**ANDERSON PAULINO DE SOUZA**

**Modelagem da Qualidade da Água de Reservatórios**

**por Meio de Técnicas de Aprendizado de Máquina.**

**Estudo de Caso: Reservatório de Três Marias/MG.**

Projeto de pesquisa de doutorado apresentado ao Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos da Universidade Federal de Minas Gerais como requisito da disciplina de Seminários.

Área de Concentração: Meio Ambiente.

Linha de Pesquisa: Avaliação e Gerenciamento de Impactos e de Riscos Ambientais.

Orientadora: Camila Costa de Amorim Amaral

Coorientador: Jefersson Alex dos Santos

**BELO HORIZONTE**

**Janeiro/2022**

**1. INTRODUÇÃO**

A Inteligência Artificial (IA) pode ser definida como um conjunto de sistemas que simulam o raciocínio e a atividade humana, sendo capaz de aprender, reconhecer problemas, analisar dados e tomar decisões (SANTOS, 2016). O termo *Machine Learning*, Aprendizado de Máquina, é um subcampo da IA, um ramo de estudo que provê aos computadores a capacidade de aprender sem ser explicitamente programados para tal (SAMUEL, 1959). Dentre os diversos tipos de aprendizado inerentes ao *Machine Learning*, destaca-se o Aprendizado Supervisionado, técnica na qual são fornecidos conjuntos de dados de exemplos ao computador, que por meio da experiência e característica dos dados, aprende e identifica padrões associativos, tornando-se apto a prever determinada categoria (classificação) ou valor (regressão) de determinado dado (Kotsiantis *et al*., 2007).

A aplicação de técnicas de *Machine Learning* normalmente requer grandes volumes de dados e exige elevadas capacidades de processamento computacional, fatores muitas vezes indisponíveis. Neste contexto, *Transfer Learning*, Transferência de Conhecimento (TC), apresenta-se como solução plausível, a técnica permite que os conhecimentos adquiridos em atividades anteriores e/ou por outros modelos de ML sejam transferidos para uma nova tarefa ou modelo (TAN *et al*., 2018).

A aplicação de técnicas de *Machine Learning* (ML), especificamente Aprendizado Supervisionado, e Transferência de Conhecimento (TC) no processo de estimativa de variáveis de qualidade da água de reservatórios constitui o eixo central desta pesquisa de doutorado.

A água é um fator crítico para o desenvolvimento e manutenção da sociedade, a utilização múltipla do recurso está atrelada à sua disponibilidade qualitativa e quantitativa. Neste contexto, grandes reservatórios têm papel fundamental, caracterizados pela retenção de substanciais volumes de água, nutrientes e sedimentos, estes detêm capacidade de transformação de habitats e são importantes provedores de serviços ecossistêmicos. Normalmente, a água armazenada nos reservatórios tem usos múltiplos, sendo os mais comuns: abastecimento público, geração de energia, irrigação e controle de inundação (Gernaat *et al*., 2017).

O nível da qualidade da água armazenada determina o estágio de conservação dos reservatórios, bem como delimita a utilização do recurso pela sociedade e ecossistemas, e pode ser estimado por meio do monitoramento de variáveis-chave, sendo as principais: clorofila-a, matéria orgânica dissolvida, profundidade do disco de *Secchi*, turbidez, sólidos suspensos totais, temperatura, fósforo total, oxigênio dissolvido, demandas química e bioquímica de oxigênio (Gholizadeh *et al*., 2016).

O monitoramento da qualidade da água em grandes reservatórios requer mensuração de elevado número de parâmetros físico-químicos e biológicos, frequentemente realizada por equipamentos especializados, apresenta altos custos operacionais (coleta/análise), elevado tempo de processamento de dados, frente à escassez de recursos humanos/financeiros (KARAOUI *et al*., 2019). Neste contexto, o uso de técnicas de ML e TC aplicadas em imagens de sensoriamento remoto provê elementos capazes de estimar variáveis de qualidade da água com bons níveis de acurácia, permitindo diagnósticos objetivos e confiáveis, possibilitando o monitoramento dos parâmetros de forma custo-efetiva e favorecendo processos de tomada de decisão, tendo em perspectiva o uso sustentável dos recursos hídricos.

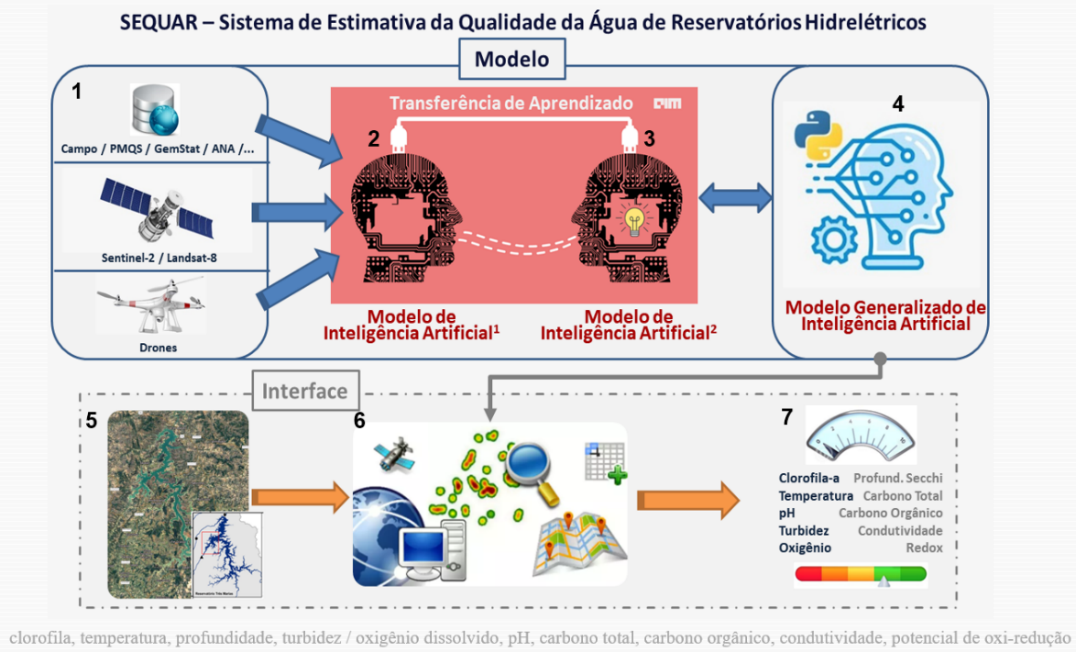
A aplicação de técnicas de ML e TC no processo de estimativa de variáveis de qualidade da água exploradas nesta pesquisa têm como objeto de análise o Reservatório de Três Marias, localizado na região central do Estado de Minas Gerais, pertencente à bacia do Rio São Francisco. O reservatório tem múltiplas finalidades, sendo a principal a acumulação de volume para geração de energia. No âmbito de integração entre tecnologia e recursos hídricos, esta pesquisa irá dispor entre seus produtos a implementação de um sistema computacional de auxílio à gestão de recursos hídricos especializado na estimativa de variáveis da qualidade da água de reservatórios fundamentado na aplicação de técnicas de ML e TC: SEQUAR – Sistema de Estimativa da Qualidade da Água de Reservatórios, no âmbito do Grupo de Pesquisa – SIMOA – Sistemas Inteligentes de Monitoramento Ambiental.

O SEQUAR consiste basicamente em um sistema disponibilizado em *nuvem* que possibilitará ao usuário estimar variáveis de qualidade da água para determinada coordenada geográfica inserida em um reservatório. A partir da seleção da data e do ponto geográfico de interesse, acionado via interação com mapas ou inserção de coordenadas, o sistema permitirá estimar valores de variáveis de qualidade da água, tais quais: clorofila-a, temperatura da superfície da água, profundidade do disco de *Sechhi*, turbidez, oxigênio dissolvido, pH, carbono total, carbono orgânico, condutividade, dentre outras.

O fluxo funcional básico de modelagem e uso do SEQUAR é apresentado na Figura 1. As variáveis de entrada correspondem às imagens de satélites, drones e dados estruturados inerentes ao(s) reservatório(s) de interesse e serão processadas por um Modelo1 de ML com capacidade para estimar variáveis de qualidade da água. Por meio de técnicas de TC, os conhecimentos assimilados pelo Modelo1 serão transmitidos para outro Modelo2 de ML, cujo algoritmo deverá avaliar um reservatório/conjunto de reservatórios que não tenha sido contemplado pelo Modelo1. Deste modo, será possível acelerar o tempo de aplicação dos algoritmos e aumentar a acurácia na estimativa dos parâmetros de qualidade da água. O conhecimento assimilado pelo Modelo2 será então transmitido ao Modelo3 (inicialmente aplicado para validação no Reservatório de Três Marias), que deverá dispor de maior capacidade de generalização e aplicabilidade em reservatórios dispostos em localidades e contextos geoambientais distintos.

O estudo se justifica devido à fragmentada difusão de trabalhos relacionados à aplicação de técnicas de TC no contexto de reservatórios, de modo a fornecer elementos que permitam melhor compreensão dos fatores potencializadores e limitantes na utilização das mesmas no processo de estimativa de variáveis de qualidade da água neste escopo. Outro fundamento concerne à utilização de múltiplas fontes de dados, conferindo aos modelos de ML aplicados maior capacidade de generalização.

Figura 1 – SEQUAR: Fluxo básico



Embora a aplicação de algoritmos tenha registrado avanços significativos nas últimas décadas no monitoramento e estimativa de parâmetros de qualidade da água (MONTANHER, 2013), os resultados obtidos pelos modelos normalmente são válidos apenas para a região objeto de estudo, área para a qual os dados foram calibrados junto ao algoritmo. Esta limitação restringe o desenvolvimento e aplicação de ferramentas de monitoramento da qualidade da água em contextos generalizados, a exploração dessa lacuna técnico-científica irá possibilitar identificar nuances e novas relações que permeiam indicadores de qualidade da água, ao passo que aplicações inovadoras poderão ser desencadeadas no contexto das ciências hidrológicas e no processo de gestão dos recursos hídricos.

A elaboração do trabalho será pautada pelas seguintes etapas: revisão de literatura, caracterização da área de estudo, seleção das variáveis de qualidade da água, identificação de reservatórios para aplicação dos modelos, criação de banco de dados, extração de dados de imagens de satélites, treinamento dos modelos de ML, aplicação de modelos de TC e criação do Modelo Generalizado.

**2. REVISÃO DE LITERATURA**

**2.1 Reservatórios e Variáveis de Qualidade da Água**

Reservatório pode ser definido como um corpo d'água contido por aterros ou barragens, tendo como objetivo principal fornecer água para a manutenção das atividades humanas, em geral são caracterizados pelo barramento de um rio, formando uma área de inundação de montante (Thorton *et al*., 1996). Os reservatórios são corpos d’água artificiais e diferem-se dos lagos (naturais) quanto à origem, fatores de forma, tempo de detenção, dentre outros aspectos. Em termos ecossistêmicos, reservatórios e lagos são classificados como ambientes lênticos, em contrapartida aos lóticos (rios, riachos, etc.), estes se caracterizam pela retenção de grandes volumes de água, nutrientes e sedimentos, bem como pela regulação de processos hidrológicos, conferindo-lhes capacidade de formação e transformação de habitats, sendo importantes provedores de serviços ecossistêmicos (SCHALLENBERG *et al*., 2013).

Naime (2012) identificou uma série de impactos socioambientais associados à construção e operação de grandes reservatórios de água, dos quais podemos destacar: deslocamento de populações; inundação de áreas agricultáveis; perdas de espécies de flora e fauna nativas; interferência em processos migratórios e reprodutivos da ictiofauna; danos ao patrimônio histórico e cultural; maior disseminação de doenças de veiculação hídrica; intensificação de atividades extrativistas; alterações das características físicas e químicas da água (temperatura, oxigênio dissolvido, pH, etc.)

