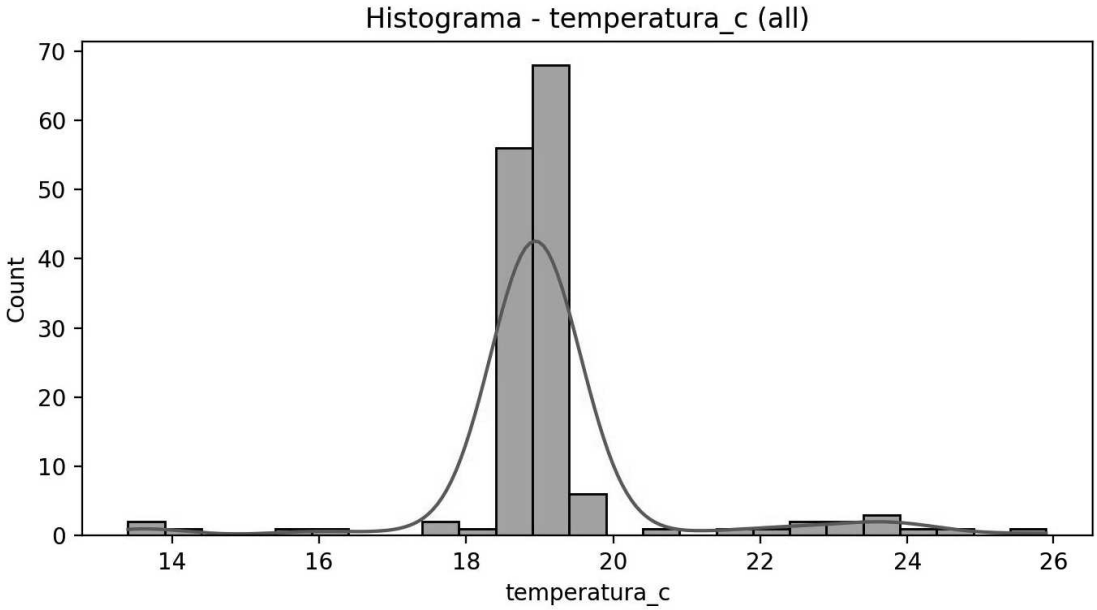

Figura 15: Histograma da temperatura

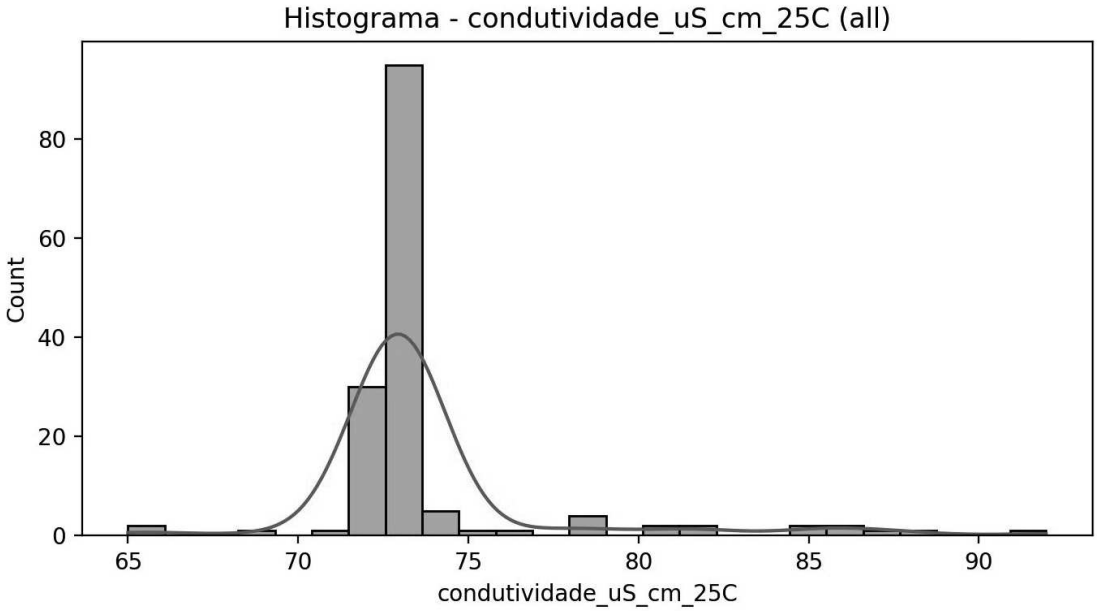

Figura 16: Histograma da condutividade

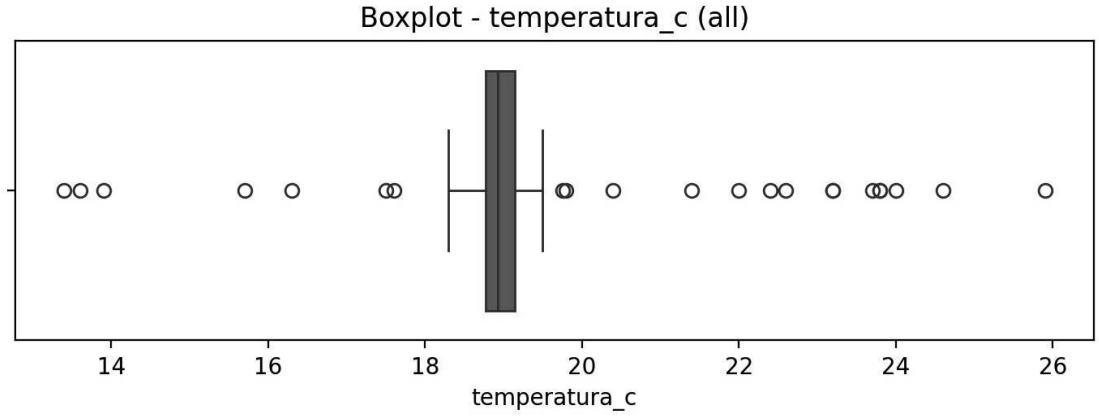

Figura 17: Boxplot para a temperatura

O boxplot indica que a maior parte das observações de temperatura está muito concentrada perto de ~19 °C: a mediana está aproximadamente nos 19 °C e o intervalo interquartil é estreito, sugerindo baixa variabilidade no "miolo" dos dados.

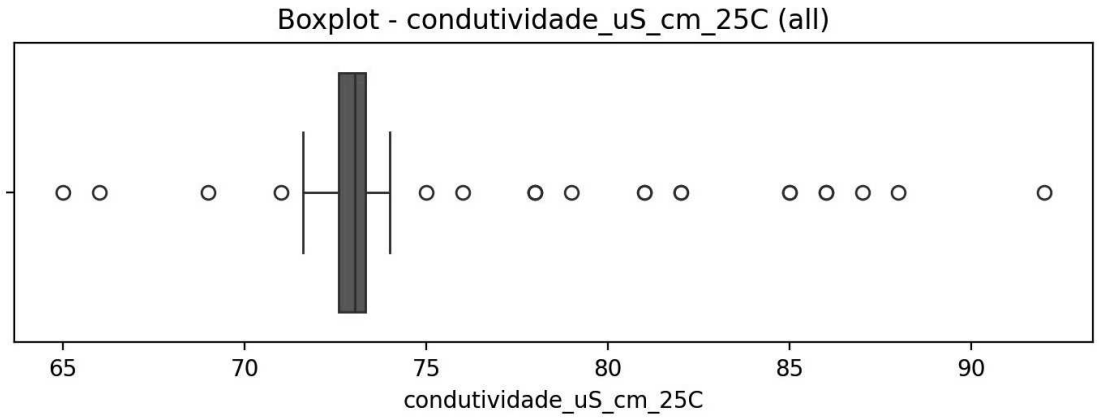

Figura 18: Boxplot para a condutividade

A figura acima mostra que a condutividade está muito concentrada em torno de ~73 µS/cm (a mediana está perto desse valor) e que o intervalo interquartil é estreito, o que indica pouca variabilidade na maioria das observações.

Isto pode refletir eventos com maior mineralização/suspensão de solutos ou diferenças espaciais e mesmo questões operacionais, mas também pode incluir leituras anómalas.

Com a metodologia traçada e o método implementado, o Capítulo 5 transita para a implementação de uma plataforma de visualização dos dados. Descrevemos a área de estudo, os procedimentos de recolha de dados, o processamento de imagens Sentinel-2, e a construção efetiva dos modelos XGBoost no contexto do Reservatório de Castelo do Bode, um caso de estudo emblemático para a monitorização ambiental em Portugal.

Capítulo 5 - Plataforma IoT para qualidade da água com Sentinel-2

5.1 Arquitetura do Sistema

O sistema proposto integra múltiplas camadas:

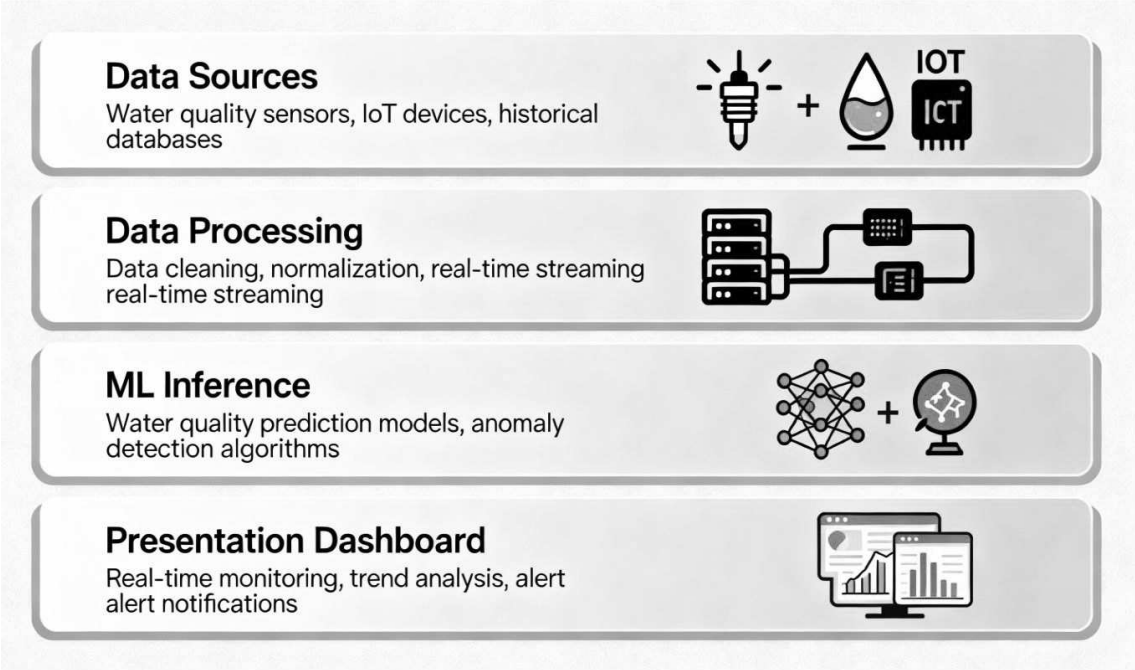

Figura 19: Arquitetura do sistema de monitorização de qualidade de água integrando Sentinel-2, IoT, modelos XGBoost e dashboard web

Fonte: imagem gerada com apoio de GPT-5 em 27/11/2025.

Componentes:

1. Fontes de Dados:
   - Imagens Sentinel-2 via Copernicus/GEE
   - Sensores IoT locais (sondas multiparamétricas)
   - Base de dados histórica

2. Processamento:
- Google Earth Engine: Extração de imagens, cálculo de índices
- Python: Pré-processamento, feature engineering
- Matchup: Sincronização temporal de dados

3. Modelação e Inferência:
- Modelos XGBoost treinados para temperatura e condutividade
- Validação cruzada e testes de robustez
- Armazenamento de modelos em formato .joblib
- API REST (Flask/FastAPI) para predições em tempo real
- Reutilização de resultados previamente calculados

4. Visualização:
- Dashboard Streamlit [24]
- Mapas de qualidade de água
- Séries temporais e alertas

5.2 Processamento em Google Earth Engine

Google Earth Engine fornece computação geoespacial em larga escala [76].

Vantagens:
- Acesso a repositório completo Sentinel-2 (~8 petabytes)
- Processamento distribuído (rápido)
- APIs em JavaScript e Python
- Integração com Google Cloud Storage

Usou-se o Google Earth Engine (https://code.earthengine.google.com/) para obter imagens Sentinel-2 sobre a albufeira de Castelo do Bode, calcular vários índices espectrais e exportar todos esses valores para um ficheiro CSV no Google Drive.
Este script no Google Earth Engine pega num conjunto de pontos com medições *in situ* (temperatura, condutividade e metadados) e, para cada ponto, procura imagens Sentinel-2 L2A (COPERNICUS/S2_SR_HARMONIZED) numa janela de ±3 dias, já filtradas por

área, datas e percentagem máxima de nuvens. Em cada imagem, aplica uma máscara de nuvens baseada na classe SCL, converte as bandas para refletância (dividindo por 10000) e calcula índices/rácios (NDVI, NDCI, NDMI, NDBI, B3/B2 e B4/B8), ficando com um conjunto final de variáveis espectrais.

Depois, faz a mediana das imagens existentes nessa janela temporal e extrai, no local do ponto (escala 20 m), os valores das bandas/índices, guardando também quantas imagens contribuíram (contagem_s2) e as datas de início/fim da janela. Por fim, exporta a tabela completa (dados in situ + variáveis Sentinel-2 + metadados e geometria) em formato csv.

