12º CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS

TÍTULO DO TRABALHO:

Application of Artificial Neural Networks in the Prediction of Petroleum Fluid Properties

AUTORES:

Daniel Matheus Fernandes Rocha
Victor Rolando Ruiz Ahón
André Ferreira Young

INSTITUIÇÃO:

Universidade Federal Fluminense

APPLICATION OF ARTIFICIAL NEURAL NETWORKS IN THE PREDICTION OF PETROLEUM FLUID PROPERTIES

Resumo

A determinação precisa de propriedades de fluidos de petróleo, tais como viscosidade, pressão de bolha e compressibilidade, é uma tarefa muito importante para diversas aplicações na indústria de petróleo e gás. Com o avanço de novas ferramentas computacionais e de técnicas de aprendizado de máquina, Redes Neurais Artificiais surgem como alternativa a correlações matemáticas para a predição dessas propriedades, devido à sua capacidade de aprender com grandes volumes de dados. Neste sentido, o objetivo deste trabalho foi avaliar o desempenho de Redes Neurais Artificiais na predição de cinco propriedades de petróleo: pressão de bolha (P_b), razão de solubilidade (R_s), viscosidade de óleo subsaturado (μ_o), viscosidade de óleo morto (μ_{od}) e compressibilidade isotérmica (C_o). A base de dados utilizada neste trabalho consistiu em 195 amostras provenientes da Bacia do Mediterrâneo, África, Golfo Pérsico e Mar do Norte, obtidas da literatura. A metodologia consistiu na análise e preparação dos dados, construção e otimização dos hiperparâmetros de Redes Neurais Artificiais em Python, a fim de encontrar a melhor configuração a ser utilizada na predição das diferentes propriedades. A base de dados foi pré-processada para eliminar linhas de dados faltantes e *outliers*. Posteriormente, foi dividida em conjuntos de treinamento, validação e teste (70%-15%-15%), garantindo que cada conjunto de dados tivesse a mesma representatividade dos diferentes tipos de óleo que a base de dados original. Os hiperparâmetros avaliados foram o número de camadas ocultas, o número de neurônios nas camadas ocultas, a função de ativação, a taxa de aprendizado e o momento. Os resultados foram avaliados em termos de Erro Relativo Absoluto Médio (RAAE) e Coeficiente de Determinação (R^2), e uma análise comparativa foi realizada com correlações matemáticas ajustadas no trabalho de referência. Os resultados mostraram que a estratégia de otimização dos hiperparâmetros foi bem-sucedida na redução do erro de predição das propriedades consideradas. As Redes Neurais Artificiais otimizadas resultaram em valores de RAAE de 6.60%, 15.96%, 12.49%, 24.48% e 7.46% para os conjuntos de teste de P_b, R_s, μ_o, μ_{do} e C_o, respectivamente, superando os resultados obtidos com as correlações da literatura para quatro das cinco propriedades (a exceção foi viscosidade de óleo subsaturado, μ_o). Os resultados obtidos confirmam o potencial de Redes Neurais Artificiais na predição de ao menos algumas propriedades de fluidos de petróleo, mas devem ser discutidos juntamente com nível de complexidade e tempo computacional para se obter conclusões assertivas com relação à real aplicabilidade dessa estratégia no âmbito industrial.

Palavras-chave: propriedades de petróleo, aprendizado de máquina, redes neurais artificiais, hiperparâmetros.

Abstract

Accurate determination of petroleum fluid properties, such as viscosity, bubble point pressure and compressibility, is a very important task for several applications in the oil and gas industry. With the advancement of new computational tools and machine learning techniques, Artificial Neural Networks have emerged as an alternative to mathematical correlations for the prediction of these properties, due to their ability to learn from large amounts of data. In this regard, the objective of this work was to evaluate the performance of Artificial Neural Networks in predicting five petroleum fluid properties: bubble point pressure (P_b), solution gas-oil ratio (R_s), unsaturated oil viscosity (μ_o), dead-oil viscosity (μ_{do}) and isothermal compressibility (C_o). The database applied in this work consisted of 195 samples from the Mediterranean Basin, Africa, Persian Gulf and North Sea, obtained from the literature. The methodology was to analyze and prepare the data, and to construct and optimize the