Ressalvados os aspectos negativos, a ampla difusão de uso destas estruturas pela sociedade é mitigada pela sua capacidade inerente de armazenamento de um dos recursos mais fundamentais à manutenção da vida, a água. Não por acaso, cientistas vêm alertando a comunidade internacional a respeito do risco global de escassez de água doce para as presentes e futuras gerações, preocupação oficializada na Agenda 2030, que contempla os 17 Objetivos de Desenvolvimento Sustentável, dos quais a água doce é item de destaque: Objetivo 6 – Assegurar a disponibilidade e gestão sustentável da água e saneamento para todas e todos (FONTES FILHO e ANJOS, 2016).

Jorgensen *et al*. (2013) identificaram mais de 60.000 represas com altura de barragem superior a 15 m distribuídas por todos os continentes. O grande número de reservatórios é por vezes justificado pelo aumento da demanda de água imposta pela sociedade. A demanda mais do que duplicou entre as décadas de 1960 e 2000, enquanto a capacidade de armazenamento dos reservatórios atingiu um nível de estabilidade a partir da década de 1990.

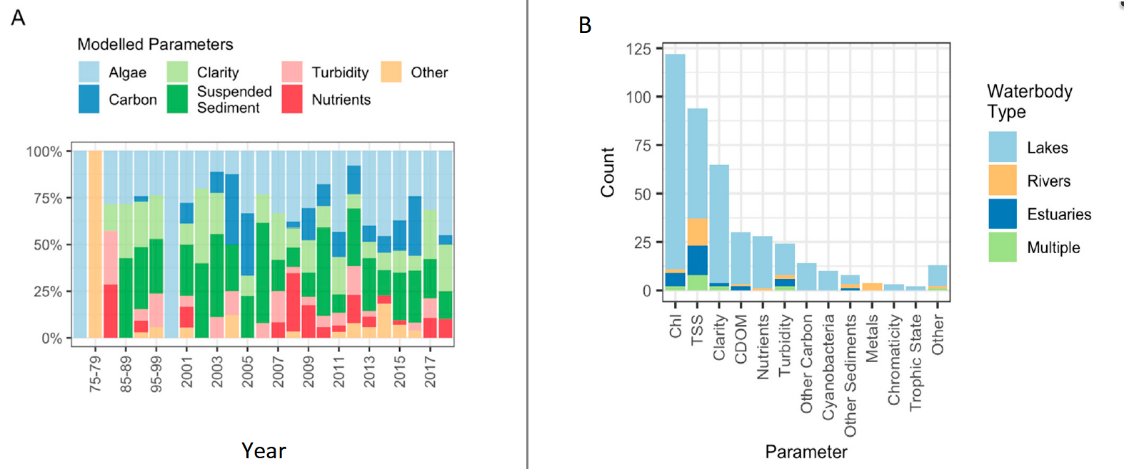
Além da disponibilidade quantitativa, o nível de qualidade da água armazenada nos reservatórios é fator crítico para a degradação da saúde humana e inviabilização da utilização do recurso em muitos processos produtivos e ecossistêmicos (TUNDISI, 2005). O termo qualidade da água refere-se ao “conjunto de concentrações, especiações e partições físicas de substâncias orgânicas, inorgânicas e a composição, diversidade e estado da biota encontrada em um determinado ecossistema aquático”. Essa qualidade apresenta variações temporais e espaciais, resultantes de processos físico-químicos e fatores externos/internos inerentes ao ecossistema aquático. Poluição do ecossistema aquático remete à introdução antrópica de substâncias ou energia que resultam em efeitos prejudiciais a: seres vivos; saúde humana; quantidade e qualidade da água disponível para uso em atividades agrícolas, econômicas e industriais. A qualidade da água indica as condições do sistema aquático e permite avaliar o estágio de degradação/conservação de rios, lagos, reservatórios, estuários, dentre outros, de modo a delimitar o uso da água de acordo com o nível de deterioração da sua qualidade (CHAPMAN, 1992).

Gholizadeh *et al*. (2016) destacam 11 onze parâmetros abordados com maior frequência em estudos relacionados à qualidade da água, a descrição dos mesmos está disposta no Quadro 1. Este conjunto de variáveis, acrescido o parâmetro pH (potencial hidrogeniônico), contempla os elementos centrais no monitoramento da qualidade da água de reservatórios (FIEDLER, 2010).

Quadro 1 – Principais Parâmetros de Avaliação da Qualidade da Água

| **Variável** | **Descrição** |
| --- | --- |
| **Clorofila** | Pigmentos esverdeados com capacidade de realizar fotossíntese, indicam o nível de eutrofização de lagos e reservatórios. Há diversos tipos de clorofila, sendo a cloforila-a a mais representativa, sua concentração está relacionada à disponibilidade de nutrientes e produção algal, o excesso de nutrientes prejudica o equilíbrio dos ecossistemas aquáticos (LIM *et al*., 2015). |
| **Oxigênio dissolvido (OD)** | Representa o nível de oxigênio dissolvido na água, é o parâmetro mais importante na avaliação da qualidade. Nível muito alto ou muito baixo de OD tem efeitos prejudiciais para a sobrevivência da biota aquática, especialmente peixes (ROUNDS *et al*., 2006). |
| **pH** | Potencial hidrogeniônico representa a concentração de íons hidrogênio H+ na água e determina a solubilidade e disponibilidade biológica dos constituintes químicos sob forma de nutrientes para o ecossistema (fósforo, nitrogênio e carbono). pH baixo pode prejudicar animais aquáticos e seres humanos, pH elevado pode resultar na formação de incrustações, água turva e reduzir a eficiência do tratamento da água (DICKSON, 1981). |
| **Salinidade** | Concentração de sais na massa de água, tendo como principais elementos responsáveis o sódio, cálcio, magnésio, potássio, cloretos, sulfatos e bicarbonatos. A salinidade afeta a densidade da água e o equilíbrio químico do ecossistema (FIEDLER, 2010). |
| **Turbidez** | Representa a alteração da capacidade de penetração da luz no corpo d’água, tendo como causa principal a presença de partículas (plâncton, bactérias, argilas, silte, poluição). Turbidez elevada interfere na oxigenação e produtividade fotossintética, obstrui os sistemas respiratório de peixes e macroinvertebrados e depositam poluentes nos corpos d’água, além de aumentar a temperatura da água (PORTO *et al*.; Jiang *et al*. 2018). |
| **Temperatura** | A variação, ainda que moderada, na temperatura pode afetar seriamente o ecossistema aquático, organismos aquáticos freqüentemente têm tolerâncias estreitas à temperatura, que também detém a capacidade de alterar a disponibilidade de gases dissolvidos, especialmente oxigênio. De forma geral, a temperatura controla as reações químicas e biológicas no corpo hídrico (Ranjeeta *et al*., 2011). |
| **Demanda bioquímica de oxigênio (DBO)** | Representa a quantidade de oxigênio dissolvido que será consumido por organismos aeróbios no processo de degradação da matéria orgânica biodegradável. A DBO sinaliza o nível de contaminação do corpo d'água, trechos com altas concentrações de poluentes apresentam níveis de DBO elevados e consequentemente menor disponibilidade de oxigênio para organismos aeróbios (VERMA *et al*., 2013). |
| **Demanda química de oxigênio (DQO)** | Representa a quantidade de oxigênio dissolvido consumido no processo de degradação de matéria orgânica, biodegradável ou não. O alto nível de DQO normalmente está associado à poluição por esgotos e defensivos agrícolas (Mustapha, 2008). |
| **Matéria orgânica dissolvida colorida** | Matérias orgânicas dissolvidas coloridas são substâncias orgânicas heterogêneas, biogênicas e solúveis em água, normalmente elas conferem à água uma cor amarela/ marrom, dependendo da concentração da matéria dissolvida, sendo predominantemente associada à presença de clorofila e sólidos suspensos (Aiken *et al*., 1985). |
| **Sólidos suspensos totais** | Correspondem à matéria em suspensão nos corpos d’água. A presença de sólidos suspensos pode provocar alterações do sabor da água, corrosão e dificultar o processo de tratamento da água (GONÇALVES, 2009). |
| **Fósforo total** | Representa a quantidade de fósforo, em suas diversas formas (especialmente ortofosfato), no corpo d’água, nutriente essencial para o desenvolvimento de plantas, animais e seres humanos. Altas concentrações de fósforo indicam a contaminação por águas residuais (esgoto doméstico) e resultam na eutrofização do ambiente aquático, causando deficiência de oxigênio, mortandade de peixes e outros organismos aquáticos (WAGLE *et al*., 2020). |
| **Profundidade do disco de *Secchi*** | Refere-se à profundidade em que o disco de *Secchi* (preto e branco) pode ser visto no corpo d’água. A transparência da água é afetada por dois fatores principais, algas e partículas em suspensão. A medida fornece indícios sobre o estado trófico da água (LEE *et al*., 2015). |

De modo complementar, Topp *et al*. (2020) identificaram as variáveis mais aplicadas em estudos relacionados estritamente a reservatórios, rios, lagos e estuários entre os anos 1974 e 2018. Conforme disposto na Figura 3 (A), modelagens atreladas aos parâmetros algas, carbono orgânico, transparência *Secchi*, sólidos suspensos, turbidez e nutrientes foram as mais representativas. Este aspecto corrobora com o levantamento disposto na Figura 3 (B), que sinaliza maior aplicação das variáveis clorofila-a, sólidos suspensos, transparência/*clarity* (*Secchi*), carbono orgânico, nutrientes e turbidez nos estudos analisados pelos autores.

Figura 3 – Variáveis mais aplicadas em modelagens hidrológicas 

Fonte: Adaptado de Topp *et al*., 2020.

O monitoramento da qualidade da água em grandes reservatórios requer a mensuração de um elevado número de parâmetros físico-químicos e biológicos, sendo, muitas vezes, realizado por medições com equipamentos especializados *in situ*. Os altos custos operacionais de coleta e análise de amostras, elevado tempo de processamento de dados, escassez de recursos humanos e financeiros, são fatores críticos no processo (KARAOUI *et al*., 2019), comprometendo o nível de efetividade da gestão dos recursos hídricos e o respectivo controle de qualidade da água, fator essencial às atividades socioeconômicas e à manutenção dos ecossistemas aquáticos. Associados à escassa rede de monitoramento hidrológico, estes aspectos configuram uma conjuntura complexa, em um cenário de informações hidrológicas restritas e, muitas vezes, de baixa qualidade, as modelagens hidrológicas têm a acurácia comprometida, resultando na perda de efetividade das políticas de gestão dos recursos hídricos (DUAN *et al*., 2021).

**2.2 Sensoriamento Remoto Aplicado aos Recursos Hídricos**

A aplicação de técnicas de sensoriamento remoto no monitoramento dos recursos hídricos tem se mostrado uma alternativa razoável frente às práticas tradicionais (*in situ*). A tecnologia está fundamentada no uso de sensores espaciais/aerotransportados, equipamentos com capacidade de receber e/ou emitir ondas eletromagnéticas, cada objeto/fenômeno observado emite ondas eletromagnéticas com frequências e comprimentos característicos, permitindo a identificação dos mesmos de acordo com os dados capturados pelos sensores (IBGE, 2001). No contexto hidrológico, a radiação refletida/emitida pela superfície da água é utilizada, ​​direta ou indiretamente, para determinar diferentes variáveis de qualidade da água, tais quais a concentração de poluentes, parâmetros físico-químicos e biológicos, fatores essenciais no monitoramento e avaliação do corpo hídrico. O fluxo de interação entre a radiação emitida pelo Sol, corpo d’água e o sensor orbital é apresentado na Figura 4 (Barbosa *et al*., 2019).

