XXIII SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

AVALIAÇÃO DO ESTADO TRÓFICO DE CORPOS HÍDRICOS DO ESTADO DE SERGIPE ATRAVÉS DO ALGORITMO RANDOM FOREST E IET

Igor Santos Silva [1]; Carlos Alexandre Borges Garcia [2]; Euler Rodrigues de Sousa Faria [3]; José do Patrocínio Hora Alves[4]; Silvânio Silvério Lopes da Costa[5]; Adnívia Santos Costa Monteiro [6]; Helenice Leite Garcia[7]

RESUMO – A avaliação dos corpos hídricos que sofrem com atividades antrópicas é preocupante. Os impactos são diversos, e dentre eles a eutrofização é um fenômeno crítico que afeta a dinâmica natural do corpo hídrico. Neste sentido, ferramentas estatísticas e de aprendizado de máquina vêm se difundindo como formas de entender as relações entre os parâmetros de qualidade de água, e no caso da eutrofização, o parâmetro concentração de clorofila-a é um forte indicador desse problema. Sendo assim, este trabalho teve como objetivo predizer a concentração de clorofila-a por meio do algoritmo Random Forest e classificar os corpos hídricos no estado de Sergipe quanto à eutrofização através do Índice de Estado Trófico (IET). Observou-se uma forte contaminação advinda de atividades agrícolas e despejos domésticos nos corpos hídricos, o que acarretou no aumento da concentração de nutrientes como fósforo e nitrogênio, e consequentemente, da clorofila-a, conforme observado na correlação de Pearson e no Random Forest, que apresentou uma boa predição da mesma. Os corpos, então, foram classificados pelo IET com valores de clorofila-a predito e experimental, e observou-se que boa parte destes está em estado de eutrofização avançado, necessitando de maior estratégia de monitoramento e de conservação destes mananciais.

ABSTRACT – The evaluation of water bodies that suffer from anthropogenic activities that affect them is worrisome. The impacts are diverse and among them eutrophication is an undesired phenomenon that affects the natural dynamics of the water body. In this sense, the use of statistical and machine learning tools has been diffused as a way of understanding the relationships between water quality parameters, and in the case of eutrophication, chlorophyll-a concentration is a strong indicator of this problem. Therefore, the main goal of this paper was to predict chlorophyll-a concentration through the Random Forest algorithm seeking to classify the water bodies in the state of Sergipe regarding eutrophication through the Trophic State Index (TSI). So, it was observed a strong contamination coming from agricultural activities and domestic dumps in the water bodies, which resulted in an increase in the concentration of nutrients such as phosphorus and nitrogen, and consequently in chlorophyll-a, as observed in the Pearson correlation and Random Forest, which had

1) Mestre em Recursos Hídricos, Programa de Pós-Graduação em Recursos Hídricos, UFS, Avenida Marechal Rondon, s/n, Jardim Rosa Elze, São Cristóvão, SE, CEP: 49100-000, igorss@academico.ufs.br
2) Professor Doutor em Química, Programa de Pós-Graduação em Recursos Hídricos, UFS, Avenida Marechal Rondon, s/n, Jardim Rosa Elze, São Cristóvão, SE, CEP: 49100-000, cgarcia@ufs.br
3) Mestre em Engenharia de Computação e Automação Industrial, UNICAMP, Cidade Universitária Zeferino Vaz - Av. Albert Einstein, 400, Distrito Barão Geraldo, Campinas, SP, CEP: 13083-852, eulerodriguesousa@gmail.com
4) Professor Doutor em Química, Programa de Pós-Graduação em Recursos Hídricos, UFS, Avenida Marechal Rondon, s/n, Jardim Rosa Elze, São Cristóvão, SE, CEP: 49100-000, jphalves@uol.com.br
5) Professor Doutor em Química, Programa de Pós-Graduação em Engenharia e Ciências Ambientais, UFS, Avenida Marechal Rondon, s/n, Jardim Rosa Elze, São Cristóvão, SE, CEP: 49100-000, silvanioslc@gmail.com
6) Professora Doutora em Química, Programa de Pós-Graduação em Recursos Hídricos, UFS, Avenida Marechal Rondon, s/n, Jardim Rosa Elze, São Cristóvão, SE, CEP: 49100-000, adniviacosta@hotmail.com
7) Professora Doutora em Engenharia Química, Departamento de Engenharia Química, UFS, Avenida Marechal Rondon, s/n, Jardim Rosa Elze, São Cristóvão, SE, CEP: 49100-000, helenice@ufs.br

a good performance. The bodies were then classified by the TSI values obtained by predicted and experimental chlorophyll a, and it was observed that a good part of them are in state of advanced eutrophication, , necessitating a greater strategy of monitoring and conservation of these sources.

