https://doi.org/10.1590/S1413-415220220099

Artigo Técnico

Inter-relações entre preditores de eutrofização em reservatórios do semiárido brasileiro: como mensurar? Uma aplicação de aprendizado de máquina por árvores de decisão

Interrelationships between eutrophication predictors in Brazilian semiarid reservoirs: how to measure? A decision tree machine learning application

Letícia Lacerda Freire[1]*, Francisco de Assis Souza Filho[1]

RESUMO

Um problema emergente para a segurança hídrica consiste nas consequências da eutrofização sobre a qualidade das águas. Metodologias de regressão convencionais não têm sido suficientes para explicar satisfatoriamente a complexidade da relação entre as variáveis hidrológicas e limnológicas desse processo. Nessa perspectiva, esta pesquisa buscou identificar preditores para variáveis indicadoras de eutrofização (cianobactérias, clorofila **a,** nitrogênio, fósforo e medição em disco de Secchi), por meio das relações destas entre si e entre 17 variáveis fisiográficas e climáticas das bacias hidrográficas de 155 reservatórios do semiárido brasileiro. Aplicou-se um método de aprendizado de máquina com o algoritmo *classification and regression trees* para árvores de decisão. Os resultados revelaram que os indicadores de eutrofização estão intrinsecamente relacionados entre si, de maneira especial as concentrações de clorofila *a* com os demais. A variabilidade da vazão afluente repercutiu no aumento da concentração de cianobactérias; a redução do volume de água armazenado gerou aumento da concentração de nitrogênio e fósforo; e a densidade de drenagem gerou aumento da concentração de nitrogênio. As concentrações de nitrogênio superiores a 5 mg.L^{-1} apresentaram consequências representativas sobre a clorofila **a**, a qual esteve fortemente associada às cianobactérias. O volume de água armazenado, a precipitação e a vazão afluente aos reservatórios também foram preditores da transparência das águas. Apesar de os índices de *performance* do modelo apontarem para margens de erro amplas para os conjuntos de dados com elevados coeficientes de variação, a aplicação de árvores de decisão pode auxiliar no entendimento de processos ocorridos e no planejamento de ações estratégicas para a governança hídrica.

Palavras-chave: nitrogênio; fósforo; cianobactérias; clorofila a; algoritmo *classification and regression trees*.

ABSTRACT

An emerging issue for water security is the consequences of eutrophication on water quality. Conventional regression methodologies have not been sufficient to satisfactorily explain the complexity of the relationship between the hydrological and limnological variables of this process. In this sense, this research aimed to identify predictors for eutrophication variables (cyanobacteria, chlorophyll-a, nitrogen, phosphorus, and Secchi disk measurement), through their relationships with each other and between 17 physiographic and climatic variables of the watersheds of 155 reservoirs in the Brazilian semiarid region. A machine learning method was applied with the classification and regression trees algorithm for decision trees. The results revealed that the eutrophication indicators are intrinsically related to each other, especially the concentrations of chlorophyll-a with the others. The variability of the inflow resulted in an increase in the concentration of cyanobacteria; the reduction in the volume of stored water generated an increase in the concentration of nitrogen and phosphorus; and, the drainage density generated an increase in the concentration of nitrogen. Nitrogen concentrations greater than 5 mg.L^{-1} had significant consequences on chlorophyll-a, which was strongly associated with cyanobacteria. The volume of stored water, precipitation and the inflow to the reservoirs were also predictors of water transparency. Although the model's performance indexes indicate wide margins of error for datasets with high coefficients of variation, decision trees can help in understanding the processes that have taken place and in planning strategic actions for water governance.

Keywords: nitrogen; phosphorus; cyanobacteria; chlorophyll-a; classification and regression trees algorithm.

[1]Departamento de Engenharia Hidráulica e Ambiental, Universidade Federal do Ceará – Fortaleza (CE), Brasil.
***Autora correspondente:** leticia.l.arquivos@gmail.com
Conflitos de interesse: os autores declaram não haver conflitos de interesse.
Financiamento: nenhum.
Recebido: 24/04/2022 – **Aceito:** 31/08/2022 – **Reg. ABES:** 20220099

INTRODUÇÃO

As alterações resultantes das atividades antrópicas no período pós-industrial estão acarretando um progressivo desequilíbrio dos ciclos biogeoquímicos de nitrogênio e fósforo, um dos nove aspectos essenciais para a dinâmica ecossistêmica configurada no holoceno (ROCKSTRÖM *et al.*, 2009).

Os reservatórios do semiárido possuem elevada vulnerabilidade à eutrofização, considerando o aumento das concentrações de nutrientes no volume de água armazenado, associado aos possíveis cenários de mudanças climáticas (LACERDA *et al.*, 2018; RAULINO; SILVEIRA; LIMA NETO, 2021). Em um período extremamente seco, entre os anos de 2008 e 2017, mais de 90% de 65 reservatórios localizados no semiárido brasileiro apresentaram aumento do índice de estado trófico e alcançaram condições hipereutróficas (WIEGAND *et al.*, 2021).

Como consequência da eutrofização, as concentrações elevadas de cianobactérias e a liberação de cianotoxinas para os recursos hídricos têm sido preocupações emergentes no âmbito da pesquisa, da gestão e da prestação de serviços de saneamento (SANTANA *et al.*, 2016). A floração de cianobactérias nos reservatórios pode exigir tratamentos de água mais complexos que ainda não estão implementados nessas regiões (BARROS *et al.*, 2017; PESTANA *et al.*, 2019; BARROS *et al.*, 2020). As cianobactérias possuem taxas de crescimento máximo em temperaturas acima de 25°C, o que gera maior fator de vulnerabilidade às águas dos reservatórios de lagos semiáridos tropicais quando comparadas àquelas de lagos temperados (CHORUS; WELKER, 2021). As concentrações de cianotoxinas detectadas em reservatórios de usos múltiplos do semiárido indicam a urgência do monitoramento de tais condições e do planejamento de ferramentas para o controle da poluição ambiental (COSTA *et al.*, 2006; FONSECA *et al.*, 2015; LORENZI *et al.*, 2018).

Muito já se conhece sobre as reações que desencadeiam a eutrofização em águas de reservatórios do semiárido (LACERDA *et al.*, 2018), no entanto a relação entre as características físicas da bacia hidrográfica e a dinâmica de desenvolvimento de algas e cianobactérias permanece pouco compreendida (ANDERSEN *et al.*, 2020). A fragilidade das correlações entre clorofila *a* e outras variáveis de qualidade da água indica que o uso de métodos de regressão tradicionais pode não ser suficiente para a explicação dos processos e a formulação de modelos preditivos, mesmo entre fósforo e nitrogênio, para os quais se tem correlações significativas (JIMENO-SÁEZ *et al.*, 2020). Em consequência disso, a inteligência artificial nas diversas áreas da ciência se tornou uma ferramenta matemática importante para a identificação de relações complexas entre diferentes variáveis (AHMED *et al.*, 2019). Em estudos no Vietnã, Pham *et al.* (2021) sugeriram que o controle de problemas de eutrofização, especialmente aqueles relacionados com florações de cianobactérias tóxicas em regiões tropicais, deve ser voltado para a avaliação de parâmetros limnológicos e hidrológicos, como a concentração de nutrientes e o nível da água armazenada nos reservatórios.

Desse modo, o presente trabalho buscou analisar a aplicação de uma metodologia de aprendizado de máquina baseada em árvores de decisão para a investigação de indicadores de eutrofização e a identificação de variáveis preditoras entre os aspectos limnológicos, fisiográficos e hidrológicos de reservatórios do semiárido brasileiro.

METODOLOGIA

Área de estudo

O presente trabalho foi realizado com base em dados de 155 reservatórios situados no estado do Ceará e cujas bacias hidrográficas estão inseridas no semiárido brasileiro. Na Figura 1 é apresentada a localização dos reservatórios e das respectivas bacias de contribuição.