Na prática, é uma forma de automatizar a criação de uma série temporal de variáveis espectrais derivadas das imagens Sentinel-2, em pontos específicos da albufeira, já pronta para análise posterior.

5.3 Desenvolvimento de Modelos em Python

O desenvolvimento dos modelos de previsão foi realizado em Python, recorrendo a bibliotecas de ciência de dados e aprendizagem automática (p. ex., pandas, NumPy, scikit-learn e XGBoost), com o objetivo de tornar o processo reprodutível desde a leitura dos "*matchups*" até à disponibilização do modelo para uso operacional. Neste trabalho são treinados dois modelos de regressão independentes: um para estimar a temperatura da água e outro para estimar a condutividade, usando como variáveis de entrada bandas e índices espectrais derivados do Sentinel-2.

O ponto de partida é um ficheiro csv com a correspondência satélite–terrestre (variáveis Sentinel-2 e medições "in situ"), obtido a partir do processamento descrito na secção anterior. As variáveis explicativas incluem bandas do Sentinel-2 e índices/razões espectrais (por exemplo NDVI, NDCI, NDMI, NDBI, B3/B2 e B4/B8), selecionados por serem descritores compactos do comportamento espectral e, por isso, potencialmente úteis como indicadores indiretos para parâmetros que não são diretamente "óticos", como a condutividade.

Antes do treino, os dados são preparados de forma consistente para os dois alvos, selecionam-se as colunas de entrada e a variável-alvo (temperatura ou condutividade), removem-se registos com valores em falta nas colunas relevantes e confirma-se a coerência de unidades e intervalos. Quando se aplica normalização/padronização por motivos de uniformização do pipeline, esta transformação é ajustada apenas com os dados

de treino e aplicada de igual modo aos conjuntos de validação e teste, para evitar fuga de informação. A avaliação assenta numa separação treino/validação/teste, mantendo o teste intocado até ao fim, de modo a obter uma estimativa mais realista do desempenho em dados não vistos.

O treino e a seleção de hiperparâmetros estão consolidados no que se descreve no apêndice C, onde é realizado um processo de afinação apenas com o subconjunto de treino, com validação cruzada, e onde o treino final beneficia de "early stopping" com base no conjunto de validação para reduzir o risco de sobreajuste. No final, os artefactos necessários para reutilização do modelo são guardados em disco em formato .joblib (modelo e transformador de escala, quando aplicável), bem como um ficheiro de resultados (por exemplo trainingresults.csv) com métricas e parâmetros usados, para garantir rastreabilidade e facilitar comparação entre experiências. A análise detalhada do desempenho (métricas por subconjunto, análise de erros/resíduos e importância de variáveis) encontra-se no apêndice D, suportando a interpretação do que o modelo está efetivamente a aprender.

Depois de treinados e persistidos em ficheiros .joblib, os modelos podem ser carregados por componentes de inferência para produzir previsões sob pedido, sem necessidade de repetir o processo de treino. A solução principal de visualização e interação com o utilizador é um dashboard desenvolvido em Streamlit, descrito no apêndice E, onde é possível explorar resultados, testar cenários e obter previsões a partir das variáveis de entrada. Como alternativa opcional, para cenários em que se pretenda integrar as previsões com outros sistemas (por exemplo, serviços externos ou automações), é disponibilizada uma API REST, descrita no apêndice F.

5.4 Integração de IoT e Tecnologias Web

No apêndice F, descreve-se duas formas de usar os modelos de aprendizagem de máquina de qualidade de água: uma API REST com Flask e um *"dashboard"* interativo com Streamlit.

Na parte da API, no apêndice F, uma aplicação Flask é descrita, carrega a partir de ficheiros dois modelos previamente treinados (um para temperatura e outro para condutividade) e expõe um endpoint /predict que recebe um pedido POST em JSON com as bandas e índices espectrais (B2–B12, NDVI, NDCI, NDBI, NDMI, B3_B2, B4_B8). O código verifica se todas as variáveis necessárias foram enviadas, organiza esses valores numa lista, na ordem correta e chama os dois modelos para obter as previsões de temperatura e condutividade. Em seguida devolve um json com uma data, as duas previsões e um campo status. Existe ainda um "*endpoint"* simples /health que apenas indica se a API está a funcionar, sendo útil para verificar o estado do serviço e para monitorização básica do sistema.

Na parte do Streamlit, apêndice E, o código constrói um "*dashboard"* web para interagir com os modelos sem necessidade de chamadas diretas à API. Primeiro configura a página, define o título e um subtítulo, e carrega os mesmos modelos a partir de ficheiros. Depois organiza a interface em três separadores: "Dashboard", "Predição" e "Análise". No separador de Dashboard, mostra alguns indicadores de exemplo (temperatura média, condutividade média, número de amostras e última atualização). No separador de Predição, disponibiliza campos para o utilizador escolher valores das bandas Sentinel-2 (B2 a B12); a partir dessas bandas, calcula automaticamente os índices NDVI, NDCI, NDBI, NDMI e as razões B3/B2 e B4/B8. Os valores são agregados numa lista e, quando o utilizador clica no botão "Prever Parâmetros", os modelos são chamados e as previsões de temperatura e condutividade são mostradas de forma destacada. No terceiro separador, "Análise", já está preparado o espaço para futuramente mostrar gráficos de importância das variáveis para cada modelo, embora os gráficos ainda não estejam implementados no código apresentado.

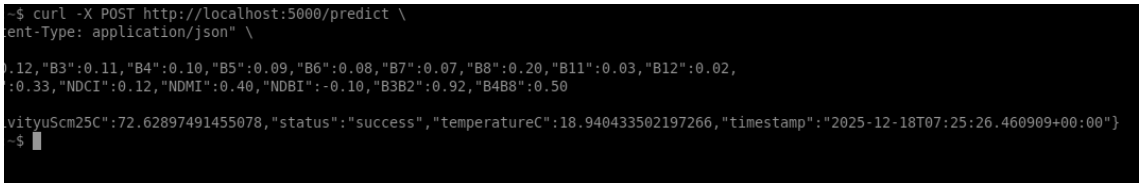

Figura 20: Chamada à API usando o endpoint /predict

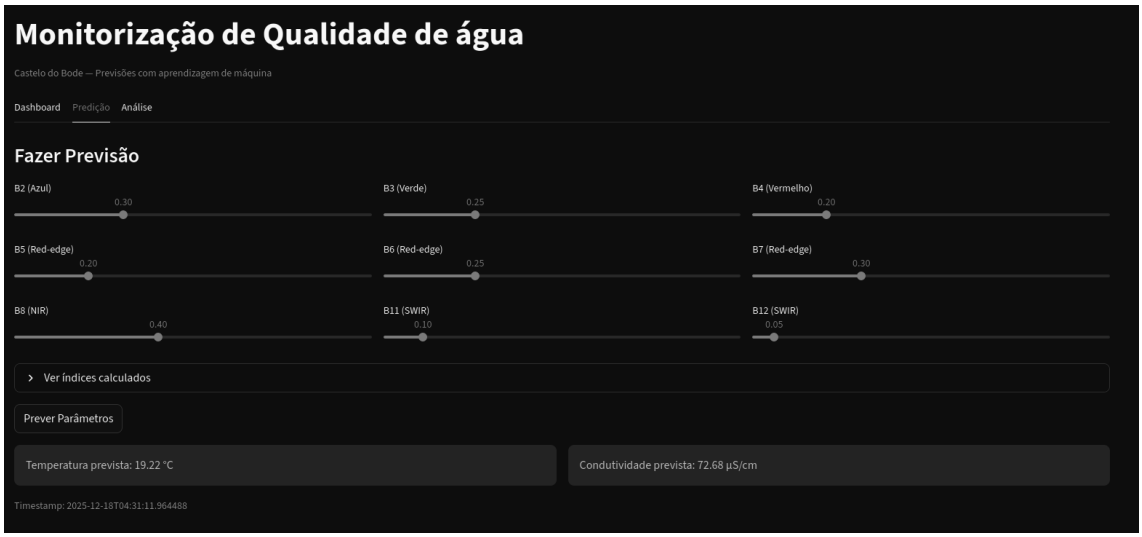

Figura 21: Captura de ecrã do protótipo do dashboard em Streamlit

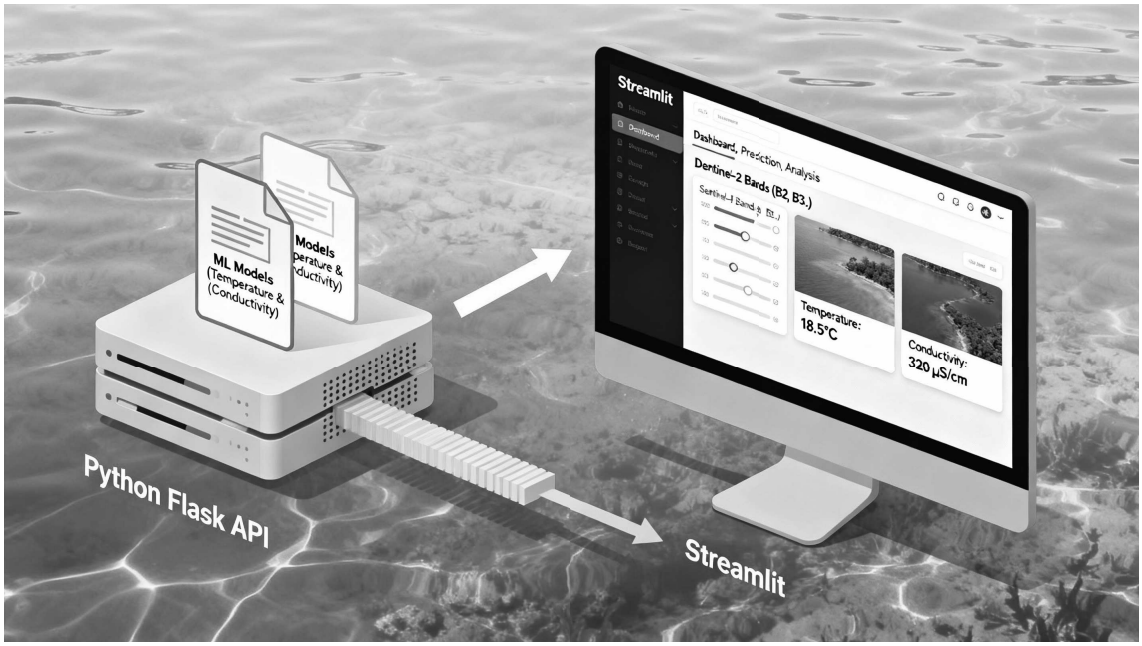

Figura 22: Arquitetura do sistema de estimativa visual de qualidade da água

Fonte: imagem gerada com apoio de GPT-5 em 12/12/2025.

5.5 Procedimentos de Matchup Satélite-Terrestre

Procedimento crítico para garantir qualidade dos dados de treino.

Critérios de "*Matchup"*:

1. Critério Temporal: Imagem Sentinel-2 no máximo ±3 dias da medição "*in situ*"
   - Justificativa: Minimiza mudanças nas condições da água

2. Critério Espacial: Pixel Sentinel-2 (20m x 20m) contém local de medição
   - Usar coordenadas GPS precisas do ponto de amostragem
   - Verificar se píxel não está parcialmente fora da água

3. Critério de Qualidade de Imagem:
   - Cobertura de nuvens < 30% no pixel específico
   - Sem artefatos de processamento
   - Ângulo de visão < 5°

4. Critério de Qualidade de Dados *In Situ*:
   - Medições replicadas (2-3 leituras) para validar
   - Sem anomalias óbvias (temperatura 5-30°C é normal)
   - Equipamento calibrado recentemente

No apêndice G, faz-se a associação entre medições "*in situ"* e dados de satélite que estejam próximos no tempo e no espaço, no caso em que tenhamos um ficheiro de extração e um ficheiro de medições, também serve para demonstrar em maior detalhe os critérios de seleção. A função recebe dois conjuntos de dados. Um com dados in situ (data, latitude, longitude, temperatura, condutividade) e outro com dados de satélite (data, coordenadas, bandas e índices). Para cada registo "*in situ"*, o código primeiro filtra as observações de satélite que estão dentro de uma tolerância temporal em dias. Depois, para esses candidatos, calcula a distância geodésica entre as coordenadas in situ e as do satélite, em metros, e mantém apenas os que estão dentro da tolerância espacial.
Quando encontra um par que respeita essas condições, cria um dicionário com as datas *in situ* e satélite, as diferenças temporal e espacial, as coordenadas, os valores de temperatura e condutividade medidos no local, e copia também todas as bandas e índices espectrais

disponíveis (B2–B12, NDVI, NDCI, etc.) da linha de satélite correspondente. Todos esses registos são acumulados numa lista e convertidos no fim para um conjunto de dados de *matchup*. Por último, o código imprime estatísticas básicas sobre o número total de "*matchups"* e as distribuições das diferenças temporal e espacial, e devolve o conjunto de dados pronto para análise ou modelação.