Palavras-Chave – Aprendizado de Máquinas; Eutrofização; Qualidade da Água.

1. INTRODUÇÃO

A escassez de água e a necessidade de uso adequado dos recursos hídricos visando uma sociedade sustentável, vem aumentando a demanda por estudos sobre a qualidade da água e os indesejados fenômenos ambientais atrelados as atividades antrópicas. Dentre esses fenômenos, destaca-se a eutrofização, gerada pelo aporte excessivo de nutrientes que tem como produto elevação da produção algal no ambiente aquático, trazendo cheiro e cor indesejados, bem como aumento do custo no tratamento da água para abastecimento, devido a necessidade de maiores concentrações de produtos de tratamento necessários para que os níveis de potabilidade sejam atingidos (Tundisi e Matsumara-Tundisi, 2011).

Nesse sentido, para a quantificação das variáveis e melhor interpretação das mesmas, estratégias estatísticas e computacionais têm surgido como ferramenta auxiliadora para o melhor entendimento dos fenômenos que ocorrem no corpo hídrico. Dentre os cálculos que estão associados a eutrofização, especificamente, tem-se o Índice de Estado Trófico (IET), proposto por Carlson (1977), e modificado por outros autores, como Lamparelli (2004). Este índice busca identificar o nível de eutrofização do corpo hídrico por meio de parâmetros que possuem oscilações significativas quando está ocorrendo esse tipo de fenômeno. Nitrogênio e fósforo são os nutrientes advindos de atividades antrópicas, como a agricultura, efluentes domésticos e industriais, ou de forma natural, Disco Secchi refere-se a transparência da água, e o produto desse fenômeno é observado pela variação da concentração de clorofila-a no corpo hídrico (Garcia et al., 2018).

Dentre os parâmetros ambientais, a clorofila-a como parâmetro de qualidade da água influenciado por outros fatores ambientais, possui um método analítico bastante demorado e custoso, em decorrência disso, sua determinação é restrita em algumas campanhas. Neste contexto, o uso de técnicas de aprendizado de máquinas que possuem capacidade de fazer o computador aprender a partir dos dados que lhe são apresentados, e assim, predizer variáveis diversas, vem sendo desenvolvida e aplicada em problemas ambientais (Hollister *et al.,* 2016; Li *et al,* 2018).

No contexto dos diversos algoritmos de aprendizagem de máquinas, uma bastante efetiva é a proposta por Breiman (2001), que é o *Random Forest,* baseado em um conjunto de árvores de decisões que isoladamente possuem baixa performance, e que juntas elevam a performance dos resultados significativamente. Essa ferramenta tem se mostrado eficiente para a predição ou classificação de parâmetros ambientais (Yuan e Pollard, 2014; Silva *et al.*, 2019). Além disso, o uso do *Random Forest*

pode ainda ampliar a observação de relações entre a variável a ser predita, bem como suas preditoras, devido a possibilidade de identificação do ranking de variáveis utilizadas como entrada (*input*) do modelo. Essas relações podem não ser tão relevantes ou podem estar eclipsadas quando são aplicadas correlações como a de Pearson ou Spearman nos dados (Ahmad et al., 2018).

Sendo assim, o uso de ferramentas computacionais de aprendizado de máquinas visa reduzir o tempo gasto em laboratório com determinadas análises, bem como os gastos em coletas e acelerar o entendimento dos analistas no que diz respeito aos fenômenos que ocorrem no corpo hídrico. Neste contexto, o presente trabalho buscou observar corpos hídricos no estado de Sergipe, nas diversas bacias existentes, calculando o IET com base nos dados mensurados nos anos 2013 e 2018, e comparando com os valores IET determinados a partir das concentrações de clorofila-a preditas pelo algoritmo *Random Forest*, classificando esses corpos hídricos.

