Bulletin of Earth Sciences of Thailand

Pay Zone Determination by Applying Automatic Machine Learning (AutoML) in the Mckee Field, Taranaki Basin, New Zealand

Natalia De Castro Arce[1*], Pongthep Thongsang[1]

[1] M.Sc. Program in Petroleum Geoscience, Faculty of Science, Chulalongkorn University, Bangkok, Thailand
*Corresponding author e-mail: natidecastroa@gmail.com

Received: 15 Jun 2021
Revised : 9 Jul 2021
Accepted: 11 Jul 2021

Abstract

Reservoir characterization is a critical objective in order to understand the subsurface geology and develop geological models that support sustainable and economic oil and gas exploitation. By measuring and analyzing electrical responses of the penetrated successions, it is possible to infer properties about rock matrix and fluid content, among others, that are indirectly related to petrophysical properties of the rocks. Machine Learning (ML) is a data science tool that fits models that are then used to identify patterns or similarities among observations in a dataset in order to make predictions about unobserved data. The main objective of this study is to build a supervised classification model applying AutoML (Automatic Machine Learning) to identify pay zones using information derived from well logs. Twenty wells from the Mckee field, Taranaki Basin, New Zealand were selected containing a basic set of electric logs. For the response variable, reservoir pay properties were estimated. The classification model was first trained with a proportion of the observed data for the independent and response variables and then tested against the remained data set not observed by the model. Model definition and implementation were done with the AutoML function from H2O open source, and model performance was assessed by means of the confusion matrix method. The results indicate that XGBoost (Extreme Gradient Boosting) is the best model for the classification of the Mckee field and that the correct classification of the pay properties is high in all tested wells. This suggests that AutoML procedures can be a valuable tool for the exploration and assessment of the geological properties of the area, helping to reduce operating costs by optimizing the decision-making time during the well evaluation phase as well as the initial petrophysical evaluation to understand reservoir characteristics.

Keywords: Automatic Machine Learning, Supervised Classification, Petrophysical Evaluation, Well Logs, Reservoir Characterization.

1. Introduction

Reservoir characterization is a crucial step during the exploration and development projects in oil and gas companies. One of the main objectives of characterizing a reservoir is to understand the subsurface geology to develop geological models that serve as support for sustainable and economic oil and gas exploitation.

Frequently, geoscientists deal with a limited amount of available information to develop geological models that allow, with tolerable uncertainty, to identify areas of interest. This shortage of information often results in a constant search for alternatives to evaluate a given formation.

Proper estimation of petrophysical variables such as porosity, clay volume, and water saturation requires that various reservoir properties must be assessed and understood. Accurate characterization of the reservoir is also fundamental to other petrophysical analyses and further exploration and hydrocarbon production (Abbey et al., 2018).

Reservoir characterization can be conducted by either direct or indirect methods. Direct methods often comprise the examination of core samples from the interval of interest by an experienced geologist. However, these methods have some associated difficulties.

On the one hand, the extraction of core samples for analysis is limited and often represents a substantial increase in the overall cost and time spent in the exploration phase (Vasini et al., 2018). On the other hand, different qualified specialists examining the core data can provide contrasting interpretations rendering the examination partially reliable. Indirect methods include the use of well logs data to derivate the geophysical properties of the formation, which are the basis of the data available for the characterization.

However, these methods underperform when compared with the former ones (Thomas et al., 1995). These differences in performance augment as the number of logs to evaluate increases (Vasani et al., 2018). Thus, the use of computational technologies to assist the specialist in characterizing the reservoir can improve the efficiency and overall accuracy of the process (Yang et al., 2017).

Recently, automated log analyses have gained increased popularity in the oil and gas industry (Bangert, 2021; Zhong et al., 2020), offering a series of advantages over traditional methods like a faster initial assessment of reservoir properties (Otchere et al., 2021; Ippolito et al., 2021; Gu et al., 2021) which, in turn, translate in reduction of operational costs and optimization of decision-making time during the well evaluation phase.

Furthermore, novel developments in automated analyses have been proved fundamental to the process as, generally, models are developed manually, a task that is both time-consuming and prone to errors (Saporetti et al., 2020).

Automated Machine Learning (AutoML) is the process of defining, building, and training a suite of models through a series of techniques to automatically select the best model (or set of models) for the data. The process is based on training scores, applying the selected model to perform regression, classification, or clustering on the newly observed data. In the context of petroleum geosciences, the main application of AutoML is lithologies identification. For it, different families of models have been employed aiming for more efficient, fast, and unbiased search for the proper variable settings. It allows to performing reservoir and rock characterization as best and consistent as possible (Min et al., 2020; Zeng et al., 2020).

Results on applying machine learning methods in reservoir characterization have resulted in a variety of models that best fit the data. For instance, Li and Anderson-Sprecher (2006) compared the performance of Discriminant Analysis (DA) against Naïve Bayes Classifier (NBC) and concluded that the latter is more suitable for facies identification. Similarly, Xie et al., (2018) presented a comparison of a variety of Machine Learning methods for lithology identification. They concluded that, of the models tested, Gradient Tree Boosting (GTB) and Random Forest Classifier (RFC) presented the lower error values for the classification, suggesting that GTB is the more suitable option.