A interação da água com a luz elucida duas características essenciais do processo: propriedades ópticas inerentes (IOPs, *Inherent Optical Properties*), que representam as magnitudes, composição e concentração de substâncias particuladas e dissolvidas presentes na água (incluída) que não são afetadas por mecanismos externos e variações da intensidade de luz; propriedades ópticas aparentes (AOPs, *Apparent Optical Properties*), representam as combinações de grandezas radiométricas para indicar cor e transparência da água, são dependentes da composição do corpo d’água (IOPs) e das condições ambientais (superfície da água, vento, variação de luz, estação do ano, etc.) (Barbosa et al., 2019).

Figura 4 - Trajetória da luz: Sol, corpo d’água e sensor orbital

|  |  |
| --- | --- |
|  | Processos (p):  (1) processo de espalhamento atmosférico;  (2) processo de reflexão especular da luz difusa do céu;  (3) processo de reflexão especular da luz solar direta;  (4) radiação emergente da coluna d’água;  (5) processo de espalhamento e absorção da luz por fitoplâncton;  (6) processo de espalhamento da luz por sedimentos;  (7) processo de absorção da luz por matéria orgânica dissolvida. |

Fonte: Adaptado Barbosa *et al*., 2019.

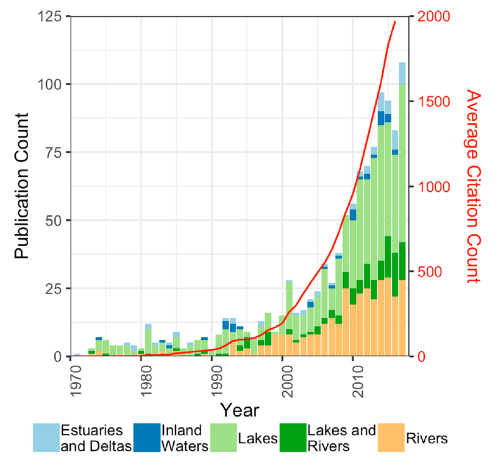
A absorção de luz no meio aquático está atrelada a quatro tipos de componentes principais: água (pura); matéria orgânica dissolvida colorida (CDOM, *collored organic dissolued matter*); pigmentos fotossintetizantes e presentes no fitoplâncton; partículas orgânicas e inorgânicas. Estes elementos são denominados componentes opticamente ativos (COAs), cada qual dispõe de propriedades ópticas específicas de absorção e de espalhamento da luz solar (IOPs). Os COAs estão presentes em proporções distintas e suas interações com a radiação solar provoca efeitos nos processos de absorção e espalhamento uns dos outros. A partir da observação dos IOPs dos COAs é possível estimar a contribuição dos mesmos na radiância/reflectância no ambiente aquático, determinando suas concentrações no corpo d’água (Meneses *et al*., 2019).

Aplicações de sensoriamento remoto em rios e reservatórios normalmente se restringem à região do espectro visível, infravermelho próximo (em função do elevado coeficiente de absorção da água comprimentos de onda maiores que 750 nm) e do infravermelho térmico (mensuração da temperatura da superfície da água). (Meneses *et al*., 2019). Para muitos autores, a identificação de parâmetros da água por meio de imagens de satélite normalmente é mais complexa do que a de outros eventos/objetos naturais pelas seguintes razões (Meneses et al., 2019; Kirk, 2011):

* A região de maior penetração de luz na água (400 nm a 500 nm) e de maior interação entre a radiação eletromagnética e os componentes da coluna d'água são bastante afetadas pelas condições atmosféricas (até 90% da energia detectada sobre um corpo d'água por um sensor orbital pode estar relacionado à interferência atmosférica);
* A transmitância da água é elevada e espectralmente variável, a energia oriunda do corpo d’água fornece dados relativos a diferentes profundidades em um mesmo recurso hídrico;
* COAs apresentam espectros de absorção semelhantes, dificultando a discriminação simultânea dos mesmos no corpo hídrico;
* A baixa reflectância da água (média 4%) compromete a capacidade de identificação de mudanças menos expressivas no corpo d’água por sensores orbitais;
* A reflectância da superfície da água é maior do que a reflectância oriunda do volume e sofre interferência de fatores externos, tais como o vento.

Além dos desafios elencados, o uso do sensoriamento remoto tem avançado de forma representativa nas últimas décadas, mostrando-se uma ferramenta indispensável para o monitoramento dos recursos hídricos, aspecto evidenciado na Figura 5. Dentre as principais vantagens proporcionadas pelo uso do sensoriamento remoto destacam-se a capacidade de mapeamento da qualidade da água em escala de tempo quase real e a abrangência de grandes extensões espaciais sem a necessidade de instrumentos locais e campanhas de campo (WAGLE *et al*., 2020).

Figura 5 – Publicações Relacionadas a Estudos Hidrológicos por meio de Sensoriamento Remoto



Fonte: Topp *et al*., 2020.

O mapeamento e monitoramento em grande escala de recursos hídricos a partir de imagens de satélite de alta resolução tem como resultado a geração de grandes volumes de dados, dificultando o processamento e análises por técnicas estatísticas tradicionais. Neste contexto, o uso de técnicas de ML aplicadas em imagens de sensoriamento remoto provê elementos capazes de estimar parâmetros de qualidade da água e suas respectivas distribuições espaciais com bons níveis de acurácia (WAGLE *et al*., 2020).

**2.3 *Machine Learning* Aplicado ao Monitoramento e Diagnóstico de Recursos Hídricos**

A integração de algoritmos de Inteligência Artificial (IA) ao processamento de dados de satélites tem aliado a computação de alto desempenho à criação de novas perspectivas de análise, quantificação e compreensão dos ecossistemas aquáticos (AHMED *et al*., 2019). Os algoritmos de IA potencializaram a capacidade de estimar os parâmetros de qualidade da água, reduzindo tempo de análise e apresentando resultados muito precisos, minimizando esforços humanos, racionalizando recursos técnicos e econômicos (WAGLE *et al*., 2020).

A IA pode ser entendida como a automatização de atividades associadas ao pensamento humano, tais como a tomada de decisões (BELLMAN, 1978). Sob esta abordagem, a IA deve ter capacidade de: (1) armazenar conhecimento; (2) aplicar o conhecimento armazenado para solucionar problemas; (3) adquirir novos conhecimentos por meio de experiências (RUSSELL *et al.*, 2013). A IA é representada por qualquer técnica computacional com a capacidade de emular o comportamento humano, ela pode ser subdividida em outras denominações, sendo as mais difundidas *Machine Learning* (aprendizado de máquina) e *Deep Learning* (aprendizado profundo).

*Machine Learning* (ML), aprendizado de máquina, é o ramo da ciência da computação que, a partir da utilização de métodos estatísticos, permite aos computadores funcionarem sem serem programados explicitamente, ao passo que são capazes de melhorar sua performance a partir das experiências. De modo geral, o modelo é treinado automaticamente utilizando grandes volumes de dados, recursos e rótulos, que são posteriormente processados para obter novos conjuntos de dados e fazer previsões (Mohri *et al*., 2012).

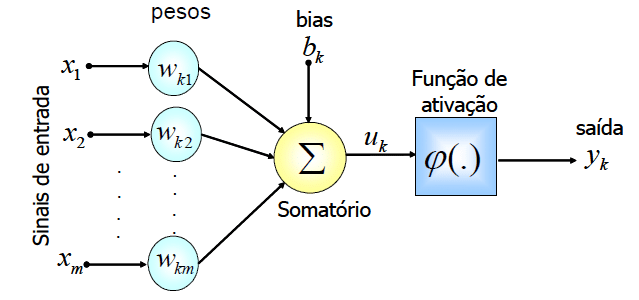
Há três formas principais de aprendizagem: **supervisionada**, onde dados utilizados para treinamento dos modelos de ML contém as soluções almejadas, aplicável em problemas de classificação e regressão (estimativa de valores); **não supervisionada**, na qual os dados utilizados para treinamento não dispõem as soluções, utilizada em *clustering* (agrupamento), segmentação e redução de dimensionalidade; **aprendizagem por reforço**, o algoritmo realiza ações a partir da observação do ambiente, recebendo punições em caso de fracasso e recompensa em caso de sucesso, aprendendo com os próprios erros, utilizada em sistemas de recompensa, jogos e robótica (GÉRON, 2017).

No âmbito do aprendizado supervisionado há dois tipos de modelos principais: **Regressão**, visa estimar o valor (numérico) de uma variável de saída de acordo com as características das variáveis de entrada conhecidas (preço de imóvel, temperatura da água, etc.); **Classificação**, especifica a classe à qual elementos de dados pertencem a partir da observação das propriedades de um conjunto de dados já categorizados (classificação de texto, reconhecimento de fala, identificação de objetos em imagens, etc.) (Pentakalos, 2019).

*Deep Learning* (DL), aprendizado profundo, é um subcampo do aprendizado de máquina, *Machine Learning* (ML), no qual um dos algoritmos mais difundidos é a Rede Neural Artificial (RNA), que tem como estrutura básica uma camada de entrada, camadas intermediárias (ocultas) e uma de saída. As RNAs são muito eficientes e suas estruturas lógicas e matemáticas normalmente são complexas, bem como os custos computacionais envolvidos nos processamentos. Contudo, isso lhes confere capacidade de aprender relações lineares e não lineares entre variáveis de entrada e saída, de modo geral, elas possibilitam a representação de recursos complexos em camadas posteriores a partir de representações mais simples dispostas em camadas anteriores na rede (Goodfellow *et al*., 2016).

As RNAs foram inspiradas no comportamento do cérebro humano, a Figura 6 apresenta uma abstração sobre seu modo básico de funcionamento. A primeira parte do neurônio artificial é responsável por receber os sinais de entrada *xi* [x1, x2,...,x j] ∈ X, que representam os dados/informações de determinado objeto/evento. Estes dados são multiplicados por pesos com valores aleatórios, *wk* [*w1,w2,...,wj*] ∈ *W*, sendo o resultado das operações somado ao viés (*bias*) *bk*, que representa um valor escalar que busca prover um melhor ajuste matemático da função aplicada. Como resultado obtém-se o valor *vk*, sendo este processado por uma função de ativação ϕ(.), que tem como objetivo restringir a amplitude de saída *yk* do neurônio *k*, frequentemente limitado ao intervalo [0,1], com exceção de problemas tipificados como regressão (COSTA, 2018). Desta forma, uma rede neural básica pode ser definida com a seguinte equação (MICHIE *et al*., 1994): yk = ϕ(bk +∑i wki x Xi).

Figura 6 – Diagrama Básico de Rede Neural Artificial



Fonte: Lima, 2005.

As RNAs são versáteis e podem ser utilizadas para solucionar uma ampla variedade de problemas, tais quais a identificação de padrões, classificação, previsões, etc. Elas também são capazes de modelar relações não lineares entre as variáveis de interesse, e, de modo adicional, possuem a capacidade de adaptar seus pesos de acordo com as modificações do meio ambiente, aumentado seu potencial de generalização. Entre os aspectos negativos, podemos destacar: muitos resultados obtidos não são explicáveis; o treinamento de modelos pode demorar de horas a dias; normalmente exige grandes volumes de dados para obtenção de resultados expressivos; requer recursos computacionais de alto desempenho para processamento dos dados (FLECK *et al*., 2016).