Capítulo 6 - Resultados e Análise

6.1 Performance do Modelo XGBoost para Temperatura

Configuração do Modelo:

Após "*grid search"*(técnica de otimização de hiperparâmetros) em 9720 combinações de hiperparâmetros, os melhores foram:

Hiperparâmetro	**Valor**
n_estimators	200
max_depth	7
learning_rate	0.03
subsample	0.8
colsample_bytree	0.8
lambda	0
alpha	0

Tabela 5: Hiperparâmetros do modelo XGBoost, utilizados nas previsões de temperatura

Métricas de Desempenho:

Conjunto	**RMSE**	**MAE**	**R^2**
Treino	0,065 °C	0,038 °C	0.912
Validação	0,104°C	0,071 °C	0.697
Teste	3,8 °C	3,337 °C	-0,147

Tabela 6: Métricas de desempenho do modelo XGBoost, para a temperatura

Interpretação:

Os resultados para temperatura mostram que o modelo funciona muito bem no treino e na validação (erros baixos e R2 elevado), o que indica que consegue aprender a relação entre as variáveis do Sentinel-2 e a temperatura quando os dados são semelhantes aos usados para treinar. No entanto, no teste a performance cai bastante (RMSE e MAE muito mais altos e R2 negativo), sugerindo que o modelo não está a generalizar bem para esse conjunto.
Isto acontece, muito provavelmente, porque o teste tem uma gama de temperaturas bem mais ampla do que o treino/validação. Assim, o modelo acaba por ser "obrigado" a extrapolar para valores que praticamente não viu durante o treino, o que aumenta o erro e faz com que o R^2 fique abaixo de zero.

O modelo de temperatura foi ajustado com procura de hiperparâmetros e treinado com "*early stopping"*, sendo depois avaliado em treino, validação e teste (Tabela 6). Antes de tirar conclusões, confirmou-se que as métricas estão corretas e coerentes entre si (por exemplo, o MAE não pode ser maior do que o RMSE), porque erros de cálculo ou de transcrição podem acontecer.

6.2 Performance do Modelo XGBoost para Condutividade

Hiperparâmetros similares ao modelo de temperatura.

Métricas de Desempenho:

Conjunto	**RMSE**	**MAE**	**R²**
Treino	0,198 µS/cm	0,133 µS/cm	0.862
Validação	0,300 µS/cm	0,240 µS/cm	0.316
Teste	8,980 µS/cm	7,269 µS/cm	-0,655

Tabela 7: Métricas de desempenho do modelo XGBoost, para a condutividade

O modelo de condutividade foi treinado da mesma forma e avaliado com as mesmas métricas (Tabela 7). Tal como na temperatura, confirmou-se a coerência das métricas e das unidades antes de interpretar os valores.

A condutividade é mais difícil de estimar a partir do satélite porque não é um parâmetro ótico. Na prática, o modelo aprende relações indiretas (por exemplo, com sedimentos e matéria dissolvida), e essas relações podem mudar com a época do ano e com o tipo de água.

Interpretação:

Os resultados para a condutividade mostram um padrão semelhante ao da temperatura: o modelo ajusta bem no treino (RMSE 0,198 e R^2 0,862) e ainda mantém alguma capacidade na validação, embora já com quebra clara (R^2 0,316). No entanto, no teste a performance degrada-se bastante (RMSE 8,980, MAE 7,269 e R^2 -0,655), o que indica que o modelo não está a generalizar bem para o conjunto de teste.

A explicação mais provável é, novamente, a diferença grande entre os dados usados para treinar/validar e os dados do teste: no treino e na validação a condutividade está muito concentrada (cerca de 71,6–73,8 µS/cm), enquanto no teste varia muito mais (65–92 µS/cm). Assim, o modelo aprende relações num intervalo estreito e, quando é avaliado num teste com valores bem mais dispersos, acaba por extrapolar e perde desempenho, levando ao R^2 negativo.

6.3 Análise Comparativa com Métodos "Tradicionais"

Alvo	Modelo	RMSE (val)	MAE (val)	R^2 (val)	RMSE (teste)	MAE (teste)	R^2 (teste)
Temperatura	Baseline (média)	1.164	0.476	-0.010	8.081	4.676	-0.410
Temperatura	Ridge	1.271	0.550	-0.203	7.945	4.375	-0.362
Temperatura	SVR (RBF)	1.643	0.498	-1.012	8.132	5.014	-0.427
Temperatura	Random Forest	1.350	0.515	-0.359	8.109	5.132	-0.419
Temperatura	Extra Trees	1.306	0.488	-0.272	8.120	4.922	-0.423
Condutividade	Baseline (média)	2.853	1.193	-0.026	236.772	235.297	-79.496
Condutividade	Ridge	2.662	1.162	0.107	238.664	237.177	-80.788
Condutividade	SVR (RBF)	2.757	1.204	0.042	238.452	237.100	-80.643
Condutividade	Random Forest	2.114	0.700	0.437	237.138	235.739	-79.745
Condutividade	Extra Trees	2.266	0.778	0.353	237.988	236.641	-80.325

Tabela 8: Análise comparativa de métodos de referência

Nos testes com os modelos "tradicionais" (Ridge, SVR, Random Forest, Extra Trees e baseline), a temperatura ficou sempre muito semelhante e fraca, com RMSE perto de 8 °C e R^2 negativo, e na condutividade o cenário foi ainda pior, com RMSE ~237 µS/cm e R^2 perto de -80, o que indica que o problema principal é a mudança de contexto/fonte e não o algoritmo em si, na medida em que os valores de teste é de uma fonte completamente distinta.

No último cenário, o XGBoost acabou por se comportar melhor, sobretudo na temperatura, onde baixou o erro do teste para RMSE 3,85 °C (ainda com R^2 negativo), mostrando que é mais robusto do que essas alternativas, embora a diferença entre fontes continue a ser o maior desafio.

6.4 Testes de Robustez

Teste de Robustez com Dados Ruidosos:

Adicionar ruído gaussiano (±5%, ±10%, ±15%) aos dados de teste:

Nível de ruído	**Degradação RMSE Temp.**	**Degradação RMSE Cond.**
0%	referência	referência
5%	+0.06%	+0.01%
10%	+0.12%	+0.07%
15%	+0.06%	+0.02%

Tabela 9: Impacto da adição de ruído gaussiano nas variáveis espectrais do Sentinel-2 sobre o erro de previsão (RMSE) da temperatura e da condutividade, expresso como variação percentual face ao cenário de referência (0% de ruído)

Para a temperatura, mesmo com ruído gaussiano multiplicativo até 15% nas variáveis espectrais, o RMSE praticamente não mudou, e as pequenas diferenças observadas são tão reduzidas que se explicam facilmente por variação aleatória do próprio teste.
Na condutividade, mesmo com ruído até 15% nas variáveis espectrais, o RMSE praticamente não aumentou, o que mostra que o modelo é estável face a perturbações moderadas nas entradas.

6.5 Importância de Variáveis e Análise de Sensibilidade

Top 10 variáveis - Modelo Temperatura:

Rank	Variável	Importância	Contributo(%)
1	NDVI	20.82	33.85
2	B2	13.85	22.52
3	B6	6.24	10.14
4	B5	5.2	8.46
5	B8	4.54	7.39
6	Outras variáveis (soma)		3.23
7	B3B2	1.96	3.19
8	B7	1.6	2.6
9	NDMI	1.39	2.26
10	B4	0.67	1.08

Tabela 10: Importância relativa das bandas espectrais e índices derivados no modelo XGBoost para previsão dos parâmetros de qualidade da água, ordenada por contributo percentual para o desempenho preditivo

Top 10 variáveis - Modelo Condutividade:

Rank	Variável	Importância	Contributo(%)
1	B5	30.88	29.02
2	B3B2	12.5	11.75
3	Outras variáveis (soma)		11.84
4	B6	10.7	10.06
5	B4	5.21	4.9
6	B8	5.13	4.82
7	NDVI	5.01	4.71
8	B2	4.69	4.41
9	NDMI	4.4	4.14
10	B3	3.38	3.18

Tabela 11: Importância relativa das bandas espectrais e índices derivados no modelo XGBoost treinado para prever a condutividade elétrica da água, ordenada por contribuição percentual para o desempenho preditivo.

As Tabelas 10 e 11 mostram quais foram as variáveis que mais contribuíram para reduzir o erro do XGBoost ao longo do treino, usando a métrica *total_gain*, ou seja, o ganho acumulado em todos os conjuntos onde cada variável é usada. No caso da temperatura, a maior parte da capacidade preditiva está concentrada no NDVI (33,85%) e na banda B2 (22,52%), seguidos das bandas B6 e B5 (red-edge) e B8 (NIR), o que indica que o modelo não está a "medir" temperatura diretamente, mas sim a explorar padrões espectrais que funcionam como *proxies* do estado da massa de água e do contexto local, como variações sazonais, presença de vegetação aquática/algas, turbidez e efeitos de mistura com margens. O facto de surgirem bandas do red-edge e do NIR com peso relevante é coerente com essa leitura, porque estas regiões espectrais tendem a responder a alterações na coluna de água e na influência de materiais biológicos ou partículas que, em certos períodos, podem estar correlacionados com o regime térmico observado.

Já na condutividade, o padrão é diferente: a banda B5 surge claramente como a variável dominante (29,02%), seguida da razão B3B2 (11,75%) e da banda B6 (10,06%), enquanto as restantes variáveis do visível e do NIR aparecem com importâncias mais distribuídas. Isto sugere que o modelo está sobretudo a capturar variações no sinal espectral associadas a condições que mudam a cor e a estrutura ótica da água — por exemplo, episódios de maior carga de partículas, entradas de água com características diferentes ou maior influência da margem — e a usar esses sinais como indicadores indiretos para prever condutividade, que por si só não é um parâmetro ótico. É também por isso que a interpretação deve ser cautelosa: importâncias altas significam que aquelas variáveis foram úteis para o modelo no teu conjunto de dados, mas não provam uma relação física direta, até porque bandas e índices podem estar correlacionados e "partilhar" informação, fazendo com que o ganho se distribua por várias variáveis relacionadas.

6.6 Exercício exploratório de forecasting com comparação a dados futuros

Foi considerada a realização de um exercício de validação temporal (treinar em anos anteriores e testar em anos mais recentes) para simular um cenário de previsão operacional. No entanto, devido ao reduzido número de "*matchups"* e à distribuição temporal assimétrica das amostras, essa avaliação temporal não é, infelizmente, estatisticamente robusta e poderia levar a conclusões enganadoras e com desvios elevados. Por esse motivo, a validação temporal e a extensão para modelos explicitamente temporais ficam propostas como trabalho futuro.

6.7 Interpretação dos Resultados

O modelo XGBoost mostrou-se bastante mais eficaz do que métodos tradicionais, tanto em termos de precisão como de capacidade de generalização, porque consegue captar relações não lineares complexas entre as bandas espectrais que modelos lineares não conseguem representar. Além disso, integra mecanismos de regularização que reduzem o sobreajuste mesmo quando se utilizam muitas árvores, o que permite ajustar modelos relativamente complexos sem perder robustez. Outra vantagem é a capacidade de interpretação, ao contrário de muitas redes neuronais tratadas como "caixa negra", o XGBoost fornece medidas claras de importância das variáveis, ajudando a perceber quais bandas e índices são mais relevantes para as previsões. Em termos práticos, o algoritmo é ainda bastante eficiente do ponto de vista computacional, permitindo treinos rápidos com validação cruzada extensa, algo particularmente útil quando se trabalha em ambiente de computação em nuvem.

No caso deste estudo, o desempenho do modelo para temperatura (R^2 de 0,912) foi superior ao obtido para condutividade (R^2 de 0,862). Uma das razões é que a temperatura da água está ligada à estratificação térmica e a processos que influenciam a distribuição de fitoplâncton, o que gera uma relação ótica indireta mas relativamente consistente com a informação espectral captada pelo satélite. Já a condutividade é um parâmetro menos "ótico": não absorve nem reflete a luz de forma direta e a sua ligação com os sólidos dissolvidos pode variar bastante consoante a composição mineral da água, o que torna a relação com as bandas espectrais mais fraca e mais ruidosa. Além disso, a temperatura tende a apresentar uma amplitude de variação maior (por exemplo, variações da ordem de dezenas de graus) do que a condutividade numa escala relativa semelhante, o que facilita ao modelo distinguir padrões e produzir previsões mais estáveis.

As bandas red-edge do Sentinel-2 (B5, B6 e B7) tiveram um peso relevante no modelo, sobretudo na estimativa da temperatura, com destaque para a B6, que surge entre as variáveis mais influentes na Tabela 10. Estas bandas situam-se na zona de transição entre o vermelho e o infravermelho próximo — aproximadamente entre 705 e 783 nm — onde a refletância é particularmente sensível a alterações associadas a pigmentos e matéria orgânica, sendo por isso frequentemente usada para detetar variações de clorofila e biomassa em vegetação e, em certos contextos, para captar sinais indiretos relacionados com a dinâmica biológica e com o estado da superfície da água. No enquadramento deste estudo, a importância do red-edge deve ser entendida como um indicador de que o modelo está a explorar *proxies* espectrais (e não uma medição direta de temperatura), isto é,

padrões que covariam com processos sazonais e biogeoquímicos no reservatório e que acabam por ajudar a reduzir o erro de previsão.

O fundamento teórico da capacidade do modelo em estimar a temperatura e a condutividade elétrica, parâmetros que não são oticamente ativos e, portanto, 'invisíveis' ao satélite explica-se pela deteção de correlações de componentes visíveis.

No caso da temperatura, o modelo utiliza a biomassa de fitoplâncton como um indicador indireto, uma vez que o calor impulsiona o crescimento de algas detetáveis nas bandas Red-Edge do Sentinel-2. Relativamente à condutividade, a estimativa apoia-se na sua forte correlação com a turbidez e os sólidos dissolvidos; o modelo aprende a associar o aumento da difração de luz nas bandas do visível a valores mais elevados de condutividade. Essencialmente, o algoritmo infere estes parâmetros através das suas relações físicas e biológicas com a clorofila e os sedimentos suspensos que covariam com o aumento da força iónica.