Finally, Imamverdiyev and Sukhostat, (2019) employed a Convolutional Neural Network (CNN) approach for the classification of facies based on conventional well logs data and compared the result with Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) outcome among others. Their results indicated that CNN performed better than the other models proposing it as a good candidate model for facies identification.

Under this premise, machine learning algorithms present themselves as a robust alternative to traditional methods for faster reservoir characterization. These algorithms take some initial training data (variables) as input to classify and predict some response variable's behavior. Defined series of models are subsequently refined by a recursive method (the defined model), and its associated parameters (Xie et al., 2018).

In the context of a classification approach, the model is first trained with a proportion of the observed data for the independent and response variables. Then, the model is tested against the unseen remaining data set. Based on the measurement of performance and cross-validation, the model parameters are adjusted to maximize execution and minimize the number of incorrect classifications to the desired level of accuracy and precision. Subsequently, the model can be applied to new data to obtain an initial rapid assessment of the reservoir properties.

This study aims to evaluate the applicability of machine learning algorithms to accurately predict the reservoir's pay properties, improving efficiency and objectivity during the reservoir characterization phase.

2. Geological Setting

This study was conducted for the oil and gas onshore Mckee field, located in the Taranaki Basin (Fig. 1). The Taranaki Basin is a Cretaceous and Tertiary sedimentary basin situated along the western side of the North Island, New Zealand. The roughly north-south trending Taranaki fault defines the eastern margin of the basin.

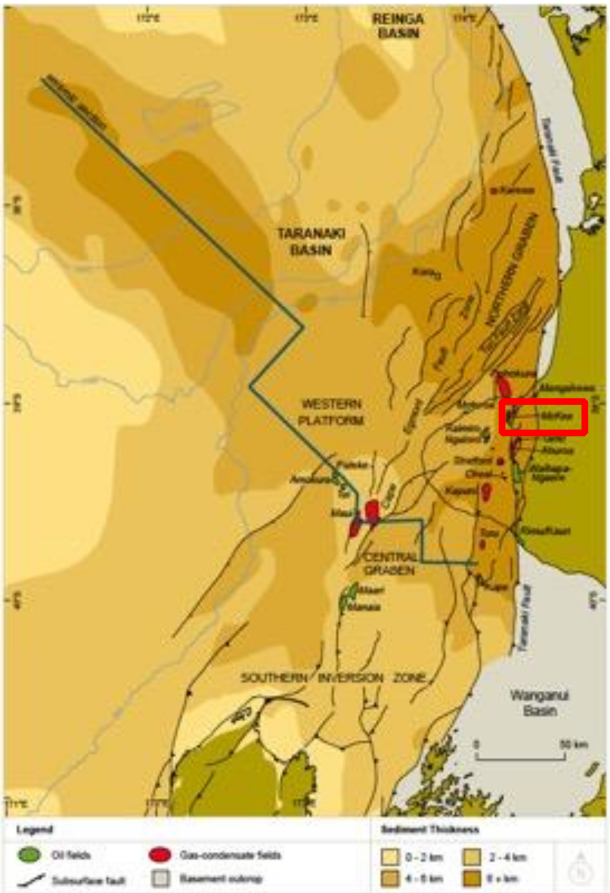

Figure 1. Mckee field location. Modified after Ministry of Business, Innovation, and Employment, New Zealand. 2014.

The basin extends westwards, underlying the onshore Taranaki Peninsula and continuing offshore beyond the continental shelf edge (Ministry of Business, Innovation, and Employment, New Zealand. 2014). The McKee Field is developed in a thrust-faulted anticline within the Tarata Thrust Zone. As a result of regional compression in the Early to Mid-Miocene, the main structure of the Mckee field evolved westward overthrusting on a low-angle fault (Dong et al., 2018).

The generalized stratigraphy of the Taranaki Basin (Fig. 2) consists of the Pakawau Group, Kapuni Group, Moa Group, Ngatoro Group, Wai-iti Group, and Rotokare Group from the Late Cretaceous to the Pleistocene (King & Thrasher, 1996).

In this study, only the stratigraphic groups comprised by the interval logged in most of the wells selected for the evaluation, and subsequent identification of pay zones, were considered. Therefore, they will be summarized in the petrophysical evaluation section.