hyperparameters of Artificial Neural Network models in Python, in order to find out the best model configuration to be applied in the prediction of the several properties. The database was pre-processed to eliminate missing-data lines and outliers. Subsequently, it was divided into training, validation and testing sets (70%-15%-15%), ensuring that each set had the same representability of the different types of oil as the original database. The network hyperparameters evaluated were the number of hidden layers, the number of neurons in the hidden layers, the activation functions, the learning rate and momentum. The results were evaluated in terms of Mean Relative Absolute Error (MRAE) and Coefficient of Determination (R^2), and a comparative analysis was carried out with the mathematical correlations adjusted in the reference work. The results showed that the hyperparameter optimization strategy was well succeeded in the reduction of the prediction error of the considered properties. The optimized Artificial Neural Network models resulted in MRAE values of 6.60%, 15.96%, 12.49%, 24.48% and 7.46% for the testing sets of P_b, R_s, μ_o, μ_{do} and C_o, respectively, surpassing the results obtained with the literature correlations for four of the five properties (the exception was unsaturated oil viscosity, μ_o). The results obtained confirm the potential of Artificial Neural Networks for predicting at least some petroleum fluids properties, but must be discussed along with complexity level and computational time for one to reach assertive conclusions about the real applicability of this strategy in the industry.

Keywords: petroleum properties, machine learning, artificial neural networks, hyperparameters.

Introduction

The properties of a petroleum fluid are preferably determined in laboratory analyses, but, for economic or operational reasons, the operationalization of experiments may be unfeasible (AL AMOUDI et al., 2019). For this reason, over the last 70 years, mathematical correlations have been proposed as an alternative to experimental methods for predicting these properties (BEGGS; ROBINSON, 1975; VASQUEZ; BEGGS, 1977; STANDING, 1981). They have the advantages of not requiring detailed information about composition and involving simple mathematical calculations, relating the properties of interest to readily available data (OLOSO et al., 2009).

One of the limitations of using correlations is due to the high degree of complexity in the composition between samples of petroleum fluids from different regions. Thus, a correlation that is developed to predict a property of a crude oil from a particular region generally fails to provide satisfactory performance when applied to oil samples from another region (OLOSO et al., 2009). Therefore, artificial intelligence techniques, such as the use of Artificial Neural Networks, emerge as a powerful tool for predicting properties with a greater degree of generalization. An Artificial Neural Network is an information processing system that has certain performance characteristics similar to a biological neural network (MOHAGHEGH, 2000). In the oil and gas industry, such algorithms assist in modelling processes with large amounts of data and can recognize highly complex patterns within an available data set (OSMAN; ABDEL-WAHHAB; AL-MARHOUN, 2001).

The present work aims to evaluate the use of Artificial Neural Networks in the prediction of five petroleum fluid properties: bubble point pressure, solution gas-oil ratio, unsaturated oil viscosity, dead-oil viscosity and isothermal compressibility - properties studied by de Ghetto, Paone and Villa (1994). The networks were evaluated with regard to their predictive power, training time and adjustment difficulty, in order to support discussions about their applicability in an industrial context.

Methodology

The first step of the methodology of this work consisted of data acquisition and preparation, also called pre-processing. This is done to avoid errors and inconsistencies in data processing. For example, if the original database had incorrect data, missing data or even data that is not well read by the codes, the algorithm could not work correctly. So, it is essential to carry out a detailed analysis and apply appropriate treatments to avoid these problems. Therefore, the data was obtained from the work of de Ghetto, Paone and Villa (1994) and consisted of a table with 195 lines and 20 columns. Each line represented a PVT (pressure-volume-temperature) report containing information on the properties of oils from the Mediterranean Basin, Africa, the Persian Gulf and the North Sea. Each column represented a property of the oils, totaling 19 properties for each oil. The data was entered into a Microsoft Excel® spreadsheet and this spreadsheet was imported into the *Jupyter Lab* programming environment. Table 1 presents a statistical summary of the data. In this table: T_r is the reservoir temperature; P_r is the reservoir pressure; R_s is the solution gas-oil ratio in the flash tank; R_{sp} is the solution gas-oil ratio in the separator; γ (av.) is the mean relative density of gases on the surface; T_{sp} is the temperature in the separator; P_{sp} is the pressure in the separator; P_b is the bubble point pressure; Y_i is the mole fraction of component i in the gases on the surface; γ (P_{sp}) is the mean relative density of gases at the separator pressure; B_{ob} is the oil formation volume factor at the bubble point; μ_{od} is the unsaturated oil viscosity; μ_{ol} is the saturated oil viscosity; μ_o is the dead-oil viscosity; and C_o is the isothermal compressibility.