Nos cenários avaliados, o cenário em que o treino foi feito com dados do drone e teste com dados da estação, foi o que apresentou melhor comportamento global, por ser o que mais se aproxima de um teste de transferência realista entre fontes de medição.

No entanto, os resultados mostram um contraste claro entre o desempenho em treino/validação e o desempenho em teste: para a temperatura, o modelo apresenta R^2 elevado em treino (0,912) e ainda positivo em validação (0,697), mas no teste cai para $R^2 < 0$, com aumento do erro (RMSE ≈ 3,85 ºC; MAE ≈ 3,34 ºC).

O mesmo padrão surge na condutividade: R^2 de treino é 0,862 e na validação 0,316, mas no teste o modelo degrada para $R^2 < 0$ (RMSE ≈ 8,89 µS/cm; MAE ≈ 7,27 µS/cm), o que indica que o modelo está a perder capacidade preditiva quando aplicado a dados “fora do domínio” onde foi treinado.

Estes resultados são coerentes com o enquadramento teórico da tese, tanto a temperatura como a condutividade não são parâmetros óticos diretos, pelo que o modelo aprende sobretudo relações indiretas que podem funcionar bem quando o contexto é semelhante ao treino, mas podem falhar quando mudam as condições ambientais e/ou o tipo de medição.

6.8 Limitações da Abordagem

Limitações Identificadas:

1. Tamanho Limitado do conjunto de dados: 98 "*matchups"* é relativamente pequeno para aprendizagem de máquina. Aumentar para 200-300 seria ideal.

2. Cobertura Espacial Reduzida: Apenas 2 estações de coleta. Reservatórios grandes têm heterogeneidade espacial que não está totalmente capturada.

3. A cobertura temporal não é homogénea, as medições da estação existem ao longo de vários anos, mas são relativamente esporádicas e com espaçamento irregular, enquanto os dados recolhidos pelo drone são muito mais densos (muitas leituras), porém concentrados num intervalo de tempo curto.

4. Mudanças Espectrais Sazonais: Assinatura espectral da água varia com composição algal e sedimentar, que muda sazonalmente. Possível que modelos treinados numa estação funcionem mal noutras.

5. Mesmo os dados L2A, já corrigidos para a superfície, ainda podem ter restos de efeito atmosférico, principalmente nas bandas azuis (B2) [77].

6. Validação com dados de referência reais, A sonda C4E utilizada no drone asv1 da Flyrobotics, é uma sonda de condutividade/salinidade com 4 elétrodos que usa corrente alternada com uma tensão fixa. Tendo um limite de precisão de aproximadamente ±1%. Isto ajuda a evitar erros causados pela sujidade nos elétrodos e pela polarização, mantendo a medição mais estável e precisa em valores de condutividade baixos e altos [20] .

7. A incorporação de dados históricos anteriores a 2017 não é possível usando exclusivamente a coleção COPERNICUS/S2_SR_HARMONIZED, cuja disponibilidade no Google Earth Engine se inicia em 2017-03-28; para estender a série temporal para 2010+ seria necessário integrar sensores alternativos (p.ex., Landsat), com harmonização radiométrica e/ou modelos específicos por sensor [21].
Recomendações para Mitigação:

- Expandir coleta de dados "*in situ*" para mínimo 200 amostras (“matchups” em treino)
- Adicionar 3-4 estações adicionais para cobertura espacial
- Incorporar dados históricos (2010+) se disponíveis, com outras constelações de satélites
- Treinar modelos sazonalmente (modelo_primavera, modelo_verão, etc.)
- Usar dados Sentinel-1 (SAR) para um melhor perfilamento do modelo hidrológico e desta forma perceber melhor a sazonabilidade dos estudos e aplicação do(s) modelo(s) resultante(s)
- Usar Sentinel-3 (SLSTR) para séries temporais de temperatura superficial, aceitando a resolução de ~1 km e o tempo de revisita diário a 2 dias, por ter bandas térmicas
- Para assegurar a correspondência espacial dos matchups, as medições in situ por drone devem ser recolhidas em zonas de água aberta, minimizando efeitos de mistura com margens e evitando pixeis parcialmente fora da massa de água.

A principal limitação evidenciada por estes testes é a robustez entre fontes (drone → estação), mesmo com bom ajuste no treino, a relação aprendida pode não se manter quando as medições provêm de outro sistema, com dinâmicas, ruído e distribuição de valores diferentes.
Isto reflete-se nos R^2 negativos no teste, indicando que o modelo pode ficar pior do que uma previsão simples baseada na média, pelo que o problema central passa pela diferença estatística entre conjuntos e pela representatividade do treino.
Acrescem limitações estruturais já discutidas na tese (*matchups*, janela temporal e restrições de observação por satélite), que reduzem a estabilidade quando o modelo é aplicado a condições não observadas.
Por fim, a condutividade é mais difícil por não ser um parâmetro ótico direto, ficando dependente de relações indiretas e variáveis no tempo, o que ajuda a explicar a degradação mais acentuada.

6.9 Implicações para Monitorização Ambiental

O sistema proposto tem um forte potencial operacional por reduzir custos, aumentar a cobertura e escalar com facilidade, tirando partido de dados Sentinel-2 e de um pipeline automatizado. Do ponto de vista económico, exige apenas um investimento inicial e uma

operação anual relativamente baixos quando comparados com a instalação e manutenção de estações *in situ*, que tipicamente implicam custos anuais de ordem muito superior. Esta diferença de custos torna plausível um retorno em horizonte curto, à medida que parte da monitorização tradicional é substituída ou otimizada. Em termos de cobertura, a monitorização deixa de estar limitada a alguns pontos por reservatório e passa a produzir mapas em grelha com milhares de píxeis (por exemplo, cerca de 3.500 píxeis válidos em Castelo do Bode, à escala de 20 m), aproximando-se de uma caracterização quase contínua da qualidade da água [22].

O tempo de revisita também melhora substancialmente, em vez de campanhas semanais ou quinzenais, a utilização combinada de diferentes satélites possibilita observações quase diárias, multiplicando a frequência de monitorização por um fator entre 7 e 50. O modelo treinado num reservatório pode ser adaptado a outros com pequenos ajustes ("*transfer know-how"*), abrindo a porta à aplicação em toda a bacia do Tejo. Operacionalmente, isto permite detetar anomalias (poluição, eutrofização) com rapidez, antecipar blooms algais, apoiar a gestão de recursos hídricos (produção hidroelétrica e abastecimento público), fornecer dados em larga escala para investigação científica e reforçar o cumprimento de diretivas ambientais europeias através de uma vigilância contínua e de grande detalhe.

6.10 Perspetivas de Transferência Tecnológica – Caso Prático

A integração de uma abordagem de monitorização avançada na ETA da Asseiceira poderia trazer benefícios claros para a previsão da qualidade de água bruta e a otimização do tratamento operado pela EPAL. Ao combinar dados de satélite (por exemplo, sobre temperatura, biomassa algal e turbidez na albufeira de Castelo do Bode) com medições em tempo quase real nos pontos de captação e ao longo da linha de tratamento, seria possível antecipar variações de qualidade e ajustar previamente dosagens de coagulantes, ativar carvão em pó, reforçar desinfeção ou otimizar tempos de filtração. Em vez de uma reação apenas após a deteção de um problema, a ETA passaria a atuar de forma preventiva, com base em previsões robustas.

Modelos de aprendizagem de máquina, treinados com históricos de operação da própria Asseiceira (caudal, qualidade da água bruta, ocorrências de *blooms* de algas/cianobactérias, consumos de reagentes, quebras de filtros, parâmetros de saída), poderiam aprender relações entre condições na albufeira e desempenho dos diferentes

processos de tratamento. Assim, seria possível prever situações de maior risco, como aumentos de matéria orgânica, episódios de sabor e odor ou flutuações de turbidez, e adaptar de forma automática ou assistida os setpoints de operação. Isto poder-se-ia traduzir numa maior estabilidade da qualidade de água tratada, redução de custos com reagentes e energia, e menor probabilidade de não conformidades.

Num cenário ideal, a ETA disporia de um painel operacional integrado onde os técnicos visualizariam, num único interface, o estado atual da albufeira, previsões de curto prazo para parâmetros críticos (temperatura, condutividade, clorofila, risco de blooms), recomendações de ajuste de operação e indicadores de desempenho dos filtros, decantadores e desinfeção. Alertas automáticos poderiam ser gerados quando o modelo previsse risco acrescido de degradação da qualidade da água bruta ou quando o sistema detetasse que a configuração atual de tratamento não é a mais eficiente para as condições previstas. Para uma empresa como a EPAL, esta abordagem significaria uma gestão mais inteligente do risco, maior resiliência face a eventos extremos e um contributo importante para garantir, de forma sustentável, água segura e de elevada qualidade à população abastecida pela ETA da Asseiceira.

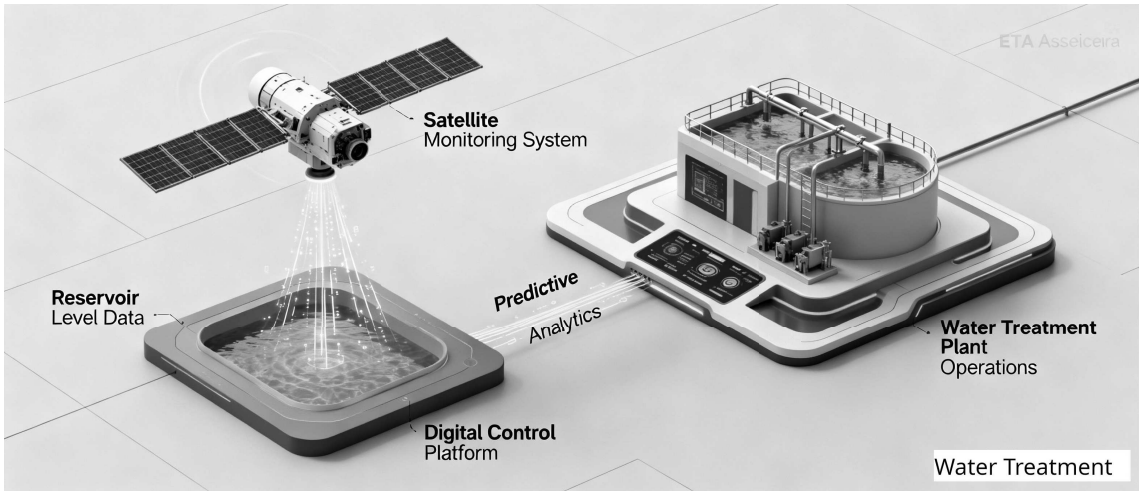

Figura 23: Monitorização por satélite e análises preditivas a suportar a operação de uma estação de tratamento de água

Fonte: imagem gerada com apoio de GPT-5 em 12/12/2025.

Capítulo 7 - Conclusões

7.1 Síntese dos Contributos Principais

Este trabalho mostra, no caso de estudo da albufeira de Castelo do Bode e com 98 *matchups* entre 2017 e 2024, que é possível estimar temperatura e condutividade com base em variáveis extraídas do Sentinel-2 e modelos XGBoost. Os resultados são mais fortes para a temperatura e mais moderados para a condutividade, o que é esperado porque a condutividade não é diretamente visível para o sensor e a sua relação com a refletância é indireta.

O contributo principal é a construção de um pipeline que pode ser repetido: extração de variáveis no Google Earth Engine, criação de *matchups* espaço-temporais, treino/validação dos modelos e análise das variáveis mais relevantes. Além disso, foi feito um protótipo de integração (API/Streamlit e *dashboard*) que mostra como este tipo de modelo pode ser usado num contexto de monitorização.

Contributos Científicos:

1. Validação de XGBoost para Qualidade de Água: Este estudo fornece evidência robusta de que XGBoost é superior a métodos tradicionais para estimação de parâmetros de qualidade de água usando dados Sentinel-2. Performance em RMSE bastante melhor que referências.

Na temperatura, comparando com o melhor modelo do ficheiro de comparação para temperatura (Ridge, RMSE=7,94), o XGBoost reduziu o RMSE em 4,09 °C, o que corresponde a cerca de 51,5% de melhoria.

Já na condutividade, comparando com o melhor do ficheiro de comparação para condutividade (baseline_mean, RMSE=236,77), o XGBoost reduziu o RMSE em 227,88 µS/cm, cerca de 96,2% de melhoria.

2. Foi identificada uma escassez de estudos que estimem, de forma consistente e no mesmo trabalho, temperatura da água e condutividade elétrica a partir de Sentinel-2 e modelos de aprendizagem automática, sendo mais comum encontrar trabalhos focados noutros parâmetros (por exemplo, clorofila-a, turbidez, DO, nutrientes) ou apenas num dos dois.

3. Importância de Bandas Red-Edge: Confirmação de que bandas red-edge do Sentinel-2 (B5, B6, B7) são críticas para estimação de parâmetros aquáticos, justificando sua inclusão em futuros satélites.

4. *Matchup* Metodologia: Desenvolvimento de procedimento rigoroso de sincronização espaço-temporal entre dados de satélite e "*in situ"*.

Contributos Técnicos:

1. Pipeline Operacional Completo: Do processamento GEE até deployment do Streamlit(ou API REST), fornecendo base para implementações futuras.