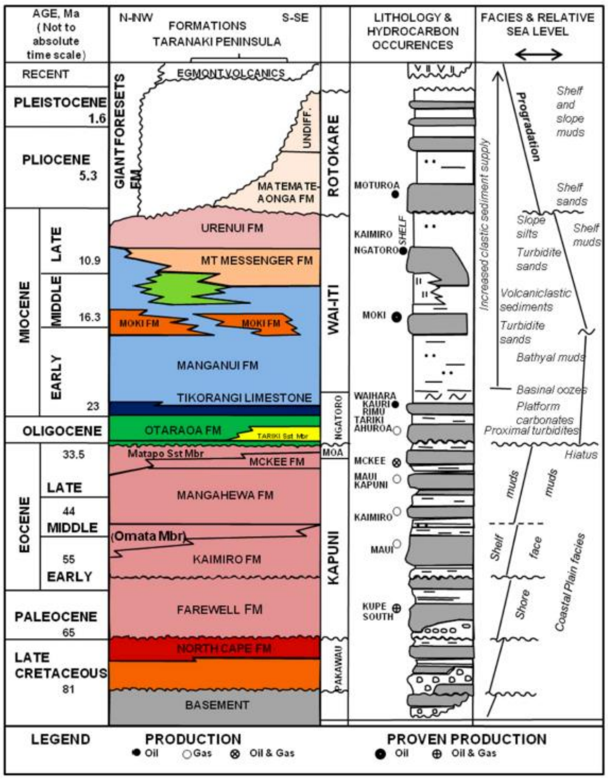

Figure 2. The generalized stratigraphy of the Taranaki Basin. Image taken from Dong et al., 2018

3. Methodology

Pay properties from petrophysical evaluation and associated well log variables from the McKee field, Taranaki Basin, New Zealand were used as the input data for the AutoML classification model. This model was then used to predict the pay zones based on the training dataset. Performance evaluation was conducted in parallel with the testing set metrics analysis.

To determine the pay reservoir properties to be used as the predicted variable, the electric logs of the 20 selected wells were analyzed by conventional log analysis.

The petrophysical evaluation was conducted for the interval between the formations Tikorangi and Mangahewa, whose general lithology characteristics are:

- Tikorangi: Sandy to silty, occasionally glauconitic limestone.
- Otaraoa: Transition from limestone to clastics- siltstone, claystone, sandstone and trace limestone.
- Turi: First appearance of non-calcareous siltstone and sandstone.
- Mckee: Sandstone, siltstone and trace coal.
- Mangahewa: Siltstone, sandstone coal, trace claystone and sandstone.

First, the shale volume (VSH) was estimated using the arithmetic mean method from the linear model, with the gamma ray (GR) and the combined use of the density-neutron (DEN-NEU) cross plot. Then, total porosity (PHIA) was calculated from the DEN-NEU curves and corrected for the shale volume content to obtain the effective porosity (PHIE). Next, the water saturation estimation was performed using the Indonesia model. Finally, all petrophysical results were calibrated with available core data and production information.

Identification of the pay properties, such as, gross rock volume (GRV), net pore volume (NPV), and hydrocarbon pore volume (HCPV) during the petrophysical evaluation, is a fundamental process for reservoir characterization, and to estimate in-place volumetrics. These properties are calculated based on specific cut-off values to each study area and operating company. However, for this study default cut-offs were used as shown in Table 1.

Table 1. Cut-off values used for the field under study.

VSH	$<50\%$
PHIE	$>10\%$
SW	$<50\%$

Pay properties parameters result in a significant impact on subsequent decisions in a field development project. For this reason, this task requires experience, many hours of work, and not only well logs but other sources of information, for instance, core data, lithological description, and production data, among others.

4. AutoML Implementation

Twenty wells were selected containing a basic set of electric logs. In addition, a proxy variable was estimated containing the target reservoir properties (classes), such as GRV, NPV, and HCPV. All analyses were conducted in Phyton 3.8.5 version. Model definition and implementation were done with the H2O function: AutoML. The model training phase was made using 80% of the total data matrix, and the initial parameters default values. Model performance was assessed by the mean per class error and other cross-validation metrics. Classification results were then visualized and interpreted in the form of a confusion matrix and log plot.

The data matrix used for this study comprises five log measurements (features): gamma-ray (GR), deep resistivity (RESD), bulk density (DEN), neutron (NEU), compressional slowness (DT), and the target variable containing the reservoir properties to classify: GRV, NPV, HCPV, and non-reservoir (SHL).

Model training was performed by dropping the well of interest from the dataset to avoid the model being exposed to the test data. The model parameters were as follow: The model searching time was set to 5 minutes, the maximum number of models to be evaluated by AutoML was set to 25, and the number of cross-validation events used to validate the trained data set was set to 10. The best model was selected following the AutoML evaluation metrics, and in all cases, the stacked ensemble models were omitted. Three wells were randomly selected for the testing phase, dropped from the training set, and used as testing datasets independently.

5. Results and Discussion

Working with a data matrix based on well logs implies a high possibility of missing values due to several operating acquisitions, and data processing issues. Handling missing values represents an essential task in the data preparation process in machine learning because many algorithms do not support missing values. For this reason, several ways to manage missing values have been reported, such as those described by Elhassan et al., (2021). However, in the present study, the missing values were replaced by (-999), an alternative that AutoML can identify and process. Figure 3 presents the distribution of missing values per feature within the study data matrix.