Além das RNAs, outros algoritmos se destacam no processo de aprendizado e generalização, especificamente no âmbito do Aprendizado Supervisionado, tópico de interesse dessa pesquisa (Quadro 2):

Quadro 2 – Exemplos de Algoritmos de Aprendizado Supervisionado

| **Algoritmo** | **Descrição** |
| --- | --- |
| **Regressão Linear** | Estabelece a relação entre variáveis independentes e dependentes em problemas de previsão, podendo ter uma (simples) ou mais variáveis independentes (múltipla). Apresenta bom nível de explicação, contudo, não é capaz de modelar relações complexas entre variáveis. |
| **Regressão Logística** | Algoritmo estatístico que tem como principal característica a capacidade de prover análises de variáveis qualitativas nominais, de modo a estimar a probabilidade de uma observação pertencer a determinada classe/grupo. De modo adicional, o método permite identificar as variáveis mais representativas, as que mais contribuem para a separabilidade entre as classes. Os coeficientes do modelo são ajustados para permitir a dependência entre os parâmetros. |
| ***Support Vector Machine* (SVM)** | Técnica de aprendizado supervisionado caracterizada por determinar um hiperplano ideal para separar duas classes de determinado objeto/evento. O SVM apresenta uma variação para modelar problemas de regressão não linear denominado *Support Vector Regression* (SVR). |
| **Árvore de Decisão** | A análise de árvore decisão é uma técnica de aprendizado supervisionado que pode ser utilizada para modelar problemas de classificação e regressão. O processo de aprendizagem envolve a construção de uma árvore de decisão a partir de dados de treinamento rotulados por classe. Árvores de decisão são estruturadas em blocos de decisão do tipo “se verdadeiro, então” e “se falso, então”. Cada decisão é representada por um nó e cada resposta à decisão é uma ramificação da árvore (verdadeiro ou falso). |
| ***Random Forest*** | Método de aprendizado supervisionado que tem como fundamento a aplicação de diversas árvores de decisão simultaneamente, sendo que estas dispõem de estruturas diferenciadas. O resultado final da previsão/classificação é obtido por um processo de “votação”/combinação de resultados entre as diversas árvores aplicadas no processo. |
| ***K-nearest Neighbours* (KNN)** | Algoritmo de classificação supervisionado que visa enquadrar uma amostra de teste não rotulada tendo como referência a maioria das amostras semelhantes entre os k-vizinhos mais próximos da respectiva amostra. As distâncias entre a amostra de teste e cada uma das amostras de dados de treinamento são determinadas por uma medida de distância específica. |
| ***Naive Bayes*** | Método que tem como referência o Teorema da Probabilidade de Bayes, no qual a probabilidade posterior de um evento (A) está atrelada à probabilidade anterior do evento (B), supõe que as variáveis são independentes. |
| ***Convolutional Neural Networks* (CNN)** | Análogas às RNAs tradicionais, a principal diferença entre CNNs e RNAs é que as CNNs são usadas principalmente no campo de reconhecimento de padrões em imagens. Isso permite codificar recursos específicos da imagem na arquitetura, tornando a rede mais adequada para tarefas focadas em imagem, ao passo em que reduz o número de parâmetros necessários para configurar o modelo e otimiza o tempo de processamento e acurácia do algoritmo. |

Fonte: Adaptado de Nogueira, 2019; Alzubi *et al*., 2018; O’Shea *et al*., 2015; Goodfellow *et al*., 2014.

No âmbito das ciências ambientais, durante muito tempo, boa parte dos problemas atrelados à análise e estudos ambientais estava relacionada à disponibilidade e qualidade de dados. O advento de novas tecnologias, maior conectividade e IOT (*Internet of Things*), propiciaram a geração de um grande volume de dados/informações (*Big Data*), transferindo parte do problema para a capacidade de análise dos dados. Cabe ainda ressaltar que dados hidrológicos e ambientais apresentam algumas características que os diferenciam dos demais: *outliers* (valores atípicos), dados faltantes, erros de análise, distribuições anormais, dentre outras (OLIVEIRA, 2018).

Neste contexto, ML é uma ferramenta poderosa, provendo análises mais precisas e prognósticos mais assertivos, dentre algumas aplicações com o foco exclusivo na análise de recursos hídricos podemos destacar: Wang *et al*. (2008) utilizou RNA em imagens de sensoriamento remoto do Landsat-7 para melhorar, avaliar e monitorar a qualidade da água de um reservatório na cidade de Taipei, Taiwan; Pu *et al*. (2019) aplicaram uma CNN para determinar os níveis de qualidade da água dos lagos Erhai e Chaohu (China) a partir de imagens do Landsat-8; Silva e Panella (2018) utilizaram imagens do satélite Sentinel-2 e RNA para mensurar florações algais (clorofila-a) no reservatório Cefni (Anglesey, Inglaterra); Barzegar *et al*. (2020) aplicaram três modelos de redes neurais (CNN, LSTM e CNN-LSTM) para estimar o nível de oxigênio dissolvido e de concentração de clorofila-a em pequenos lagos na região de Prespa, Grécia.

Sit *et al*. (2020) identificaram 129 estudos publicados entre 2018 e 2020 relacionados à aplicação de *Deep Learning* na avaliação, monitoramento e mensuração de parâmetros de recursos hídricos em diferentes escopos: enchentes, superfície da água, água subterrânea, qualidade da água, gestão, dentre outros. Topp *et al*. (2020) realizaram um levantamento referente ao uso de imagens de sensoriamento remoto em algoritmos (empíricos, semi-analíticos e *Machine Learnig*) com foco exclusivo em recursos hídricos, especificamente águas interiores (água doce). Foram identificados 416 estudos, contemplando o período entre os anos 1974 e 2018, nos quais o sensoriamento remoto propiciou diagnósticos e análises de corpos d’água em suas mais amplas variações (reservatórios, rios, lagos, estuários, etc.).

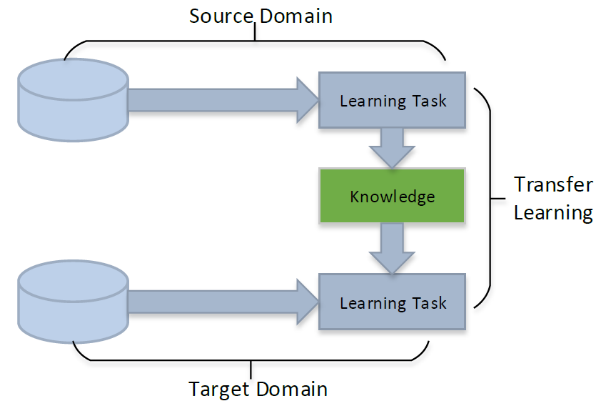
Embora a área de recursos hídricos tenha avançado no uso de técnicas de ML, Shen (2018) destaca que algoritmos de *Deep Learning* têm registrado aplicações na hidrologia com certo atraso quando comparado às demais ciências ambientais. Esta lacuna quanto às outras frentes de estudo têm acelerado a produção técnica e científica sobre o tema, um levantamento realizado por Ayezel (2021) no site *Web of Science* identificou que o número de artigos publicados em 2019 com foco de abordagem *Deep Learning* e recursos hídricos aumentou seis vezes em comparação a 2016, tendência que deve ser mantida nos próximos anos.

Os tópicos apresentados anteriormente ratificam as potencialidades do uso integrado de ML e do sensoriamento remoto no escopo de recursos hídricos, especialmente na delimitação de estimativa de variáveis da qualidade da água de reservatórios, objeto de interesse desta pesquisa.

**2.4 Transferência de Conhecimento**

*Transfer Learning*, Transferência de Conhecimento (TC), é uma técnica desenvolvida para mitigar três problemas recorrentes na aplicação de modelos de ML: dados insuficientes, limitação na capacidade de processamento computacional e incompatibilidade na distribuição de frequências. Técnicas tradicionais de ML buscam aprender as tarefas a partir de um estágio de conhecimento nulo, onde o aprendizado é acrescentado gradativamente ao longo do processo, enquanto TC permite a transferência do conhecimento adquirido em atividades anteriores para uma nova tarefa (alvo). O fluxo básico do processo de transferência de conhecimento está disposto na Figura 8 (TAN et al. 2018).

Figura 8 – *Transfer Learning*, Fluxo Básico



Fonte: TAN *et al*., 2018.

Três itens básicos fundamentam a aplicação de TC: **domínio**, ramo do conhecimento do objeto de pesquisa (medicina, segurança, recursos hídricos, etc.); **tarefa**, ação executada no processo de obtenção do conhecimento pelo modelo (classificação de imagens, identificação de padrões, etc.); **modalidade**, natureza estrutural do dado utilizado (imagem, áudio, texto, etc.) (NIU *et al*., 2020).

No contexto de aplicação de TC, destaca-se uma técnica denominada *Domain Adaptation* (DA), Adaptação de Domínio, que consiste basicamente num processo de adaptar um ou mais domínios de origem por meio de transferência de conhecimento, a fim de melhorar a performance/acurácia na avaliação/estimativa de uma variável alvo. No processo de DA tem-se um cenário onde os domínios de origem e de destino apresentam distribuições diferentes, neste sentido, a técnica é caracterizada pela tentativa de alteração de um domínio de origem, visando reduzir as variações entre a distribuição dos dados de determinada fonte e um conjunto de dados alvo (Mehrparvar *et al*., 2018).

A TC é subdividida em cinco categorias, delimitadas de acordo com a natureza do objeto de estudo e as características dos dados (in)disponíveis: indutiva, transdutiva, não supervisionada, negativa e modalidade cruzada. Cada categoria de transferência de aprendizado é organizada em quatro tipos em quatro tipos de aprendizagem: por instâncias, recursos, parâmetros ou aprendizado por relações (NIU *et al*., 2020).

Ainda que as pesquisas relacionadas à transferência de aprendizagem entre algoritmos de estimativa de parâmetros de qualidade da água estejam em estágio incipiente, alguns estudos identificados confirmam a demanda e relevância desse campo de pesquisa. O trabalho desenvolvido por Chen *et al*. (2021) apresentou um novo algoritmo, denominado TrAdaBoost-LSTM, que contempla a transferência de aprendizagem entre os modelos LSTM e TrAdaBoost a fim de estimar as concentrações de oxigênio dissolvido na água, a precisão do modelo foi de 15% a 25% superior aos modelos tradicionais. Zhu *et al*. (2021) estimaram a concentração de oxigênio dissolvido no lago Yixing (China), os autores implementaram um modelo híbrido de ML que incorporou os algoritmos ResNets e BiLSTM, sendo o modelo aplicado para análise do lago Taihu (China), o aprendizado assimilado neste lago foi transferido para então estimar o oxigênio dissolvido no lago Yixing (pequena quantidade de dados) com bom nível de assertividade.

Zhou *et al*. (2021) utilizaram o resultado do processamento de dados de monitoramento *in situ* em um modelo de regressão como forma de transferência de aprendizado para o algoritmo *Echo State Networks* (ESN), em seguida o modelo foi aplicado para estimar a qualidade da água de uma região em Hong Kong, apresentando boas métricas de validação. Willard *et al*. (2021) aplicaram uma técnica de aprendizagem de meta-transferência na qual o conhecimento assimilado por modelos de ML aplicados em 145 lagos bem monitorados foi utilizado/transferido para estimar a temperatura de 305 lagos-alvo (não monitorados) na região Centro-Oeste dos Estados Unidos, o erro médio para os lagos alvo oscilou entre 2,43 °C e 2,52 °C, resultado promissor para estimar parâmetros de sistemas não monitorados e variáveis ​​ambientais.

**3. OBJETIVOS**

**3.1 Objetivo Geral**

* Investigar a aplicação de técnicas de aprendizado de máquina baseadas em transferência de conhecimento para os estimar parâmetros de qualidade de água de reservatórios.

**3.2 Objetivos Específicos**

* Analisar estatisticamente dados de variáveis de qualidade da água de reservatórios distribuídos em bacias geográficas e biomas distintos, buscando identificar padrões, fatores de divergência e evidenciar as variáveis com maior potencial para serem utilizadas em modelos de inteligência artificial e transferência de aprendizado.
* Identificar algoritmos de *Machine Learning* que apresentem bons níveis de acurácia na estimativa de variáveis de qualidade da água por meio da aplicação de diferentes técnicas e comparação de seus respectivos resultados.
* Aplicar técnicas de transferência de aprendizado, por meio das quais modelos de estimativa de variáveis de qualidade da água implementados em diferentes contextos geoambientais serão utilizados para estimar variáveis de qualidade da água do Reservatório de Três Marias.
* Identificar potencialidades e limitações do uso de técnicas de transferência de aprendizado na estimativa de variáveis de qualidade da água de reservatórios.
* Implementar a estrutura lógica e funcional de um sistema computacional, disponibilizado em *nuvem*, especializado na estimativa de variáveis de qualidade da água de reservatórios a partir do processamento de imagens de satélites.