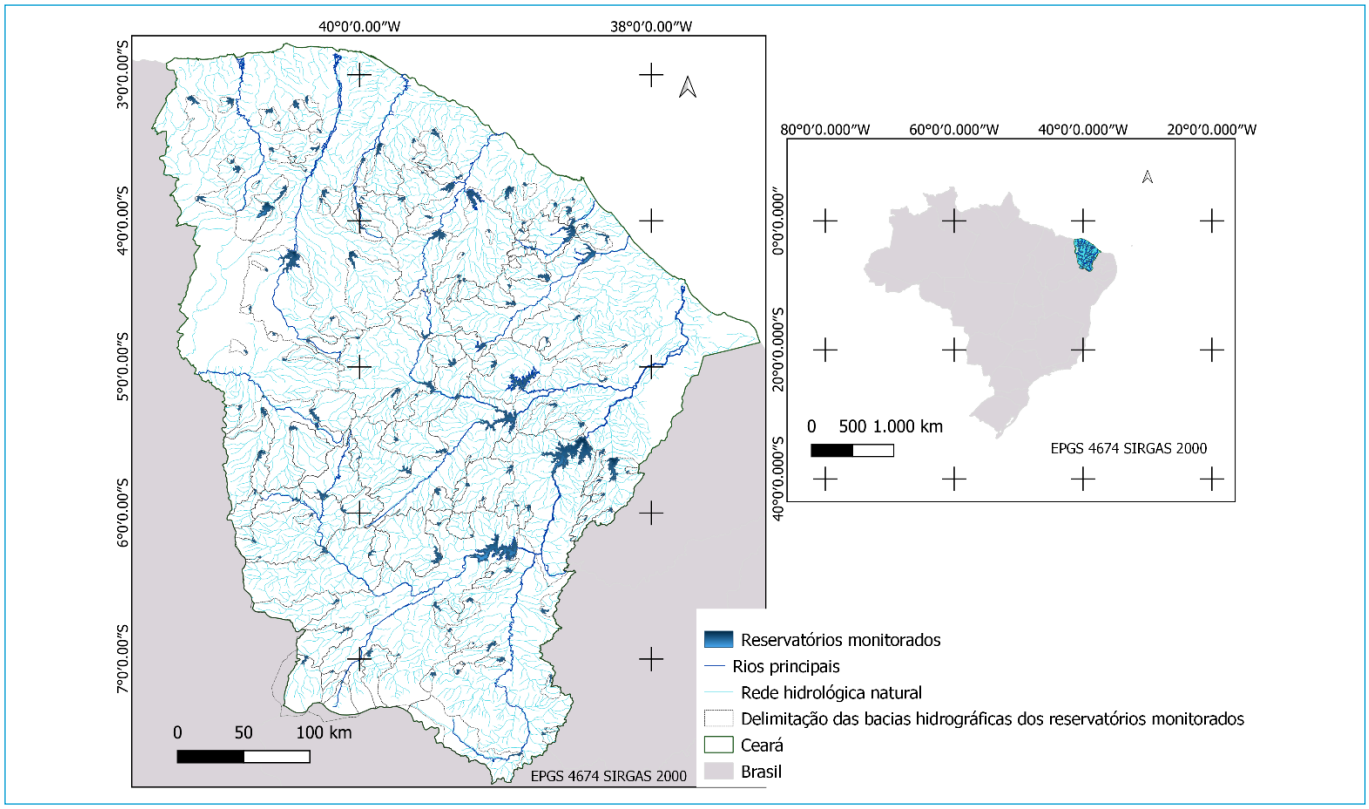


Figura 1 – Localização dos reservatórios monitorados no estado do Ceará e as respectivas bacias hidrográficas.

No Brasil, as áreas semiáridas são classificadas por resolução normativa federal, como aquelas com precipitação média anual igual ou inferior a 800 mm, índice de aridez de Thornthwaite igual ou inferior a 0,50 e percentual de déficit hídrico igual ou superior a 60% (BRASIL, 2017). Elevadas taxas de evaporação (> 2.000 mm) e solos rasos e cristalinos também são características do semiárido brasileiro (ALVALÁ *et al*., 2017).

A zona de convergência intertropical (ZCIT) atua como o principal sistema de formação de chuvas nessa região, na quadra chuvosa entre fevereiro e maio (FUNCEME, 2021). A precipitação local também sofre influência de fatores orográficos e de fenômenos oceânicos-atmosféricos, como o El-Niño, o qual pode inibir a formação de nuvens, mudando a posição da célula Walker e transferindo a ZCIT para a faixa mais ao sul, enquanto são esperadas precipitações acima da média com a La-Niña (ZANELLA, 2014; COSTA *et al*., 2017). As vazões anuais dos mananciais no semiárido brasileiro apresentam coeficientes de variação superiores a 1 (GÜNTNER; BRONSTERT, 2004). Além da variabilidade temporal de precipitação e de vazões, há variabilidade espacial. Em razão disso, Xavier *et al*. (1999) propuseram oito regiões pluviometricamente homogêneas para o Ceará, classificando as áreas centrais do estado (sertões de Crateús, Inhamuns e Jaguaribe) com precipitação anual por volta de 400 a 600 mm, o sul e o nordeste do estado (Cariri e maciço de Baturité) com médias normais entre 600 e 800 mm e aquelas localizadas na faixa litorânea e extremo noroeste (faixa litorânea e serra da Ibiapaba) com precipitações médias anuais variando entre 800 e 1.000 mm.

Obtenção de dados e variáveis analisadas

Os dados utilizados para a presente pesquisa foram disponibilizados pelo monitoramento realizado pela Fundação Cearense de Meteorologia e Recursos Hídricos e pela Companhia de Gestão dos Recursos Hídricos. O estado do Ceará possui sistemas de informações específicos para a obtenção de dados hidrológicos e meteorológicos. No Portal Hidrológico do Ceará (http://www.hidro.ce.gov.br/), é possível obter a série histórica de volume de água armazenado nos reservatórios, a capacidade de armazenamento dos reservatórios, os dados dos postos pluviométricos, as concentrações de nitrogênio, fósforo e clorofila a, a contagem de cianobactérias, a transparência da água dos reservatórios e as informações fisiográficas da bacia hidrográfica dos reservatórios. Os dados dos postos fluviométricos podem ser consultados no Portal Hidroweb da Agência Nacional das Águas (https://www.snirh.gov.br/hidroweb/apresentacao). O *shapefile* dos arquivos para determinação de características físicas, não expressas diretamente no banco de dados, pode ser acessado no Atlas dos Recursos Hídricos do Estado do Ceará (http://atlas.cogerh.com.br/). As informações de monitoramento dos recursos hídricos também podem ser obtidas por solicitação no Portal da Transparência do Estado do Ceará (https://ceara-transparente.ce.gov.br/).

Utilizaram-se as variáveis fisiográficas e climáticas (Quadro 1) como preditoras nos conjuntos de dados. As variáveis de qualidade da água — cianobactérias (Cél.mL^{-1}), clorofila *a* ($\mu g.L^{-1}$), nitrogênio total (mg.L^{-1}), fósforo total (mg.L^{-1}) e transparência da água por medição em disco de Secchi (m) — foram avaliadas como variáveis-alvo, uma por vez, a cada processamento do modelo de classificação adotado (ver Figura 2). Assim, organizou-se um banco de dados coletados e analisados entre os anos de 2008 e 2019, composto de 4.967 linhas e reduzido para 2.426 linhas, excluindo-se aquelas com falhas em pelo menos uma das variáveis de interesse.

Com base nos valores organizados por reservatórios, foram definidos dois conjuntos de dados, sendo o primeiro com a totalidade de informações para cada variável (Conjunto A), e o segundo com 70% dos dados subdivididos para treino e 30% para teste, formando o Conjunto B. Para a seleção de dados de treino e teste, foram segregados, de forma aleatória, os percentuais em cada subconjunto de linhas de variáveis para os reservatórios, com a finalidade de abranger as distintas condições (por exemplo, áreas das bacias hidrográficas, capacidades de armazenamento, padrões de variação de qualidade da água). Uma esquematização da formação dos subconjuntos está ilustrada na Figura 2.

Árvore de decisão

A árvore de decisão consiste em uma técnica de aprendizado de máquina supervisionada (Figura 3) em que são selecionadas variáveis preditoras e uma variável-alvo (TANGIRALA, 2020). A construção de uma árvore de decisão depende de vários estágios decisórios para a determinação de classes hierárquicas, com base em uma estrutura representada por um nó raiz, nós de decisão, nós terminais e arestas direcionadas (XU *et al*., 2021). Para que o nó raiz ou nó pai possa gerar nós de decisão ou nós filhos, um problema de maximização de impureza é resolvido.

Utilizou-se o algoritmo *classification and regression trees* (CART) em linguagem computacional R (BREIMAN *et al*., 1984; TIMOFEEV; HÄRFLE, 2004), em que a métrica para definição da classe é determinada pelo coeficiente de Gini. Em cada nó, o algoritmo busca minimizar o coeficiente de Gini e maximizar a homogeneidade na classe de subdivisão. Considerando a impureza do nó pai (i(Np)), a homogeneidade máxima ocorre quando $\Delta i\ (C)$, (Equação 1), ou seja, a diferença entre a impureza do nó pai e os nós filhos (esquerdo $\left(i(Ne)\right)$ e direito $\left(i(Nd)\right)$ para a classe —) for maximizada (Equação 3). Desse modo, o algoritmo realiza uma busca em toda a matriz

Quadro 1 – Variáveis fisiográficas e climáticas das bacias hidrográficas dos reservatórios avaliados.