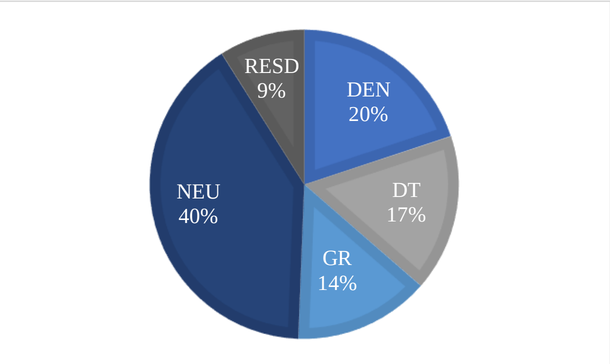

Figure 3. Missing values per feature contained in the complete data set.

Another critical factor considered in this study was the data set imbalance. This condition has been extensively studied, searching for an optimum technique to handle the class imbalance classification problems (Japkowicz, 2000; Li et al., 2006; Vasini, 2018). In brief, it refers to the fact that the classes are not represented equally. This condition causes a direct impact on the selection of the metric to assess classification results. For example, Figure 4 shows a reduced number of data for the GRV class, in contrast with the amount of data collected for the other classes (NPV, HCPV, and SHL).

Results were analyzed considering three performance factors: the model, the classification, and the prediction.

According to the data set and the defined task (multi-class classification), the best AutoML model was selected using the mean per class error metric. It represented the error average of each class and was interpreted as the ability of the algorithm to correctly identify any data point as belonging to one class or another. Under the described parameters, the extreme gradient boosting (XGBoost) was identified as the best algorithm model (Fig. 5).

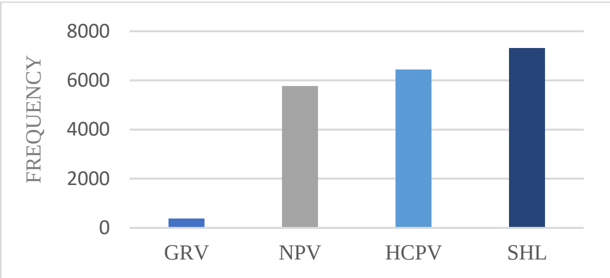

Figure 4. Data distribution per class in the complete data set.

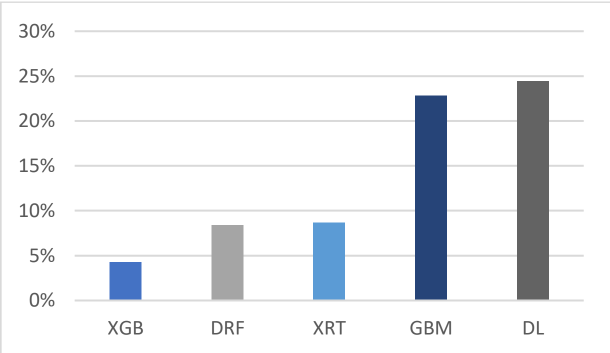

Figure 5. Average mean per class error from the selected wells under the study area.

The unusual high values observed for the accuracy suggested that class imbalance classification gave inexact and misleading information about the classifier performance, as shown by Aida et al., (2013). Therefore, classification performance was analyzed based on the confusion matrix outcomes obtained from the test set, as in Figures 6 and 7 In this approach, classifier performance can be assessed based on the element's distribution per class within the data matrix. In a confusion matrix, the values showed on the diagonal represent correct predictions, whereas off-diagonal represent incorrect predictions.

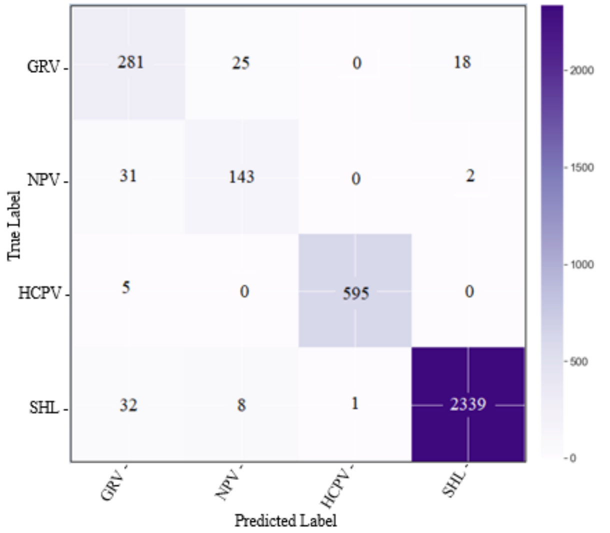

Figure 6. Confusion matrix of the test data representing true versus predicted pay properties for the well Mckee-10.

