USO DE *MACHINE LEARNING* APLICADO A MODELAGEM DE DADOS AMBIENTAIS

PRATA, Kaio, F.C.[1]; SCHIMIDT, Fernando[1,*]

[1]Instituto Federal de Goiás, Câmpus Goiânia,* fernando.quimica@ifg.edu.br

A modelagem de dados consiste em estabelecer hipóteses sobre a estrutura ou o comportamento de um sistema físico e através dela procura-se explicar as propriedades do sistema e prever suas reações a estímulos. A modelagem da qualidade das águas superficiais traz como resultado um melhor conhecimento dos mecanismos e interações que justificam os variados comportamentos da qualidade das águas e constitui uma base racional para tomada de decisões no controle de qualidade das mesmas. A biblioteca de *machine learning* PyCaret desenvolvida em linguagem Python, é de fácil instalação e uso, permitindo a construção e gerenciamento de modelos diversos (de regressão de dados contínuos e também classificação de dados discretos) com poucas linhas de código e de rápida execução, não precisando de *hardware* avançado. Toda a instalação e execução/cálculos da biblioteca foi feita na plataforma Google Colab. Os modelos matemáticos multivariados foram desenvolvidos e aprimorados através da inserção de dados (através de planilhas), com separação dos conjuntos de calibração e treinamento (a partir de conjuntos de dados referenciais), visualização das diferenças entre configurações, cálculo de valores de saída dos modelos com determinação de métricas (erros de previsão e calibração). Foi utilizada a plataforma *Google Colab* com o teste de diversos modelos baseados em métodos referenciados. Neste trabalho utilizamos os dados de análise da recuperação da Bacia Hidrográfica do rio Doce, disponibilizados pelo Ministério Público mineiro, correspondentes à zona costeira e estuarina adjacentes, por meio da avaliação sistemática da qualidade da água e dos sedimentos, devido ao rompimento da barragem de Fundão, no município de Mariana-MG, em 5 de novembro de 2015. Como resultado de acordo com o ministério público, há uma rede de monitoramento com 56 pontos na bacia do rio Doce. Nos pontos de monitoramento foram medidos os seguintes parâmetros: (Água): oxigênio dissolvido, pH, turbidez, sólidos suspensos totais, ferro dissolvido, alumínio dissolvido, manganês total, arsênio total, cádmio total, chumbo total, Escherichia coli, níquel total, zinco total, cromo total e mercúrio total. (Sedimentos): ferro, alumínio, manganês, cádmio, chumbo e arsênio. Escolhemos o ponto RDO-016 em função das disponibilidades dos dados (diversos metais e parâmetros físico químicos de análise de água) com datas de análise química coincidentes, pois a coleta de amostras no rio Doce não necessariamente foram feitas nos mesmos dias. Utilizando os dados disponíveis no site de monitoramento do rio Doce, coletados pelo sensor da estação automática RDO-016, localizada na cidade de Linhares-ES, entre 2018 e 2023

(177 amostras no total), e também pelas análises das concentrações dos metais feitas em laboratórios, foi realizado o pré-processamento de dados e escolhas dos métodos de regressão no PyCaret. Foram selecionados dados para dias que apresentavam as medições completas a respeito das análises dos metais na água, para serem comparados com as medições de oxigênio dissolvido (previsão) como variável de saída dos modelos. Após os primeiros testes e problemas de super ajuste, obteve-se os coeficientes de correlação do melhor modelo de regressão *Extra Tree*: conjunto de treinamento 0,738 e conjunto teste 0,679. O trabalho terá prosseguimento para um melhor aperfeiçoamento das análises matemáticas.

Palavras-chave: big data; análise de água; modelagem; machine learning, parâmetros ambientais.

Agradecimentos: O presente trabalho foi realizado com apoio do Instituto Federal de Goiás (n° 19/2023). Prata, Kaio agradece ao CNPq pela bolsa concedida.