**4. HIPÓTESES**

Ao longo das fases de elaboração, fundamentação técnica-científica, implementação, testes e validação deste projeto, pretende-se verificar e elucidar as seguintes hipóteses:

* O conhecimento obtido por meio de técnicas de *Machine Learning* aplicadas à estimativa de parâmetros físico-químicos de qualidade da água de determinado reservatório pode ser transferido para mensurar a qualidade da água de um reservatório em um contexto diferenciado, independentemente das variações geoambientais características entre os reservatórios.
* As técnicas de aprendizado de máquina aplicadas na estimativa de valores de diferentes variáveis de qualidade da água de determinado reservatório apresentam níveis de acurácia distintos, sendo dependentes das características físicas e biológicas locais, níveis de sensibilidade óptica do satélite e dispersão dos dados das variáveis de interesse.

**5. MATERIAL E MÉTODOS**

O desenvolvimento desta tese está alinhado a um projeto de pesquisa e desenvolvimento denominado “GT0607 – Monitoramento Inteligente da Qualidade da Água em Reservatórios Hidrelétricos Através do Desenvolvimento de um Algoritmo Fotogramétrico”, sob gestão da CEMIG e execução em forma de convênio com a UFMG. O fluxo básico de ações a serem contempladas pelos métodos propostos nesta pesquisa é apresentado na Figura 9.

Figura 9 – Ações do Fluxo Metodológico Básico



Fonte: Autor, 2022.

**5.1 Área de Estudo: Reservatório de Três Marias, MG**

Inaugurado em 1962, o Reservatório de Três Marias foi primeiro empreendimento hidráulico de grande porte de usos múltiplos da água implementado no Brasil, tendo como uso principal a geração de energia elétrica. O reservatório está localizado na Bacia Hidrográfica do Rio São Francisco, Minas Gerais, região caracterizada por apresentar duas estações bem definidas, inverno seco e verão úmido. A Figura 10 apresenta a disposição geográfica do reservatório e as características básicas do mesmo.

Figura 10 – Reservatório de Três Marias e Distribuição dos Pontos de Amostragem



Fonte: Adaptado de Pizani *et al*., 2020; CEMIG, 2021.

**5.2 Seleção de Parâmetros de Qualidade da Água e Definição de Escopo**

A seleção dos parâmetros de qualidade da água que serão estimados pelos modelos de ML empregados nesta pesquisa teve como critério de escolha a capacidade de estimativa dos mesmos por imagens obtidas por meio dos satélites Sentinel-2 e Landsat-8, disponibilidade dos parâmetros de interesse nas bases de dados avaliadas e nas mensurações registradas em campanhas *in situ*. Inicialmente foram selecionados para análise dez parâmetros, quatro opticamente ativos (clorofila-a, temperatura da superfície da água, profundidade do disco de *Sechhi* e turbidez) e seis não opticamente ativos (oxigênio dissolvido, pH, carbono total, carbono orgânico, condutividade e potencial de oxi-redução).

As variáveis de qualidade da água serão abordadas sob dois eixos: **regressão**, previsão de um valor numérico de determinada variável; **classificação**, enquadramento de determinada variável em classes (bom, ruim, médio, etc.). Em ambas as abordagens serão utilizadas variáveis explicativas (independentes) para prever o comportamento/valor de uma determinada variável (dependente). No caso específico desta pesquisa, serão utilizadas 12 variáveis de entrada (referentes aos valores dos pixels contidos nas bandas espectrais do satélite Sentinel-2) para estimar os valores das variáveis dependentes (turbidez, clorofila-a, etc.).

**5.3 Banco de Dados de Variáveis de Qualidade da Água e Sensoriamento Remoto**

Serão mobilizados diversos conjuntos de dados/imagens para o desenvolvimento desta pesquisa, subsídios essenciais na concepção dos algoritmos que serão implementados. Basicamente, dados estruturados (Campo, GemStat, PMQS, etc.) compõem as variáveis dependentes (que serão previstas/estimadas) e imagens (valor de pixels extraídos de imagens de satélites e drones) as variáveis independentes (utilizadas para prever as variáveis dependentes).

*5.2.1 Campanhas in situ (Campo)*

As amostragens de parâmetros de qualidade da água serão obtidas por meio de campanhas *in situ* realizadas entre 2019 e 2022, em períodos de cheia e seca. Foram selecionados 78 pontos de amostragem, distribuídos ao longo do reservatório. Dentre os dez parâmetros selecionados, seis foram mensurados por sensores especializados: oxigênio dissolvido, pH, temperatura, profundidade do disco de Secchi, turbidez, cor, condutividade e potencial de oxidação-redução. As análises dos quatro parâmetros restantes foram realizadas em laboratório: clorofila-a, carbono total, carbono orgânico, nitrogênio, fósforo, cloreto, sulfato, ferro total e sólidos totais em suspensão. As datas de coleta das amostragens *in situ* estão em consonância com as datas de operação dos demais mecanismos de obtenção de dados: satélites e drones.

*5.2.2 GemStat*

O *GemSta*t (*Global Freshwater Quality Database*) é um banco de dados abertos e gratuitos que disponibiliza informações cientificamente confiáveis sobre o estado e a tendência de parâmetros de qualidade da água de água doce de reservatórios, lagos, rios, áreas alagadas e águas subterrâneas em nível global. O portal disponibiliza dados de corpos hídricos de dezenas de países, obtidos de 13.009 estações de monitoramento de qualidade da água, sendo mapeados centenas de parâmetros de qualidade da água, distribuídos em dez classes (ópticos, orgânicos, dentre outros). *GemStat* é mantido pelo *Water Data Center* (GWDC), vinculado ao Centro Internacional de Recursos Hídricos e Mudanças Globais, sediado em Koblenz, Alemanha.

*5.2.3 PMQS*

O Programa Especial de Monitoramento da Qualidade das Águas e dos Sedimentos do Reservatório de Três Marias e Entorno (PMQS Três Marias) tem como objetivo avaliar e acompanhar a evolução espaço-temporal dos parâmetros determinantes da qualidade da água e dos sedimentos nos cursos de água da região por meio de coletas regulares em dez pontos amostrais fixos.

*5.2.4 Bases de Dados Complementares*

A partir da disponibilidade/necessidade de maior volume de dados, outras fontes de dados estruturados/não-estruturados e/ou geoespaciais poderão serão utilizados nos modelos de ML a partir de plataformas representativas e confiáveis: ANA, IGAM, entidades públicas estaduais relacionadas à gestão de recursos hídricos (CETESB, etc.), dentre outras fontes.

*5.2.5 Satélites: Sentinel-2 e Landsat-8*

Serão utilizadas imagens gratuitas de nível 2A oriundas dos satélites Sentinel-2 (Agência Espacial Europeia - ESA) e Landsat-8 (Serviço Geológico dos Estados Unidos - USGS). As imagens deverão ser processadas a fim de prover correções radiométricas e geométricas, buscando disponibilizar produtos de reflectância com menores níveis de interferência. Todas as bandas espectrais do satélite Sentinel-2 serão reamostradas (quando necessário) para a resolução espacial de 60 m, as bandas espectrais do satélite Landsat-8 terão a resolução espacial mantida, 30 m.

Para extração dos valores contidos nos pixels das bandas espectrais das imagens de satélite será desenvolvido um algoritmo em linguagem *Python*, por meio do qual os dados serão obtidos diretamente da plataforma *Google Earth Engine* e consistidos em formato de planilhas (sem necessidade de *download* das imagens).

**5.3 Técnicas de *Machine Learning***

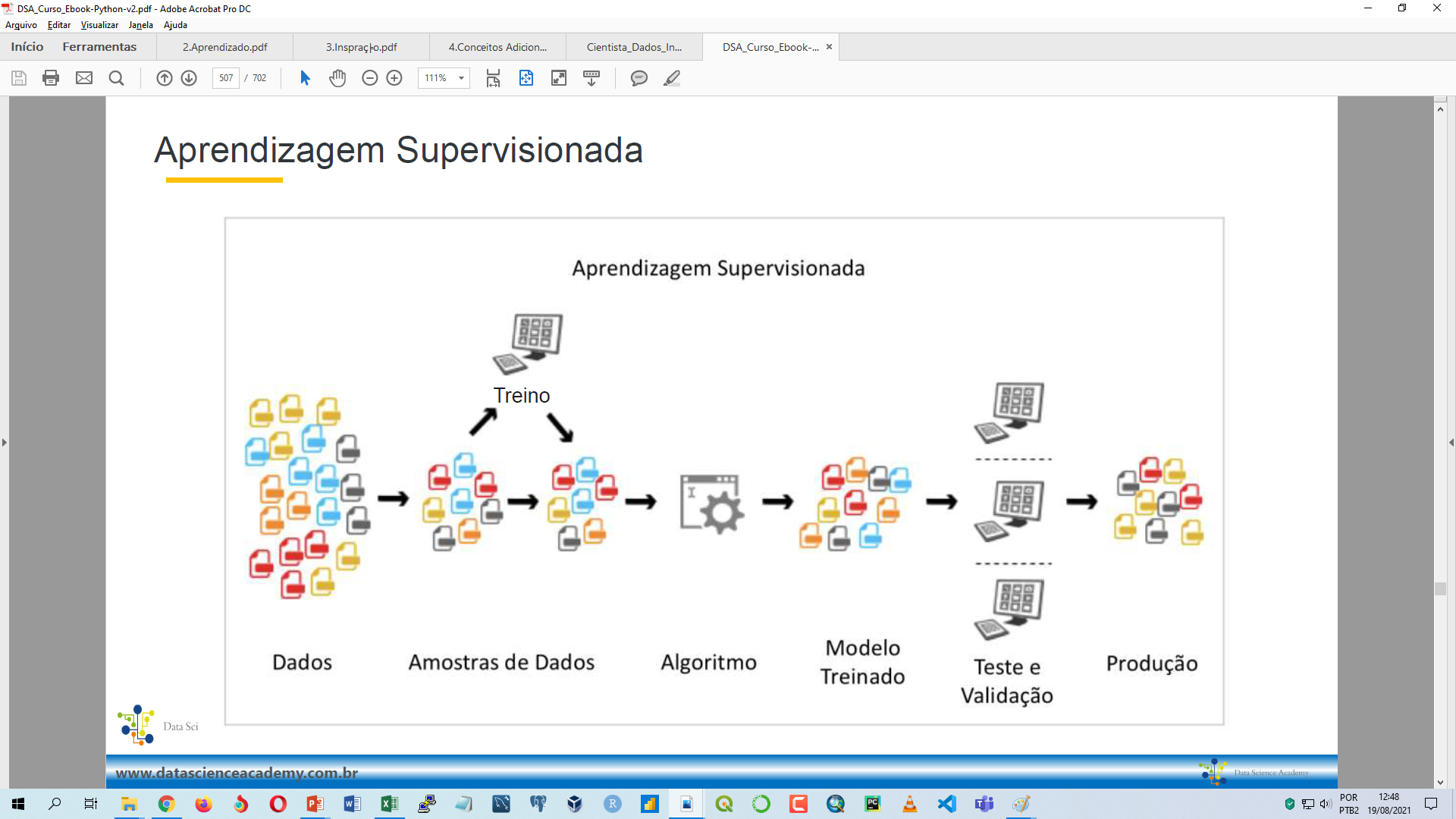
Nesta etapa da pesquisa serão aplicadas 27 técnicas genéricas de IA (20 delas por meio de Aprendizado Automatizado). O Quadro 3 apresenta uma descrição sintética dos métodos e funções aplicadas.