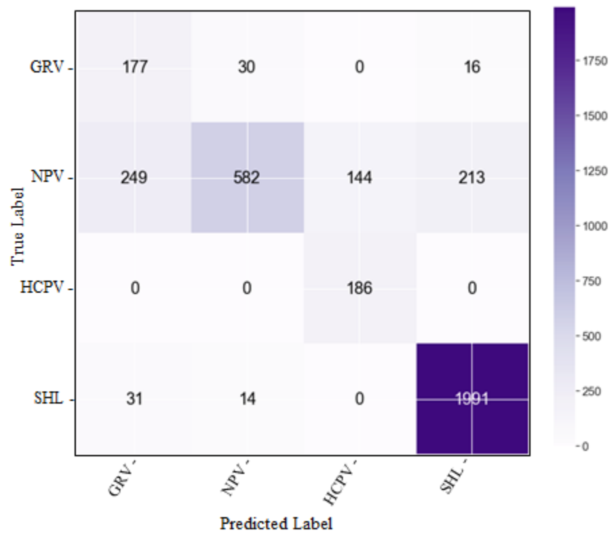

Figure 7. Confusion matrix of the test data representing true versus predicted pay properties for the well Mckee-2A

The classification performance was satisfactory, considering that the error rate was under 20% in all test sets (Fig. 8). Furthermore, it is relevant that the most important class, the HCPV, related directly to the possible presence of hydrocarbons was classified with considerable high precision in all cases.

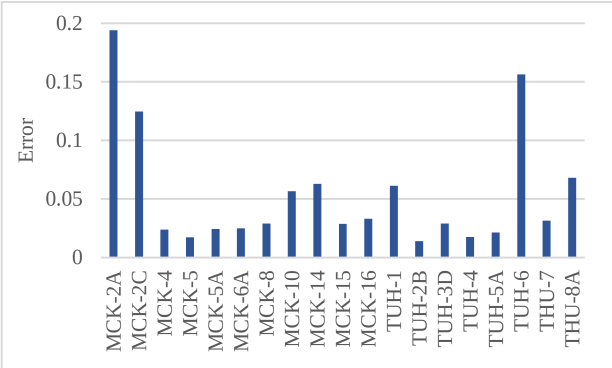

Figure 8. Graphic shown the classification error per well.

Prediction results are illustrated in Figures 9 and 10 The log plot in depth shows the true vs. predicted pay properties allowing the identification of incorrect predictions intervals and the possible causes for the misclassified observations.

The Mckee-10, for example, presents an outstanding relationship between the true pay properties compared with the predicted one, with an error of 5%. In contrast, the well Mckee 2A exhibits a prediction error of 20%. To understand the possible causes for the significant difference between the wells, it was noticed that for the same depth interval (1800 to 1900 m.), the behavior of the neutron log does not follow the typical pattern. In other words, it presents values slightly higher than what was learned by the algorithm.

This behavior underlines the importance of understanding, consider, and manage the automatic machine learning limitations in predicted pay zones. It will mitigate potential errors when using a limited source of information, allowing a better interpretation of what the model can predict (Malik, 2020).

Cuddy (2021) stated that machine learning cannot generate information by itself. They only do what it is told to do.

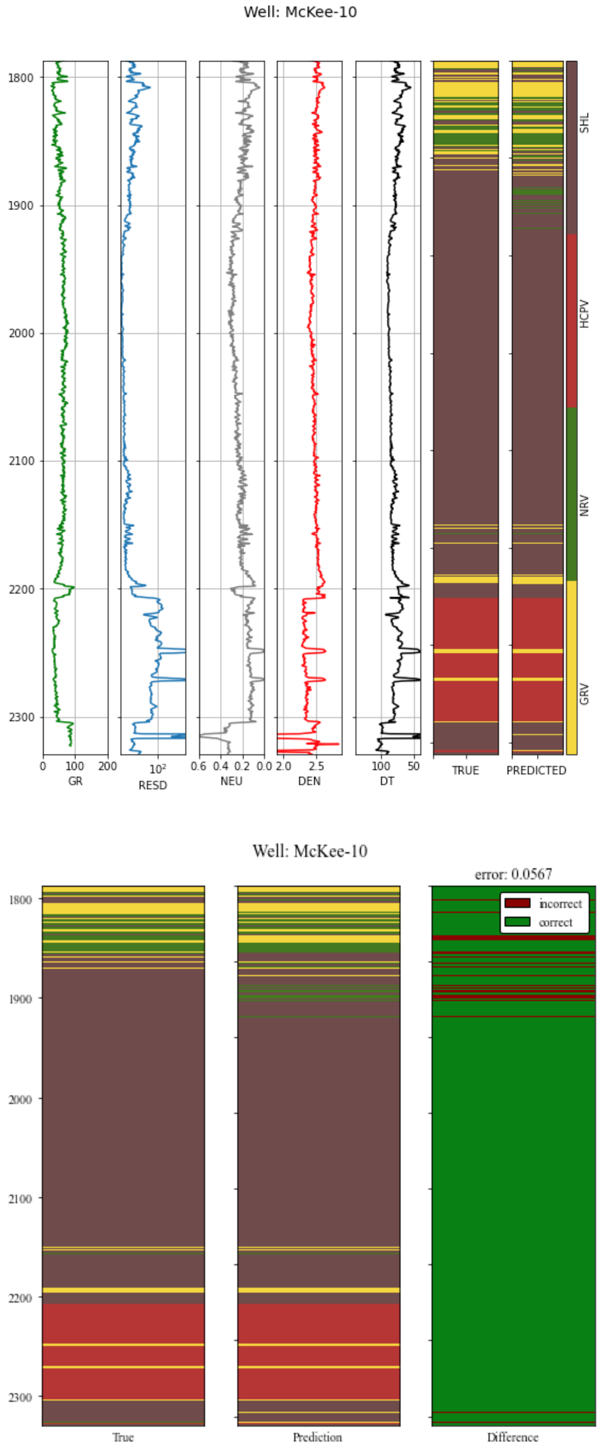

Figure 9. Top, Mckee-10 log plot in depth showing true vs. prediction pay properties. Bottom, Mckee-10 prediction error between true vs. prediction.