1 Área de drenagem (A)	km^2
2 Perímetro (P)	Km
3 Percentual de solo cristalino (C%)	%
4 Coeficiente de compacidade (Kc)	Adimensional
5 Densidade de drenagem (DD)	km.km^{-2}
6 Capacidade de armazenamento do solo (CAD)	Mm
7 *Curve number* (CN)	Mm
8 Comprimento do talvegue (CTD)	Km
9 Declividade percentual (D%)	%
10 Precipitação média anual (Prec/ano)	mm.ano^{-1}
11 Evaporação média anual (Etp/ano)	mm. ano^{-1}
12 Capacidade de armazenamento do reservatório (Cap)	hm^3
13 Volume de água armazenado no reservatório (estoque)	
14 Afluência total (Aflu)	m^3.s^{-1}
15 Coeficiente de variação da afluência total anual (CV Aflu)	
16 Afluência incremental (Aflu inc)	m^3.s^{-1}
17 Coeficiente de variação da afluência média incremental (CV Aflu inc)	

do critério de classe que maximizará essa função. Como mencionado, nesse caso, a função de impureza foi determinada pelo índice de Gini (Equação 4), ao dividir os dados em classes com os subconjuntos de decisão mais puros, em que $p\left(\frac{|C_{i,D}|}{N}\right)$ é a probabilidade condicional de classe ($C_{i,D}$) em determinado nó (N). Substituindo o índice de Gini na Equação 3, obtém-se a equação resolvida pelo algoritmo CART em cada nó (Equação 5).

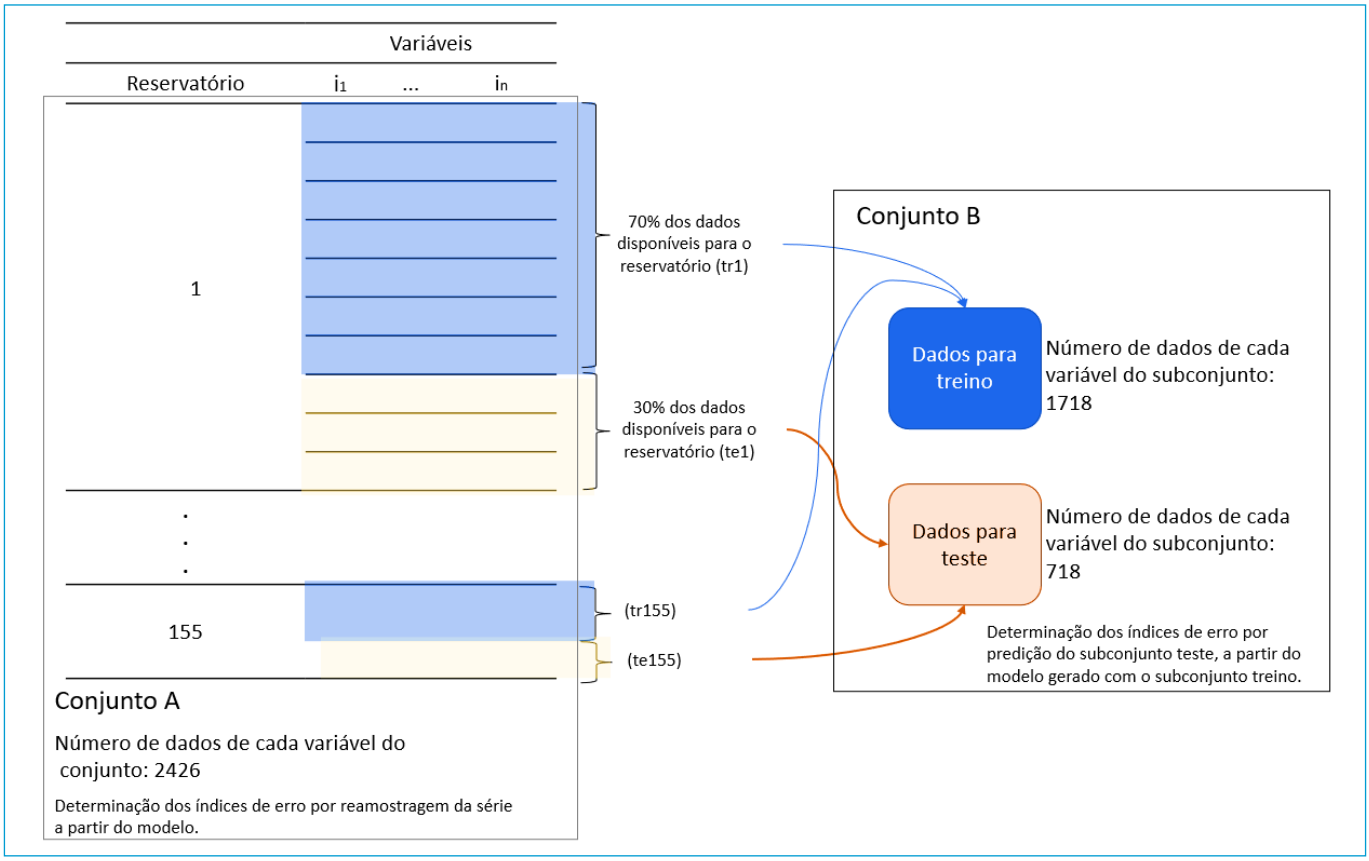
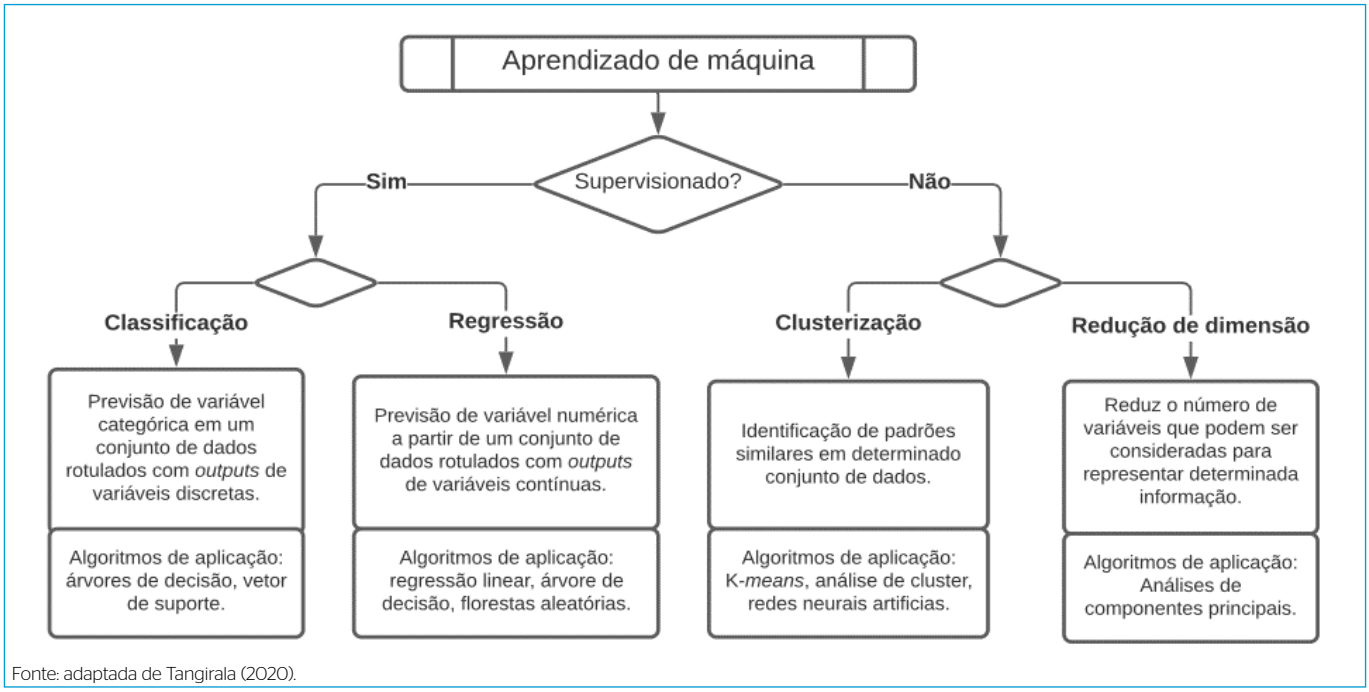


Figura 2 – Conjuntos de dados utilizados para o aprendizado de máquina.



Fonte: adaptada de Tangirala (2020).