2. METODOLOGIA

O presente trabalhou buscou avaliar parâmetros de qualidade da água de corpos hídricos no estado de Sergipe, nos anos de 2013 e 2018, visando predição da clorofila-a e classificação dos corpos hídricos. Alguns filtros para pré-processamento dos dados antes da aplicação do algoritmo *Random Forest*, bem como observação de correlações entre as variáveis, foram aplicados. Neste sentido, foram eliminados dos conjuntos de dados campanhas com dados faltantes ou que estivessem com registro incorreto, e por fim, buscou-se ainda a maior quantidade de coletas que haviam registros de concentração da clorofila-a com o maior número de variáveis possíveis, identificando dessa forma, as variáveis que serviram de entrada para a predição.

Os corpos hídricos avaliados são apresentados na Figura 1. Esses corpos hídricos estão em sua maioria em uma região semi-árida, que além da escassez de água sofre com a contaminação de atividades antrópicas diversas, dentro elas o mal-uso do solo que acaba promovendo assoreamento e lixiviação em direção aos corpos hídricos, potencializando o início da eutrofização.

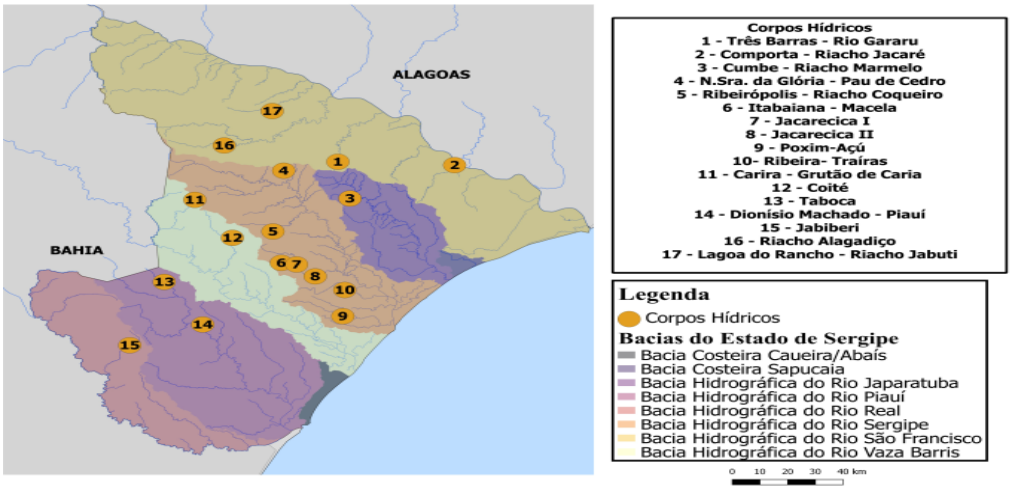

Figura 1 - Corpos hídricos de estudo

Nesse sentido, para enquadramento dos corpos hídricos foi realizado o cálculo do IET proposto por Lamparelli (2004), em que avalia-se a concentração de Fósforo Total (Equação 1) e de Clorofila-a (Equação 2), então, calcula-se a média dos índices (Equação 3) relativos a esses dois parâmetros para classificação do grau de eutrofização do corpo hídrico.

$$IET(PT) = 10x(6 - \left(1{,}77 - \frac{0{,}42x(ln\ PT)}{ln2}\right) \quad (1)$$

$$IET(Cl-a) = 10x(6 - \left(0{,}92 - \frac{0{,}34x(\ln Cl-a)}{ln2}\right) \quad (2)$$

$$IET = \frac{IET(PT)+IET(Cl)}{2} \quad (3)$$

Sendo: PT: concentração de fósforo total medida à superfície da água, em µg/L; Cl-a: concentração de clorofila a medida à superfície da água, em µg/L.

A classificação do corpo hídrico quanto a seu grau de eutrofização é apresentada na Tabela 1.

