Lithofacies Classification using Supervised Machine Learning Techniques

João Lucas M. da Silva* (Federal University of Pará, Brazil), José Jadsom S. De Figueiredo (Federal University of Pará, Brazil INCT-GP, Brazil) & Pamela Carolayne R. Bolsem (Federal University of Pará, Brazil)

This paper was prepared for presentation during the 18th International Congress of the Brazilian Geophysical Society held in Rio de Janeiro, Brazil, 16-19 October 2023.

Abstract

Lithofacies classification is vital in hydrocarbon exploration, involving the interpretation of rock layer characteristics. However, this process can be complex and subjective. Advancements in Artificial Intelligence (AI) technology present an opportunity for optimization through machine learning algorithms. This study utilized supervised methods (Random Forest, XGBoost, LightGBM, CatBoost) on Kansas, USA gas field datasets. The methodology involved preprocessing well data, implementing algorithms, and evaluating results. Machine learning models successfully recognized and differentiated lithofacies based on their characteristics. Evaluation metrics (Accuracy, Precision, Recall, F1) indicated XGBoost outperformed other models with: Accuracy (0.729341), Precision (0.753512), Recall (0.706092) and F1 (0.720048). These results highlight the promise of machine learning for automated lithofacies classification, providing accurate predictions and efficient classifications. These techniques can optimize geological interpretation in hydrocarbon exploration and apply to other scenarios.

Introduction

The different layers of rocks found in the subsurface of the Earth are known as lithofacies. Their classification plays a crucial role in hydrocarbon exploration, involving the interpretation of the main physical and chemical characteristics of rock layers through geophysical well logs (Serra, 1983). However, this interpretative classification is often complex and subjective, depending on the interpreter's perspective, which can lead to inaccurate and time-consuming results. With advancements in artificial intelligence technology, an opportunity has emerged to optimize and enhance this classification process through machine learning techniques. The use of machine learning algorithms can assist in identifying complex patterns in geophysical and geological datasets, making the classification more accurate and efficient (Ma and Zhang, 2019). Moreover, models can be trained on large labeled datasets, allowing for detailed and precise analysis of subsurface rock characteristics (Raschka, 2015).

The lithofacies classification process follows a sequence of well-defined steps. Initially, data collection takes place, seeking relevant information to predict rock characteristics. This data may include geophysical well logs, which provide physical and chemical data about rock layers (Nery, 2013). In this study, we explore the use of machine learning techniques to achieve accurate and automated lithofacies classification.

It is important to note that machine learning does not rely on explicitly programming rock characteristics but rather on the ability to recognize patterns. This allows models to be applied to new scenarios, making predictions or classifications based on previous training experiences (Raschka, 2015). In this specific study, supervised methods such as RF (Random Forest), XGBoost (Extreme Gradient Boosting), LightGBM (Light Gradient Boosting Machine), and Categorical Boosting (CatBoost) were utilized for lithofacies classification. All these methods were applied without hyperparameter tuning. These algorithms can learn from labeled examples and apply that knowledge to classify new geophysical datasets, increasing the accuracy and efficiency of the classification process.

To assess the efficiency of the algorithms used in lithofacies classification, commonly adopted metrics such as accuracy, precision, recall, and F1-score were employed (Hossin and Sulaiman, 2015). In order to evaluate the algorithms' performance in a real-world scenario, a prediction was made on a blind well that was not part of the machine learning model's training dataset. The application and visualization of this prediction allowed for an analysis of the algorithms' efficiency in a practical context.

The combined analysis of the metrics and the prediction revealed that the XGBoost algorithm demonstrated the best overall efficiency. It yielded the highest results in terms of accuracy, recall, and F1-score. The LightGBM algorithm also exhibited satisfactory performance, particularly in the precision metric. On the other hand, the CatBoost and Random Forest algorithms displayed slightly lower efficiency, but still achieved acceptable classification results.

These findings underscore the importance of selecting the appropriate algorithm for lithofacies classification, taking into consideration the evaluation metrics and the specific characteristics of the dataset at hand. Furthermore, the analysis of the prediction on a blind well provided additional validation of the algorithms' efficiency under real-world conditions.