Figura 3 – Técnicas de aprendizado de máquina.

$$\Delta i\ (C) = i(Np) - E(i(Nf) \quad (1)$$

$$\Delta i\ (C) = i(Np) - Pd\big(i(Nd)\big) - Pe(i(Ne)) \quad (2)$$

$$\max_{XJ \le x_j^R, J=1,\ldots,M} \Delta i\ (C) = i(Np) - Pd\big(i(Nd)\big) - Pe(i(Ne))] \quad (3)$$

$$i(N) = \sum_{i=1}^{n} p\left(\frac{|C_{i,D}|}{N}\right)\left(1 - p\left(\frac{|C_{i,D}|}{N}\right)\right) \quad (4)$$

$$\max_{XJ \le x_j^R, J=1,\ldots,M} - \sum_{i=1}^{n} p^2\ (Ci/Np) + Pd \sum_{i=1}^{n} p^2\ (Ci/Np) + Pd \sum_{i=1}^{n} p^2\ (Ci/Ne)] \quad (5)$$

O erro médio absoluto (MAE), o erro percentual médio absoluto (MAPE), a raiz do erro quadrático médio (RMSE) e o coeficiente de determinação (R^2) foram as métricas de desempenho avaliadas para a reamostragem utilizada no Conjunto A (com base no modelo de árvore de regressão obtido com os dados gerais) e para o conjunto de teste do modelo desenvolvido com base nos dados de treino do Conjunto B. O RMSE (Equação 6) possui ampla aplicação em pesquisas climáticas e ambientais, no entanto pode ser influenciado pela magnitude da amostra e pela raiz quadrada no número de erros, sendo o MAE (Equação 7) uma resposta mais fidedigna ao erro identificado (WILLMOTT; MATSUURA, 2005). O MAPE (Equação 8) determina o percentual de erro médio em relação aos valores reais, enquanto o R^2 varia de 0 a 1 e indica a parcela da variância dos dados que consegue ser explicada pelo modelo (Equação 9) (LIAO *et al.*, 2021; SINGH *et al.*, 2021). Nas Equações de 6 a 9, as variáveis n, yi, $\hat{y}$ e $\bar{y}$ representam o tamanho da amostra, os valores previstos, os valores observados e o valor médio observado, respectivamente.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y})^2} \quad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}| \quad (7)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|yi - \hat{y}|}{yi}\ 100\% \quad (8)$$

$$R^2 = 1 - \frac{\sum(yi - \hat{y})^2}{\sum(yi - \bar{y})^2} \quad (9)$$

RESULTADOS E DISCUSSÃO

A estatística descritiva para as médias de dados para cada reservatório está apresentada na Tabela 1. A maior heterogeneidade de concentrações foi identificada para as cianobactérias, com coeficiente de variação máximo superior a 4, considerando a estatística descritiva individual em cada reservatório. As concentrações de nitrogênio, fósforo, clorofila *a* e as medidas de transparência também apresentaram coeficientes de variação superiores a 1, indicando elevada variabilidade tanto para o conjunto amostral total quanto para as amostragens em um mesmo reservatório. Comparando os grupos amostrais entre si, a variável mais homogênea consistiu na transparência da água, cujos valores médios nos reservatórios variaram entre 0,20 e 2,49 m. Por meio dos resultados, observou-se ainda que há reservatórios mais fortemente impactados por processos de eutrofização.

A árvore de decisão para a qual a contagem de cianobactérias foi considerada como variável-alvo está na Figura 4. Dentro das caixas do nó raiz (primeira caixa da figura) e dos nós de decisão (caixas subsequentes) estão o valor médio do conjunto e o percentual de dados contidos nele. Vê-se nessa primeira árvore (Figura 4) que a classe mais representativa para as cianobactérias ocorre para a condição cujo coeficiente de variação de afluência é inferior a 1,8 (98% dos dados). Nesse subconjunto, o valor médio da contagem de cianobactérias corresponde a 1,96e + 5. Essa lógica interpretativa pode ser adotada no entendimento das demais árvores de decisão. Os coeficientes de variação maiores que 1,8 ocorreram em reservatórios que apresentaram concentrações de cianobactérias mais elevadas quando comparadas ao outro subconjunto.

Segundo Wu *et al.* (2016), a variabilidade da clorofila *a*, a qual também esteve relacionada com as concentrações de cianobactérias, pode estar associada ao transporte de diferentes tipos de sedimento aos recursos hídricos, o que pode ser alterado com a variabilidade da vazão afluente. As reduções de vazão afluente podem impulsionar o desenvolvimento de cianobactérias, em razão do aumento das concentrações de fósforo reativo, concomitantemente ao desenvolvimento de macrófitas e à redução da transparência e da profundidade da água (DALU; WASSERMAN, 2018). Esse fato também corrobora o conjunto de classes da Figura 5.

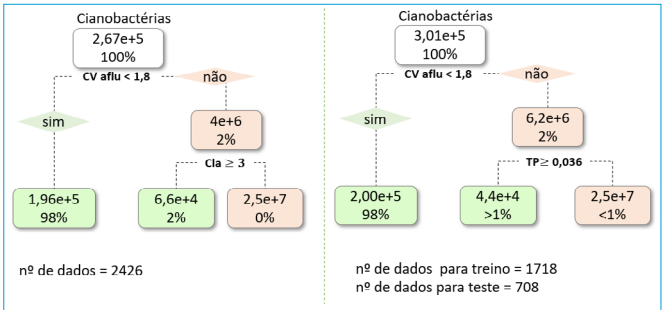
Em relação à clorofila *a*, o fator de decisão foi preponderante para as concentrações de nitrogênio total. Nesse sentido, Wiegand *et al.* (2020) identificaram a importância do nitrogênio como nutriente limitante em reservatórios do semiárido brasileiro, por meio de modelos empíricos e de regressão linear. Além disso, Andersen *et al.* (2020) verificaram aumento de cianobactérias fortemente influenciado pelas concentrações de nitrogênio amoniacal. Embora o nitrogênio não tenha sido detectado como variável preditora direta das concentrações de cianobactérias, estas se apresentaram como parâmetro de definição das subclasses preditoras de clorofila *a* e estiveram diretamente relacionadas a uma classe precedente definida pelas concentrações de nitrogênio.

Ademais, nessa variável-alvo, houve consistência ao comparar a classificação dos resultados para a totalidade dos dados e o conjunto de treino e teste no que se refere às seguintes variáveis preditoras: nitrogênio total, transparência

Tabela 1 – Estatística descritiva para os valores médios em cada reservatório das variáveis-alvo avaliadas.

	Cianobactérias	Fósforo total	Nitrogênio total	Transparência	Clorofila a
Média	248.700,29	0,16	2,45	0,80	62,63
Mínimo	2.239,73	0,03	0,55	0,20	3,63
Máximo	9.386.639,16	1,23	9,21	2,49	338,29
Desvio padrão	764.380,36	0,13	1,56	0,42	54,11
Coeficiente de variação	3,07	0,82	0,64	0,53	0,86

e cianobactérias. Todavia, cabe destacar que uma maior quantidade de ramificações e classificações quanto ao nitrogênio e à concentração de cianobactérias foi identificada para o conjunto da totalidade de dados. Tal fato indica homogeneidade na capacidade representativa do modelo, uma vez que as árvores de decisão tendem a ser inconstantes quando a amostra apresenta coeficientes de variação elevados, tratando-se de uma metodologia cujos resultados estão bastante associados a um determinado conjunto de dados avaliados (TIMOFEEV; HÄRFLE, 2004; TANGIRALA, 2020). A capacidade representativa também foi evidenciada pelas métricas de *performance* (Tabela 2), em que o R^2 foi superior a 0,5 para os dois conjuntos amostrais e os valores do MAPE foram reduzidos (< 3%) para ambos os casos.

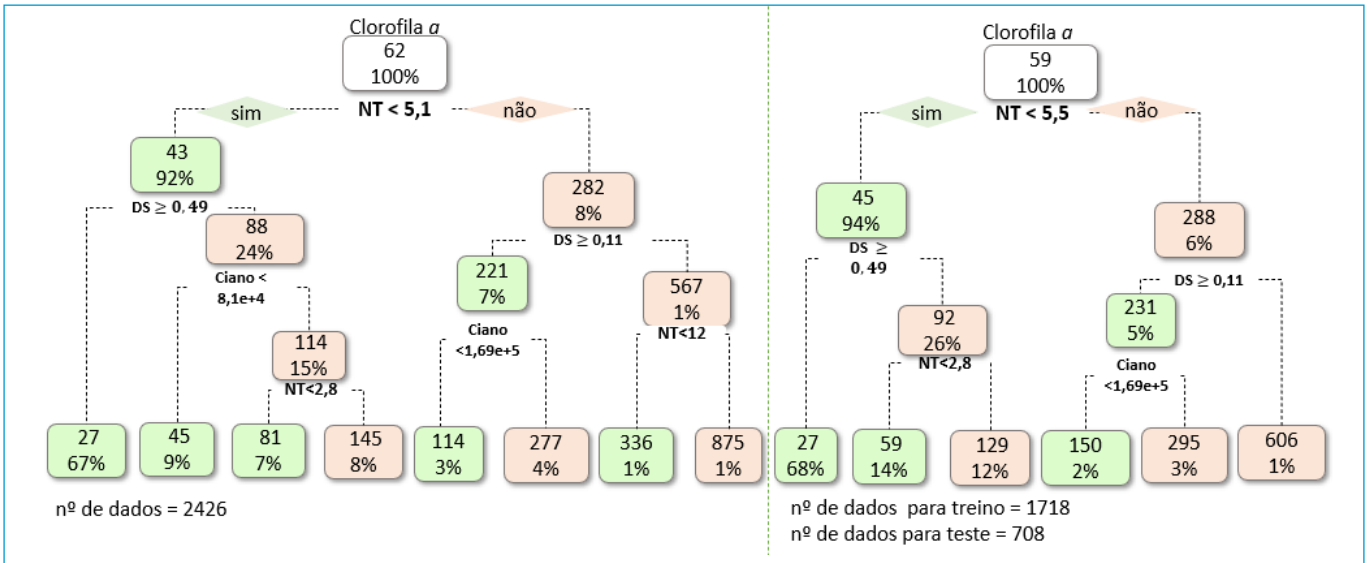


CV aflu: coeficiente de variação da vazão afluente total; Cla: concentração de clorofila ($\mu g.L^{-1}$); TP: concentração de fósforo total (mg.L

Figura 4 – Árvore de decisão para os dois agrupamentos de dados avaliados considerando as concentrações de cianobactérias (Cél.mL^{-1}) como variável-alvo.