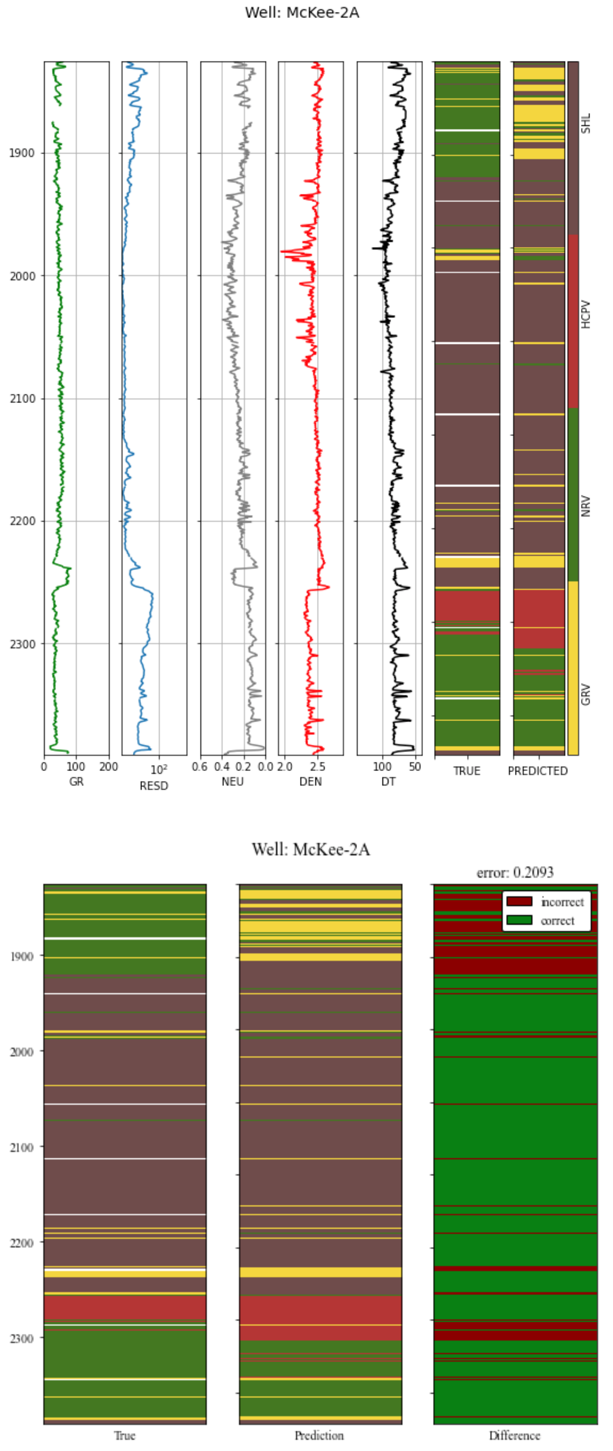

Figure 10. Top, Mckee-2A log plot in depth showing true vs. prediction pay properties. Bottom, Mckee-2A prediction error between true vs. prediction.

Making decisions from machine learning models implies a complete understanding of how the training data contributes to the final model. In AutoML, the model characteristics and description provide a table with the percentage of importance for the features (or explanatory variables), allowing the identification of the feature with the higher impact in the training of the model.

According to the nature of the classified properties, the distribution of the features' importance in Figure 11 was as expected, considering the physical principle of the logging tools and the primary uses of the properties they measure.

Otchere et al. (2021) reported that the density and the resistivity features were the most important in their study. This observation is in accordance with our results, where both features represent around the 50% of the classification power. This result could be interpreted in terms of the possible presence of pore volume, and its corresponding saturating fluid.

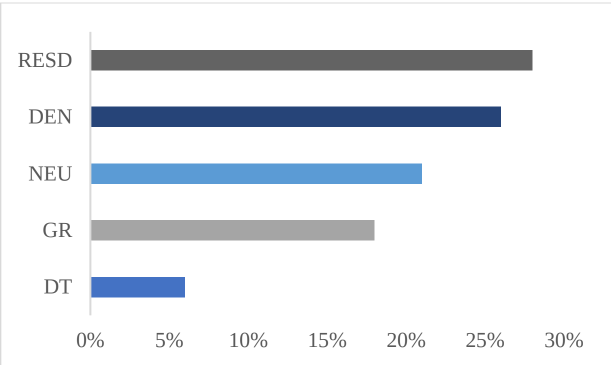

Figure 11. Input variables average percentage of importance for the classification of the pay properties.