Theorethical Background

To achieve the proposed objectives, the adopted methodology consisted of several main steps. The first step involved preprocessing the well data to make it suitable for implementing the classification algorithms. In this step, well data visualization, target and feature selection, data normalization, outlier removal, and categorical variable encoding, among other minor procedures, were performed.

Additionally, the data were trained.

The second step involved the implementation of the aforementioned classification algorithms. For this purpose, the Scikit-learn and CatBoost libraries were used for machine learning, along with the LightGBM library. These classification models have a similar usage as regression models, aiming to identify groups in the input data, with a categorical variable as the output. Random Forest is an ensemble machine learning algorithm that combines multiple individual decision trees to form a robust and accurate classifier. It was introduced by Breiman (2001) and has since been widely used in various application fields.

The basic idea behind the model is to create a set of decision trees, where each tree is trained with a random sample of the training data and uses a random subset of features to make node splits. In other words, each decision tree is built with a random selection of observations and variables, preventing a specific set of data or features from dominating the learning processes.

During the testing phase, the algorithm classifies a new example by counting the votes from the results of each decision tree. The class with the highest number of votes is chosen as the final class for the input example. Studies have shown that Random Forest is capable of handling a variety of machine-learning problems, including classification, regression, and anomaly detection. It also has the advantage of being relatively insensitive to outliers, noise, and missing data.

Random Forest can be defined as:

$$f(x) = \sum_{k=1}^{L} h_k(x)$$

where $h_k(x)$ is a decision tree trained on T_k independent random samples that are equally distributed, and each decision tree determines the class with the highest popularity for an input (Breiman, 2001). On the other hand, XGBoost is an open-source machine learning library that utilizes algorithms based on a gradient boosting framework. Gradient boosting is a technique that allows the use of regression and classification in prediction problems and produces a model in the form of a decision tree, which can be interpreted as an optimization algorithm on a suitable cost function. This method was developed by Chen et al. (2015) in a research project that used boosted gradient decision trees and observed several advantages over other methods, such as parallel data processing, regularization to reduce the chance of overfitting, and flexibility in hyperparameter tuning.

Its numerical definition involves the minimization of a regularized loss function, defined as the sum of the model's losses on the training instances plus a penalty for the model's complexity controlled by the hyperparameter λ:

$$\mathscr{L}(\phi) = \sum_i l(y_i, \hat{y}i) + \sum k\Omega(f_k)$$

where l is a loss function that measures the discrepancy between the model's predictions $\hat{y}_i$ and the true responses y_i on the training instances, f_k is a decision tree that models the relationship between the features and responses, Ω is a penalty on the complexity of the trees, and ϕ is the set of all trees (Chen et al., 2015).

The algorithm starts with a single simple decision tree and, in each iteration, adds a new tree that focuses on the residual errors of the previous prediction. The contribution of the new tree is controlled by the hyperparameter η, which determines the learning rate or step size. The addition of new trees continues until the loss function can no longer be reduced or until a stopping criterion is reached. During tree construction, a technique called pruning is applied, which removes irrelevant branches from the tree and helps prevent overfitting. The final result of the algorithm is a collection of decision trees that can be used to predict new instances or understand the relationship between features and responses.

The model provides flexible and efficient solutions, capable of solving various problems involving databases more quickly and accurately. This justifies its reputation as one of the beloved algorithms used in competitions, such as the renowned Kaggle platform. Regarding geophysical data, particularly well data, XGBoost can be used in the classification of facies and other problems. Additionally, other machine learning libraries such as Random Forests, LightGBM, and CatBoost are also applicable to this type of problem. These tools can help obtain more accurate and efficient models for the classification of geological facies, which is a fundamental step in the process of hydrocarbon reservoir exploration.

LightGBM is a machine learning algorithm based on decision trees that utilize gradient boosting techniques. It has a vertical growth structure, allowing a leaf-wise growth of the tree, which differentiates it from other algorithms that have a level-wise horizontal structure as mentioned by Ke et al. (2017). It was created by Microsoft in 2017 and has since been widely used in classification problems due to its high efficiency and training speed, as well as its lower computational memory usage and high accuracy.