Tabela 1- Classificação dos corpos hídricos quanto ao nível trófico

Valor do IET	Classes de Estado Trófico	Características
= 47	Ultraoligotrófico	Corpos d'água limpos, de produtividade muito baixa e concentrações de nutrientes que não acarretam em prejuízos aos usos da água.
47<IET= 52	Oligotrófico	Corpos d'água limpos, de baixa produtividade, em que não ocorrem interferências indesejáveis sobre os usos da água, decorrentes da presença de nutrientes.
52 <IET= 59	Mesotrófico	Corpos d'água com produtividade intermediária, com possíveis implicações sobre a qualidade da água, mas em níveis aceitáveis, na maioria dos casos.
59<IET=63	Eutrófico	Corpos d'água com alta produtividade em relação às condições naturais, com redução da transparência, em geral afetados por atividades antrópicas, nos quais ocorrem alterações indesejáveis na qualidade da água decorrentes do aumento da concentração de nutrientes e interferências nos seus múltiplos usos.
63<IET=67	Supereutrófico	Corpos d'água com alta produtividade em relação às condições naturais, de baixa transparência, em geral afetados por atividades antrópicas, nos quais ocorrem com freqüência alterações indesejáveis na qualidade da água, como a ocorrência de episódios florações de algas, e interferências nos seus múltiplos usos
> 67	Hipereutrófico	Corpos d'água afetados significativamente pelas elevadas concentrações de matéria orgânica e nutrientes, com comprometimento acentuado nos seus usos, associado a episódios florações de algas ou mortandades de peixes, com consequências indesejáveis para seus múltiplos usos, inclusive sobre as atividades pecuárias nas regiões ribeirinhas.

Fonte: CETESB (2007); Lamparelli (2004).

O algoritmo Random Forest possui em seu funcionamento uma avaliação de análise de performance conhecida como Out-of-bag score (Oob score), em que 1/3 dos dados totais, não utilizados para treinamento são comparados com os utilizados para treinamento, e suas predições são comparadas ao fim da construção do modelo, foi aplicada (Ahmad et al., 2018) . Além disso, o mesmo possui a possibilidade de ajuste dos seus hiperparâmetros como número de arvores e nodos. Neste trabalho, observou-se o número de 300 árvores como de melhor desempenho do modelo.

3. RESULTADOS E DISCUSSÕES

3.1 Correlação de Pearson

Os parâmetros avaliados dos corpos hídricos do estado de Sergipe foram avaliados pela correlação de Pearson conforme Figura 2. Essa análise estatística apresentou correlações altas entre Sólidos Totais Dissolvidos (STD) e Turbidez indicando presença de sólidos durante os períodos analisados. Esses sólidos são lixiviados para o mesmo nas épocas de chuva, e advém, principalmente, das culturas agrícolas nas imediações dos mesmos, sendo que estes foram em sua maioria, construídos para atender essa demanda. O mal uso da terra, acaba tornando o solo mais suscetível a essa lixiviação necessitando, então, um aperfeiçoamento na forma em que estes são manejados.