Quadro 3 – Métodos de IA Aplicados no Estudo

| **Método/Função Aplicado(a)** | **Descrição** |
| --- | --- |
| Ordinary least squares **(OLS)** | A regressão de mínimos quadrados ordinários é um método estatístico de análise que estima a relação entre uma ou mais variáveis independentes e uma variável dependente; o método estima a relação minimizando a soma dos quadrados na diferença entre os valores observados e previstos da variável dependente configurada em linha reta. |
| Regressão Linear  **- Linear Regression** | Modelo de regressão linear múltipla baseado na OLS, disponível no pacote scikit-learn do *Python*, diferenciando-se do mesmo ao incluir de forma explícita variáveis de treino e teste. |
| Generalized Linear Model **(GLM)** | Generalização da regressão linear ordinária que permite processar variáveis de resposta que têm formas de distribuição de erro diferentes de uma distribuição normal. |
| Árvore de Decisão baseada em Regressão**:**  **- Decision Tree Regressor** | Modelo de Árvore de Decisão especializado na estimativa de valores contínuos de determinada variável, a partir da combinação dos parâmetros nas diversas ramificações de uma estrutura em “árvore” permite prever dados com saída contínua significativa (números reais). Disponível no pacote scikit-learn do Python. |
| Redes Neurais Artificiais  **Arquiteturas aplicadas:**  **- 4 camadas, 6 neurônios**  **- 2 camadas. 64 neurônios**  **- 3 camadas. 7 neurônios** | Modelos computacionais capazes de realizar o aprendizado de máquina bem como o reconhecimento de padrões, a partir de redes e células de processamento interconectadas, apresentado ampla gama de aplicações na resolução de problemas complexos de classificação e regressão. |
| Aprendizado Automatizado  **- AutoKeras (10 modelos)**  **- H2O (10 modelos)** | AutoKeras e H2O são pacotes computacionais que possibilitam a implementação simultânea de diversos algoritmos de IA onde o “sistema” busca selecionar o modelo aplicado que obtiver a menor métrica de erro. |

Os algoritmos serão implementados em linguagem de programação *Python*, o fluxo básico de aplicação utilizado nos mesmos encontra-se sintetizado na Figura 11. Para validação dos modelos será utilizado o método *hold-out* (Devroye e Wagner, 1979), no qual uma amostra é dividida em duas: uma destinada à implementação dos modelos (treino) e outra à validação (testes). Nesta etapa foram utilizados 70% dos dados disponíveis para treino e 30% para testes.

Figura 11 – Fluxo Básico de Aplicação dos Modelos de ML Aplicados no Estudo



Fonte: Medium, 2021.

**5.4 *Transfer Learning***

*Transfer Learning*, Transferência de Conhecimento (TC), consiste basicamente na aplicação de técnicas de ML que propiciam que o conhecimento obtido por determinado modelo seja disponibilizado para uso em outros modelos, potencializando a capacidade de generalização dos algoritmos de ML, especificamente nas tarefas de estimativa de parâmetros de qualidade da água de reservatórios. A transferência de aprendizado do tipo Transdutiva será o método inicialmente empregado nesta pesquisa, uma vez que as tarefas nos domínios fonte e alvo são as mesmas e pode haver diferenças entre os domínios. A técnica terá como referência o aprendizado baseado em características (*Learning on Features*), aplicação mais popular na literatura devido ao fato de seu uso requerer apenas que as distribuições de dados nos domínios fonte e alvo sejam semelhantes (NIU *et al*., 2020).

**5.5 Métricas de Avaliação, Testes e Calibração dos Modelos de Machine Learning**

As técnicas de ML serão aplicadas de forma efetiva na construção de modelos de estimativa de parâmetros de qualidade da água de reservatórios, a análise da aplicabilidade dos modelos terá como referência o uso de métricas de avaliação com foco em minimizar a função de determinação dos erros, adotando-se Coeficiente de Determinação (R²), Erro Médio Absoluto (MAE, *Mean Absolute Error*), Raiz do Erro Médio Quadrático (RMSE, *Root Mean Squared Error*) e Erro Médio Quadrático (MSE, *Mean Squared Error*) para algoritmos de regressão numérica; as métricas acurácia, *F1-Score*, precisão e *recall* serão utilizadas em casos de algoritmos de classificação; avaliação do nível de entropia será aplicada para algoritmos de agrupamento, caso necessário.

Os conjuntos de dados que serão utilizados para treinar e validar os modelos de IA serão preparados por uma técnica de validação cruzada denominada *K-fold Cross-validadtion*, que otimiza o tempo de treinamento, teste e o nível de acurácia dos modelos. Caso os resultados não sejam satisfatórios, pode ser necessário acrescentar maior volume de dados e/ou adotar um modelo que se ajuste melhor às características dos dados.

De acordo com o nível de acurácia obtidos pelos modelos de ML, deverão ser efetuados ajustes pontuais nos parâmetros funcionais dos algoritmos, denominados hiperparâmetros, sendo que estes podem ser identificados e otimizados automaticamente por bibliotecas inerentes à linguagem de programação que será utilizada (*Python*). Caso a calibração dos modelos não apresente resultados satisfatórios, será necessário optar por outro modelo/algoritmo de ML ou implementar modelos híbridos.

**6. RESULTADOS PRELIMINARES E DISCUSSÃO**

**6.1 Delimitação do Escopo de Aplicação dos Algoritmos de *Machine Learning***

A etapa inicial desta pesquisa contempla a aplicação de algoritmos de regressão (previsão de valores) das variáveis turbidez e clorofila-a, selecionadas em virtude da disponibilidade de dados *in situ*, bancos de dados abertos e imagens de satélites.

Para extração dos dados (valores de reflectância dos pixels) das imagens do satélite Sentinel-2A foi considerada a diferença máxima de dois dias em relação à data de obtenção de dados nos conjuntos Campo, GemStat e PMQs. Quanto ao uso das bandas espectrais das imagens orbitais, para a variável turbidez foram realizados testes utilizando todas as bandas e o conjunto de bandas 1, 3, 4, 6, 7, 8, 8A, 11. Para a variável clorofila aplicou-se testes com todas as 12 bandas disponíveis e com o conjunto de bandas 6, 7, 9, 11. Os valores limites considerados para as variáveis turbidez e clorofila foram 100 NTU e 500 µg, respectivamente.

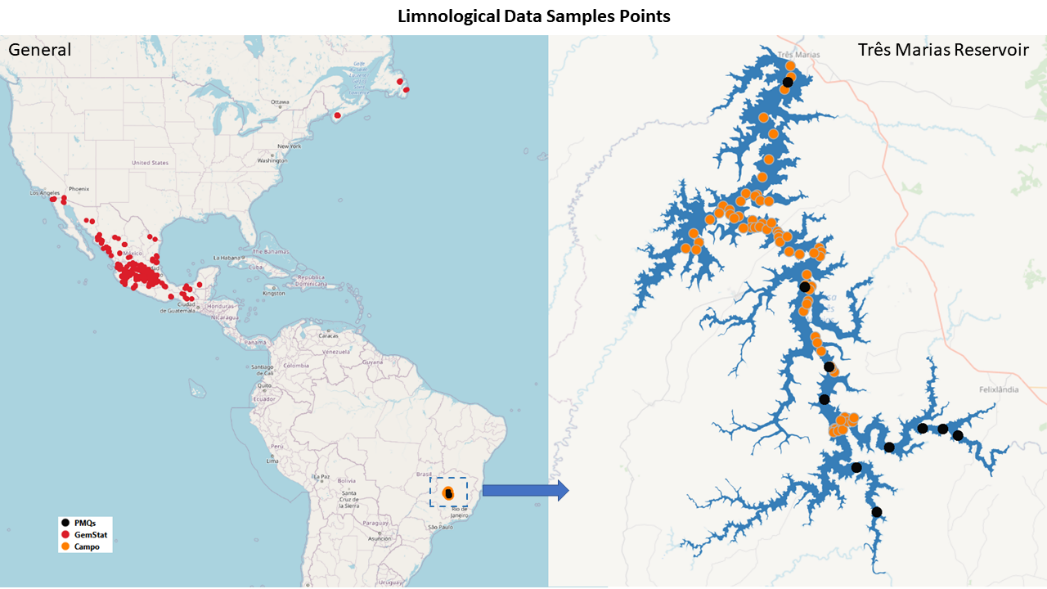
A aplicação dos filtros citados anteriormente tomou como referência o trabalho de Maillard *et al*. (2021), no qual os resultados com maior nível de acurácia foram obtidos após a aplicação dos respectivos filtros e seleção das bandas específicas. O Quadro 4 apresenta a quantidade de amostras disponíveis para os conjuntos de dados de interesse.

Quadro 4 – Variáveis de Interesse e Fontes de Dados

| **Origem/Fonte** | **Turbidez**  **Número de Amostras** | **Clorofila**  **Número de Amostras** |
| --- | --- | --- |
| Campo (in situ) | 78 | 78 |
| GemStat | 718 | 636 |
| PMQs | 1.609 | 1.609 |

A disposição geográfica dos pontos de amostragens dos conjuntos de dados utilizados é apresentada na Figura 12.

Figura 12 – Distribuição dos Pontos de Amostragem dos Conjuntos de Dados Campo, GemStat e PMQs



Nesta etapa inicial de pesquisa não será contemplada a aplicação direta de algoritmos de transferência de aprendizado, deste modo, optou-se por aplicar algoritmos genéricos para estimar os valores das variáveis turbidez e clorofila em dois blocos (por cada variável de interesse, turbidez e clorofila-a):

* **Mesmo Grupo**: modelos de regressão aplicados para estimar as variáveis turbidez e clorofila considerando o uso de dados referente apenas ao domínio/origem ao qual a variável pertence. Neste escopo os dados de treino e teste têm a mesma origem: PMQs (treino) >> PMQS (teste); *GemStat* (treino) >> *GemStat* (teste); Campo (treino) >> Campo (teste).
* **Entre Diferentes Grupos**: modelos de regressão aplicados para estimar as variáveis turbidez e clorofila considerando o uso de dados de diferentes domínios/origens aos que a variável pertence. Neste escopo os dados de treino e teste têm origens distintas: PMQs (treino) >> Campo (teste); GemStat (treino) >> Campo (teste); PMQs (treino) >> *GemStat* (teste).

**6.2 Aplicação dos Algoritmos de *Machine Learning***

A aplicação das técnicas de ML e funções matemáticas para estimar os valores das variáveis turbidez e clorofila foram obtidos por meio da codificação de algoritmos em linguagem de programação *Python*, responsável por selecionar os dados, bandas espectrais de interesse e executar os modelos computacionais implementados. Em relação às técnicas (20) de Aprendizado Automatizado, serão apresentadas apenas as duas que obtiveram melhores resultados (uma relacionada à biblioteca *AutoKeras* e outra à *H2O*). Ao todo foram considerados 72 cenários de simulação, distribuídos entre as duas variáveis de interesse (turbidez e clorofila-a), variando também quanto ao número de registros e bandas espectrais utilizadas, conforme está disposto nos tópicos 6.3 e 6.4.

Algumas variações de testes em relação ao número de dias (diferença entre a data de coleta do dado em campo e a data da imagem de satélite) foram omitidos dos resultados, buscando exibir apenas os resultados mais representativos. Os resultados mais expressivos obtidos para as variáveis turbidez e clorofila são apresentados nos Quadros 5 e 6, respectivamente. Os registros completos dos valores obtidos em todos os experimentos realizados estão dispostos nos Apêndices A (turbidez) e B (clorofila). Na apresentação dos resultados, dados sinalizados com sombreamento em tonalidade cinza e branco referem-se às análises realizadas dentro dos mesmos conjuntos de dados, dados sombreados em verde referem-se às análises realizadas entre diferentes grupos de dados.