De modo geral, quando as concentrações de nitrogênio total são superiores a 5 mg.L^{-1}, as concentrações médias de clorofila *a* superam 100 $\mu g.L^{-1}$, característica de ambientes eutrofizados (PACHECO; LIMA NETO, 2017; NGUYEN *et al.*, 2019). Ainda que o nitrogênio total não atinja essa concentração, com transparência superior a 0,50 m, a concentração de 27 $\mu g.L^{-1}$ foi identificada em mais de 60% do conjunto amostral. Se a transparência da água for inferior ao valor mencionado, o que aconteceu em mais de 20% dos casos, a clorofila *a* ainda pode alcançar 80 $\mu g.L^{-1}$, estando relacionada com as concentrações de cianobactérias.

Quanto aos resultados obtidos para as variáveis preditoras de nitrogênio total (Figura 6), a clorofila *a* esteve presente na definição da maior parte das subdivisões de nós, nos conjuntos A e B. Apesar de o fósforo total aparecer como preditor para os dados gerais, essa subdivisão não diferencia um percentual significativo de dados. Duas variáveis de influência hidrológica foram identificadas: a densidade de drenagem e o estoque do reservatório (Figura 6).



DS: transparência da água pela medição em disco de Secchi (m); Ciano: contagem de cianobactérias (Cél.mL^{-1}); NT: concentração de nitrogênio total (mg.L^{-1}).

Figura 5 – Árvore de decisão para os dois agrupamentos de dados avaliados considerando as concentrações de clorofila a ($\mu g.L^{-1}$) como variável-alvo.

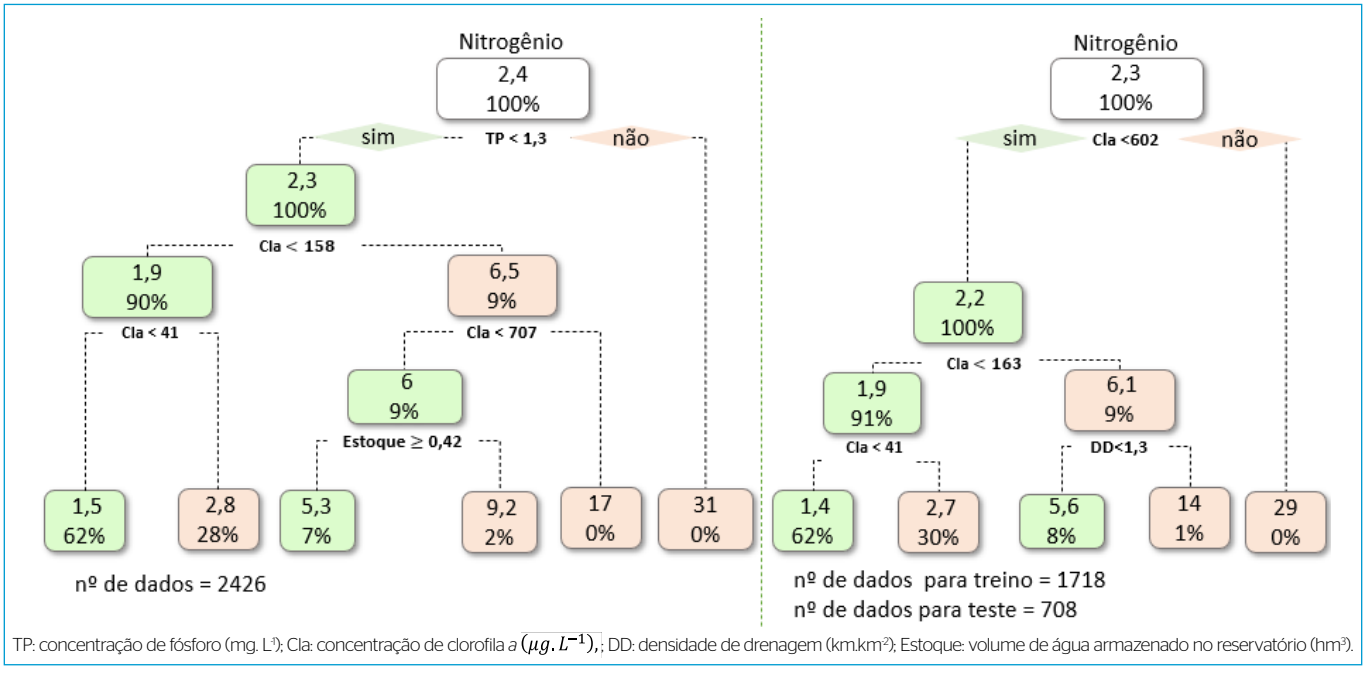
Tabela 2 – Resultados das métricas de performance aplicadas.

	Cianobactérias		Clorofila a		Nitrogênio total		Fósforo total		Transparência	
	A	B	A	B	A	B	A	B	A	B
MAE	337.864,50	431.723,20	35,53	42,35	1,07	1,29	0,09	0,09	0,32	0,31
MAPE	55,73	30,26	2,67	2,73	0,88	0,74	0,98	1,01	*	0,58
RMSE	3.401.137	2.540.355	70,81	77,62	2,66	2,87	0,30	0,23	0,77	0,63
R^2	0,14	*	0,62	0,61	0,46	0,34	0,28	0,07	0,29	0,22

MAE: erro médio absoluto; MAPE: erro percentual médio absoluto; RMSE: raiz do erro quadrático médio; R^2: coeficiente de determinação; A: conjunto de dados gerais (avaliação por reamostragem); B: conjunto de dados divididos em treino e teste; *indeterminado.

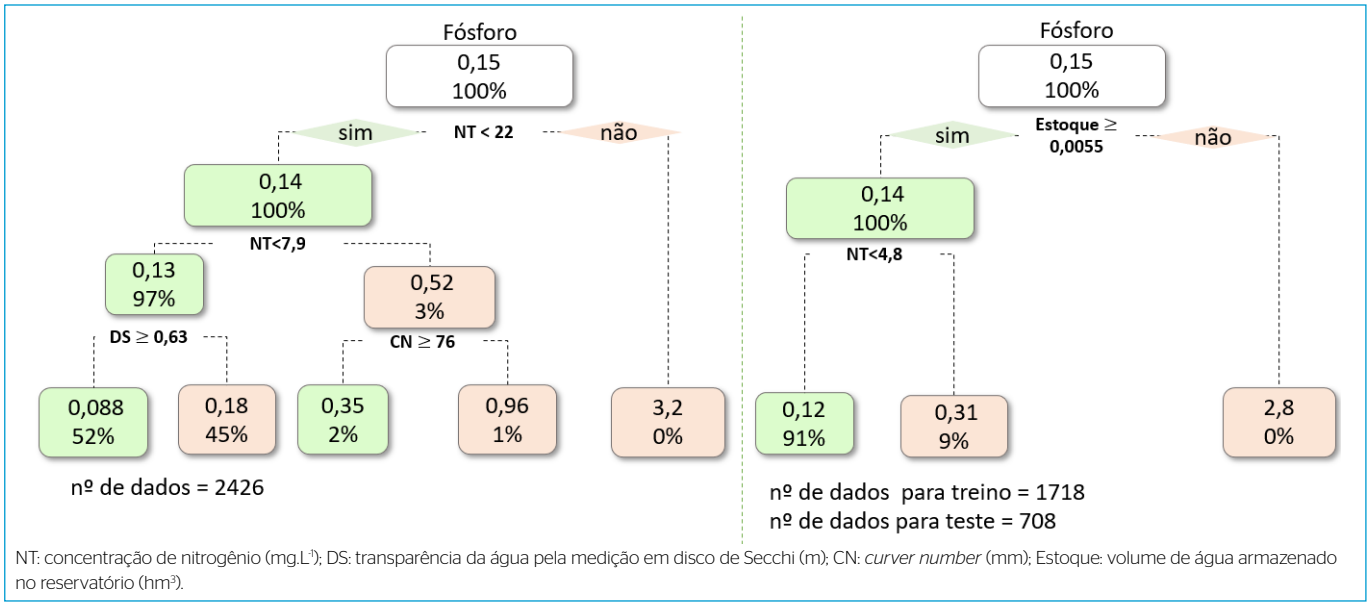
Observou-se que uma maior densidade de drenagem (> 1,3) influenciou no aumento da concentração de nitrogênio total, enquanto o inverso ocorreu em relação à influência do estoque sobre a concentração de nitrogênio. Para as concentrações de fósforo total (Figura 7), os fatores preditivos mais influentes foram as concentrações de nitrogênio total, a transparência e o *curve number*. Quanto menor a transparência (< 0,63), maiores foram as concentrações de fósforo. O inverso ocorreu para o *curve number*.

