Aplicação do Algoritmo Random Forest na Avaliação de Corpos Hídricos no Estado de Sergipe

Igor Santos Silva[1], Carlos Alexandre Borges Garcia[2], Euler Rodrigues de Sousa Faria[3], José do Patrocínio Hora Alves[4] & Helenice Leite Garcia[5]

RESUMO: *O uso de técnicas estatísticas e de aprendizado de máquina tem se tornado ferramenta fundamental na tomada de decisão para avaliação dos problemas referentes a qualidade da água. Dentre esses problemas, a eutrofização tem sido referenciada como um dos maiores em corpos hídricos lênticos e lóticos. Para auxiliar o entendimento desse problema, prejudicial aos diversos usos da água, as correlações estatísticas e de aprendizagem de máquinas como o Random Forest surgem como facilitadora. Neste trabalho, foi utilizada a técnica de Random Forest com a clorofila-a como variável alvo de predição, já que a mesma é um grande indicador de eutrofização e está associada a custos significativos na análise laboratorial. Para tal, foram usados os dados de qualidade dos corpos hídricos do estado de Sergipe, obtidos entre os anos de 2013 e 2018. Os resultados da aplicação do Random Forest apontam para a contribuição da agricultura, efluentes domésticos e industriais como principais responsáveis para o fenômeno de eutrofização. Neste sentido, o modelo de predição usando o Random Forest apresentou MAE igual a 4,1, o que significa uma boa performance do modelo levando em consideração o uso de poucos dados. Sendo assim, o Random Forest pode ser considerado uma ferramenta de auxílio a tomada de decisão que pode contribuir para mitigação de eutrofização nestes mananciais.*

Palavras-chave: Aprendizado de Máquina; Clorofila-a; Qualidade da Água.

INTRODUÇÃO

Dentre os problemas relacionados aos corpos lênticos, destaca-se a eutrofização. Este fenômeno é caracterizado pelo aporte excessivo de nutrientes advindos de esgotos, efluentes industriais e, principalmente, da agricultura. Um dos parâmetros para identificação de alterações no corpo hídrico devido a este fenômeno, é a concentração de clorofila-a. No entanto, este parâmetro possui um oneroso custo laboratorial e é dispendido bastante tempo na análise do mesmo. Neste sentido, pesquisas têm sido realizadas buscando meios mais precisos e céleres de predição da mesma, visando auxiliar a tomada de decisão dos órgãos gestores (Zhang et al, 2015; Loucks e Van Beek, 2017; Garcia et al, 2018)

Inserida neste cenário, a necessidade de avaliação de qualidade de água, e de seus mais diversos parâmetros, é facilitada pelo uso de ferramentas computacionais e estatísticas como o algoritmo de aprendizagem de máquinas Random Forest. Essas ferramentas quando associadas em uma análise permitem uma ampla interpretação dos fenômenos que estejam ocorrendo em um corpo hídrico. O algoritmo Random Forest é utilizado tanto para regressão quanto para classificação, e é um *ensemble*, ou seja, uma associação de algoritmos que isoladamente são considerados fracos e que juntos conseguem obter uma performance significativa com bons resultados. Além disso, o Random Forest é baseado em árvores de decisão e possui uma

[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 (apresentador do trabalho);
[2] Professor Doutor, 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, 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] Professora Doutora, Departamento de Engenharia Química, UFS, Avenida Marechal Rondon, s/n, Jardim Rosa Elze, São Cristóvão, SE, CEP: 49100-000, helenice@ufs.br;

adaptabilidade a escassez de dados e consegue ser alimentado pelos mesmo sem a necessidade de qualquer tipo de normalização (Breiman, 2001; Hollister et al., 2016; Yajima e Derot, 2018).

Buscando, então, uma melhor avaliação dos corpos hídricos para os diversos usos e interesses, contribuindo para elaboração de planos de ações dos órgãos responsáveis, este algoritmo vem sendo aplicado para a predição de variáveis ambientais, bem como classificação de corpos hídricos.

Sendo assim, buscando auxiliar na diminuição de impactos no ecossistemas, ecológicos, sociais e na saúde pública em relação a problemas devido à eutrofização, este trabalho analisou dados ambientais de corpos hídricos no estado de Sergipe, aplicando correlações estatísticas e a aprendizagem de máquina Random Forest para a predição da clorofila-a e identificação dos parâmetros que mais se relacionam com a mesma.