2. Os hiperparâmetros do XGBoost foram escolhidos com validação cruzada a 5 partes apenas dentro do conjunto de treino e, no fim, confirmou-se que o modelo funciona bem com dados novos ao avaliá-lo num conjunto de teste que ficou totalmente de fora do ajuste.

3. Código Reprodutível: Scripts Python completamente disponíveis para comunidade científica.

Implicações Práticas

Para Gestão de Recursos Hídricos:

- Monitorização contínua de parâmetros críticos de qualidade com custo reduzido
- Capacidade de detecção precoce de eventos adversos (poluição, eutrofização)
- Suporte a decisões informadas sobre operação de reservatórios

Para Investigação Ambiental:

- Base de dados sem precedentes de qualidade de água em resolução espaço-temporal alta
- Oportunidades para investigação em ecologia aquática, limnologia

- Validação de modelos biogeoquímicos

Para Política Ambiental:

- Cumprimento mais eficiente de diretivas europeias (Água, Biodiversidade)
- Dados para planeamento de investimentos em tratamento água
- Comunicação melhorada com público sobre estado dos ecossistemas aquáticos

Este trabalho demonstrou a viabilidade de usar Sentinel-2 e modelos de aprendizagem automática (XGBoost) para estimar temperatura e condutividade, integrando dados multiespectrais com medições *in situ* e operacionalizando um pipeline de treino/validação/teste.
Nos cenários testados, o cenário em que o treino foi feito com dados do drone asv1 e os dados de teste foram os da estação hidrológica, revelou-se o mais relevante do ponto de vista prático, por avaliar explicitamente a transferência entre fontes, embora os resultados em teste indiquem que a robustez ainda é limitada, em que os R^2_{test} <0 da temperatura e da condutividade.
Assim, o contributo central não é apenas o desempenho num conjunto "semelhante ao treino", mas a evidência experimental de que a robustez inter-plataforma é o principal desafio a resolver para tornar a solução estável em contexto real, reforçando a necessidade de mais dados representativos e estratégias de validação/treino orientadas a cenários de transferência inter-domínio.

7.2 Trabalhos Futuros

O trabalho desenvolvido nesta dissertação pode ser aprofundado e melhorado em várias frentes, sobretudo para tornar os modelos mais robustos e mais úteis num cenário real de monitorização. A prioridade mais imediata passa por aumentar e equilibrar o conjunto de dados, porque o número de "*matchups"* ainda é reduzido e a distribuição por anos não é uniforme, o que limita a confiança na generalização do modelo. Para além de recolher mais amostras ao longo do tempo, também seria importante aumentar o número de pontos de medição no reservatório (mais do que duas estações), de forma a captar melhor a variabilidade espacial.

Num horizonte de curto prazo (1 a 2 anos), faz sentido aplicar a mesma metodologia a outros parâmetros relevantes de qualidade da água, como clorofila-a, turbidez, pH e oxigénio dissolvido. Em paralelo, a abordagem pode beneficiar de uma estratégia multi-satélite, juntando dados de outras missões para melhorar a frequência de observações e reduzir falhas por nuvens. Neste ponto, torna-se especialmente interessante evoluir para um modelo híbrido Sentinel-2 + Sentinel-3: o Sentinel-2 mantém o detalhe espacial e a riqueza espectral (muito útil para padrões ópticos e variabilidade dentro do reservatório), enquanto o Sentinel-3 (SLSTR), apesar da resolução mais grosseira, traz bandas térmicas e uma cadência mais regular, permitindo séries temporais de temperatura superficial mais contínuas e estáveis [26,78,80].

À medida que o conjunto de dados for crescendo, torna-se viável avançar para uma validação temporal mais realista, isto é, treinar com períodos passados e testar com períodos futuros, sem misturar aleatoriamente as amostras. Neste momento, essa análise temporal ainda não é suficientemente sólida devido ao baixo número de observações e ao facto de grande parte dos dados estarem concentrados num período recente. Como trabalho futuro, recomenda-se recolher dados mais distribuídos por vários anos e estações e, só depois, aplicar técnicas de validação cruzada temporal (janelas deslizantes ou em expansão) para medir de forma honesta se o modelo mantém desempenho quando "avança no tempo". Com essa base, passa a fazer sentido testar modelos explicitamente temporais, como LSTMs (tipo de rede neuronal recorrente (RNN)), e avaliar previsões de curto prazo (por exemplo, 1 a 7 dias) com recalibração periódica [79,81].

Outra linha de evolução importante é a transferência de conhecimento para outros locais. Em vez de treinar sempre modelos de raiz, pode explorar-se *"transfer know-how"* e adaptação de modelos para outros reservatórios nacionais (por exemplo, Alqueva ou Alto Rabagão), usando um modelo base e afinando-o com um conjunto reduzido de dados locais. Esta estratégia pode reduzir tempo e custos de implementação e aproximar a solução de um uso prático em mais bacias hidrográficas.

A médio prazo (2 a 5 anos), a metodologia pode ganhar detalhe com dados hiperespectrais (por exemplo, PRISMA ou EnMAP), que oferecem maior resolução espectral para distinguir melhor componentes complexos da água. Para além disso, recomenda-se integrar dados SAR do Sentinel-1 não tanto como "substituto" do Sentinel-2 na qualidade da água, mas como uma fonte complementar para construir um perfil hidrológico do reservatório (por exemplo, variações de área inundada, padrões de margem e sinais associados a alterações de nível), ajudando a interpretar a sazonalidade e o contexto hidrológico em que os modelos estão a ser aplicados. Esta componente pode ser

particularmente útil para enquadrar a leitura dos resultados e para suportar decisões sobre quando recalibrar modelos ou quando aplicar modelos sazonais.

Como trabalho futuro, recomenda-se também avaliar abordagens de AutoML para automatizar a escolha de modelos e a afinação de hiperparâmetros, reduzindo o esforço manual e garantindo uma comparação mais sistemática entre alternativas. Em vez de depender apenas de configurações selecionadas manualmente, um processo AutoML pode testar diferentes pipelines (pré-processamento, seleção de variáveis/índices e modelos), registando métricas de validação e escolhendo a solução com melhor compromisso entre erro, estabilidade e complexidade. Esta estratégia é particularmente útil quando o objetivo é operacionalizar o sistema e manter desempenho ao longo do tempo, uma vez que permite repetir o processo de otimização sempre que o conjunto de dados cresce ou muda [47].

Finalmente, do ponto de vista de engenharia e disponibilização da solução, faz sentido consolidar e especializar as ferramentas de visualização e operação. O Grafana é mais adequado quando o objetivo é ter "*dashboards"* "sempre ligados" para acompanhar sensores/IoT em tempo quase real, com métricas e séries temporais, e com alertas operacionais (por exemplo, avisos automáticos quando um parâmetro ultrapassa um limiar), sobretudo quando existe ou se pretende montar uma solução típica de observabilidade/telemetria com bases de dados de séries temporais e outras fontes próprias. O Streamlit é mais indicado quando se pretende construir rapidamente uma aplicação em Python com interação (widgets, filtros e botões) e lógica personalizada, especialmente útil para "servir" modelos de ML e permitir ao utilizador testar cenários, fazer upload de ficheiros, comparar outputs e explorar resultados de forma mais analítica. Como trabalho futuro, a adoção combinada destas duas abordagens pode separar claramente o que é operação contínua (Grafana) do que é exploração e validação de modelos (Streamlit) [23-24].

Referências Bibliográficas

1. NGAMILE, S.; MADONSELA, S.; KGANYAGO, M. Trends in remote sensing of water quality parameters in inland water bodies: a systematic review. Frontiers in Environmental Science, v. 13, art. 1549301, 2025. DOI: 10.3389/fenvs.2025.1549301. Disponível em: <https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2025.1549301/full>. Acesso em: 17 out. 2025.
2. JAYWANT, S. A.; ARIF, K. M. Remote Sensing Techniques for Water Quality Monitoring: A Review. Sensors (Basel), v. 24, n. 24, art. 8041, 17 dez. 2024. DOI: 10.3390/s24248041. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11679694/>. Acesso em: 17 out. 2025.
3. EUROPEAN SPACE AGENCY (ESA). Sentinel-2: Facts and figures. Disponível em: <https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2/Facts_and_figures>. Acesso em: 17 out. 2025.
4. EUROPEAN COMMISSION (CORDIS). Global Lakes Sentinel Services (GLaSS) – Project reporting (FP7). Disponível em: <https://cordis.europa.eu/project/id/313256/reporting>. Acesso em: 17 out. 2025.
5. CUKJATI, J. et al. IoT and Satellite Sensor Data Integration for Assessment of Environmental Variables: A Case Study on NO_2. Sensors, v. 22, n. 15, art. 5660, 28 jul. 2022. DOI: 10.3390/s22155660. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9371219/>. Acesso em: 17 out. 2025.
6. U.S. GEOLOGICAL SURVEY (USGS). Use of multiparameter instruments for routine field measurements. In: National Field Manual for the Collection of Water-Quality Data (Techniques and Methods 9–A6.8). Versão 1.1, jun. 2025. Disponível em: <https://pubs.usgs.gov/tm/09/a6.8/tm9a6.8.pdf>. Acesso em: 17 out. 2025.
7. U.S. GEOLOGICAL SURVEY (USGS). What is continuous real-time water quality (RTWQ)? (Water Quality Watch – FAQ). Disponível em: <https://waterwatch.usgs.gov/wqwatch/faq?faq_id=1>. Acesso em: 17 out. 2025.
8. UNITED STATES GEOLOGICAL SURVEY. National Field Manual for the Collection of Water-Quality Data (NFM). Version 3.0. Reston, VA: USGS, 2019. Capítulo A6: Field Measurements . Disponível em: <https://www.usgs.gov/mission-areas/water-resources/science/national-field-manual-collection-water-quality-data-nfm>. Acesso em: 17 out. 2025.

9. SOUSA, D.; HERNANDEZ, D.; OLIVEIRA, F.; LUÍS, M.; SARGENTO, S. A Platform of Unmanned Surface Vehicle Swarms for Real Time Monitoring in Aquaculture Environments. Sensors, v. 19, n. 21, 4695, out. 2019. DOI: 10.3390/s19214695. Disponível em: <https://doi.org/10.3390/s19214695>. Acesso em: 19 out. 2025.

10. KUMAR, M. et al. In-situ optical water quality monitoring sensors—applications, challenges, and future opportunities. Frontiers in Water, v. 6, art. 1380133, 21 abr. 2024. DOI: 10.3389/frwa.2024.1380133. Disponível em: <https://www.frontiersin.org/journals/water/articles/10.3389/frwa.2024.1380133/full>. Acesso em: 10 dez. 2025.

11. HARRISON, S. et al. Unlocking the global benefits of Earth Observation to address the SDG 6 in situ water quality monitoring gap. Frontiers in Remote Sensing, v. 6, art. 1549286, 2025. DOI: 10.3389/frsen.2025.1549286. Disponível em: <https://www.frontiersin.org/journals/remote-sensing/articles/10.3389/frsen.2025.1549286/full>. Acesso em: 12 dez. 2025.

12. BESMER, M. D.; HAMMES, F.; SIGRIST, J. A.; ORT, C. Evaluating Monitoring Strategies to Detect Precipitation-Induced Microbial Contamination Events in Karstic Springs Used for Drinking Water. Frontiers in Microbiology, v. 8, 2229, 22 nov. 2017. DOI: 10.3389/fmicb.2017.02229. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5703154/>. Acesso em: 12 dez. 2025.

13. COPERNICUS (European Union). How to access data. Disponível em: <https://www.copernicus.eu/en/terms-use/how-access-data>. Acesso em: 10 dez. 2025.

14. NASA Applied Sciences (ARSET). What is the spatial resolution of Landsat? Disponível em: <https://appliedsciences.nasa.gov/arset-ecological-conservation-faq>. Acesso em: 09 dez. 2025.

15. NASA. Landsat 8. NASA Science. Disponível em: <https://science.nasa.gov/mission/landsat-8/>. Acesso em: 09 dez. 2025.

16. U.S. GEOLOGICAL SURVEY (USGS). Landsat Missions. Disponível em: <https://www.usgs.gov/landsat-missions>. Acesso em: 09 dez. 2025.

17. U.S. GEOLOGICAL SURVEY (USGS). Temperature and Water. Disponível em: <https://www.usgs.gov/water-science-school/science/temperature-and-water>. Acesso em: 09 dez. 2025.

18. EUROPEAN SPACE AGENCY. S3 SLSTR Instrument. SentiWiki - Copernicus Sentinel-3. Disponível em: <https://sentiwiki.copernicus.eu/web/s3-slstr-instrument>. Acesso em: 07 dez. 2025.

19. ASTROCAST. Connecting the Dots: Satellite IoT Bridges the Gap for Crucial Environmental Use Cases. Astrocast News, 26 jul. 2022. Disponível em:

<https://www.astrocast.com/news/connecting-the-dots-satellite-iot-bridges-the-gap-for-crucial-environmental-use-cases/>. Acesso em: 09 dez. 2025.

20. AQUALABO. C4E Conductivity Sensor. Disponível em: <https://www.aqualabo.fr/en/produit/c4e-conductivity-sensor/>. Acesso em: 07 dez. 2025.

21. GOOGLE EARTH ENGINE. Sentinel-2 Surface Reflectance Collection. Google Earth Engine Data Catalog, [S. l.] . Disponível em: <https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED>. Acesso em: 12 dez. 2025.