Several studies have pointed out the advantages of the model compared to other classification algorithms (Yan et al., 2021). When applying LightGBM to a dataset, methods such as adjusted decision trees are used to avoid overfitting the training data, integrated gradient optimization to improve the classifier's robustness, and gradient-based one-sided sampling (Jabeur et al., 2021). According to Jabeur et al., the estimated function of LightGBM integrates multiple regression decision trees, and it is defined as follows:

$$f(x) = \sum_i t_i(x)$$

where t_i is a regression decision tree.

Finally, like the other methods mentioned, CatBoost also utilizes a gradient-boosting algorithm in the form of a decision tree. It was responsible for solving a common problem in LightGBM and XGBoost, known as missing target. This problem causes the trained algorithm to depend on the targets present during the training data, leading to compromised results in the absence of some targets in case of data changes. The general equation of CatBoost can be expressed as follows:

$$F(x) = \sum_{i=1}^{N} \eta \cdot h(x, \theta_i)$$

where $F(x)$ represents the predictive model constructed by

CatBoost, N is the number of base estimators, η denotes the learning rate, $h(x, \theta_i)$ is a weak function (decision tree) with parameters θ_i (Dorogush et al., 2018).

This formulation allows CatBoost to iteratively combine multiple weak estimators to form a more powerful predictive model. CatBoost's characteristic is to use ordered boosting during processing as one of the algorithm's basic predictors, employing the same splitting criterion at every tree level. The trees are balanced between those less and more prone to overfitting.

It has advantages such as the ability to use the entire dataset for training, incorporating all classification features into the current tree of the dataset, employing a wide range of permutations of training data for increased robustness in the results, and utilizing binary features stored in continuous vectors to calculate the model's predictions. In the third step, the effectiveness of the algorithms on well data was analyzed. Cross-validation techniques were employed to evaluate the model's accuracy and prevent overfitting. Furthermore, the effectiveness of the classification algorithms was assessed using metrics such as accuracy, recall, precision, and F1-score.

Metrics for classification analysis are a way to evaluate the performance quality of a model. They are important to understand how a model is performing and make adjustments to improve the accuracy of predictions. Among the metrics used during the research, accuracy, recall, precision, F1-score, and confusion matrix can be mentioned. The confusion matrix aims to present the model's performance in a tabular format, showing the data that was correctly and incorrectly classified, including true positives, true negatives, false positives, and false negatives (Saito and Rehmsmeier, 2015). It is often used as a tool to understand how the model is classifying each class and can also be used to calculate other evaluation metrics. It is applied to both binary and multiclass classification. Finally, the predicted wells were visualized and compared with the lithology of real wells. It is of paramount importance to mention that the classification, up to this point, was performed without hyperparameter optimization, which should be employed in future research to further refine the results.

Methodology

The data used in this study pertains to the Hugoton and Panoma gas fields located in Kansas, United States, as mentioned by Dubois et al. (2003; 2006; 2007). These fields are situated in the Anadarko Basin, bounded by the Las Animas arch and the Central Kansas uplift. The basin is classified as a foreland type and is associated with the Ouchita-Marathon Pennsylvanian orogeny, as described by Kluth (1986) and Perry (1989). The main reservoirs consist of carbonate rocks, with secondary high-permeability and high-porosity sandstone reservoirs. The predominant sealing rocks are fine to coarse-grained silstones, as well as evaporites, as reported by Heyer (1999) and Dubois et al. (2003).

The Hugoton and Panoma fields are concentrated within the Chase and Council Grove groups, which consist of vertical successions of lithofacies with a well-established cyclical pattern. These facies successions exhibit an upward pattern resulting from depositional environments influenced by rapid changes in relative sea level, as discussed by Olson et al. (1997). More information about the depositional model attributed to these groups can be found in the study by Dubois et al. (2006) and the cited references.

The data used in this study were provided by the University of Kansas and were obtained through a challenge organized by the Society of Exploration Geophysicists for lithology prediction. It is important to note that these data were used in their raw form without undergoing any preprocessing, as mentioned by Dubois et al. (2003; 2006; 2007). The dataset consists of 3232 records from eight distinct wells: SHRIMPLIN, SHANKLE, LUKE G U, CROOS H CATTLE, NOLAN, Recruit F9, NEWBY, and CHURCHMAN BIBLE. These records contain measurements at 0.15 m intervals of various variables from well logs, such as gamma-ray (GR), resistivity (ILD_log10), photoelectric effect (PE), average porosity and neutron density (PHIND). Additionally, the dataset also includes the difference between porosity and neutron density (DeltaPHI) and information about relative position (RELPOS). It is worth noting that all records are associated with specific lithologies, which include: Nonmarine sandstone (SS), Nonmarine coarse siltstone (CSiS), Nonmarine fine siltstone (FSiS), Marine siltstone and shale (SiSh), Mudstone (limestone) (MS), Wackestone (limestone) (WS), Dolomite (D), Packstone-grainstone (limestone) (PS), Phylloid-algal bafflestone (limestone) (BS).