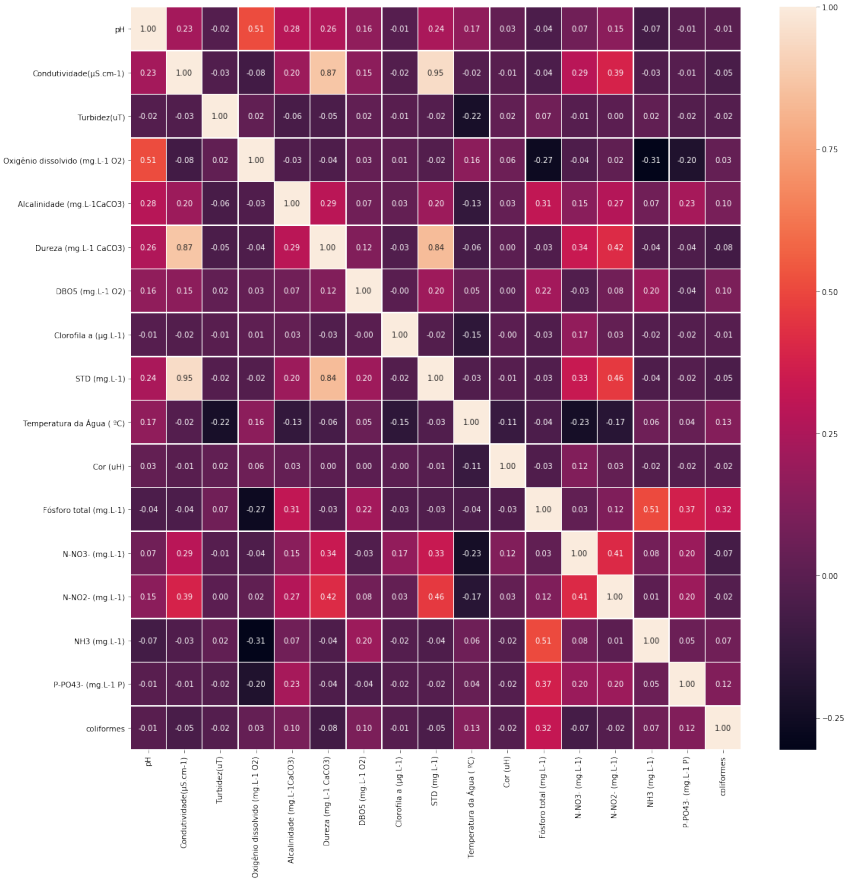

Figura 2 - Correlação de Pearson

Além disso, no contexto dos sólidos presentes na água, observa-se, ainda, uma forte correlação também entre dureza e condutividade. Essa correlação indica aumento da salinidade do corpo hídrico, o que aumenta a condutividade elétrica do mesmo, devido as atividades desenvolvidas nas proximidades, como a agricultura.

No caso da eutrofização, observa-se correlação significativa, entre Oxigênio Dissolvido e Amônia, a presença desta última acaba levando a depleção de oxigênio devido as reações nitrogenadas que se sucedem, o que acaba afetando a dinâmica natural dos corpos. Essa correlação observa-se principalmente em reservatórios em áreas urbanas suscetíveis a contaminação da ureia. Além disso, em relação ao nível trófico há uma correlação observada entre fósforo total e amônio, indicando contaminação de esgoto doméstico, servindo este de aporte de nutrientes aos corpos hídricos tornando este mais propício a eutrofização. A clorofila-a apresentou maior correlação com o nitrato, indicando este como grande contribuinte no aparecimento algal dos corpos hídricos (Lucas *et al*., 2014; Melo *et al*., 2015; Sena *et al*., 2015; Garcia *et al*., 2017; Santos *et al*., 2017)

3.2 Random Forest

O algoritmo Random Forest foi utilizado para a predição de clorofila-a e identificação das variáveis que mais contribuíram para a predição de clorofila-a. A Figura 3 apresenta o ranking de variáveis mais tiveram correlação durante a tentativa de englobar o maior número de variáveis preditoras que menos influenciassem na performance do modelo. Pode-se observar uma maior importância do Fósforo Total, pH e DBO para a predição de clorofila-a. A atividade algal influencia no pH do meio, alterando o mesmo, bem como a sua produção é influenciada pelo fósforo total que é um nutriente que contribui diretamente para o aparecimento de algas no corpo hídrico. Em relação a DBO, essa maior importância pode ser explicada pelos despejos de esgoto nos corpos hídricos do estado que acaba acarretando sua elevação. Esse mesmo despejos possuem fósforo que quando em grandes concentrações acarretam o aparecimento de algas e a clorofila-a é um indicador chave nesse fenômeno.

Outro ponto importante a se notar na aplicação do Random Forest são algumas relações da variável preditora com a predita que são mais preponderantes que na correlação de Pearson. DBO, pH e Fósforo Total não possuíam as maiores correlações na correlação como no Random Forest, o que mostra que algumas relações que não são tão percebidas na estatística de correlações podem aparecer de forma mais preponderante no *ensemble* de árvores de decisão (Yajima e Derot, 2017; Li *et al.*, 2018).

A performance do modelo foi avaliada pelo Oobscore que apresentou valor baixo, 0,8848, conferindo ao modelo uma boa predição de concentração de clorofila-a.

