International Journal of Petroleum and Petrochemical Engineering (IJPPE)
Volume 10, Issue 1, 2025, PP 1-9
ISSN 2454-7980 (Online)
DOI: https://doi.org/10.20431/2454-7980.1001001
www.arcjournals.org

Enhancing Hydrocarbon Recovery Factor Prediction Using Ensemble Machine Learning Workflow

Eric M. Amarfio[1] , Mary Aboagye[1], Daniel Asante-Otchere[2], Harrison Osei[1]

[1]*Department, of Petroleum and Natural Gas Engineering, University of Mines and Technology*

[2]*Center for Subsurface Imaging, Universiti Teknologi Petronas*

***Corresponding Author: Eric M. Amarfio.,** *Department, of Petroleum and Natural Gas Engineering, University of Mines and Technology*

Abstract: *This paper seeks to develop a machine learning workflow for accurately predicting hydrocarbon recovery factor (RF), a crucial property for exploration and production strategies. The study employed agnostic method and Bayesian Optimized Decision Tree (DT) model to assess causal relationships between input and target variables. The decision tree model yielded the lowest error metrics (0.1 MAE and 0.132 RMSE). To determine how much each input reservoir parameter contributes to the target (RF), the permutation feature importance and Shapley Value metrics were used to analyze the input features. The decision tree model was optimized and trained using the chosen parameters, yielding better results of 0.02 MAE and 0.03 RMSE. The features selected by Shapley values greatly improved the predictive model outcome because it demonstrated a causal relationship with recovery factor estimation. The proposed method outperformed several reported models, demonstrating the value of machine learning in petroleum engineering for reservoir characterization. Improved data pre-processing, analysis, and visualization techniques were also utilized in this study.*

Keywords: *hydrocarbon reserve estimation; recovery factor; machine learning; artificial intelligence.*

1. INTRODUCTION

The Recovery Factor (RF) is a crucial parameter in hydrocarbon reservoir management, defined as the ratio of ultimately recovered oil to the initial amount of oil in place (Gulstad, 1995). Estimating the recovery factor accurately is essential for making sound financial decisions, as it relies on estimated reserve quantities, production rates, and efficient management of the oil reservoir. However, the estimation of RF is challenging due to several geological and engineering factors (Holdaway, 2009). Various methods are used to calculate RF, including volumetric and analogy methods, material balance calculations, decline curve analysis, and numerical reservoir simulation. Empirical methods such as material balance equations and numerical simulations have yielded reliable results. Still, they require significant efforts and detailed reservoir descriptions to build an accurate model and conduct uncertainty qualifications (Arps, 1955).

The advent of artificial intelligence has revolutionized the petroleum engineering industry, particularly in the domain of reservoir characterization. Machine learning techniques have gained significant attention for the potential to improve recovery factor estimation. Several research attempts have been made to use machine learning to build recovery factors estimation models, such as multiple linear regression (MLR), artificial neural networks (ANN), Bayesian networks, support vector machines (SVM), and other techniques. The choice of model and input variables affects the performance of the ML models, and selecting the best features is crucial for the accuracy of the predictions. In the research conducted by Lake and Lee (2015), a range of techniques were employed to estimate the oil and gas recovery factors (RFs) were multiple linear regression (MLR) with sequential feature selection, artificial neural networks (ANN), and Bayesian network. The findings of the study revealed that the ANN and MLR with sequential feature selection exhibited superior performance in comparison to the remaining two approaches. In another study done by Aliyuda and Howell (2019), a combination of multiple linear regression (MLR) and support vector machine (SVM) techniques utilizing the Gaussian kernel was employed to analyze 93 reservoirs situated on the Norwegian Continental Shelf. Among these reservoirs, 75 were located in the Norwegian Sea, the Norwegian North Sea, and the Barents Sea, while the remaining 18 were found in the Viking Graben within the UK sector of the North Sea. The research

findings indicated that the SVM model exhibited superior performance compared to the MLR model, as reported by Aliyuda and Howell (2019). Lastly, Chen in 2019 utilized the Artificial Neural Network (ANN) technique to construct predictive oil Random Forest (RF) models. The models were developed by employing various sets of input data sourced from the TORIS database. Throughout their investigation, the researchers identified a total of 19 principal features from the initial 70 variables to be included in the construction of their machine learning (ML) model. These works highlight the potential of ML in RF estimation, and our study builds on their findings by using a broader range of ML models and assessing the importance of input features.