Results and Discussions

After applying the lithofacies classification methods using the Random Forest, XGBoost, LightGBM, and CatBoost algorithms, the following through metrics performance is shown in Table 1.

Table 1: Metrics performance on the lithofacies classification using different ML models: RF, XGB, LGBM, and CatBoost. Results in terms of Acc, Prec, Rec, and F1.

Model	Acc	Prec	Rec	F1
RF	0.716168	0.748162	0.688896	0.703561
XGB	**0.729341**	0.753512	**0.706092**	**0.720048**
LGBM	0.724551	**0.758990**	0.689090	0.709863
CatBoost	0.720958	0.746437	0.691803	0.707078

Analyzing these results, we can highlight that all models achieved relatively good performance in lithofacies classification. The accuracy, which measures the proportion of correctly classified examples, ranged from 0.716168 to 0.729341, indicating an overall accuracy rate of over 70% (see Figure 1). Looking at precision, which represents the proportion of correctly classified positive examples, all models obtained values above 0.74, indicating a consistent ability to correctly identify positive lithofacies. The recall, which measures the proportion of correctly identified positive examples relative to the total number of truly positive examples, also showed positive results for all models, with values above 0.68. This suggests that the models were able to identify a significant percentage of positive lithofacies present in the data. The F1-score, which combines precision and recall into a single metric, had values above 0.7 for all models. This result indicates a satisfactory balance between the ability to

