Geochemists Need Supervision: A Case for Expanding the Use of Supervised Machine Learning in Petroleum Geochemical Interpretation

Benjamin Kirkland[1] and Bao Nguyen[2]

[1]CNOOC Energy USA
[2]StudioX

Abstract

Petroleum geochemical data is well-suited to machine learning workflows—both unsupervised and supervised. Because there can be hundreds of compounds or ratios that may give unclear or even contradictory indications alone, the ability to integrate numerous variables in one analysis is critical for an accurate characterization of the petroleum. This is not a new finding of course: multivariate statistical analyses of geochemical data have been employed regularly for at least 35 years. Traditional multivariate methods have generally included either qualitative interpretation of visualized data (e.g., radar and cross plots) or unsupervised learning techniques such as clustering and principal component analysis. While these methods are very effective for exploring and classifying geochemical datasets, many geologic inferences are faster, easier, and more confident following a supervised analysis. One key reason is because supervised techniques such as regressions or neural networks can be trained and then used for new data without affecting the original model. A second key reason is that continuous target variables can be quantified more accurately with supervised models than with unsupervised. These supervised models are currently being used for production allocation and a handful of other exploration and production applications, but the vast majority of academic and industry geochemical interpretation still relies on traditional workflows. Supervised learning should be adopted and deployed for many more applications in geochemical analysis as it will increase credibility, accuracy, and efficiency of the entire discipline.