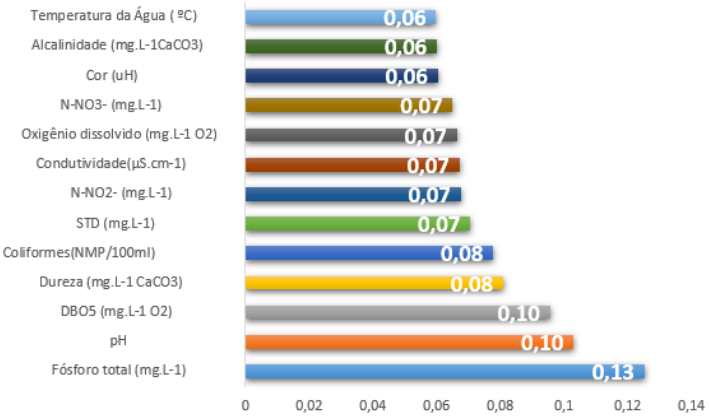

Figura 3 - Ranking de variáveis preditoras da clorofila-a pelo algoritmo Random Forest

3.2 IET

Valores preditos de clorofila-a foram utilizados para o cálculo do $IET_{predito}$ e comparado ao $IET_{experimental}$ para os corpos hídricos avaliados, conforme apresentado na Figura 4. Observa-se pelo gráfico que os dados gerados pelo modelo preditor consegue obter corpos hídricos na mesma faixa de classificação do IET utilizando dados experimentais. Corpos hídricos como os da Macela, Carira, Pau de Cedro e Dionísio Machado obtiveram em sua maioria resultados experimentais e preditos para o IET bem próximos um dos outros. Observa-se ainda que boa parte destes mananciais já estão sofrendo com a eutrofização de forma parcial e outros já bem avançados como é o caso do reservatório da Macela.

Nesse sentido, observa-se que a médio-longo prazo, boa parte dos corpos hídricos do estado de Sergipe estão suscetíveis a eutrofização de forma bastante avançada, classificados como Hipereutrófico, o que acende um grande alerta para que os órgãos gestores possam tomar decisões para a conservação dos mesmos e aumentar a fiscalização em relação as atividades desenvolvidas nas regiões próximas, além de um melhoramento na educação ambiental da população (Sena *et al*., 2015; Silva *et al*., 2016; Garcia *et al*., 2017; Santos *et al*., 2017).

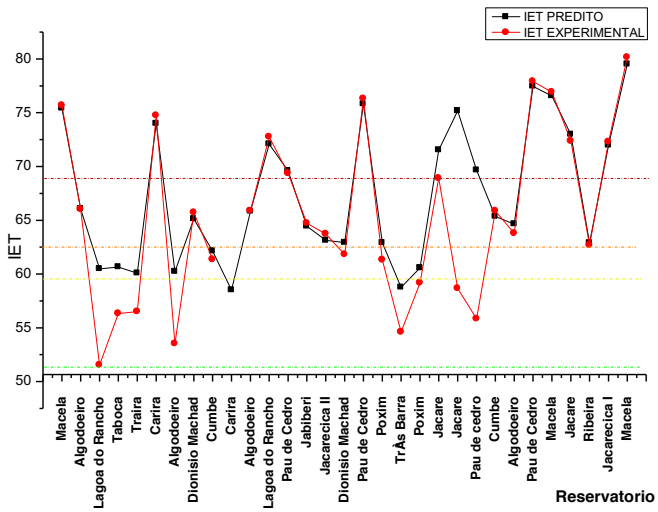

Figura 4 - Comparação de IET nos corpos hídricos em análise

4. CONCLUSÃO

A análise estatística e de aprendizado de máquinas possibilitaram meios de interpretação das relações entre os parâmetros de qualidade água neste trabalho. O entendimento do fenômeno de eutrofização e como este vem evoluindo é fundamental para a garantia de qualidade da água para toda população. Neste sentido, esse estudo possibilitou a identificação de corpos hídricos eutrofizados e em processo de eutrofização no estado.

A análise de correlação Pearson e o algoritmo *Random Forest* apontaram para contaminação advinda de atividades agrícolas nas imediações dos corpos hídricos, bem como despejos domésticos diretamente nos mesmos. Neste sentido, observou-se a complementariedade das duas análises e a eficiência de predição da concentração de clorofila-a do *Random Forest.*

Por fim, o estudo do IET com dados experimentais e preditos pelo *Random Forest* identificou corpos hídricos em estado de eutrofização avançado, classificados como Hipereutróficos, apresentado a necessidade de maiores políticas de conservação e proteção destes corpos hídricos para que estes corpos não se degradem ainda mais e se tornem irrecuperáveis.

AGRADECIMENTOS

Ao fomento da CAPES e da FAPITEC para o desenvolvimento deste trabalho (Edital CAPES/FAPITEC/SE N° 11/2016), por meio do PRORH, e a SEDURBS/SE pelos dados cedidos.