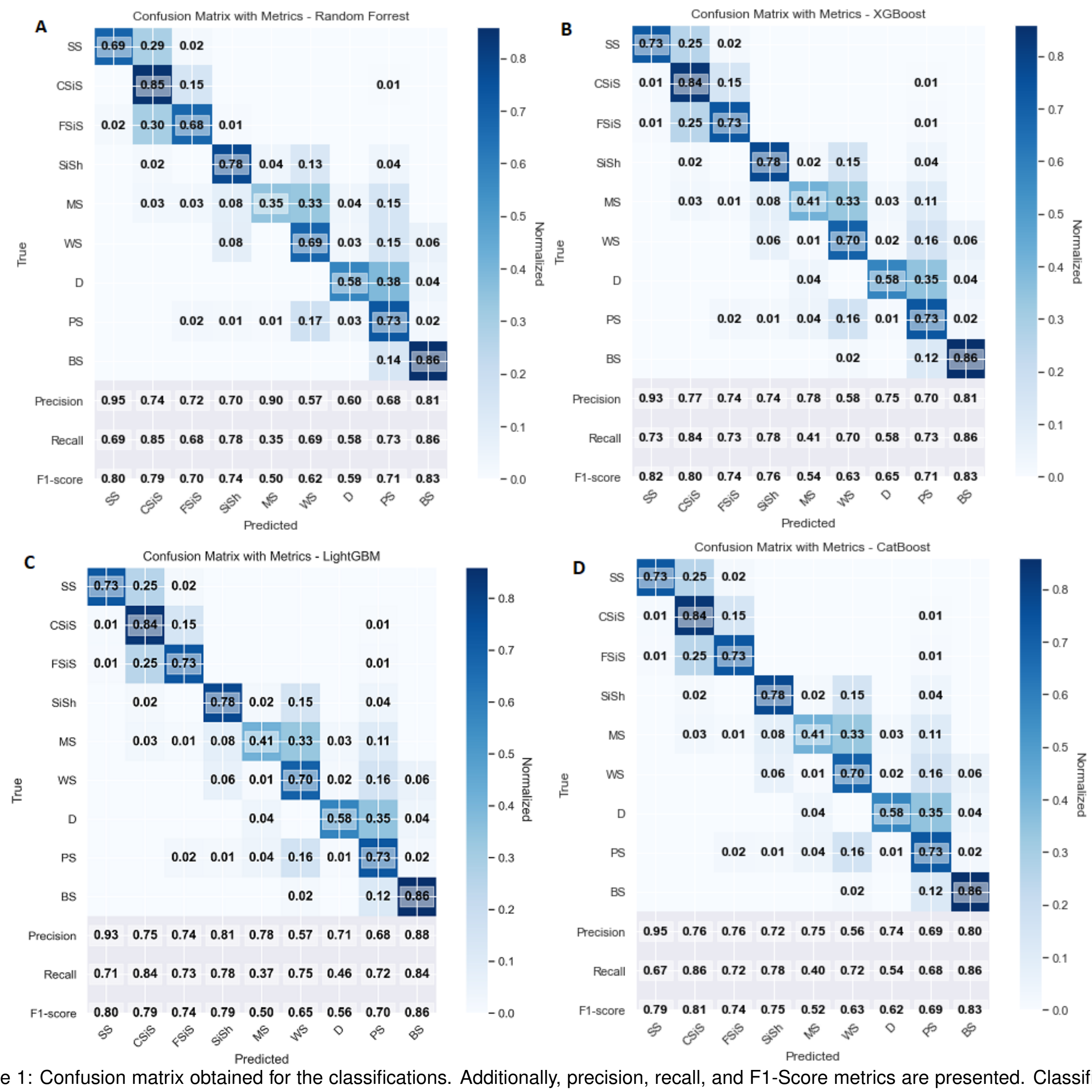

Figure 1: Confusion matrix obtained for the classifications. Additionally, precision, recall, and F1-Score metrics are presented. Classifications A, B, C, and D correspond, respectively, to the Random Forest, XGBoost, LightGBM, and CatBoost methods.

correctly classify positive lithofacies and the ability to avoid false positives (see Figure 1).

The Figure 1 present the confusion matrix of the Random Forest, XGBoost, LightGBM, and CatBoost models used in this study. By considering the confusion matrix along with the evaluation metrics, it is possible to obtain a more complete and detailed view of the classification models' performance. This in-depth analysis assists in identifying areas where the models may be experiencing difficulties and suggests possible improvements or adjustments increase the accuracy and reliability of lithofacies classifications.

From Figure 1, in A, we can observe that the Random Forest model achieved a precision of 0.748162, indicating that 74.82 of lithofacies were correctly classified. The recall, which measures the model's ability to correctly identify lithofacies, was 0.688896, representing an accuracy rate of 68.89. The F1-score, which is a weighted average of precision and recall, was 0.703561. Similarly, B, C, and D show the confusion matrices and performance metrics for the XGBoost, LightGBM, and CatBoost approaches, respectively.

By analyzing the confusion matrices, it is possible to identify areas where the models encounter difficulties, such as confusion between similar lithofacies. This detailed analysis can be used to enhance the models by adjusting parameters, selecting more relevant features, or even collecting more data to improve the representativeness of the training set.

Each cell represents the number of instances corresponding to a specific combination of true class and predicted class. The main diagonal indicates the correct classifications, while the cells off the main diagonal indicate incorrect classifications. Analyzing the confusion matrix provides insights into the models' performance for each class and assists in identifying error patterns.

A more comprehensive way to evaluate the results is through the analysis of predicted well profiles, as shown in Figure 2. This visualization allows for a direct observation of the lithofacies of the SHANKLE well, as well as the predicted profiles using their respective classification methods. Additionally, well profiles for the GR, (ILD_log10), DeltaPHI, PHIND, and PE properties have also been plotted. It is essential to highlight that the visual analysis of the predicted well profiles plays a crucial role in the intuitive understanding of the results, enabling the identification of possible patterns or discrepancies in the classification methods. This approach, combined with the evaluation metrics analysis, provides a comprehensive and detailed insight into the performance of the classification models. From this, it is possible to identify areas where the models may encounter difficulties and propose improvements or adjustments to increase the accuracy and reliability of lithofacies classifications.