22. DRESSING, S. A.; MEALS, D. W. Capítulo 9: Monitoring Costs. In: Monitoring and Evaluating Nonpoint Source Watershed Projects. Washington, DC: U.S. Environmental Protection Agency (EPA), 2016. Disponível em: <https://www.epa.gov/sites/default/files/2016-06/documents/chapter_9_may_2016_508.pdf>. Acesso em: 11 dez. 2025

23. GRAFANA LABS. Grafana: The open and composable observability platform. [S. l.]. Disponível em: <https://grafana.com/>. Acesso em: 17 out. 2025.

24. STREAMLIT. Streamlit: A faster way to build and share data apps. Disponível em: <https://streamlit.io/>. Acesso em: 17 out. 2025.

25. LAL, K.; MENON, S.; NOBLE, F.; ARIF, K. M. Low-cost IoT based system for lake water quality monitoring. PLoS ONE, v. 19, n. 3, e0299089, mar. 2024. DOI: 10.1371/journal.pone.0299089. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10977749/>. Acesso em: 17 out. 2025.

26. NIKOO, M. R. et al. Mapping reservoir water quality from Sentinel-2 satellite imagery: chlorophyll-a and dissolved oxygen. Scientific Reports, v. 14, 16699, jul. 2024. DOI: 10.1038/s41598-024-66699-2. Disponível em: <https://www.nature.com/articles/s41598-024-66699-2>. Acesso em: 17 out. 2025.

27. GUNIA, M.; LAINE, M.; MALVE, O.; KALLIO, K.; KERVINEN, M.; ANTTILA, S.; KOTAMÄKI, N.; SIIVOLA, E.; KETTUNEN, J.; KAURANNE, T. Data fusion system for monitoring water quality: Application to chlorophyll-a in Baltic sea coast. Environmental Modelling & Software, v. 155, 105465, set. 2022. DOI: 10.1016/j.envsoft.2022.105465. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1364815222001645>. Acesso em: 17 out. 2025.

28. SANTOS, V.; ROCHA, P. A.; THÉ, J.; GHARABAGHI, B. Enhancing Turbidity Modeling in the Mississippi River Using Machine Learning and Sentinel-2 Satellite Remote Sensing Data: A Generalizability Analysis. SSRN Electronic Journal, 2024. Disponível em: <https://ssrn.com/abstract=4980387>. DOI: 10.2139/ssrn.4980387. Acesso em: 12 dez. 2025.

29. OUMA, Y. O. et al. Prediction of Turbidity and TDS in Dam Reservoir from Multispectral UAV-Drone and Sentinel-2 Image Sensors Using Machine Learning Models. In: PROCEEDINGS OF THE 10th INTERNATIONAL CONFERENCE ON GEOGRAPHIC INFORMATION SYSTEMS THEORY, APPLICATIONS AND MANAGEMENT (GISTAM 2024), 2024. DOI: 10.5220/0012545600003696. Disponível em: <https://doi.org/10.5220/0012545600003696>. Acesso em: 17 out. 2025.

30. Tian, S., Guo, H., Xu, W., Zhu, X., Wang, B., Zeng, Q., Mai, Y., & Huang, J. J. (2023). Remote sensing retrieval of inland water quality parameters using Sentinel-2 and multiple machine learning algorithms. Environmental Science and Pollution Research, 30(7), 18617–18630. https://doi.org/10.1007/s11356-022-23431-9 Acesso em: 17 out. 2025.

31. Amorim, F. d. L. L. d., Rick, J., Lohmann, G., & Wiltshire, K. H. (2021). Evaluation of Machine Learning Predictions of a Highly Resolved Time Series of Chlorophyll-a Concentration. Applied Sciences, 11(16), 7208. https://doi.org/10.3390/app11167208 Acesso em: 17 out. 2025.

32. MAIER, P. M.; KELLER, S. Estimating chlorophyll a concentrations of several inland waters with hyperspectral data and machine learning models. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, v. IV-2/W5, p. 609-614, 2019. DOI: 10.5194/isprs-annals-IV-2-W5-609-2019. Disponível em: <https://doi.org/10.5194/isprs-annals-IV-2-W5-609-2019>. Acesso em: 17 out. 2025.

33. ILIC, V.; TURK SEKULIC, M.; BRBORIC, M.; RADONIC, J.; DMITRASINOVIC, S.; STOJKOVIC, M. Enhancing the monitoring system for river water quality: harnessing the power of satellite data and machine learning. Blue-Green Systems, v. 7, n. 2, 338-352, 2025. DOI: 10.2166/bgs.2025.006. Disponível em: <https://doi.org/10.2166/bgs.2025.006>. Acesso em: 15 dez. 2025.

34. NASA EARTHDATA. Moderate Resolution Imaging Spectroradiometer (MODIS) – Specifications. Disponível em: <https://www.earthdata.nasa.gov/data/instruments/modis>. Acesso em: 15 dez. 2025.

35. EUROPEAN SPACE AGENCY. 10 ways Sentinel-1 data lets us 'see' our world. Disponível em: <https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1/10_ways_Sentinel-1_data_lets_us_see_our_world>. Acesso em: 15 dez. 2025.

36. ENMAP PROJECT. Mission. Environmental Mapping and Analysis Program (EnMAP). Disponível em: <https://www.enmap.org/mission/>. Acesso em: 15 dez. 2025.

37. EUROPEAN SPACE AGENCY. PRISMA (Hyperspectral). eoPortal. Disponível em: <https://www.eoportal.org/satellite-missions/prisma-hyperspectral>. Acesso em: 12 dez. 2025.

38. MISHRA, S.; MISHRA, D. R. Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. Remote Sensing of Environment, v. 117, p. 394-406, fev. 2012. DOI: 10.1016/j.rse.2011.10.016. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0034425711003737>. Acesso em: 12 dez. 2025.

39. O'REILLY, J. E.; WERDELL, P. J. Chlorophyll algorithms for ocean color sensors - OC4, OC5 & OC6. Remote Sensing of Environment, v. 229, p. 32-47, ago. 2019. DOI: 10.1016/j.rse.2019.04.021. Disponível em: <https://ntrs.nasa.gov/api/citations/20190025308/downloads/20190025308.pdf>. Acesso em: 12 dez. 2025.

40. BROCKMANN, C. et al. Evolution of the C2RCC Neural Network for Sentinel 2 and 3 for the Retrieval of Ocean Colour Products in Normal and Extreme Optically Complex Waters. In: LIVING PLANET SYMPOSIUM, 2016, Praga. Proceedings... Edited by L. Ouwehand. Noordwijk: European Space Agency, 2016. p. 54. (ESA Special Publication, SP-740). ISBN 978-92-9221-305-3. Disponível em: <http://step.esa.int/docs/extra/Evolution%20of%20the %20C2RCC_LPS16.pdf>. Acesso em: 12 dez. 2025.

41. EUROPEAN SPACE AGENCY. MSI (Sentinel-2 Multispectral Instrument). Copernicus Sentinel-2. Disponível em: <https://s2.pages.eopf.copernicus.eu/pdfs-adfs/MSI/index.html>. Acesso em: 12 dez. 2025.

42. LOUIS, J. et al. Sentinel-2 Sen2Cor: L2A processor for users. In: LIVING PLANET SYMPOSIUM, 2016, Praga. Proceedings... Edited by L. Ouwehand. Noordwijk: European Space Agency, 2016. p. (ESA Special Publication, SP-740). ISBN 978-92-9221-305-3. Disponível em: <https://elib.dlr.de/107381/1/LPS2016_sm10_3louis.pdf>. Acesso em: 12 dez. 2025.

43. SUN, Y. et al. Application of remote sensing technology in water quality monitoring: From traditional approaches to artificial intelligence. Water Research, v. 267, 122546, dez. 2024. DOI: 10.1016/j.watres.2024.122546. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0043135424014453>. Acesso em: 14 dez. 2025.

44. TIAN, S. et al. Remote sensing retrieval of inland water quality parameters using Sentinel-2 and multiple machine learning algorithms. Environmental Science and Pollution Research, v. 30, n. 7, p. 17944-17959, fev. 2023. DOI: 10.1007/s11356-022-23431-9. Epub 14 nov. 2022. PMID: 36217046. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/36217046/>. Acesso em: 14 dez. 2025.

45. CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL

CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016, San Francisco. Proceedings... New York: ACM, 2016. p. 785-794. DOI: 10.1145/2939672.2939785. Disponível em: <https://dl.acm.org/doi/10.1145/2939672.2939785>. Acesso em: 14 dez. 2025.

46. U.S. GEOLOGICAL SURVEY. How does data from Sentinel-2A's MultiSpectral Instrument compare to Landsat data? Disponível em: <https://www.usgs.gov/faqs/how-does-data-sentinel-2as-multispectral-instrument-compare-landsat-data>. Acesso em: 14 dez. 2025.

47. PRASAD, D. V. V. et al. Automating water quality analysis using ML and auto ML techniques. Environmental Research, v. 202, 111720, 2021. DOI: 10.1016/j.envres.2021.111720. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0013935121010148>. Acesso em: 20 nov. 2025.

48. CALLEJAS, I. A. et al. A GEE toolkit for water quality monitoring from 2002 to 2022 in support of SDG 14 and coral health in marine protected areas in Belize. Frontiers in Remote Sensing, v. 3, 1020184, 22 nov. 2022. DOI: 10.3389/frsen.2022.1020184. Disponível em: <https://www.frontiersin.org/journals/remote-sensing/articles/10.3389/frsen.2022.1020184/full>. Acesso em: 14 dez. 2025.

49. *U.S. GEOLOGICAL SURVEY. What is remote sensing and what is it used for? Disponível em: <https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used>. Acesso em: 14 dez. 2025.*

50. *NASA EARTHDATA. Operational Land Imager (OLI) – especificações. Disponível em: <https://www.earthdata.nasa.gov/data/instruments/oli>. Acesso em: 14 dez. 2025.*

51. *KRITTEN, L. et al. Water remote sensing reflectance from radiative transfer simulations on a global scale of Inherent Optical Properties. Earth System Science Data Discussions (preprint), essd-2018-5, 16 abr. 2018. Disponível em: <https://essd.copernicus.org/preprints/essd-2018-5/essd-2018-5.pdf>. Acesso em: 14 dez. 2025.*

52. *MOBLEY, C. D. Shallow-water Remote Sensing: Lecture 1 – Overview. In: INTERNATIONAL OCEAN COLOUR COORDINATING GROUP (IOCCG). Shallow-water Remote Sensing Summer Lecture Series (SLS-2012). 2012. Disponível em: <https://www.ioccg.org/training/SLS-2012/Mobley_Lect1.pdf>. Acesso em: 14 dez. 2025.*

53. *COPERNICUS DATA SPACE ECOSYSTEM. Sentinel-2 – Documentation. Disponível em: <https://documentation.dataspace.copernicus.eu/Data/SentinelMissions/Sentinel2.html>. Acesso em: 14 dez. 2025.*

54. *EUROPEAN SPACE AGENCY (ESA) / COPERNICUS. Sentinel-2 User Handbook. ESA Standard Document, 24 jul. 2015. Issue 1, Rev 2. Disponível*

em: <https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook>. Acesso em: 14 dez. 2025.

55. *COPERNICAL. Copernicus: Sentinel-2 – The Optical Imaging Mission for Land Services. Disponível em: <https://www.copernical.com/projects-public/item/20279-copernicus-sentinel-2-the-optical-imaging-mission-for-land-services>. Acesso em: 15 dez. 2025.*
56. *COPERNICUS SENTINEL ONLINE. Copernicus Sentinel-2 Collection 1 MSI Level-2A (L2A). Disponível em: <https://sentinels.copernicus.eu/sentinel-data-access/sentinel-products/sentinel-2-data-products/collection-1-level-2a>. Acesso em: 15 dez. 2025.*
57. *LLODRÀ-LLABRÉS, J. et al. Retrieving water chlorophyll-a concentration in inland waters from Sentinel-2 imagery: Review of operability, performance and ways forward. International Journal of Applied Earth Observation and Geoinformation, v. 126, 103642, dez. 2023. DOI: 10.1016/j.jag.2023.103642. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1569843223004296>. Acesso em: 15 dez. 2025.*
58. *KATLANE, R. et al. Estimation of Chlorophyll and Turbidity Using Sentinel 2A and EO1 Data in Kneiss Archipelago Gulf of Gabes, Tunisia. Journal of Geoscience and Environment Protection, v. 8, n. 11, p. 95-110, nov. 2020. DOI: 10.4236/gep.2020.811008. Disponível em: <https://www.scirp.org/journal/paperinformation?paperid=103902>. Acesso em: 15 dez. 2025.*
59. *GOOGLE EARTH ENGINE. Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A (Surface Reflectance). Earth Engine Data Catalog. Disponível em: <https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED>. Acesso em: 15 dez. 2025.*
60. *COPERNICUS OPEN ACCESS HUB. User Guide (Rev. 125). Disponível em: <https://colhub.copernicus.eu/userguide/>. Acesso em: 16 dez. 2025.*
61. *POPE, R. M.; FRY, E. S. Absorption spectrum (380–700 nm) of pure water. II. Integrating cavity measurements. Applied Optics, v. 36, n. 33, p. 8710-8723, 20 nov. 1997. DOI: 10.1364/AO.36.008710. PMID: 18264420. Disponível em: <https://opg.optica.org/ao/abstract.cfm?uri=ao-36-33-8710>. Acesso em: 16 dez. 2025.*
62. *INTERNATIONAL OCEAN COLOUR COORDINATING GROUP (IOCCG). Remote Sensing of Inherent Optical Properties: Fundamentals, Tests of Algorithms, and Applications. Lee, Z.-P. (Ed.). IOCCG Report Number 5, 2006. Dartmouth, Canada: IOCCG. ISSN: 1098-6030. ISBN: 978-1-896246-56-7. Disponível em: <https://ioccg.org/reports/report5.pdf>. Acesso em: 16 dez. 2025.*