The model showed a reduction in its classification power when one or more explanatory variables were dropped. However, acceptable error values were obtained. This observation suggests that for small datasets, the number of features to consider is more flexible than for the larger datasets. However, it is important to note the impact in the performance results according to the variable importance distribution for the reservoir properties classification. For instance, the density and neutron features were dropped from the well Tuhua-6, as shown in Figure 12. The high error obtained could be due to the fact that these two features are two of the three more important in the ranking distribution.

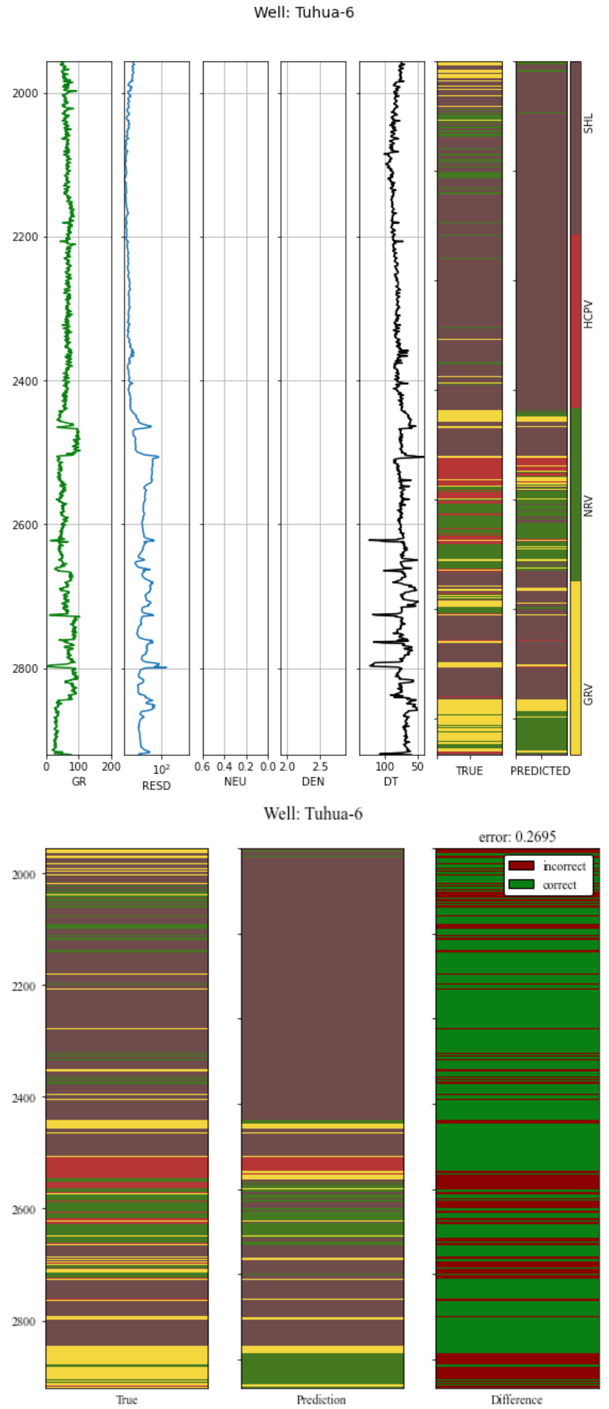

Figure 12. Top, Tuhua-6 log plot in depth showing true vs. prediction pay properties. Bottom, Tuhua-6 prediction error between true vs. prediction.

6. Conclusions

This study suggests that automatic machine learning is a promising tool in reservoir characterization since it does not require prior knowledge of reservoir properties, needs minimum intervention from the specialist, and works faster than conventional interpretation methods.

Automatic machine learning methods provide many benefits to reservoir characterization projects by automating tasks that are both time consuming and prone to subjectivity, helping to reduce operating costs by optimizing the decision-making time during the well evaluation phase and the initial petrophysical evaluation to understand reservoir characteristics.

For the pay properties prediction of the Mckee field, the performance of the automatic machine learning method could be limited in those cases in which logs (features) present changes in the curves' behavior used to train the model, such as unexpected fractures, mineralogy changes, formation damage during drilling operations, drilling fluids additives and logging operation/processing issues, among others., resulting in high prediction errors.

This study demonstrates that the XGBoost algorithm works well to predict reservoir pay properties in the Mckee field. However, it was not established that the algorithm could work in other reservoirs.

7. Acknowledgments

The first author would like to thank Department of Geology, Faculty of Science, Chulalongkorn University for the outstanding opportunity. An anonymous reviewer is thanked, whose comments substantively improved the paper.

