Semi-Supervised Machine Learning for Predicting Lacustrine Carbonate Facies in the Barra Velha Formation, Santos Basin

Pedro Vitor Abreu Affonso¹, Ana Luiza Spadano Albuquerque², and André Luiz Durante Spigolon³

¹Universidade Federal Fluminense, Geociências, Geoquímica Orgânica, Rio de Janeiro, Brazil (pvabreu7@gmail.com)
²Universidade Federal Fluminense, Geociências, Dinâmica dos Oceanos e da Terra, Brazil (ana_albuquerque@id.uff.br)
³Petrobras, Brazil (andrespigolon@petrobras.com.br)

There is an increasing availability of geoscientific exploration data for the oil and gas industry. Supporting data-driven tools have become important for the optimization and geoscientific information gain from this kind of data and thus allowing a fastest and more trustable decision making. Nonetheless, the development of this kind of technology depends on the standardization of the data and its descriptive methodologies, that many times diverges between the geoscientists and its many data sources, that recurrently comes from different scales of samples. The complexity of non-conventional reservoir, like the ones from brazilian pre-salt, increases those pre-existing difficulties. In this sense, this work evaluates the results of a semi-supervised Machine-learning methodology that was applied to the aptian carbonates of Barra Velha formation, from the Santos Basin pre-salt. This methodology follows a PU-learning approach with the utilization of the Random-forest algorithm based on public data from geological cores, side samples and geophysical data from the corresponding depths of the Barra Velha carbonates. A team of geoscientists provided a carbonate facies grouping, and this work regrouped it based on quantitative and qualitative descriptions, and in depositional criteria related for those samples, aiming to better utilize this data for Machine-learning. To deal with the fact that the samples belong for different scales and data-sources, the classified samples from geological cores were select as “labeled”, and the rest of it was defined as “unlabeled”, establishing a criteria for description of the samples and that fits the workflow for semi-supervisioned Machine-learning. Model evaluation metric were analyzed and compared to results of a regular supervisioned model approach. The results show that the overall precision of the semi-supervisioned model has increased significantly by 10% in relation to the supervisioned methodology, and critical suggestions were made based on the results for motivation of future researches from this topic.