REFERÊNCIAS

AHMAD, M. W.; REYNOLDS, J.; REZGUI, Y. (2018) Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of Cleaner Production*, v. 203, p. 810-821.

BREIMAN, L. (2001) Random forests. *Machine learning*, v. 45, n. 1, p. 5-32.

CARLSON, R. E. (1977). A trophic state index for lakes. *Limnology and oceanography*, *22*(2), 361-369.

GARCIA, C.; GARCIA, H. L.; SILVA, I. S.; MENDONÇA, M. C. S. (2018) "*Evaluation of Water Quality Indices:Use, Evolution and Future Perspectives*", Intechopen, Environmental Monitoring and Assessment.

GARCIA, C.A.B; GARCIA, H.L.; MENDONÇA, M.C.S.; SILVA, A.F.; ALVES, J.P.H.; COSTA, S.S.L.; ARAÚJO, R.G.O.; SILVA, I.S. Assessment of Water Quality Using Principal Component Analysis: A Case Study of the Açude da Macela, Sergipe, Brazil.(2017) *Modern Environmental Science and Engineering*, V.3, No. 10, pp. 690-700.

HOLLISTER, J. W.; MILSTEAD, W. B.; KREAKIE, B. J. (2016) Modeling lake trophic state: a Random Forest approach. *Ecosphere*, v. 7, n. 3.

LAMPARELLI, M. C. (2004). *Graus de trofia em corpos d'água do estado de São Paulo: avaliação dos métodos de monitoramento* (Doctoral dissertation, Universidade de São Paulo).

LI, X.; SHA, J.; WANG, Z.L. Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake.(2018) *Environmental Science and Pollution Research,* p. 1-11.

LUCAS, A.A.T.; MOURA, A.S.A; NETTO, A.O.; FACCIOLI, G.G.; SOUSA, I.F. (2014) Qualidade da água no riacho Jacaré, Sergipe, Brasil usada para irrigação. *Revista Brasileira de Agricultura Irrigada* v.8, no2, p.98-105.

MELO, A. P. S.; GARCIA, H.L.; MENDONÇA, M.C.S.; BARRETO, V.L.; GARCIA, C.A.B. "*Qualidade da água dos reservatórios Algodoeiro e Glória através do índice de qualidade de água de reservatório*". (2015) In Anais do XXI Simpósio Brasileiro de Recursos Hídricos, Brasília, Brasil, 2015.

SANTOS, C.E.O; PEIXOTO, J.S.; ALVES, J.P.H. (2017) Geoquímica das águas do reservatório Poção da Ribeira, Agreste Central de Sergipe. *Scientia Plena*, v. 13, n. 10, 2017.

SENA, I. M. N.; MACEDO, L. C. B.; ALVES, J.P.H. "*Qualidade Da Água Do Reservatório Macela/Itabaiana-Sergipe 2004-2014*". (2015) In Anais do 2o Congresso Internacional - Resag, 2015.

SILVA, I.S.; GARCIA, C.A.B.; FARIA, E.R.S.; ALVES, J.P.H.; GARCIA, H.L. (2019) "*Aplicação do Algoritmo Random Forest na Avaliação de Corpos Hídricos no Estado de Sergipe*". In Anais do XII Encontro de Recursos Hídricos em Sergipe.

SILVA, I.S.; MENDONÇA, M.C.S.; GARCIA, C.A.B.; BARRETO, V.L.; GARCIA, H.L."*Predição Da Qualidade Da Água Do Reservatório Da Macela Através Da Lógica Fuzzy*" (2016). In anais do XIII Simpósio de Recursos Hídricos do Nordeste, Dez 2016, Aracaju-SE.

TUNDISI, J.G.; MATSUMURA-TUNDISI, T. (2011) *Recursos hídricos no século XXI*. Oficina de Textos.

YAJIMA, H.; DEROT, J. (2018) Application of the *Random Forest* model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. *Journal of Hydroinformatics,* v. 20, n. 1, p. 206-220.

YUAN, L. L.; POLLARD, A. I (2014). Classifying lakes to improve precision of nutrient–chlorophyll relationships. *Freshwater Science*, v. 33, n. 4, p. 1184-1194.