Table 1 - Statistical summary of the database

Property	Valid data points	Mean value	Standard deviation	Minimum value	25%	50%	75%	Maximum value
°API	195	29.68	12.17	6.00	19.50	31.60	40.00	56.80
T_r (°F)	195	200.9	53.92	80.60	161.15	192.70	233.60	341.60
P_r (psia)	195	4185.	2384.79	242.22	2876.15	3726.08	5299.76	15304.62
R_s (scf/STB)	195	605.3	583.77	8.61	141.69	417.00	878.12	3298.66
R_{sp}	190	513.3	524.31	8.33	103.61	318.49	736.17	2985.87
R_{st} (scf/STB)	186	112.1	98.88	4.39	29.86	93.61	171.42	527.43
γ (av.)	187	1.05	0.23	0.62	0.88	1.00	1.23	1.79
T_{sp} (°F)	195	107.8	33.27	59.00	83.75	98.60	128.55	194.00
P_{sp} (psia)	195	250.2	184.01	14.50	85.57	227.71	440.92	868.79
P_b (psia)	195	2323.	1624.30	107.33	868.07	2233.62	3373.63	6613.82
Y_{co2} (%mol)	192	18.03	26.28	0.00	1.16	3.78	27.42	98.88
Y_{n2} (%mol)	192	3.03	5.49	0.02	0.71	1.51	2.91	63.32
Y_{h2s}(%mol)	192	0.26	0.87	0.00	0.00	0.00	0.00	5.65
γ (P_{sp})	192	0.90	0.24	0.61	0.71	0.79	1.08	1.53
B_{ob}	195	1.37	0.35	1.03	1.11	1.27	1.51	2.89
μ_{od} (cp)	195	50.20	148.05	0.50	1.40	2.40	35.20	1386.90
μ_{ol} (cp)	195	17.24	41.09	0.10	0.40	1.00	13.70	295.90
μ_o (cp)	195	25.40	61.56	0.10	0.40	1.10	16.20	354.60
C_o (1/psia)	195	1.17	7.66E-06	3.02E-06	5.85E-06	8.54E-06	1.53E-05	4.30E-05

After importing the data, the second step consisted of defining the input variables necessary to calculate the five studied properties. For this, we chose to maintain the same relationship between input and output variables as found in the correlations of de Ghetto, Paone and Villa (1994), which can be seen in Equations 1 – 5. In these equations, a_i are the parameters adjusted by the authors (they were not re-estimated in this work).

$$P_b = a_1 \left[\left(\frac{R_s}{\gamma_g} \right)^{a2} \cdot \frac{10^{a3.T}}{10^{a4.API}} \right] \quad (1)$$

$$R_s = \frac{\gamma_{gcorr} \cdot {P_b}^{a3}}{a_2} \cdot 10^{a4.10/(10+460)} \quad (2)$$

where $\gamma_{gcorr} = \gamma_{Psp} \cdot \left[1 + a_1 \cdot API \cdot T_{sp} \cdot Log\left(\frac{P_{sp}}{114.7} \right) \cdot 10^{-4} \right]$

$$C_o = \frac{-a_1 + a_2 \cdot R_s + a_3 \cdot T_R - a_4 \cdot \gamma_g + a_5 \cdot API}{P_R \cdot 10^5} \quad (3)$$

$$\mu_o = \mu_{ol} \left[\left(1 - \frac{p}{p_b} \right) \cdot \left(\frac{10^{-a_1} \cdot \mu_{od}^{a_2} \cdot p_b^{a_3}}{10^{a_4 \cdot API}} \right) \right] \quad (4)$$

$$loglog(\mu_{od} + 1) = a_1 - a_2 \cdot API - a_3 \cdot \log(T_R) \quad (5)$$

After that, data processing was carried out for missing data and outliers. Belyadi and Haghighat (2021) suggest two options for dealing with missing data: eliminating rows with missing values or replacing missing fields with values generated by basic mathematical functions. In this work, it was decided to eliminate all lines with missing values. Detecting and eliminating outliers in the data processing stage is recommended as it helps to reduce noise in the training data (GERON, 2019). In this work, outliers in the input and output properties were considered to be all those values that were below a lower limit and above an upper limit stipulated at a certain distance from the first and third quartiles indicated in Table 1.