8. References

Abbey, C. P., Okpogo, E. U., & Atueyi, I. O. 2018. Application of rock physics parameters for lithology and fluid prediction of 'TN' field of Niger Delta basin, Nigeria. Egyptian Journal of Petroleum, 27(4), 853–866. https://doi.org/10.1016/j.ejpe.2018.01.001

Bangert, P. 2021. Introduction to Machine Learning in the Oil and Gas Industry. In Machine Learning and Data Science in the Oil and Gas Industry (pp. 69–81). Elsevier. https://doi.org/10.1016/b978-0-12-820714-7.00004-2

Cuddy, S. 2021. The benefits and dangers of using artificial intelligence in petrophysics. Artificial Intelligence in Geosciences, 2, 1–10. https://doi.org/10.1016/j.aiig.2021.04.001

Dong, S. P., Shalaby, M. R., & Islam, M. A. 2018. Integrated reservoir characterization study of the McKee formation, Onshore Taranaki Basin, New Zealand. Geosciences (Switzerland), 8(4). https://doi.org/10.3390/geosciences8040105

Elhassan, A., Abu-Soud, S. M., Alghanim, F., & Salameh, W. 2021. ILA4: Overcoming missing values in machine learning datasets – An inductive learning approach. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2021.02.011

Gu, Y., Zhang, D., & Bao, Z. 2021. Lithological classification via an improved extreme gradient boosting: A demonstration of the Chang 4+5 member, Ordos Basin, Northern China. Journal of Asian Earth Sciences, 215. https://doi.org/10.1016/j.jseaes.2021.104798

Imamverdiyev, Y., & Sukhostat, L. 2019. Lithological facies classification using deep convolutional neural network. Journal of Petroleum Science and Engineering, 174, 216–228. https://doi.org/10.1016/j.petrol.2018.11.023

Ippolito, M., Ferguson, J., & Jenson, F. 2021. Improving facies prediction by combining supervised and unsupervised learning methods. Journal of Petroleum Science and Engineering, 200. https://doi.org/10.1016/j.petrol.2020.108300

Japkowicz, N. 2000. Learning from Imbalanced Data Sets: A Comparison of Various Strategies *. www.aaai.org

King, P.R. & Thrasher, G.P. 1996. Cretaceous-Cenozoic geology and petroleum systems of the Taranaki Basin, New Zealand. Institute of Geological and Nuclear Sciences monograph 13. 243 p, 6 enclosures. Lower Hutt, New Zealand: Institute of Geological & Nuclear Sciences Limited

Li, Y., & Anderson-Sprecher, R. 2006. Facies identification from well logs: A comparison of discriminant analysis and naïve Bayes classifier. Journal of Petroleum Science and Engineering, 53(3–4), 149–157. https://doi.org/10.1016/j.petrol.2006.06.001

Malik, M. 2020. A Hierarchy of Limitations in Machine Learning. http://arxiv.org/abs/2002.05193

Min, X., Pengbo, Q., & Fengwei, Z. 2020. Research and application of logging lithology identification for igneous reservoirs based on deep learning. Journal of Applied Geophysics, 173. https://doi.org/10.1016/j.jappgeo.2019.103929

Ministry of Business, Innovation, and Employment, New Zealand. 2014. New Zealand Petroleum Basins. New Zealand Petroleum & Minerals. Report. ISNN 2324-3988.

Otchere, D. A., Ganat, T. O. A., Gholami, R., & Lawal, M. 2021. A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction. Journal of Natural Gas Science and Engineering, 91. https://doi.org/10.1016/j.jngse.2021.103962

Saporetti, C.M., Goliatt, L. and Pereira, E. 2021. Neural network boosted with differential evolution for lithology identification based on well logs information. Earth Science Informatics,14(1), 133-140.

Thomas, E., Sneider, R., Neasham, J, and Vinegar, H. 1995. A catalog of petrophysical and geological properties of typical reservoir rocks. Update Published as Technical Progress Report BTC, pp 12–95

Vasini, E. M., Battistelli, A., Berry, P., Bonduà, S., Bortolotti, V., Cormio, C., & Pan, L. 2018. Interpretation of production tests in geothermal wells with T2Well-EWASG. Geothermics, 73, 158–167. https://doi.org/10.1016/j.geothermics.2017.06.005

Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., & Tu, M. 2018. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. Journal of Petroleum Science and Engineering, 160, 182–193. https://doi.org/10.1016/j.petrol.2017.10.028

Yang, H., Xu, Y., Peng, G., Yu, G., Chen, M., Duan, W., Zhu, Y., Cui, Y., & Wang, X. 2017. Particle swarm optimization and its application to seismic inversion of igneous rocks. International Journal of Mining Science and Technology, 27(2), 349–357. https://doi.org/10.1016/j.ijmst.2017.01.019

Zeng, L., Ren, W., & Shan, L. 2020. Attention-based bidirectional gated recurrent unit neural networks for well logs prediction and lithology identification. Neurocomputing, 414, 153–171. https://doi.org/10.1016/j.neucom.2020.07.026

Zhong, R., Johnson, R., & Chen, Z. 2020. Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost). International Journal of Coal Geology, 220. https://doi.org/10.1016/j.coal.2020.103416