



Algumas aplicações de calibração multivariada com modelos de *Machine Learning* utilizando a biblioteca *PyCaret*

Fernando Schimidt (PQ)^{a*}, Marcela H. Miyabara (PG)^a, Alexandre Pancotti (PQ)^b, Mariângela F. Santiago (PQ)^c

^a Instituto Federal de Goiás, Mestrado prof. em Tecnologia, Gestão e Sustentabilidade, Goiânia -GO, Brazil, 74050-110

^b Universidade Federal de Jataí, Central analítica, Jataí-GO, Brazil, 75801-615

^c Universidade Federal de Goiás, Faculdade de Farmácia, Goiânia-GO, Brazil, 74605-170

*e-mail: schimidt99@gmail.com

Palavras Chave: *Machine Learning*, calibração multivariada, Python, *PyCaret*.

INTRODUÇÃO

Neste trabalho desenvolvemos modelos de calibração multivariada para três conjuntos de amostras diferentes (análises de cafeína em água, benzeno no ar atmosférico, e matéria orgânica em solos) utilizando uma biblioteca de aprendizado de máquina (*Machine Learning*) de código aberto em linguagem Python¹. Esta biblioteca 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.

RESULTADOS E DISCUSSÃO

Os resultados dos modelos de calibração multivariados obtidos através do uso da biblioteca estão mostrados nas tabelas 1 e 2 e consistem de: (1) análise de cafeína em água através da pré-concentração de 56 amostras em fibra vegetal e determinação por espectroscopia de fluorescência de raios X de energia dispersiva²; (2) análise de matéria orgânica em 108 amostras de solos por espectroscopia de infravermelho próximo³; (3) análise de benzeno em 820 amostras de ar atmosférico coletadas utilizando um dispositivo multissensor de gás⁴. Estes dois últimos foram obtidos de repositórios e o primeiro de uma dissertação de mestrado. Todos os três conjuntos de amostras (matrizes de dados) foram pré-tratados com Análise de Componentes Principais (PCA) e os *scores* obtidos foram utilizados como dados de entrada dos modelos. Todos os espectros foram previamente centrados na média e também tratados com segunda derivada. Os dados das amostras de ar foram auto escalonados previamente. A biblioteca *PyCaret* possui atualmente (versão 2.3.10) 25 métodos de regressão para cálculos de calibração multivariada. Toda a instalação e execução/cálculos da biblioteca foi feita na plataforma *Google Colab*. O conjunto de treinamento selecionado é testado com todos os modelos da biblioteca, sendo gerado uma pontuação usando validação cruzada k-fold para avaliação das métricas. A saída imprime uma grade de pontuação que mostra para cada modelo de calibração: erro médio absoluto (MAE), erros quadráticos médios (MSE e RMSE), coeficiente de correlação R2 (entre os dados reais e os

previstos), erro logarítmico quadrado médio (RMSLE), erro percentual absoluto médio (MAPE) nos subconjuntos (*fold*), e tempo de treinamento (em segundos). O modelo PLS apresentou piores resultados que os mostrados nas tabelas.

Tabela 1. Resultados obtidos usando biblioteca *PyCaret*.

Amostras	Melhor modelo	total amos.	Faixa	CPs	Var. total
Cafeína/água	Extra Trees Regressor	56	1 a 50 mg/L	6	98,5%
MO/solos	Bayesian Ridge	108	42 a 95%	10	89,7%
C ₆ H ₆ /ar	Bayesian Ridge	820	0,5 a 39 µg/m ³	9	99,6%

Tabela 2. Resultados obtidos usando biblioteca *PyCaret*.

Amostras	RMSE cal	R2 cal	RMSE pred	R2 pred
Cafeína/água	8,11	0,6084	3,26	0,9670
MO/solos	3,17	0,8502	3,44	0,8323
C ₆ H ₆ /ar	0,563	0,9921	0,562	0,9935

CONCLUSÕES

Os resultados resumidos mostrados nas tabelas 1 e 2 acima contém os melhores erros RMSE de calibração e previsão (conjunto teste) obtidos pelo melhor modelo encontrado para cada conjunto de amostras utilizado, selecionado na tabela geral de modelos da biblioteca. Os modelos permitem otimização de seus parâmetros de ajuste, armazenar as configurações ideais obtidas e fazer previsões com outros conjuntos de dados.

AGRADECIMENTOS

Empresa Google Inc, desenvolvedores da biblioteca *PyCaret*, website de hospedagem GitHub

[1] <https://pycaret.gitbook.io/docs/>

[2] Miyabara MH, Dissertação de mestrado, Mestrado profissional em Tecnologia de Processos Sustentáveis, Instituto Federal de Goiás - IFG, Goiânia-GO, 2021.

[3] Rinnan and Rinnan, *Soil Biology & Biochemistry*, **2007**, 39, 1664.

[4] De Vito S; Fattoruso G; Pardo M; Tortorella F; Di Francia G; *IEEE Sensors Journal*, **2012**, 12(11), 3215.