MATERIAL E MÉTODOS

Os dados de qualidade da água analisados neste trabalho foram cedidos pelo Instituto Tecnológico e de Pesquisas do Estado de Sergipe (ITPS) e correspondem aos anos de 2013 e 2018. Estes dados são dos principais corpos hídricos de Sergipe abrangem diversas bacias e estão apresentados na Tabela 1. É importante ainda comentar que a avaliação dos dados contemplou tanto o período chuvoso quanto o período seco.

Tabela 1. Corpos Hídricos avaliados

Corpo Hídrico	Cidade	Bacia
Algodoeiro - Riacho Alagadiço	Nossa Senhora da Glória	Rio Vaza Barris
Lagoa do Rancho – Riacho Jabuti	Porto da Folha	Porto da Folha
Três Barras – Rio Gararu	Graccho Cardoso	Rio São Francisco
Comporta – Riacho Jacaré	Propriá	Rio São Francisco
Cumbe – Riacho Marmelo	Cumbe	Rio Japaratuba
Riacho Pau de Cedro	Nossa Senhora da Glória	Rio Vaza Barris
Riacho Coqueiro	Ribeirópolis	Rio Jacarecica
Riacho Macela	Itabaiana	Rio Sergipe
Jacarecica I	Itabaiana	Rio Jacarecica
Jacarecica II	Malhador	Rio Jacarecica
Poxim – Poxim Açú	São Cristóvão	Rio Poxim
Ribeira – Traíras	Campo do Brito	Rio Traíras
Grutão de Carira	Carira	Rio Sergipe
Coité	Frei Paulo	Rio Vaza Barris
Taboca	Simão Dias	Rio Vaza Barris
Donísio Machado	Lagarto	Rio Piauí
Jabiberi	Tobias Barreto	Rio Jabiberi

Visando uma melhor performance do algoritmo Random Forest, alguns filtros foram aplicados como retirada de células vazias dos parâmetros avaliados ou que estivessem com o registro com discrepância significativa, como por exemplo medidas de pH maior que 14. Dessa forma, evitou-se o sobreajuste do modelo com a retirada dos outliers. O software utilizado foi o Jupyter que utiliza linguagem Python em sua programação. A métrica utilizada para verificação de erros no algoritmo foi o valor médio do erro absoluto (MAE), conforme equação 1.

$$MAE = \frac{\sum_{i=1}^{n} \left| y_{i(observado)} - y_{i(predito)} \right|}{n} \quad (1)$$

RESULTADOS E DISCUSSÃO

Os parâmetros mensurados nestes corpos hídricos foram avaliados por meio da correlação de Pearson, conforme Figura 1. Esta correlação apresenta a análise da correlação especificamente da variável clorofila-a com os demais parâmetros ambientais. As demais correlações entre as variáveis avaliadas estão na Figura 2.

Na Figura 1 observa-se uma correlação maior positivamente da clorofila-a com nitrato, a cor

e a alcalinidade. A relação da clorofila-a com o nitrato advém da lixiviação de fertilizantes utilizados na agricultura que acabam sendo arrastados para estes corpos hídricos, principalmente no período chuvoso, fato este observado nos reservatórios da Macela e Jacarecica I, conforme citam Sena *et al,* (2015), Garcia *et al,* (2017) e Santos *et al,* (2017). A presença de algas devido o processo de eutrofização acaba interferindo na dinâmica natural do meio como o aumento da alcalinidade devido aos processos fotossintéticos que serão aumentados, além disso a presença de algas acaba alterando a cor natural do corpo hídrico.

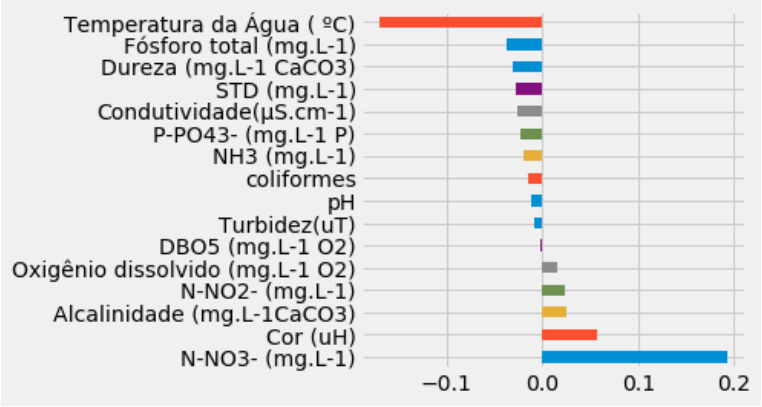

Figura 1. Correlação de Pearson da Clorofila-a e como os demais parâmetros avaliados.