Next, the data was divided into training, validation and testing sets. This followed the data distribution suggested by Al-Marhoun (2021), of 70% for training, 15% for validation and 15% for testing. A low number of extra-heavy oil samples was detected in comparison to the other samples, as shown in Table 2. So that the data in each set could be representative of the original set, a Python function was created to divide it randomly into the sets, but keeping the same proportion of light, medium, heavy and extra-heavy oils in each set. Additionally, several trials were made with different groups of training, validation and testing sets, as indicated in the following optimization sequence.

Table 2 - Proportion of oils by °API range in the database

Classification	Points	Proportion (%)
Light (°API >= 31.1)	98	53.3
Medium (22.3 to 31.1)	31	16.8
Heavy (10.0 to 22.3)	42	22.8
Extra-heavy (°API < 10.0)	13	7.1
Total	184	100

Finally, the Artificial Neural Network models were developed using the *Keras* library, according to the following procedure:

• Generate an initial model with initial hyperparameter settings;

- Initialize the parameters that will be used by the model;
- Train the model using the training set with normalized data to adjust the parameters and the validation set with normalized data for cross-validation;
- Use the trained model to make the prediction for the testing set;
- Calculate evaluation metrics with denormalized data;
- Optimize network hyperparameters;
- Select the best model based on the evaluation metrics.

After generating the initial networks for each property, the hyperparameters were adjusted according to the following sequence:

- Specification of the activation function;
- Specification of optimization algorithm, learning rate and momentum;
- Specification of the best training set (several trials);
- Specification of the number of neurons for one hidden layer;
- Specification of the number of hidden layers.

The intention of carrying out the hyperparameter optimization sequentially was to reduce the number of required training steps, since there are many model possibilities. The computer used for training had the following specifications: Intel Core™ i5-8265U processor, CPU 1.60 GHz - 1.80 GHz and 8 GB RAM.

Results and Discussion

Table 3 presents a summary of the optimized neural network configurations adjusted to calculate each property, as well as the results obtained for the training and testing sets. Figure 1 shows the format of each network, with the corresponding input data necessary to predict the properties.

Table 3 - Optimized neural networks' configurations and respective results

	P_b	R_s	C_o	μ_o	μ_{od}
Activation in hidden layer	tanh	relu	relu	relu	sigmoid
Activation in output layer	relu	relu	relu	linear	linear
Net configuration	(4, 8, 12, 4, 1)	(5, 16, 8, 4, 1)	(5, 10, 1)	(4, 4, 1)	(2, 16, 4, 1)
Learning rate	0,01	0,005	0,0075	0,0025	0,0075
Momentum	-	-	-	-	-
Training duration (min)	1.47	1.31	1.90	1.23	1.99
Prediction results					
MRAE - training set	0.0901	0.1557	0.0779	0.1474	0.3372
R^2 - training set	0.9804	0.9340	0.9322	0.9757	0.8609
MRAE - testing set	0.0659	0.1596	0.0746	0.1249	0.2448
R^2 - testing set	0.9896	0.9519	0.9597	0.9699	0.9386