As concentrações de clorofila *a* foram as preditoras para a classificação inicial dos dados de transparência nos reservatórios avaliados (Figuras 8 e 9). Concentrações de clorofila *a* superiores a 23 $\mu g.L^{-1}$ indicaram média de 0,5 m de transparência da água no manancial, e a transparência da água reduziu consideravelmente quando a concentração de 98 $\mu g.L^{-1}$ foi superada, limitando-se a um valor médio de 0,30 m. Por outro lado, quando as concentrações de clorofila *a* foram inferiores a 23 $\mu g.L^{-1}$, pôde ser identificada uma zona eufótica de 1,2 m, reduzida gradativamente pelo aumento das concentrações de fósforo e da vazão afluente para os dados gerais (Conjunto A). As concentrações de clorofila *a*, fósforo total e nitrogênio total, o volume acumulado no reservatório e o coeficiente de compacidade da bacia hidrográfica foram as variáveis preditoras para o conjunto de dados selecionados para teste e treino (Conjunto B). A afluência total também foi identificada como variável preditora



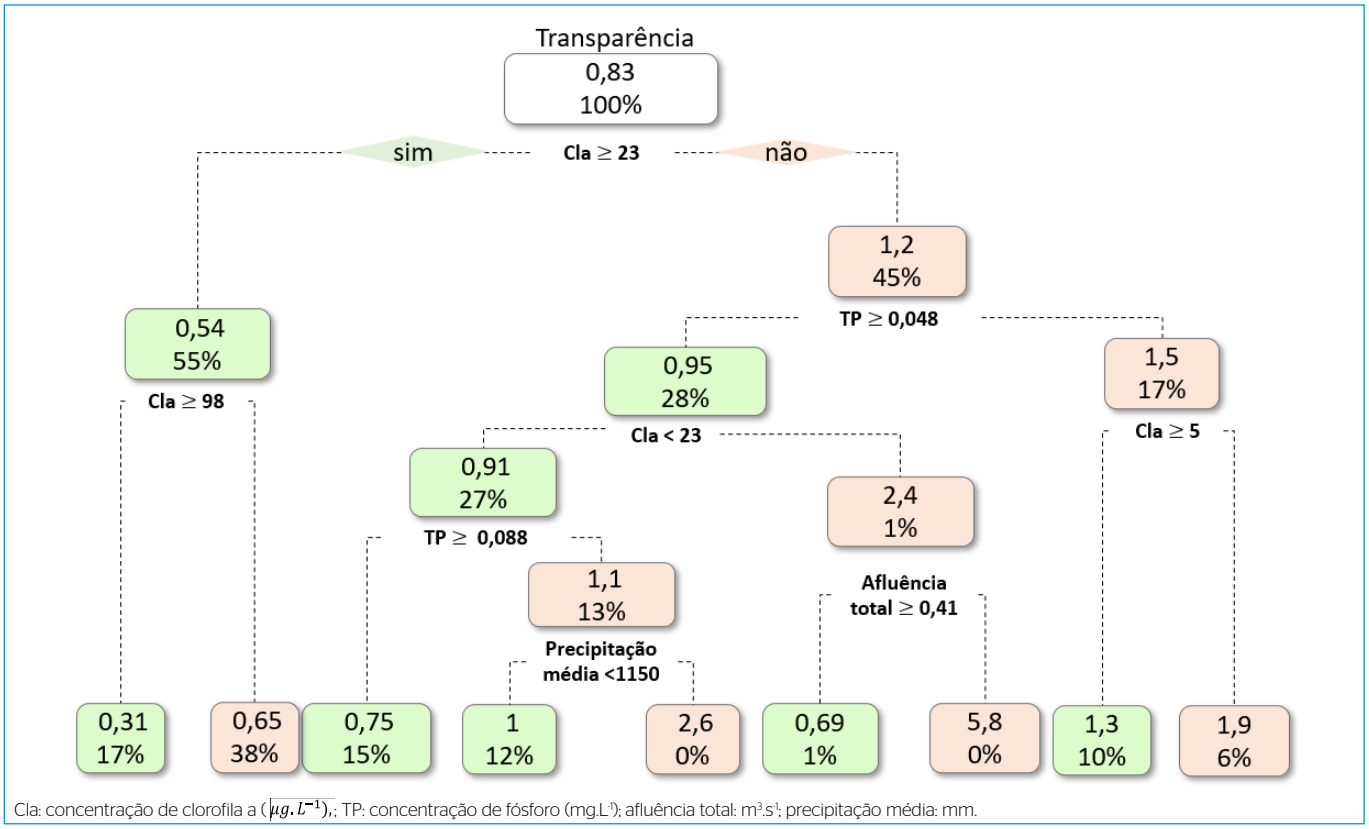
TP: concentração de fósforo (mg. L^{-1}); Cla: concentração de clorofila *a* $(\mu g.L^{-1})$,; DD: densidade de drenagem (km.km^{-2}); Estoque: volume de água armazenado no reservatório (hm^3).

Figura 6 – Árvore de decisão para os dois agrupamentos de dados avaliados considerando as concentrações de nitrogênio total (mg. L^{-1}) como variável-alvo.



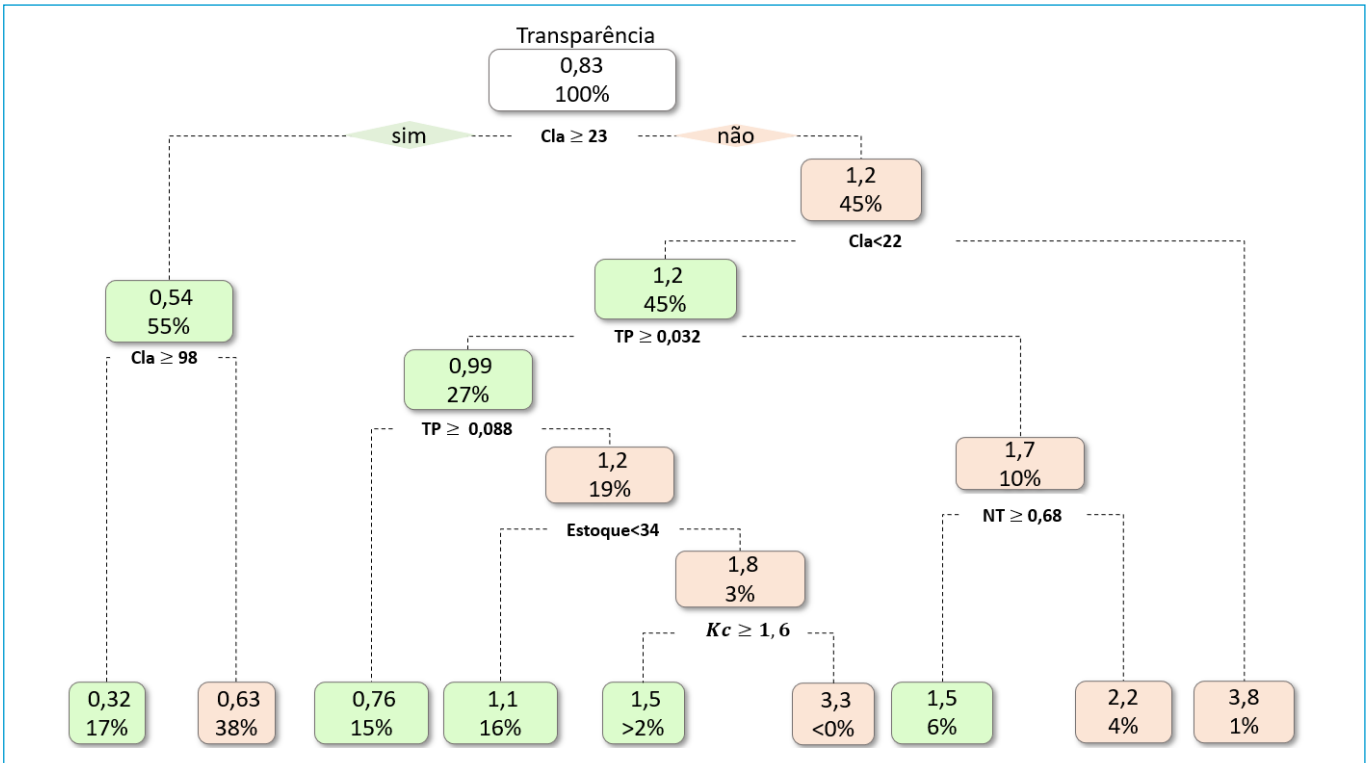
NT: concentração de nitrogênio (mg.L^{-1}); DS: transparência da água pela medição em disco de Secchi (m); CN: *curver number* (mm); Estoque: volume de água armazenado no reservatório (hm^3).