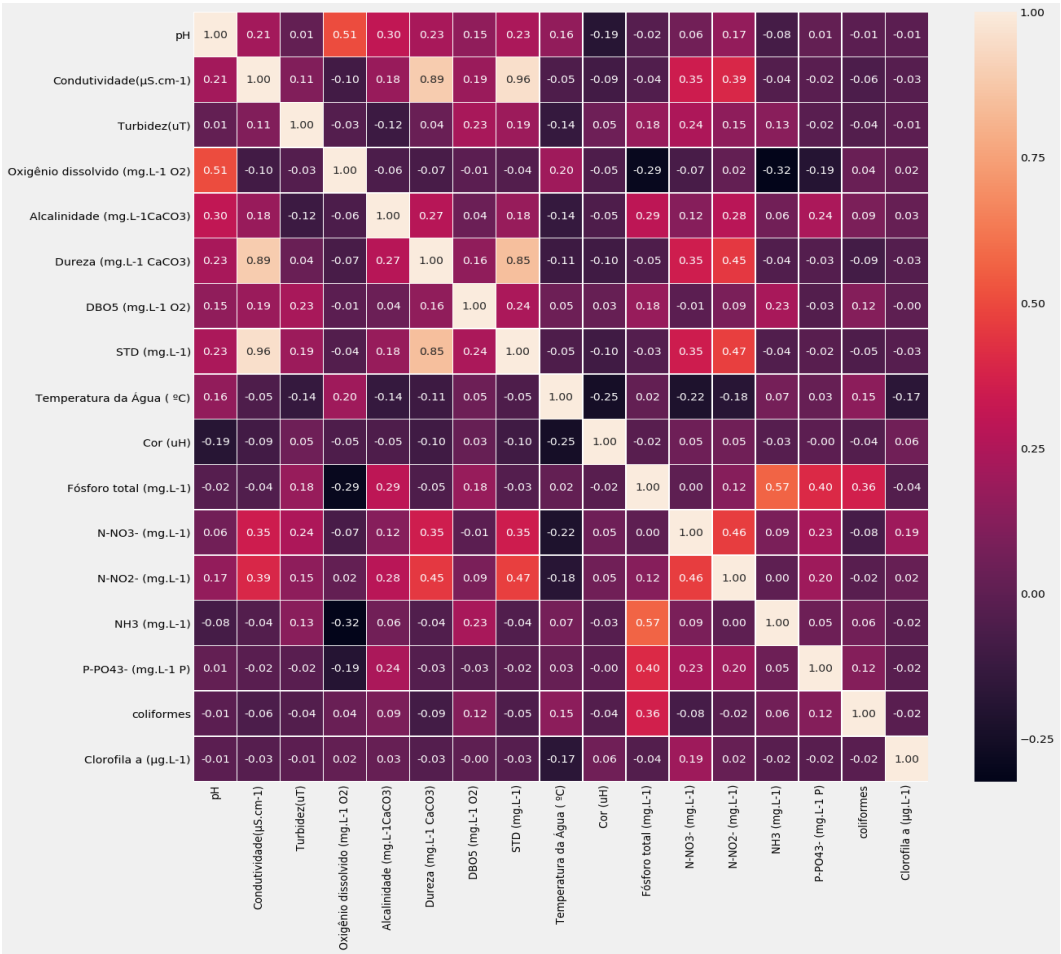

Figura 2. Correlação de Pearson entre as variáveis ambientais mensuradas

Segundo Yajima e Derot (2017), uma das grandes potencialidades do uso do Random Forest

juntamente com técnicas estatísticas tradicionais, como a correlação de Pearson, é a percepção de relações entre as variáveis que acabam não sendo apresentadas. Essa identificação permite entender e explicar melhor as relações de causa-efeito em corpos hídricos eutrofizados. Neste contexto, a Figura 3 apresenta o ranqueamento de importância das variáveis (*Feature Importance*) na predição de clorofila-a, identificando as que mais impactam nessa predição. Observa-se que o fósforo total e DBO5 tem maior correlação na predição da clorofila-a, o que indica que estes corpos hídricos possuem grande contaminação de esgoto e que além da eutrofização, os mesmos acabam sofrendo com outras formas de poluição.