**6.2.1 Turbidez: Regressão**

O Quadro 5 apresenta os resultados mais expressivos obtidos no processo de estimativa de valores da variável turbidez pelos modelos de ML aplicados nesta etapa. Para o conjunto de dados GemStat, obteve-se valor do coeficiente de determinação (R²) igual a 0,86 e Erro Médio Absoluto (MAE, *Mean Squared Error*) de 6,34 NTU; dados de Campo registrou valor de R² igual a 0,62 e MAE de 0,17 NTU; o conjunto de dados PMQs apresentou valor de R² igual a 0,89 e MAE de 2,37 NTU. Os modelos baseados em redes neurais foram os responsáveis por obter os resultados mais promissores na estimativa da turbidez, com destaque ao conjunto de dados Campo, que embora tenha apresentado valor de R² inferior aos demais, registrou pequena margem de erro na predição dos valores da variável.

Quanto aos experimentos de verificação da aplicabilidade de modelos treinados em determinado conjunto de dados para estimar variáveis em um contexto distinto, modelos treinados no conjunto de dados PMQs e validados no conjunto de dados de Campo tiveram como melhor resultado o valor de R² igual a 0,86 e MAE de 2,15 NTU; modelos treinados no conjunto de dados GemStat e validados no conjunto Campo registraram valor de R² igual a 0,73 e MAE de 5,54 NTU; modelos treinados no conjunto de dados PMQs e validados no conjunto GemStat obtiveram valor de R² igual 0,85 e MAE 2,01 NTU. Neste contexto, os modelos híbridos registraram os valores mais representativos e sinalizaram o potencial uso de transferência de conhecimento para estimar valores de turbidez de reservatórios em diferentes contextos.

Quadro 5 – Turbidez: Resultados de Estimativa



Fonte: Autor, 2022.

Durante os experimentos, os métodos atrelados às funções lineares e árvores de decisão apresentaram resultados menos qualificados ao longo dos testes efetuados. Ressalta-se que quantidade e qualidade dos dados, empregadas de formas variadas neste experimento, sinalizaram que os modelos de ML aplicados têm grande potencial de melhoria, desde que disponham de dados em quantidade e qualidade adequados.

**6.2.2 Clorofila-a: Regressão**

Os resultados mais expressivos obtidos no processo de estimativa de valores da variável clorofila pelos modelos de ML são apresentados no Quadro 6. Para o conjunto de dados GemStat, obteve-se valor do R² igual a 0,47 e MAE de 3,86 µg; dados de Campo registrou valor de R² igual a 0,91 e MAE de 0,29 µg; o conjunto de dados PMQs apresentou valor de R² igual a 0,32 e MAE de 0,872 µg. Os modelos de redes neurais foram os responsáveis por obter os resultados mais promissores na estimativa da clorofila, seguido por Árvore de Decisão (Regressão), com destaque ao conjunto de dados de Campo, que obteve valor de R² elevado e pequena margem de erro na predição dos valores da variável. Embora a quantidade de registros de Campo seja pequena (78), eles detêm padrão de distribuição próximo à normalidade, permitindo aos algoritmos evoluir de forma mais consistente no processo de aprendizagem.

No escopo de experimentos de modelos treinados em determinado conjunto de dados para estimar variáveis em um contexto distinto, modelos treinados no conjunto de dados PMQs e validados no conjunto de dados Campo tiveram como melhor resultado o valor de R² igual a 0,22 e MAE de 1,13 µg; modelos treinados no conjunto de dados GemStat e validados no conjunto Campo não obtiveram ajuste; modelos treinados no conjunto de dados PMQs e validados no conjunto GemStat obtiveram valor de R² igual 0,23 e MAE 0,86 µg. Neste contexto, os modelos híbridos registraram os valores mais representativos. Embora os valores de R² tenham se mostrado pouco expressivos, baixos valores de MAE nos modelos aplicados sobre o conjuntos de dados PMQs validados nos conjuntos de dados Campo e GemStat sinalizaram o potencial uso de transferência de conhecimento para estimar valores de clorofila de reservatórios em diferentes contextos, para isso deverão ser melhorados processos de preparação das variáveis de entrada.

Quadro 6 – Turbidez: Resultados de Estimativa

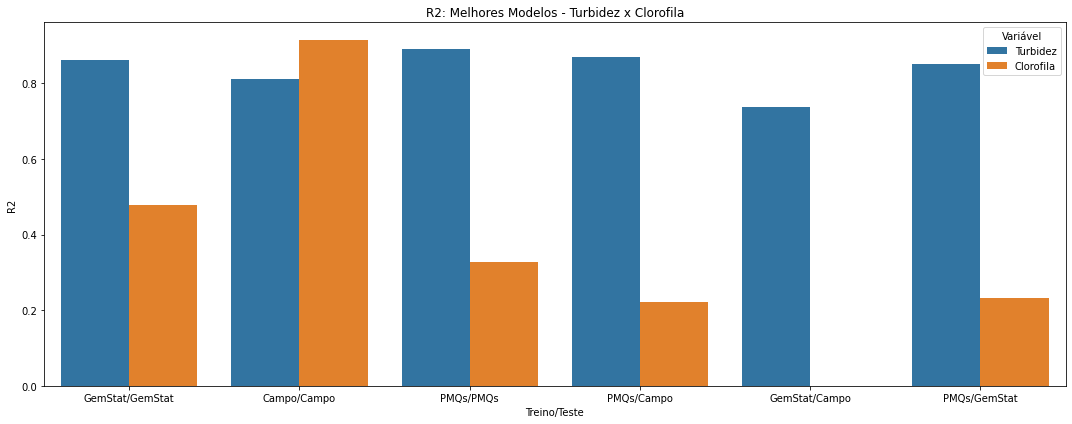


De modo geral, com exceção ao conjunto de dados de Campo, a maioria dos modelos de ML aplicados para estimar clorofila-a não obtiveram boas métricas de acurácia/erros, especialmente modelos baseados em funções lineares. Nestes casos, o uso de maior número de registros como variáveis de entrada não resultou em melhorias de performance dos modelos, sendo que em muitos casos foi identificada a situação inversa. Este fator pode estar atrelado às dispersões e comportamento de variáveis, que não contemplam uma distribuição normalizada.

**6.5 Síntese: Turbidez x Clorofila**

A Figura 13 apresenta a síntese dos resultados obtidos pelos melhores modelos aplicados para estimar as variáveis turbidez e clorofila, especificamente a métrica R². Os resultados demonstram que a aplicação de técnicas de ML pode apresentar acurácia distinta para diferentes conjuntos de dados e variáveis distintas, sinalizando correlação com o nível de dispersão dos dados e a complexidade das interações das variáveis com fatores ambientais. Modelos aplicados para estimar a variável turbidez registraram melhor nível de ajuste (R²) quando comparados à variável clorofila em todos os conjuntos avaliados, exceto para o conjunto de dados de Campo. As menores distorções foram observadas no conjunto de dados Campo, as maiores, nos conjuntos GemStat e PMQs, respectivamente.

Figura 13– R2: Melhores Modelos – Turbidez x Clorofila



A variação dos ajustes e erros registrada pelos modelos é justificada, em parte, pela diferente distribuição dos dados nos domínios explorados (Figuras 14 e 15), onde a maior amplitude/dispersão cooptou com erros mais representativos.

Figura 14 – Distribuição de Dados de Turbidez

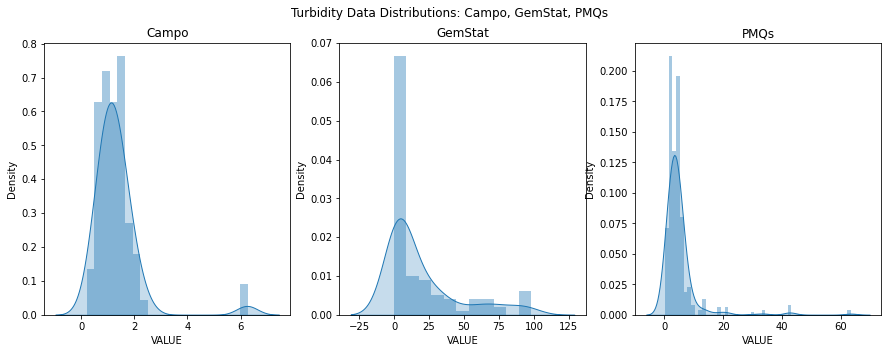
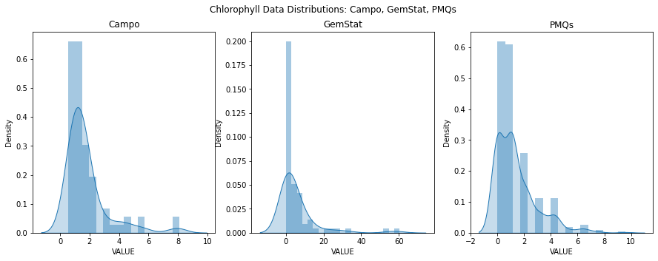


Figura 15 – Distribuição de Dados de Clorofila



Entre os produtos da aplicação dos algoritmos de ML apresentam-se os mapas de turbidez e clorofila, Figuras 16 e 17, respectivamente, obtidos por meio do processamento de uma imagem de satélite Sentinel-2A do dia 20/07/2021 via algoritmo de regressão linear múltipla (uma das técnicas aplicadas nesta etapa da pesquisa).

|  |  |
| --- | --- |
| Figura 16 – Turbidez: Mapa | Figura 17 – Clorofila: Mapa |
|  |  |

**7. CONSIDERAÇÕES PARCIAIS**

Os experimentos aplicados de forma expedita nesta etapa da pesquisa permitiram identificar de forma sucinta os grupos de algoritmos com maior potencial e aplicabilidade na abordagem do problema, estimativa de parâmetros de qualidade da água de reservatórios, caso em que as Redes Neurais Artificiais e modelos híbridos apresentaram os melhores resultados. Em contrapartida, modelos lineares e estruturados em árvores de decisão tiveram desempenhos menos encorajadores sob as condições de testes vigentes. Os resultados mais expressivos computados por alguns modelos de Redes Neurais foram obtidos quando os mesmos tiveram como entrada dados de melhor qualidade (Campo), salientando que as técnicas de ML têm um bom potencial para estimar parâmetros de qualidade da água em experimentos futuros.

Os resultados preliminares obtidos registraram para a variável turbidez, no melhor cenário, coeficiente de determinação (R²) igual a 0,89 e Erro Médio Absoluto (MAE) igual a 2,37 NTU; clorofila-a, no melhor cenário, obteve valor de R² igual 0,91 e MAE igual a 0,29 µg. No contexto dos experimentos realizados, os algoritmos aplicados para estimar a variável turbidez apresentaram resultados mais promissores, fundamentados em maior potencial de adequação/generalização dos modelos aos dados disponíveis. Em contrapartida, a aplicação dos modelos para estimar a variável clorofila pode ter sido prejudicada em alguns experimentos em virtude da dispersão de alguns dados disponíveis. Contudo, os resultados sinalizaram o potencial do uso de técnicas de ML e transferência de aprendizado para estimar variáveis de qualidade da água de reservatórios em contextos distintos.

Dentre os fatores limitantes identificados, evidencia-se a pequena quantidade de dados disponíveis frente ao número de parâmetros utilizados para estimar as variáveis de qualidade da água, este aspecto tende ter contribuído para que diversos modelos de ML aplicados não terem sido bem sucedidos nesta etapa, uma vez que o uso de técnicas de aprendizado de máquina normalmente requer grandes volumes de dados, ao passo que sob as condições atuais muito modelos não foram capazes de identificar padrões representativos nos dados, e, consequentemente, não assimilaram aprendizado de modo satisfatório.

De modo geral, observou-se um potencial de melhoria de performance e acurácia nos modelos aplicados neste estudo atrelado à disponibilização de maiores volumes de dados, neste contexto, a variação dos métodos de seleção dos registros e variáveis independentes para treinamento e validação dos modelos são elementos a serem explorados/verificados, nos quais as técnicas de validação cruzada (*K-fold Cross-validadtion*) e identificação de componentes mais representativos (*PCA – Principal Components Analysis*) tendem a otimizar os resultados alcançados pelos algoritmos. A partir das experiências iniciais poderão ser realizados ajustes e testes de novos algoritmos, bem como a implementação de modelos híbridos.