Figura 7 – Árvore de decisão para os dois agrupamentos de dados avaliados considerando as concentrações de fósforo (mg.L^{-1}) como variável-alvo.



Cla: concentração de clorofila a ($\mu g.L^{-1}$),; TP: concentração de fósforo (mg.L^{-1}); afluência total: m^3.s^{-1}; precipitação média: mm.

Figura 8 – Árvore de decisão para os dois agrupamentos de dados avaliados considerando as medições do disco de Secchi (m) como variável-alvo (conjunto A).



Cla: concentração de clorofila a ($\mu g.L^{-1}$),; TP: concentração de fósforo (mg.L^{-1}); NT: nitrogênio total; Estoque: volume de água armazenado no reservatório (hm^3); kc: coeficiente de compacidade.

Figura 9 – Árvore de decisão para os dois agrupamentos de dados avaliados considerando as medições do disco de Secchi (m) como variável-alvo (Conjunto B).

da transparência, em que vazões superiores a 0,41 $m^3.s^{-1}$ resultaram em valores maiores de transparência da água.

Nesse sentido, Silva *et al.* (2019) relacionaram o aumento da afluência com o aumento da biomassa de cianobactérias em reservatórios de bacias hidrográficas urbanas, no entanto, em grandes reservatórios rurais, Rocha e Lima Neto (2020) identificaram um modelo de diluição simples das concentrações de fósforo com o aumento da afluência. De acordo com Costa *et al.* (2019), a redução do estoque de água nos reservatórios durante os períodos de seca favorece o aumento das concentrações de cianobactérias e o desprendimento de sedimentos para a coluna d'água, o que pode justificar as relações identificadas.

Quanto às métricas de *performance* (Tabela 2), observou-se que o MAPE foi obtido para as medições de transparência, no entanto o maior R^2 ocorreu para o modelo de árvores formulado para a clorofila *a*. Para as concentrações de cianobactérias, a elevada variabilidade e a ordem de grandeza de variação dos dados podem ter prejudicado as métricas mais expressivas, o que já era esperado pela quantidade de variáveis influentes identificadas e os seus respectivos percentuais de representação. Os modelos propostos para nitrogênio e fósforo também apresentaram o MAPE com valores reduzidos, embora as indicações de R^2 não sugiram boa correlação entre valores previstos e valores reais.

As variáveis fisiográficas identificadas como preditoras com percentuais de representação mais expressivos consistiram na variabilidade da vazão afluente, no volume de água armazenado no reservatório e na densidade de drenagem. A variabilidade da vazão afluente repercutiu no aumento da concentração de cianobactérias; a redução do volume de água armazenado gerou aumento da concentração de nitrogênio e fósforo; e a densidade de drenagem esteve relacionada ao aumento da concentração de nitrogênio. Em reservatórios de regiões semiáridas, as concentrações de nitrogênio superiores a 5 $mg.L^{-1}$ podem implicar maiores concentrações de clorofila *a*, o que pode reduzir a zona eufótica para cerca de 0,30 m, acelerando o agravamento dos processos de eutrofização.

A aplicação de árvores de decisão deve permanecer cautelosa para objetivos de predição, uma vez que conjuntos de dados com elevados coeficientes de variação podem levar a erros percentuais de reamostragem e previsão superiores a 50%. No entanto, para o entendimento de processos e para a inferência de medidas e padrões a serem aplicados com base em cenários ocorridos, a ferramenta apresentou-se como uma alternativa importante, podendo contribuir para a tomada de decisões em ações de controle de poluição e da gestão dos recursos hídricos. Trabalhos futuros podem investigar tais relações com parâmetros de uso e ocupação do solo.

CONCLUSÃO

Os indicadores de eutrofização avaliados apresentaram relações frequentes entre si, principalmente no que se refere às concentrações de clorofila *a*, a qual esteve relacionada com a maioria das classes de decisão resultantes dos modelos.

CONTRIBUIÇÕES DOS AUTORES

Freire, L.L.: Primeira Redação, Curadoria de Dados, Escrita — Revisão e Edição. Souza Filho, F.A.: Conceituação, Curadoria de Dados, Supervisão, Escrita — Revisão e Edição.

REFERÊNCIAS

AHMED, A.N.; OTHMAN, F.B.; AFAN, H.A.; IBRAHIM, R.K.; FAI, C.M.; HOSSAIN, M.S.; EHTERAM, M.; ELSHAFIE, A. Machine learning methods for better water quality prediction. *Journal of Hydrology*, v. 578, 124084, 2019. https://doi.org/10.1016/j.jhydrol.2019.124084

ALVALÁ, R.C..; CUNHA, A.P.M.A.; BRITO, S.S.B.; SELUCHI, M.E.; MARENGO, J.A.; MORAES, O.L.L.; CARVALHO, M.A. Drought monitoring in the Brazilian Semiarid region. *Anais da Academia Brasileira de Ciências*, v. 91, supl. 1, p. 1-15, 2017. https://doi.org/10.1590/0001-3765201720170209

ANDERSEN, I.; WILLIAMSON, T.J.; GONZÁLEZ, M.J.; VANNI, M.J. Nitrate, ammonium, and phosphorus drive seasonal nutrient limitation of chlorophytes, cyanobacteria, and diatoms in a hyper-eutrophic reservoir. *Limnology and Oceanography*, v. 65, n. 5, p. 962-978, 2020. https://doi.org/10.1002/lno.11363

BARROS, M.U.G.; LEITÃO, J.I.R.; ARANHA, T.R.B.T.; SIMSEK, S.; BULEY, R.P.; FERNANDEZ-FIGUEROA, E.G.; GLADFELTER, M.F.; WILSON, A.E.; CAPELO-NETO, J. Icyano: a cyanobacterial bloom vulnerability index for drinking water treatment plants. *Water Supply*, v. 20, n. 8, p. 3517-3530, 2020. https://doi.org/10.2166/ws.2020.239

BARROS, M.U.G.; LOPES, I.K.C.; CARVALHO, S.M.C.; CAPELO-NETO, J. Impact of filamentous cyanobacteria on the water quality of two tropical reservoirs. *Revista Brasileira de Recursos Hídricos*, v. 22, e6, 2017. https://doi.org/10.1590/2318-0331.011716072

BRASIL. Resolução n° 107, de 13 de setembro de 2017. Estabelece critérios técnicos e científicos para delimitação do Semiárido Brasileiro e procedimentos para revisão de sua abrangência. *Diário Oficial da União*, Seção 1, edição 176, p. 48, 2017.

BREIMAN, L.; FRIEDMAN.; OLSHEN, R.A.; STONE, C.J. *Classification and Regression Trees*. Chapman & Hall/CRC, 1984.

CHORUS, I.; WELKER, M. Toxic cyanobacteria in water: a guide to their public health consequences, monitoring and management. 2. ed. World Health Organization, Taylor & Francis Group, 2021. 859 p.

COSTA, C.R.; COSTA, M.F.; BARLETTA, M.; ALVES, L.H.B. Interannual water quality changes at the head of a tropical estuary. *Environmental Monitoring and Assessment*, v. 189, n. 12, p. 1-13, 2017. https://doi.org/10.1007/s10661-017-6343-2

COSTA, I.A.S.; AZEVEDO, S.M.F.O.; SENNA, P.A.C.; BERNARDO, R.R.; COSTA, S.M.; CHELLAPPA, N.T. Occurrence of toxin-producing cyanobacteria blooms in a Brazilian semiarid reservoir. *Brazilian Journal of Biology*, v. 66, n. 1b, p. 211-219, 2006. https://doi.org/10.1590/S1519-69842006000200005

COSTA, M.R.; MENEZES, R.F.; SARMENTO, H.; ATTAYDE, J.L.; STERNBERG, L.S.L.; BECKER, V. Extreme drought favors potential mixotrophic organisms in tropical semi-arid reservoirs. *Hydrobiologia*, v. 831, p. 43-54, 2019. https://doi.org/10.1007/s10750-018-3583-2

DALU, T.; WASSERMAN, R.J. Cyanobacteria dynamics in a small tropical reservoir: Understanding spatio-temporal variability and influence of environmental variables. *Science of the Total Environment*, v. 643, p. 835-841, 2018. https://doi.org/10.1016/j.scitotenv.2018.06.256

FONSECA, J.R.; VIEIRA, P.C.; KUJBIDA, P.; COSTA, I.A.S. Cyanobacterial occurrence and detection of microcystins and saxitoxins in reservoirs of the Brazilian semi-arid. *Acta Limnologica Brasiliensia*, v. 27, n. 1, p. 78-92, 2015. https://doi.org/10.1590/S2179-975X2814

FUNDAÇÃO CEARENSE DE METEOROLOGIA E RECURSOS HÍDRICOS (FUNCEME). *Para entender melhor a previsão meteorológica para a estação chuvosa no Ceará*. Nova York: Programa para a América Latina do International Research Institute for Climate and Society. Disponível em: http://www.funceme.br/produtos/manual/clima/Clima/boletins_clima_alerta/EntenderPrevisaoQuadraChuvosa.pdf. Acesso em: 1º set. 2021.