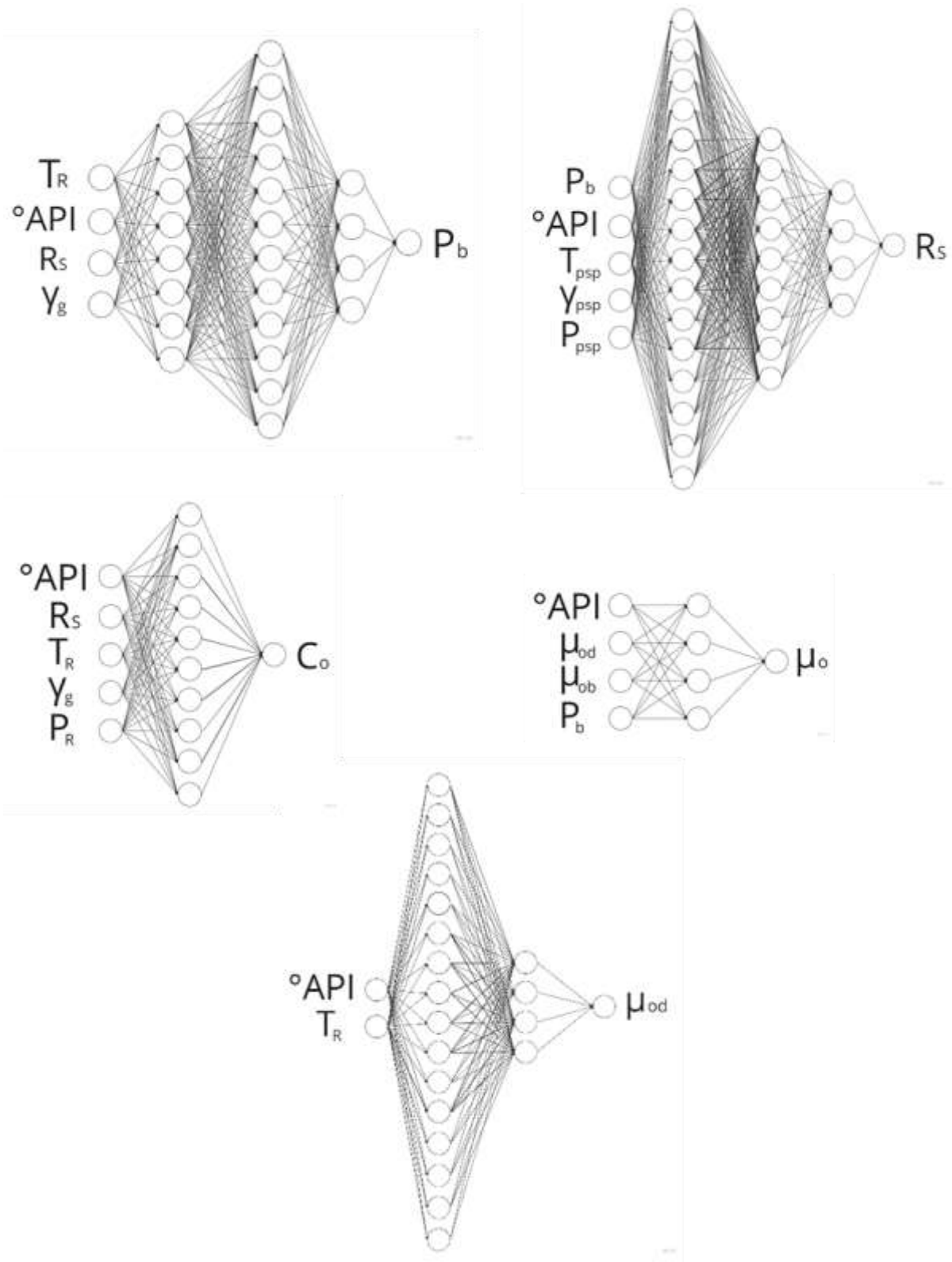

Figure 1 - Neural networks' formats

The results reported in Table 3 indicate that the model adjustment and optimization procedure was successful. The computational time required to train the networks, of less than 2 min in all cases, suggests - despite the complexity associated with this type of model, as can be seen in Figure 1 - the applicability of these models in the industrial context, where re-training may be indicated as new data is added to the initial database. In the figure, it is also possible to notice how the relationship between the output and input variables is more or less complex between the properties. For example, the relationship between unsaturated oil viscosity and ºAPI, μ_o, μ_{do} and P_b seems much less complex than the relationship between solution gas-oil ratio and P_b, ºAPI, T_{psp}, Y_{psp} and P_{psp}.