Conclusions

When comparing the results of the machine learning models applied to automated lithofacies classification, it was observed that XGBoost showed the highest accuracy. However, the LightGBM and CatBoost models also demonstrated similar performance, with closely evaluated metrics. These results highlight the potential of these algorithms when applied to geophysical and geological datasets, allowing for the precise identification of lithofacies. The use of machine learning techniques enables the detection of complex patterns in the data and learning from labeled examples, providing a detailed analysis of subsurface rock characteristics and contributing to a better understanding of geological processes.

However, it is important to note that the obtained results can be improved through hyperparameter tuning of the models. Hyperparameter tuning involves searching and optimizing the best values for the algorithm parameters, aiming to further enhance performance and classification accuracy. Therefore, future studies can focus on this refinement step, exploring different combinations of hyperparameters and advanced optimization techniques. This will enable a more precise adjustment of the models to specific geophysical and geological data, potentially resulting in higher accuracy and robustness in lithofacies classifications.

In summary, the results of this study demonstrate that machine learning models such as XGBoost, LightGBM, and CatBoost perform well in automated lithofacies classification. Hyperparameter tuning represents a promising opportunity to further improve these results, contributing to significant advancements in the understanding and interpretation of subsurface geological characteristics. We will continue to explore and refine these techniques, aiming to contribute increasingly to the advancement of geophysical and geological science.

Acknowledgments

Acknowledgments to the UFPA Faculty of Geophysics and CNPq for financing the creation of this study with a Scientific Fund (Initiation Scholarship).

References

Breiman, L., 2001, Random forests: Machine learning, **45**, 5–32.

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, et al., 2015, Xgboost: extreme gradient boosting: R package version 0.4-2, **1**, 1–4.

Dorogush, A. V., V. Ershov, and A. Gulin, 2018, Catboost: gradient boosting with categorical features support: arXiv preprint arXiv:1810.11363.

Dubois, M., G. Bohling, A. Byrnes, and S. Seals, 2003, Extracting lithofacies from digital well logs using artificial intelligence, panoma (council grove) field, hugoton embayment, southwest kansas.

Dubois, M. K., G. C. Bohling, and S. Chakrabarti, 2007, Comparison of four approaches to a rock facies classification problem: Computers & Geosciences, **33**, 599–617.

Dubois, M. K., A. P. Byrnes, G. C. Bohling, and J. H. Doveton, 2006, Multiscale geologic and petrophysical modeling of the giant hugoton gas field (permian), kansas and oklahoma, usa.

Heyer, J. F., 1999, Reservoir characterization of the council grove group, texas county, oklahoma.

Hossin, M., and M. N. Sulaiman, 2015, A review on evaluation metrics for data classification evaluations: International journal of data mining & knowledge management process, **5**, 1.

Jabeur, S. B., R. Khalfaoui, and W. B. Arfi, 2021, The effect of green energy, global environmental indexes, and stock markets in predicting oil price crashes: Evidence from explainable machine learning: Journal of Environmental Management, **298**, 113511.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, 2017, Lightgbm: A highly efficient gradient boosting decision tree: Advances in neural information processing systems, **30**.

Kluth, C., 1986, Plate tectonics of the ancestral rocky mountains: part iii. middle rocky mountains.

Ma, Y. Z., and X. Zhang, 2019, Quantitative geosciences: Data analytics, geostatistics, reservoir characterization and modeling: Springer.

Nery, G. G., 2013, Perfilagem geofísica em poço aberto, 1 ed.: Sociedade Brasileira de Geofísica (SBGf).

Olson, T. M., J. Babcock, K. Prasad, S. Boughton, P. Wagner, M. Franklin, and K. Thompson, 1997, Reservoir characterization of the giant hugoton gas field, kansas: AAPG bulletin, **81**, 1785–1803.

Perry, W. J., 1989, Tectonic evolution of the anadarko basin region, oklahoma: Department of the Interior, US Geological Survey.

Raschka, S., 2015, Python machine learning, 1 ed.: Packt Publishing.

Saito, T., and M. Rehmsmeier, 2015, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets: PloS one, **10**, e0118432.

Serra, O., 1983, Fundamentals of well-log interpretation.

Yan, J., Y. Xu, Q. Cheng, S. Jiang, Q. Wang, Y. Xiao, C. Ma, J. Yan, and X. Wang, 2021, Lightgbm: accelerated genomically designed crop breeding through ensemble learning: Genome Biology, **22**, 1–24.