GÜNTNER, A.; BRONSTERT, A. Representation of landscape variability and lateral redistribution processes for large-scale hydrological modeling in semi-arid areas. *Journal of Hydrology*, v. 297, n. 1-4, p. 136-161, 2004. https://doi.org/10.1016/j.jhydrol.2004.04.008

HOANG, T.T.; NGUYEN, V.D.; VAN, A.D.; NGUYEN, H.T.T. Decision tree techniques to assess the role of daily DO variation in classifying shallow eutrophicated lakes in Hanoi, Vietnam. *Water Quality Research Journal*, v. 55, n. 1, p. 67-78, 2020. https://doi.org/10.2166/wqrj.2019.105

JIMENO-SÁEZ, P.; SENENT-APARICIO, J.; CECILIA, J.M.; PÉREZ-SÁNCHEZ, J. Using machine-learning algorithms to eutrophication modeling: case study of Mar Menor Lagoon (Spain). *International Journal of Environmental Research and Public Health*, v. 17, n. 4, 1189, 2020. https://doi.org/10.3390/ijerph17041189

LACERDA, L.D.; SANTOS, J.A.; MARINS, R.V.; SILVA, F.A.T.F. Limnology of the largest multi-use artificial reservoir in NE Brazil: The Castanhão Reservoir, Ceará State. *Anais da Academia de Ciências*, v. 90, n. 2, supl. 1, p. 2073-2096, 2018. https://doi.org/10.1590/0001-3765201820180085

LIAO, Z.; ZANG, N.; WANG, X.; LI, C.; LIU, Q. Machine learning-based prediction of chlorophyll-a variations in receiving reservoir of world's largest water transfer project – a case study in the Miynun Reservoir, North China. *Water*, v. 13, n. 17, 2406, 2021. https://doi.org/10.3390/w13172406

LORENZI, A.S.; CORDEIRO-ARAÚJO, M.K.; CHIA, M.A.; BITTENCOURT-OLIVEIRA, M.C. Cyanotoxin contamination of semiarid drinking water supply reservoirs. *Environmental Earth Sciences*, *v.* 77, 595, 2018. https://doi.org/10.1007/s12665-018-7774-y

NGUYEN, T.T.N.; NÉMERY, J.; GRATIOT, N.; STRADY, E.; TRAN, V.Q.; NGUYEN, A.T.; AIMÉ, J.; PEYNE, A. Nutrient dynamics and eutrophication assessment in the tropical river system of Saigon – Dongnai (southern Vietnam). *Science of the Total Environment*, v. 653, p. 370-383, 2019. https://doi.org/10.1016/j.scitotenv.2018.10.319

PACHECO, C.H.A.; LIMA NETO, I.E. Effect of artificial circulation on the removal kinetics of cyanobacteria in a hypereutrophic shallow lake. *Journal of Environmental Engineering*, v. 143, n. 12, 2017. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001289

PESTANA, C.J.; CAPELO-NETO, J.; LAWTON, L.; OLIVEIRA, S.; CARLOTO, I.; LINHARES, H.P. The effect of water treatment unit processes on cyanobacterial trichome integrity. *Science of the Total Environment*, v. 659, p. 1403-1414, 2019. https://doi.org/10.1016/j.scitotenv.2018.12.337

PHAM, T.; TRAN, T.H.Y.; SHIMIZU, K.; LI, Q.; UTSUMI, M. Toxic cyanobacteria and microcystin dynamics in a tropical reservoir: assessing the influence of environmental variables. *Environmental Science and Pollution Research*, v. 28, p. 63544-63557, 2021. https://doi.org/10.1007/s11356-020-10826-9

RAULINO, J.B.S.; SILVEIRA, C.S.; LIMA NETO, I.E. Assessment of climate change impacts on hydrology and water quality of large semi-arid reservoirs in Brazil. *Hydrological Sciences Journal*, v. 66, n. 8, p. 1321-1336, 2021. https://doi.org/10.1080/02626667.2021.1933491

ROCHA, M.J.D.; LIMA NETO, I.E. Modeling flow-related phosphorus inputs to tropical semiarid reservoirs. *Journal of Environmental Management*, v. 295, 113123, 2021. https://doi.org/10.1016/j.jenvman.2021.113123

ROCHA, M.J.D.; LIMA NETO, I.E. Relação entre fósforo total e vazão afluente nos principais reservatórios rurais do Estado do Ceará no semiárido brasileiro. *Revista AIDIS*, v. 13, n. 3, 2020. https://doi.org/10.22201/iingen.0718378xe.2020.13.3.68153

ROCKSTRÖM, J., STEFFEN, W., NOONE, K.; PERSSON, A.; CHAPIN III, F.S.; LAMBIN, E.F.; LENTON, T.M.; SCHEFFER, M.; FOLKE, C.; SCHELLNHUBER, H.J.; NYKVIST, B.; WIT, C.A.; HUGHES, T.; LEEUW, S.; RODHE, H.; SÖRLIN, S.; SNYDER, P.K.; COSTANZA, R.; SVEDIN, U.; FALKENMARK, M.; KARLBERG, L.; CORELL, R.W.; FABRY, V.J.; HANSEN, J.; WALKER, B.; LIVERMAN, D.; RICHARDSON, K.; CRUTZEN, P.; FOLEY, J.A. A safe operating space for humanity. *Nature*, v. 461, p. 472-475, 2009. https://doi.org/10.1038/461472a

SANTANA, L.M.; MORAES, M.E.B.; SILVA, D.M.L.; FERRAGUT, C. Spatial and temporal variation of phytoplankton in a tropical eutrophic river. *Brazilian Journal of Biology*, v. 76, n. 3, p. 600-610, 2016. https://doi.org/10.1590/1519-6984.18914

SILVA, T.G.; VINÇON-LEITE, B.; LEMAIRE, B.J.; PETRUCCI, G.; GIANI, A.; FIGUEIREDO, C.C.; NASCIMENTO, N.O. Impact of urban stormwater runoff on cyanobacteria dynamics in a tropical urban lake. *Water*, v. 11, n. 5, p. 946, 2019. https://doi.org/10.3390/w11050946

SINGH, U.; RIZWAN, M.; ALARAJ, M.; ALSAIDAN, I. A machine learning-based gradient boosting regression approach for wind power production forecasting: a step towards smart grid environments. *Energies*, v. 14, n. 16, p. 5196, 2021. https://doi.org/10.3390/en14165196

TANGIRALA, S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, v. 11, n. 2, p. 612-619, 2020. https://doi.org/10.14569/IJACSA.2020.0110277

TIMOFEEV, R.; HÄRFLE, W. *Classification and regression trees (CART):* theory and applications. Berlim: Center for Applied Statistics and Economics, 2004. 41 p.

WIEGAND, M.C.; NASCIMENTO, A.T.P.; COSTA, A.C.; LIMA NETO, I.E. Evaluation of limiting nutrient of algal production in reservoirs of the Brazilian semiarid. *Brazilian Journal of Environmental Sciences*, v. 55, n. 4, p. 456-478, 2020. https://doi.org/10.5327/Z2176-947820200681

WIEGAND, M.C.; NASCIMENTO, A.T.P.; COSTA, A.C.; LIMA NETO, I.E. Trophic state changes of semi-arid reservoirs as a function of the hydro-climatic variability. *Journal of Arid Environments*, v. 184, 104321, 2021. https://doi.org/10.1016/j.jaridenv.2020.104321

WILLMOTT, C.J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, v. 30, n. 1, p. 79-82, 2005.

WU, X.; DUAN, H.; BI, N.; YUAN, P.; WANG, A.; WANG, H. Interannual and seasonal variation of Chlorophyll-a off the Yellow River Mouth (1977-2012): dominance of river inputs and coastal dynamics. *Estuarine, Coastal and Shelf Science*, v. 183, parte B, p. 402-412, 2016. https://doi.org/10.1016/j.ecss.2016.08.038

XAVIER, T.M.V.; XAVIER, A.F.S.; DIAS, P.L.S.; DIAS, M.A.F.S. "Tempos de chuva": Avaliação da previsão para a quadra chuvosa nas regiões pluviometricamente homogêneas no estado do Ceará, em 1997, 1998 e 1999. *In:* SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS, 12., 1999, Belo Horizonte. *Anais*... 1999.

XU, J.; XU, Z.; KUANG, J.; LIN, C.; XIAO, L.; HUANG, X.; ZHANG, Y. An alternative to laboratory testing: random forest-based water quality prediction framework for inland and nearshore water bodies. *Water*, v. 13, n. 22, p. 3262, 2021. https://doi.org/10.3390/w13223262

ZANELLA, M.E. Considerações sobre o clima e os recursos hídricos do semiárido nordestino. *Caderno Prudentino de Geografia*, v. 1, n. 36, p. 126-142, 2014.