O Random Forest obteve uma métrica de performance, MAE, de 4,1. Esse valor é menor, por exemplo, do que o obtido no trabalho de Li et al (2018) para predição de clorofila-a em um lago na China, que variou entre 7 e 11. Além disso, é importante comentar que nas simulações realizadas no trabalho de Li et al (2018), a quantidade de dados foi bem superior a utilizada neste trabalho, o que mostra a adaptabilidade do algoritmo Random Forest a situações de escassez de dados.

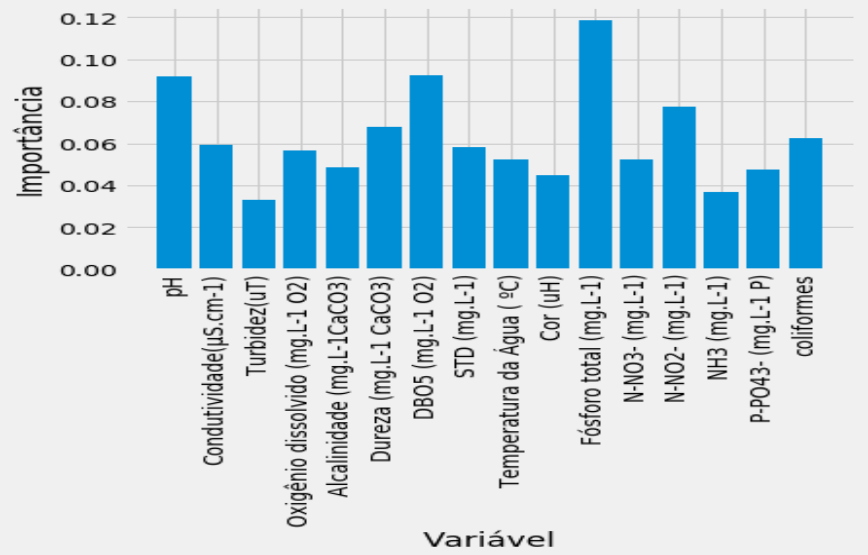

Figura 3. *Feature Importance* do algoritmo Random Forest

CONCLUSÕES

1. A eutrofização é um fenômeno recorrente nos corpos hídricos de Sergipe tendo os reservatórios da Macela e Jacarecica I considerados extremamente eutrofizados;
2. A correlação de Pearson apresentou a maior correlação da clorofila-a com o nitrato, apresentando a contribuição do uso do solo pela agricultura como maior fator de floramento algal, principal característica do fenômeno de eutrofização;
3. O Random Forest mostrou que os corpos hídricos sofrem descargas de fósforo por meio de esgoto e que tanto a clorofila-a quanto a DBO quando alteradas no corpo hídrico propiciam condições para o desenvolvimento algal.
4. É importante enfatizar a necessidade de uma alimentação com uma maior quantidade de dados para que a performance do algoritmo seja ainda melhor, e consequentemente, este possa ser usado como uma excelente ferramenta de auxílio na tomada de decisões para órgãos governamentais.

AGRADECIMENTOS

Ao fomento da CAPES e da FAPITEC para o desenvolvimento deste trabalho, por meio do PRORH, e ao ITPS pelos dados cedidos.

REFERÊNCIAS BIBLIOGRÁFICAS

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

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. Modern Environmental Science and Engineering, V.3, No. 10, pp. 690-700, 2017.

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

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

LOUCKS, D. P.; VAN BEEK, E. . Water resources systems planning and management: an introduction to methods, models and applications. Springer, 2017

SANTOS, C.E.O; PEIXOTO, J.S.; ALVES, J.P.H. 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 2004-201. 2º Congresso Internacional - Resag, 2015.

YAJIMA, H.; DEROT, J. 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, 2018.

ZHANG, Y.; HUANG, J. J.; CHEN, L.; QI, L. Eutrophication forecasting and management by artificial neural network: a case study at Yuqiao Reservoir in North China. Journal of Hydroinformatics, *17*(4), 679-695, 2015.

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