63. *GHOLIZADEH, M. H.; MELESSE, A. M.; REDDI, L. A comprehensive review on water quality parameters estimation using remote sensing techniques. Sensors, v. 16, n. 8, 1298, 16 ago. 2016. DOI: 10.3390/s16081298. PMID: 27537894. PMCID: PMC5017463. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5017463/>. Acesso em: 16 dez. 2025.*

64. *MINNERICK, R. J. Michigan lakes: An assessment of water quality. USGS Fact Sheet 2004-3048, out. 2004. Disponível em: <https://pubs.usgs.gov/fs/2004/3048/pdf/FS2004-3048.pdf>. Acesso em: 16 dez. 2025.*

65. *U.S. ENVIRONMENTAL PROTECTION AGENCY (US EPA). Indicators: Conductivity. National Aquatic Resource Surveys. Atualização: 9 jan. 2025. Disponível em: <https://www.epa.gov/national-aquatic-resource-surveys/indicators-conductivity>. Acesso em: 16 dez. 2025.*

66. *AQION. Temperature Compensation for Conductivity. 2023. Normalização EC 25°C (~2%/°C). Disponível em: <https://www.aqion.de/site/112>. Acesso em: 16 dez. 2025.*

67. *INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). Machine learning (ML): All there is to know. 2024. Visão geral normas ISO/IEC ML/AI. Disponível em: <https://www.iso.org/artificial-intelligence/machine-learning>. Acesso em: 16 dez. 2025.*

68. *WOLFRAM RESEARCH. Machine Learning Paradigms. In: Introduction to Machine Learning in Wolfram Language. Disponível em: <https://www.wolfram.com/language/introduction-machine-learning/machine-learning-paradigms/>. Acesso em: 17 dez. 2025.*

69. *XGBoost DEVELOPERS. XGBoost Parameters. XGBoost Documentation, v2.1.1 (stable). Disponível em: <https://xgboost.readthedocs.io/en/stable/parameter.html>. Acesso em: 17 dez. 2025.*

70. *NIAZKAR, M. et al. Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). Environmental Modelling & Software, v. 174, 105971, fev. 2024. DOI: 10.1016/j.envsoft.2024.105971. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S136481522400032X>. Acesso em: 17 dez. 2025.*

71. *BANERJEE, P. A Guide on XGBoost hyperparameters tuning. Kaggle, 14 jul. 2020. Notebook tutorial. Disponível em: <https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning>. Acesso em: 10 dez. 2025.*

72. COMISSÃO NACIONAL PARA AS BARRAGENS (CNPG). Castelo do Bode. Disponível em: <https://cnpgb.apambiente.pt/content/castelo-do-bode>. Acesso em: 11 dez. 2025.

73. AGÊNCIA PORTUGUESA DO AMBIENTE (APA). Estação pt16h12c - Rede Nacional de Monitorização da Qualidade da Água. Sistema Nacional de Informação de Recursos Hídricos (SNIRH). Disponível em: <https://snirh.apambiente.pt/snirh/_dadosbase/site/simplex.php?FILTRA_COVER=5453&FILTRA_SIMBOLO=16H/12C>. Acesso em: 10 dez. 2025.

74. FLYROBOTICS, LDA. Drone ASV1 - Especificações Técnicas. Site oficial. Disponível em: <https://flyrobotics.pt/>. Acesso em: 17 dez. 2025. Comunicação Eng.º Hugo Magalhães.

75. KOMOROWSKI, M.; MARSHALL, D. C.; SALCICCIOLI, J. D.; CRUTAIN, Y. Chapter 15: Exploratory Data Analysis. In: SALCICCIOLI, J. D. et al. (Eds.). Secondary Analysis of Electronic Health Records. Cham: Springer, 2016. p. 189-208. (Open Access). Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK543641/>. Acesso em: 17 dez. 2025

76. GOOGLE EARTH ENGINE TEAM. Google Earth Engine: Planetary-scale platform for Earth science data & analysis. Disponível em: <https://earthengine.google.com/>. Acesso em: 10 dez. 2025.

77. MARUJO, R. F. B.; FRONZA, J. G.; SOARES, A. R.; QUEIROZ, G. R.; FERREIRA, K. R. Evaluating the impact of LaSRC and Sen2Cor atmospheric correction algorithms on Landsat-8/OLI and Sentinel-2/MSI data over AERONET stations in Brazilian territory. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, v. V-3-2021, p. 271-277, 2021. DOI: 10.5194/isprs-annals-V-3-2021-271-2021. Disponível em: <https://isprs-annals.copernicus.org/articles/V-3-2021/271/2021/>. Acesso em: 10 out. 2025.

78. WARREN, M. A.; SIMIS, S. G. H.; SELMES, N. Complementary water quality observations from high and medium resolution Sentinel sensors by aligning chlorophyll-a and turbidity algorithms. Remote Sensing of Environment, v. 252, 112136, 2021. DOI: 10.1016/j.rse.2020.112136. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8507437/>. Acesso em: 20 nov. 2025.

79. PYO, J.; PACHEPSKY, Y.; KIM, S.; ABBAS, A.; KIM, M.; KWON, Y. S.; LIGARAY, M.; CHO, K. H. Long short-term memory models of water quality in inland water environments. Environmental Advances, v. 13, 100318, 2023. DOI: 10.1016/j.envadv.2023.100318. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2589914723000439>. Acesso em: 20 nov. 2025.

80. RODRIGUES, G. et al. The Use of Sentinel-3/OLCI for Monitoring the Water Quality Parameters in Alqueva Reservoir. Remote Sensing, v. 14, n. 21, 5514,

2022. DOI: 10.3390/rs14215514. Disponível em: <https://www.mdpi.com/2072-4292/14/21/5514>. Acesso em: 20 nov. 2025.

81. HYNDMAN, R. J.; ATHANASOPOULOS, G. Forecasting: Principles and Practice. 3. ed. Melbourne: OTexts, 2021. Disponível em: <https://otexts.com/fpp3/>. Acesso em: 18 dez. 2025.

Nota de Transparência: Durante a elaboração do presente trabalho, recorreu-se a ferramentas de inteligência artificial generativa como apoio à elaboração, organização e revisão do conteúdo.

Apêndices

Apêndice A - Especificações Técnicas Detalhadas

Servidor de Processamento:

- Python 3.12+
- XGBoost 2.0+
- Google Earth Engine API
- O desenvolvimento foi realizado em Python, utilizando as bibliotecas pandas, NumPy, scikit-learn, Matplotlib, Seaborn, Streamlit, Flask e XGBoost, bem como todas as demais dependências necessárias e declaradas no ambiente do projeto

Armazenamento de Dados:

- conjunto de dados de matchup: 98 amostras × 17 variáveis = ~2 MB
- Modelos treinados: ~5 MB cada (2 modelos)
- Imagens Sentinel-2 histórico: ~500 GB (processadas em cloud)

Apêndice B - UsarGoogle Earth Engine para Pré-processamento e Extração de Índices Sentinel-2

NOTA / ATRIBUIÇÃO

Este script foi desenvolvido no Google Earth Engine para extrair bandas e índices espectrais do Sentinel-2 SR em janelas temporais centradas em observações in situ, e exportar os resultados para CSV.
Implementação inspirada/adaptada de documentação oficial do Google Earth Engine:
- Dataset Sentinel-2 SR Harmonized (bandas SR escaladas; inclui a banda SCL):
https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED
- Exemplo de cálculo de índices e mapeamento de funções sobre coleções (map/addBands):
https://developers.google.com/earth-engine/tutorials/tutorial_api_06
Notas:
- O método de máscara por SCL, a escolha de bandas/índices e os parâmetros (janela temporal, escala, limiar de nuvens, seletores de exportação) foram ajustados para este trabalho.
Fonte original: Copernicus Sentinel data (2017–2025).
Atribuição (processado): Contains modified Copernicus Sentinel data (2017–2025).

PROCEDIMENTO sentinel2castelobode_gerar_matchups_S2_insitu()

Integrar observações in situ com variáveis espectrais Sentinel-2, obtidas numa janela temporal centrada no instante de observação, e exportar numa tabela de matchups.

```
# 1) Dados in situ e domínio espacial
insitu <- carregar_FeatureCollection(INSITU_ASSET)
roi <- obter_geometria_global(insitu)

# 2) Coleção Sentinel-2 e critérios de seleção
colecao_S2 <- ImageCollection("COPERNICUS/S2_SR_HARMONIZED")
```

```
        .filtrar_por_geometria(roi)
        .filtrar_por_data(START, END)
        .filtrar_por_metadado("CLOUDY_PIXEL_PERCENTAGE" <= MAX_CLOUD)

  # 3) Pré-processamento radiométrico e temático (por imagem)
  FUNÇÃO sentinel2castelobode_filtrar_pixéis_por_SCL(imagem)
    scl <- obter_banda(imagem, "SCL")

    # Excluir classes não fiáveis (ex.: nuvens/sombras/neve-gelo)
    mascara_fiabilidade <- (scl ≠ 3) E (scl ≠ 7) E (scl ≠ 8) E (scl ≠ 9) E (scl ≠ 10) E (scl ≠
11)

    SE AGUA_APENAS:
      mascara_fiabilidade <- mascara_fiabilidade E (scl = 6)   # restringir a água

    DEVOLVER aplicar_mascara(imagem, mascara_fiabilidade)

  FUNÇÃO sentinel2castelobode_derivar_variaveis(imagem)
    imagem_f <- sentinel2castelobode_filtrar_pixéis_por_SCL(imagem)

    # Reflectância de superfície normalizada
    sr <- selecionar_bandas(imagem_f, BASE_BANDS) / 10000

    # Índices/racios espectrais
    ndvi <- (B8 - B4)  / (B8 + B4)
    ndci <- (B6 - B4)  / (B6 + B4)
    ndmi <- (B8 - B11) / (B8 + B11)
    ndbi <- (B11 - B8) / (B11 + B8)
    r_b3b2 <- B3 / B2
    r_b4b8 <- B4 / B8

    imagem_vars <- juntar_bandas(sr, [ndvi, ndci, ndmi, ndbi, r_b3b2, r_b4b8])
    DEVOLVER manter_bandas(imagem_vars, FINAL_BANDS)

  # 4) Imagem nula
```

```
FUNÇÃO sentinel2castelobode_imagem_nula()
    img0 <- criar_imagem_constante(0, bandas=FINAL_BANDS)
    DEVOLVER mascarar_totalmente(img0)

# 5) Conversão de tempo (in situ)
FUNÇÃO sentinel2castelobode_ler_valor_tempo
    valor_iso <- normalizar_iso_millis(valor)
    DEVOLVER converter_para_dados(valor_iso)

# 6) Matchup por observação (janela temporal + composição + estatística zonal)
FUNÇÃO sentinel2castelobode_matchup_por_observ
    t0 <- sentinel2castelobode_ler_tempo(obs[col_tempo])
    t_ini <- t0 - WINDOW_DAYS dias
    t_fim <- t0 + WINDOW_DAYS dias

    janela <- colecao_S2.filtrar_por_data(t_ini, t_fim)
    n <- tamanho(janela)

    SE n > 0:
        # Composição temporal robusta (mediana) sobre variáveis pré-processadas
        img_comp <- mediana( janela.map(sentinel2castelobode_derivar_variaveis) )
    SENÃO:
        img_comp <- sentinel2castelobode_imagem_nula()

    # Estatística zonal na geometria da observação (um valor por variável)
    z <- reduzir_regiao(
        imagem      = img_comp,
        estatistica = "primeiro_valor",
        area        = geometria(obs),
        resolucao   = SCALE,
        ajuste_auto = verdadeiro,
        limite_pix  = 1e9
    )
```

```
  DEVOLVER obs
          .set(z)
          .set("contagem_s2", n)
          .set("window_start", formatar_data(t_ini))
          .set("window_end",   formatar_data(t_fim))

    # 7) Execução em lote e exportação
    tabela_matchups <- insitu.map(sentinel2castelobode_matchup_por_obs)

    Export.table.toDrive(
       collection = tabela_matchups,
       fileFormat = "CSV",
           selectors   = [campos_insitu, "contagem_s2", "window_start", "window_end",
FINAL_BANDS,
                  opcional ".geo" se EXPORTAR_GEOMETRIA]
    )
```