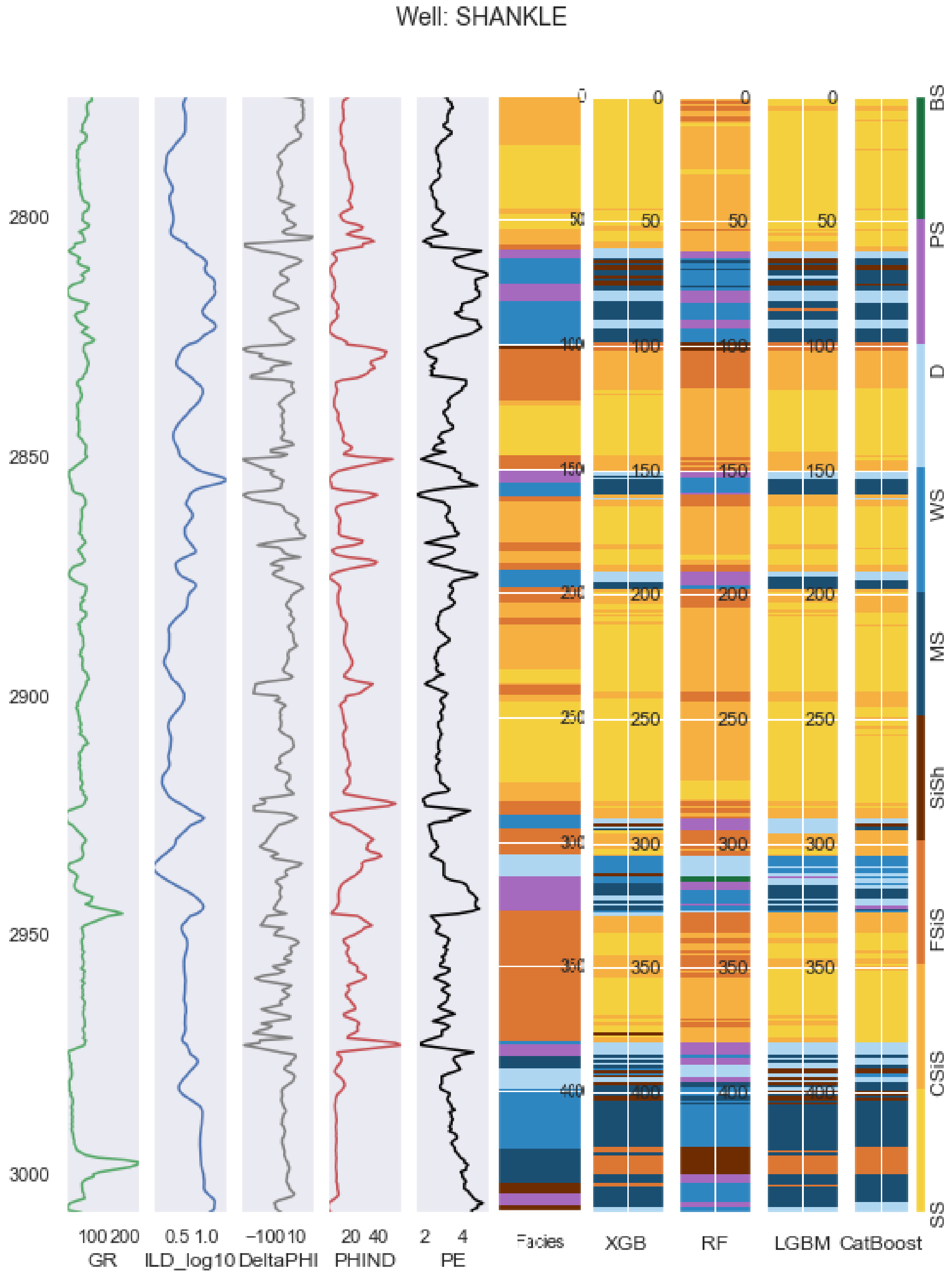

Figure 2: Comparison between the actual facies and the facies predicted by the classification methods used in this research is presented. The gamma-ray curves (GR), logarithmic scale resistivity on base 10 ((ILD_log10)), the difference between porosity and neutron density (DeltaPHI), the average porosity and neutron density (PHIND), and photoelectric effect (PE) is also illustrated.