Outro fator que tem comprometido a acurácia dos modelos está relacionado ao processo de extração dos valores dos pixels das imagens de satélite, distorções foram observadas em diversos registros, tendo como origem a presença de nuvens, proximidade da margem do reservatório, baixa profundidade do corpo d'água e coordenadas dispostas em solo/estradas, tratamento a ser otimizado futuramente.

Na próxima etapa desta pesquisa serão abordadas outras variáveis de interesse (profundidade de *Secchi*, sólidos suspensos, dentre outras) e explorada a transformação do problema de regressão em classificação, na qual ao invés de estimar os valores numéricos das variáveis de interesse, estas seriam rotuladas em classes de acordo com os níveis de qualidade dos parâmetros identificados. Neste contexto, portarias e índices de qualidade da água, tais quais RESOLUÇÃO CONAMA N° 357/2005, RESOLUÇÃO CONAMA N° 430/2011, Índice de Qualidade Das Águas – IQA, dentre outros, seriam os referencias de classificação empregados e parametrizados na implementação dos algoritmos de ML.

**8. CRONOGRAMA**

Abaixo detalhamos o cronograma das atividades previstas para a conclusão do presente projeto de pesquisa, dividindo-o em etapas semestrais consecutivas:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades** | **2/2020** | **1/2021** | **2/2021** | **1/2022** | **2/2022** | **1/2023** | **2/2023** | **1/2024** |
| 1. Revisão do Projeto de Pesquisa |  |  |  |  |  |  |  |  |
| 1. Cumprimento dos créditos referentes a disciplinas obrigatórias e opcionais |  |  |  |  |  |  |  |  |
| 1. Especialização técnica em algoritmos de ML |  |  |  |  |  |  |  |  |
| 1. Construção da fundamentação teórica |  |  |  |  |  |  |  |  |
| 1. Aquisição de dados e imagens de satélite |  |  |  |  |  |  |  |  |
| 1. Aplicação dos modelos de ML para dados estruturados |  |  |  |  |  |  |  |  |
| 1. Calibração e Validação dos modelos de ML para dados estruturados |  |  |  |  |  |  |  |  |
| 1. Aplicação de Modelos de Transferência de Aprendizado |  |  |  |  |  |  |  |  |
| 1. Qualificação (banca) |  |  |  |  |  |  |  |  |
| 1. Implementação de Modelo Generalizado de Estimativa de Variáveis de Qualidade da Água |  |  |  |  |  |  |  |  |
| 1. Sequência de análises e escrita da tese |  |  |  |  |  |  |  |  |
| 1. Calibração e Validação do modelo generalizado |  |  |  |  |  |  |  |  |
| 1. Revisão da tese |  |  |  |  |  |  |  |  |
| Defesa da tese |  |  |  |  |  |  |  |  |

**9. REFERÊNCIAS BIBLIOGRÁFICAS**

Aiken, G.R.; McKnight, D.M.;Wershaw, R.L.; MacCarthy, P. Humic Substances in Soil, Sediment, and Water: Geochemistry, Isolation and Characterization; John Wiley & Sons: New York, NY, USA, 1985.

Ayzel, G.V. (2021). Deep neural networks in hydrology: the new generation of universal and efficient models. Vestnik of Saint Petersburg University. Earth Sciences, 66 (1), 5–18. https://doi.org/10.21638/spbu07.2021.101

BARBOSA, C. C. F., NOVO, E. M. L. M., MARTINS, V. S. Introdução ao sensoriamento remoto de sistemas aquáticos: princípios e aplicações. São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2019, 178p.

BELLMAN, R. E. An introduction to artificial intelligence - Can Computers Think? San Francisco: Boyd & Fraser Publishing Company, 1978.

Boddula, V., Ramaswamy, L., e Mishra, D. (junho de 2017). A Spatio-Temporal Mining Approach for Enhancing Satellite Data Availability: A Case Study on Blue Green Algae. En G. Karypis e J. Zhang (Presidencia), IEEE 6th International Congress on Big Data, Honolulu, Hawaii, USA.

Chapman, Deborah. (1992). Water Quality Assessments - A Guide to Use of Biota, Sediments and Water in Environmental Monitoring - Second Edition. 10.4324/9780203476710.

Costa, Jean Carllo Jardim. Uma análise sobre o impacto de Dados Faltantes no desempenho de métodos de Aprendizado de Máquina / Jean Carllo Jardim Costa. – 2018. 55 f.

D. Jiang, J. Li, Y. Zhou, J. Wang, Y. Chen, and W. Xiao: Water Sci. Technol. Water Supply 18 (2018) 1173. https://doi.org/10.2166/ws.2017.189.

DUTTA, I. S.; MIA, I. The Global Information Technology Report 2010 – 2011. V. 24. Geneva: SRO- -Kundig, 2011.

G. Phillips, A. Lyche-Solheim, B. Skjelbred, U. Mischke, S. Drakare, G. Free, M. Järvinen, C.D. Hoyos, G. Morabito, S. Poikane A phytoplankton trophic index to assess the status of lakes for the Water Framework Directive Hydrobiologia, 704 (2013), pp. 75-95.

Dickson, A. G.: Deep Sea Res. Part A, Oceanogr. Res. Pap. 28 (1981) 609. https://doi.org/10.1016/0198-0149(81)90121-7

Gernaat, D.E.H.J., Bogaart, P.W., Vuuren, D.P.v. et al. High-resolution assessment of global technical and economic hydropower potential. Nat Energy 2, 821–828 (2017). https://doi.org/10.1038/s41560-017-0006-y

Gholizadeh MH, Melesse AM, Reddi L. A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques. Sensors (Basel). 2016;16(8):1298. Published 2016 Aug 16. doi:10.3390/s16081298

Jafar Alzubi et al 2018 J. Phys.: Conf. Ser. 1142 012012. Menard SW. Applied logistic regression analysis. Thousand Oaks: Sage Publications; 1995.

KARAOUI, I., ARIOUA, A., BOUDHAR, A., HSSAISOUNE, M., EL MOUATASSIME, S., OUHAMCHICH, K. A., ELHAMDOUNI, D., IDRISSI, A. A., NOUAIM, W. Evaluating the potential of Sentinel-2 satellite images for water quality characterization of artificial reservoirs: The Bin El Ouidane Reservoir case study (Morocco). Meteorology Hydrology and Water Management: Research and Operational Applications, v.7, n.1, p.31–39, 2019.

Kotsiantis, Sotiris B, I Zaharakis e P Pintelas: Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160:3–24, 2007.

Lee, Z.; Shang, S.; Hu, C.; Du, K.;Weidemann, A.; Hou,W.; Lin, J.; Lin, G. Secchi disk depth: A new theory and mechanistic model for underwater visibility. Remote Sens. Environ. 2015, 169, 139–149. [CrossRef]

Lim, J.; Choi, M. Assessment of water quality based on Landsat 8 operational land imager associated with human activities in Korea. Environ. Monit. Assess. 2015, 187, 1–17. [CrossRef] [PubMed]

Lima, I. V. (2005). Uma abordagem simplificada de detecção de instrusão baseada em Redes Neurais Artificiais. Universidade Federal de Santa Cataria.

Machine Learning with Neural Networks An Introduction for Scientists and Engineers. AUTHOR: Bernhard Mehlig, Göteborgs Universitet, SwedenPUBLICATION PLANNED FOR: January 2021

M. Lewis-Beck, A. Bryman, and T. F. Liao: SAGE Encycl. Soc. Sci. Res. Methods (2012) 1. https://doi.org/10.4135/9781412950589.n974

MENESES, P.R.; ALMEIDA, T.; BAPTISTA, G.M.M. (Org.). Reflectância dos materiais terrestres; análise e interpretação. Oficina de Textos, 2019.

Medium, 2021. Machine Learning: Conceitos e Modelos. Disponível em: https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-f0373bf4f445.

MICHIE, D.; SPIEGELHALTER, D.; TAYLOR, C.; CAMPBELL, J. Machine learning, neural and statistical classification. [S.l.]: Ellis Horwood, 1994.

Mustapha M.K., “Assessment of the Water Quality of Oyun Reservoir, Offa, Nigeria, Using Selected Physico-Chemical Parameters”, Turkish Journal of Fisheries and Aquatic Sciences, Vol. 8, 2008, pp.309-319.

Naime, R. 2012. Impactos socioambientais de hidrelétricas e reservatórios nas bacias hidrográficas brasileiras. Revista Eletrônica em Gestão, Educação e Tecnologia Ambiental, 7(7): 1409-1422.

Nogueira, Carolina Calió. Previsibilidade no mercado acionário utilizando machine learning / Carolina Calió

Nogueira. - 2019. 48 f. Orientador: Emerson Fernandes Marçal. Dissertação (mestrado profissional MPFE) – Fundação Getulio Vargas, Escola de Economia de São Paulo.

O. Pentakalos: Introduction to Machine Learning (MIT Press, 2019).

P.Y. Li, H. Qian, K.W.F. Howard, J. Wu Building a new and sustainable Silk Road economic belt Environ. Earth Sci., 74 (2015), pp. 7267-7270.

Ranjeeta C., Ratwani P. and Vishwakarma M., “Comparative study of Drinking Water Quality Parameters of three Manmade Reservoirs i.e. Kolar, Kaliasote and Kerwa Dam,” Current World Environment, Vol. 6(1), 2011, pp.145-149.

RUSSELL, STUART E NORVIG, PETER. 2013. Inteligência Artificial. Rio de Janeiro: Elsevier Editora Ltda, 2013.

S. A. Rounds, F. D. Wilde, and G. F. Ritz: Dissolved Oxygen (2006) Chap. A6, Sect. 6.2.

Schallenberg, M., de Winton, M. D., Verburg, P., Kelly, D. J., Hamill, K. D., & Hamilton, D. P. (2013). Ecosystem services of lakes. Ecosystem services in New Zealand: conditions and trends. Manaaki Whenua Press, Lincoln, 203-225.

Soares, P.L. e da Silva, J. 2011. Aplicação de Redes Neurais Artificiais em Conjunto com o Método Vetorial da Propagação de Feixes na Análise de um Acoplador Direcional Baseado em Fibra Ótica. Revista Brasileira de Computação Aplicada. 3, 2 (nov. 2011), 58-72. DOI:https://doi.org/10.5335/rbca.2013.1803.

SAMUEL, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development.

SANTOS, Dennis dos Santos. Inteligência Artificial: Conceitos e Aplicações. Revista Olhar Científico; Faculdade Associadas de Ariquemes, v. 01, n.2, Ago./Dez. 2010

Tan, Chuanqi & Sun, Fuchun & Kong, Tao & Zhang, Wenchang & Yang, Chao & Liu, Chunfang. (2018). A Survey on Deep Transfer Learning: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018, Proceedings, Part III. 10.1007/978-3-030-01424-7\_27.

TIBERIUS, Jose; O cérebro e os Computadores Modernos: a teoria cognitiva global – MOLWICK, v. 1, p. 16, 2016.

Thornton, K.W., Kimmel, B.L. and Payne, F.F. 1990 Reservoir Limnology: Ecological Perspectives. Wiley, New York, 246 pp.

Tundisi JG (2018) Reservoirs: New challenges for ecosystem studies and environmental management. Water Security 4: 1-7.

Verma A.K. and Singh T.N., “Prediction of water quality from simple field parameters”, Springer Environmental Earth Science, Vol. 69, 2013, pp.821–829 DOI-10.1007/s12665-012-1967-6.

**APÊNDICE A**

**Quadro A – Turbidez: Regressão, Métodos Aplicados e Resultados**



**APÊNDICE B**

**Quadro B – Clorofila: Regressão, Métodos Aplicados e Resultados**