Apêndice C – Implementação dos Modelos XGBoost para Estimação da Qualidade da Água

```
CONSTANTES sentinel2castelobode
    sentinel2castelobode_COL_PARTICAO_OPCOES <- ["conjunto_final", "conjunto"]
    sentinel2castelobode_ENTRADAS <- FEATURE_COLUMNS
    sentinel2castelobode_FRACAO_VALIDACAO <- VALIDATION_FRACTION
    sentinel2castelobode_GRELHA_HIPERPARAM <- HYPERPARAM_GRID

FUNÇÃO sentinel2castelobode_obter_coluna_particao(df)
    PARA cada nome EM sentinel2castelobode_COL_PARTICAO_OPCOES:
        SE nome existe em df.colunas: DEVOLVER nome

PROCEDIMENTO sentinel2castelobode_treinar_avaliar(df, alvo)
    col_particao <- sentinel2castelobode_obter_coluna_particao(df)

    treino_total <- filtrar(df, df[col_particao] == "treino")
    teste        <- filtrar(df, df[col_particao] == "teste")

    treino_total <- remover_nan(treino_total, sentinel2castelobode_ENTRADAS + [alvo])
    teste        <- remover_nan(teste,        sentinel2castelobode_ENTRADAS + [alvo])

treino, validacao <- dividir_aleatorio(treino_total,
sentinel2castelobode_FRACAO_VALIDACAO)

    melhores_hp <- sentinel2castelobode_selecionar_hiperparametros(treino, alvo,
    grelha=sentinel2castelobode_GRELHA_HIPERPARAM, k=5)

    modelo <- sentinel2castelobode_treino_final(treino, validacao, alvo, melhores_hp)

    m_treino <- sentinel2castelobode_avaliar(modelo, treino, alvo)
    m_val    <- sentinel2castelobode_avaliar(modelo, validacao, alvo)
    m_teste  <- sentinel2castelobode_avaliar(modelo, teste, alvo)
```

DEVOLVER (melhores_hp, m_treino, m_val, m_teste)

Apêndice D - Avaliação do Desempenho dos Modelos XGBoost (Previsões, Erros e Importância das Variáveis)

```
ENTRADAS
  modelo
  (X_treino, y_treino), (X_val, y_val), (X_teste, y_teste)
  nomes_entradas

SAÍDAS
  tabela_metricas
  fig_residuos_hist
  fig_residuos_vs_ajustado
  p_normalidade
  fig_importancias
  tabela_importancias

FUNÇÃO sentinel2castelobode_metricas_regressao(y_real, y_ajustado):
  RMSE <- sqrt( media( (y_real - y_ajustado)^2 ) )
  MAE  <- media( abs(y_real - y_ajustado) )
  R2   <- 1 - soma( (y_real - y_ajustado)^2 ) / soma( (y_real - media(y_real))^2 )
  DEVOLVER (RMSE, MAE, R2)

PROCEDIMENTO sentinel2castelobode_avaliar_modelo(modelo):
  # 1) Valores ajustados por conjunto
  y_ajustado_treino <- prever(modelo, X_treino)
  y_ajustado_val    <- prever(modelo, X_val)
  y_ajustado_teste  <- prever(modelo, X_teste)

  # 2) Métricas (treino/validação/teste)
  m_treino <- sentinel2castelobode_metricas_regressao(y_treino, y_ajustado_treino)
  m_val    <- sentinel2castelobode_metricas_regressao(y_val,    y_ajustado_val)
  m_teste  <- sentinel2castelobode_metricas_regressao(y_teste,  y_ajustado_teste)

  tabela_metricas <- criar_tabela(
```

```
    linhas=["Treino","Validação","Teste"],
    colunas=["RMSE","MAE","R2"],
    valores=[m_treino, m_val, m_teste]
  )
  guardar_csv(tabela_metricas, "sentinel2castelobode_metricas.csv")

  # 3) Resíduos no teste
  residuos <- y_teste - y_ajustado_teste   # resíduos = observado - ajustado

  fig_residuos_hist <- histograma(residuos, titulo="Resíduos (teste)")
  fig_residuos_vs_ajustado <- dispersao(
    x=y_ajustado_teste,
    y=residuos,
    linha_y0=verdadeiro,
    titulo="Resíduos vs ajustado (teste)"
  )
guardar_figura(fig_residuos_hist, "sentinel2castelobode_residuos_hist.png")
guardar_figura(fig_residuos_vs_ajustado,"sentinel2castelobode_residuos_vs_ajustado.p
ng")

  # 4) Normalidade dos resíduos (Shapiro–Wilk)
  (_, p_normalidade) <- shapiro_wilk(residuos)

  # 5) Importância das variáveis
  importancias <- extrair_importancia(modelo)
  ordem <- ordenar_indices(importancias, desc=True)

  fig_importancias <- barras_horizontais(
    valores=importancias[ordem],
    etiquetas=nomes_entradas[ordem],
    titulo="Importância das variáveis"
  )
  guardar_figura(fig_importancias, "sentinel2castelobode_importancias.png")

  tabela_importancias <- criar_tabela(
```

```
    colunas=["Variável","Importância","Importância_%"],
    valores=para cada i em ordem:
            (nomes_entradas[i], importancias[i], 100*importancias[i]/soma(importancias))
)
guardar_csv(tabela_importancias, "sentinel2castelobode_importancias.csv")

DEVOLVER (tabela_metricas, p_normalidade, tabela_importancias)
```

Apêndice E – Dashboard Streamlit

Nota de referência: A aplicação foi desenvolvida em Python usando Streamlit (https://streamlit.io) para a interface; Streamlit é distribuído sob Apache License 2.0; foram utilizadas, entre outras, as bibliotecas scikit-learn, XGBoost e SciPy.

```
PROCEDIMENTO sentinel2castelobode_app_streamlit()

  # 0) Inicialização da interface
  definir_pagina(titulo=SENTINEL2CASTELOBODE_PAGE_TITLE,
          layout=SENTINEL2CASTELOBODE_PAGE_LAYOUT)

  escrever_titulo(SENTINEL2CASTELOBODE_APP_TITLE)
  escrever_subtitulo(SENTINEL2CASTELOBODE_APP_CAPTION)

  # 1) Estrutura de navegação (abas)
  (aba_predicao, aba_sobre) <- criar_abas(["Predição", "Sobre"])

  # 2) Aba: Predição
  COM aba_predicao:

    escrever_texto("Previsão a partir de bandas Sentinel-2")

    # 2.1) Entrada de dados (valores normalizados 0..1)
    b2  <- entrada_slider("B2",  intervalo=[0,1], valor_inicial=0.40)
    b3  <- entrada_slider("B3",  intervalo=[0,1], valor_inicial=0.25)
    b4  <- entrada_slider("B4",  intervalo=[0,1], valor_inicial=0.30)
    b5  <- entrada_slider("B5",  intervalo=[0,1], valor_inicial=0.30)
    b6  <- entrada_slider("B6",  intervalo=[0,1], valor_inicial=0.35)
    b7  <- entrada_slider("B7",  intervalo=[0,1], valor_inicial=0.30)
    b8  <- entrada_slider("B8",  intervalo=[0,1], valor_inicial=0.40)
    b11 <- entrada_slider("B11", intervalo=[0,1], valor_inicial=0.20)
    b12 <- entrada_slider("B12", intervalo=[0,1], valor_inicial=0.15)
```

```
    # 2.2) Derivar índices espectrais
                                (ndvi, ndci, ndmi, ndbi, r_b3b2, r_b4b8) <-
sentinel2castelobode_calcular_indices(b2,b3,b4,b6,b8,b11)

    # 2.3) Construir vetor de entrada na ordem do modelo
    mapa <- {
        "B2":b2, "B3":b3, "B4":b4, "B5":b5, "B6":b6, "B7":b7, "B8":b8, "B11":b11,
"B12":b12,
         "NDVI":ndvi, "NDCI":ndci, "NDMI":ndmi, "NDBI":ndbi, "B3B2":r_b3b2,
"B4B8":r_b4b8
    }
X <- linha_1xN( para cada f em SENTINEL2CASTELOBODE_FEATURE_COLUMNS: mapa[f] )

    # 2.4) Inferência (ação do utilizador)
    SE botao("Prever") for acionado:
      X_temp <- aplicar_scaler(scaler_temp, X)
      X_cond <- aplicar_scaler(scaler_cond, X)

      temp_prev <- prever(modelo_temp, X_temp)
      cond_prev <- prever(modelo_cond, X_cond)

      mostrar_valor("Temperatura prevista (°C)", temp_prev)
      mostrar_valor("Condutividade prevista (µS/cm)", cond_prev)

  # 3) Aba: Sobre
Com aba_sobre, escrever_texto("Os modelos são carregados a partir do diretório de
artefactos e reutilizados durante a execução.")
```

Apêndice F – API REST

CONSTANTES:
- SENTINEL2CASTELOBODE_HOST = "localhost"
- SENTINEL2CASTELOBODE_PORT = 8443

-SENTINEL2CASTELOBODE_FEATURE_COLUMNS=
["B2","B3","B4","B5","B6","B7","B8","B11","B12",
"NDVI","NDCI","NDMI","NDBI","B3B2","B4B8"]

- SENTINEL2CASTELOBODE_PATH_SCALER_TEMP = "scaler_temp.*"
- SENTINEL2CASTELOBODE_PATH_MODEL_TEMP = "model_temp.*"
- SENTINEL2CASTELOBODE_PATH_SCALER_COND = "scaler_cond.*"
- SENTINEL2CASTELOBODE_PATH_MODEL_COND = "model_cond.*"

PROCEDIMENTO sentinel2castelobode_iniciar_api():
app <- criar_app_web()

# carregar uma vez no arranque (evita reload a cada pedido)
scaler_temp <- load(SENTINEL2CASTELOBODE_PATH_SCALER_TEMP)
model_temp <- load(SENTINEL2CASTELOBODE_PATH_MODEL_TEMP)
scaler_cond <- load(SENTINEL2CASTELOBODE_PATH_SCALER_COND)
model_cond <- load(SENTINEL2CASTELOBODE_PATH_MODEL_COND)

registar_GET(app, "/v2/health", sentinel2castelobode_health)
registar_POST(app, "/v2/valor", sentinel2castelobode_valor)
executar_servidor(app,host=SENTINEL2CASTELOBODE_HOST,
port=SENTINEL2CASTELOBODE_PORT, debug=false)

FUNÇÃO sentinel2castelobode_health(pedido):
DEVOLVER JSON({"OK"}), HTTP 200 # health check simples

```
FUNÇÃO sentinel2castelobode_valor(pedido):
    data <- pedido.json

    # validar campos obrigatórios
    falta_dados <- [f para f em SENTINEL2CASTELOBODE_FEATURE_COLUMNS se f não está em data]
    SE falta_dados não vazio:
        DEVOLVER JSON({"erro":"campos_em_falta"}), HTTP 400

    # construir vetor
    tentar:
valores_matriz←valor_num([float(dados[f]) para f em SENTINEL2CASTELOBODE_FEATURE_COLUMNS])
    se falhar:
        DEVOLVER JSON({"erro":"erro_numerico"}), HTTP 400

    # prep + inferência
    tentar:
        x_temp <- scaler_temp.transform(valores_matriz)
        x_cond <- scaler_cond.transform(valores_matriz)

        temp_pred <- primeiro_valor(model_temp.valor(matriz_temp))
        cond_pred <- primeiro_valor(model_cond.valor(matriz_cond))
    se falhar:
        DEVOLVER JSON({"erro":"previsao_falhada"}), HTTP 500

    DEVOLVER JSON({
        "temperaturaC": temp_val,
        "condutividadeuScm25C": cond_val
    }), HTTP 200
```

Apêndice G – Análise exploratória de dados (EDA)

```
DEFINIÇÕES (Castelo do Bode)
   FONTE_CSV   <- "matchupdataCasteloBode.csv"
   PASTA_EDA   <- "eda_outputs/sentinel2castelobode"
                                          ENTRADAS                                          <-
["B2","B3","B4","B5","B6","B7","B8","B11","B12","NDVI","NDCI","NDMI","NDBI
","B3B2","B4B8"]
   SAIDAS      <- ["temperatura_c", "condutividade_uS_cm_25C"]
   CHAVES_conjunto <- ["conjunto_final", "conjunto"]
   Subconjuntos     <- "total"     # "total" | "treino" | "teste"

PROCEDIMENTO varrer_dados_castelobode()
   tabela <- ler_csv(FONTE_CSV)

   SE existe_coluna(tabela, "time"):
      tabela["time"] <- tentar_converter_data(tabela["time"])

   # escolher a coluna de partição (se existir)
   coluna_particao <- NULO
   PARA cada k EM CHAVES_conjunto:
    SE existe_coluna(tabela, k): coluna_particao <- k; PARAR
   campos <- manter_existentes(tabela, ENTRADAS + SAIDAS)

   # 1) radiografia de falhas
   n_faltas  <- contar_vazios(tabela[campos])
   perc_faltas <- percentagem(n_faltas, total=numero_linhas(tabela))
   guardar_csv(n_faltas, "/falhas_n.csv")
   guardar_csv(perct_faltas, "/falhas_perc.csv")

   # 2) versão pronta para análise
   dados <- remover_linhas_com_vazios(tabela, campos)

#Segmentar o subconjunto de dados de treino e/ou teste e analisar o mesmo
 # 3) sinais dos alvos (números + forma)
```

```
guardar_csv(resumo_basico(dados[SAIDAS]), "/alvos_resumo.csv")
guardar_csv(assimetria(dados[SAIDAS]), "/alvos_assimetria.csv")

# 4) ligações entre variáveis (correlação + mapa)
vars_corr <- manter_existentes(dados, ENTRADAS + SAIDAS)
M <- correlacao(dados[vars_corr])
guardar_csv(M, "/mapa_c.csv")
guardar_heatmap(M, "/mapa_c.png")

# 5) gráficos rápidos dos alvos
PARA cada y EM SAIDAS:
   guardar_histograma(dados[y],"/distribuicao.png")
   guardar_boxplot(dados[y],"/caixa.png")

imprimir("Pronto: varredura EDA)
```