Table 4 presents a comparison between the results obtained by de Ghetto, Paone and Villa (1994) and the results obtained with the Artificial Neural Networks generated in this work. The three-layer models for bubble point pressure were successful in fitting the data, surpassing the MRAE of 12.80% found by de Ghetto et al (1994). For the solution gas-oil ratio, there was also an improvement from the MRAE obtained by the authors of 3 percentage points, demonstrating that the addition of more hidden layers served to improve the Artificial Neural Network models in adjusting the points for this property. For isothermal compressibility, Artificial Neural Networks were also very successful in adjusting the points for the testing and training sets, with MRAE values of 7.79% and 7.46%, surpassing the 17.90% obtained by de Ghetto, Paone and Villa (1994). For unsaturated oil viscosity, no model was able to approach the 6.40% MRAE found by the authors. Some hypotheses may be the failure to perform the outlier elimination procedure and also the low quantity and quality of the original data, with a very high number of outliers. And for dead-oil viscosity, MRAE was 24.48% for the testing set, lower than the error of 30.70% found by de Ghetto, Paone and Villa (1994), although it should be noted that, for the training set, MRAE was worse, indicating that the network may have adjusted points that were not representative of the total database.

Table 4 - Comparison between the results obtained in this work and literature results

	de Ghetto, Paone e Villa (1994)	Neural networks (training set)	Neural networks (testing set)
Bubble point pressure (P_b)	12.80	9.01	6.60
Solution gas-oil ratio (R_s)	18.30	15.57	15.96
Unsaturated oil viscosity (μ_o)	6.40	14.74	12.49
Dead-oil viscosity (μ_{od})	30.70	33.72	24.48
Isothermal compressibility (C_o)	17.90	7.79	7.46

Conclusion

Although the adjustment of Artificial Neural Networks has many possibilities and there is not exactly a best combination of hyperparameters for a specific network, as pointed out by Passos and Mishra (2022), the adoption of a strategy for optimizing hyperparameters in sequential steps, as done in this work, contributes to reducing the number of training steps carried out until the optimal configurations are reached. At the end of this work, it was possible to obtain networks that could be re-trained in acceptable times for practical application in an industrial environment.

The models optimized in this work surpassed the results obtained by literature mathematical correlations in four of the five properties: bubble point pressure, solution gas-oil ratio, isothermal

compressibility and dead-oil viscosity. On the other hand, the greater complexity associated with these models was not only unable to guarantee better results for all properties, but in some cases the improvement obtained was not so significant. Therefore, it must be considered whether the adoption of this type of model is really compensatory, given the difficulty of obtaining sufficient data and adjusting it.

References

AL-AMOUDI L. A.; GERI, B. S., PATIL S.; BAARIMAH, S. O. Development of Artificial Intelligence Models for Prediction of Crude Oil Viscosity. In: SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain: 2019.

AL-MARHOUN, M. A. Single Artificial Neural Network Model Predicts Bubble Point Physical Properties of Crude Oils. In: SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain: 2021.

BEGGS, H. D.; ROBINSON, J. R. Estimating the Viscosity of Crude Oil Systems. Journal of Petroleum Technology, v. 27, n. 9, p. 1140–1141, 1975.

BELYADI, H.; HAGHIGHAT, A. Machine Learning Guide for Oil and Gas Using Python. Gulf Professional Publishing, 2021.

DE GHETTO, G.; PAONE, F.; VILLA, M. Reliability Analysis on PVT Correlations. In: European Petroleum Conference, London, United Kingdom: 1994.

GÉRON, A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed. O'Reilly Media, 2019.

MOHAGHEGH, S. Virtual-Intelligence Applications in Petroleum Engineering: Part 1 - Artificial Neural Networks. Journal of Petroleum Technology, West Virginia, United States, v. 52, n. 9, p. 64–73, 2000.

OLOSO, M. A.; KHOUKHI, A.; ABDULRAHEEM, A.; ELSHAFEI, M. Prediction of Crude Oil Viscosity and Gas/Oil Ratio Curves Using Recent Advances to Neural Networks. In: SPE/EAGE Reservoir Characterization and Simulation Conference, Abu Dhabi, United Arab Emirates: 2009.

OSMAN, E.; ABDEL-WAHHAB, O.; AL-MARHOUN, M. Prediction of Oil PVT Properties Using Neural Networks. In: SPE Middle East Oil Show, Manama, Bahrain: 2001.

PASSOS, D.; MISHRA, P. A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. Chemometrics and Intelligent Laboratory Systems, v. 223, 2022.

STANDING, M. B. Volumetric and Phase Behavior of Oil. 9th ed. Dallas, United States: Millet the Printer, 1981.

VAZQUEZ, M.; BEGGS, H. D. Correlations for Fluid Physical Property Prediction. Journal of Petroleum Technology, v. 32, n. 6, p. 968–970, 1980.