The effectiveness of machine learning models and their ability to estimate reservoir recovery factors depends on the type of model used and the input variables chosen. When irrelevant data is fed into these models, their performance can significantly decrease. Therefore, developing a procedure to identify and select the best features and create new ones from existing variables is crucial. By including these new features, the model's performance should improve compared to when they are inputted individually. Combining several robust models into one supermodel can enhance its overall performance. For our investigation, a dataset comprising of 3420 data points was sourced from published literature to evaluate the importance of input features in predicting recovery factors. We employed model-agnostic metrics to identify a relevant dataset with the most valuable features. Next, we tested various machine learning models to predict recovery factors, and the model with the best results was optimized using the Bayesian Optimization (BO) algorithm. By following this approach, we can develop a reliable model for estimating reservoir recovery factors, which can significantly affect the oil and gas industry.

2. METHODS AND MATERIALS

This research employed Python programming to investigate the relationship between crucial input variables relevant to recovery factor prediction. Various regression models were used to identify the model with the lowest error rate. The relevant features identified by model agnostic techniques were selected as input features. Several machine learning models were then utilized to predict recovery factors based on their learning theory and capacity to work with high-dimensional and complex data. A significant challenge with machine learning models and their varying performance is the issue of data quality. Thus, this study aimed to explain the input features, not only in terms of correlation but also in causation, by developing new methodologies. The goal was to achieve even the slightest increase in accuracy, which is essential in enhancing the decision-making process in the petroleum industry.
The summary analysis of the reviewed models used in this study is summarized in Table 1.

Table1. *Summary of Algorithms and Corresponding Authors*

SN	Model	Description	Authors
1.	Ridge Regression.	Minimizes standard errors by applying a bias to model estimates	Hoerl and Kennard (1970).
2.	Lasso Regression	Zeroes coefficients of unimportant input variables	Tibshirani (1996)
3.	Support Vector Machine (SVM)	Produces better generalized, sparse, and unique solutions	Vapnik and Lerner (1963)
4.	Decision Tree	Builds hierarchical decision boundaries and removes unnecessary structures	Gordon et al. (1984)
5.	Extreme Gradient Boosting	Maximizes the loss function with an extra regularization term	Chen and Guestrin (2016)
6.	Random Forest	Robust against overfitting and offers easy interpretability	Breiman (2001)

2.1 Data Collection and Description

A comprehensive collection of data from 139 sandstone reservoirs was utilized to develop a recovery efficiency bulletin. The dataset comprises 3197 data points encompassing 23 distinct features, which were consolidated as reports submitted to the API subcommittee. Descriptive statistics of a subset of these features are presented in Table 2.

	h	p	K	k/Uob	Sw	T	API	Pi	Boi
count	139	139	139	139	139	139	139	139	139
mean	63.0187	0.1928	198.461	233.997	0.3009	150.86	36.02	843.90	0.528
std	122.523	0.1928	354.98	528.217	0.1012	44.91	7.54	631.85	23.77
min	4.5	0.065	0.100	0.200	0.000	75.00	14.00	101.00	0.754
25%	15.00	0.14	25.50	15.17	0.235	122.00	33.00	449.00	31.0
50%	25.00	0.180	66.80	58.86	0.300	144.00	38.00	786.00	68.40
75%	50.00	0.231	239.0	187.55	0.363	183.00	41.20	922.40	84.92
max	1100.0	0.354	2970.0	4500.0	0,600	270.00	49.70	3630	404.40

1. Energy: initial pressure, bubble point pressure, reservoir pressure at the end of primary, reservoir temperature
2. Fluid: initial oil viscosity, oil viscosity at bubble point, viscosity at abandonment of primary, water viscosity
3. Rock: Porosity, Saturation, Gas Saturation, wettability, salinity
4. Asset size: net pay zone, well spacing
5. Recovery: API gravity, initial gas-oil ratio, gas-oil ratio at the bubble point, oil formation volume factor,

Data analytics and machine learning enabled to uncover patterns and extract hidden insights from high-dimensional data. However, some input data may be missing, distorted, or irrelevant in predicting recovery factors. The issue of missing, distorted, irrelevant data is addressed using several strategies, including missing value imputation. The dataset was evaluated to identify missing or malformed data by counting the rows and columns in the data frame, and it was found that there were no missing or distorted data in any of the parameters. A statistical analysis of the input was performed to calculate the sum of data points, mean, standard deviation, minimum and maximum values of all parameters, and (25%, 50%, 75%) distribution of the data points. The statistical analysis revealed that the data were normally distributed.

The Spearman rho covariance matrix was used to quantify the degree of correlation between all input features and the target variable (RF). The Spearman rho covariance matrix helped to identify the correlations' strength and direction between the input features and the target variable.

Finally, irrelevant parameters, such as calculated OOIP and produced, were dropped from the input dataset due to their high potential of correlation to the prediction target (RF). By employing these strategies, we can effectively tackle incomplete datasets and uncover valuable insights from high-dimensional data.

2.2 Explainable AI

The data were split randomly into two groups to build the machine learning (ML) models: 80% for training the model and 20% for testing it. This split ratio was chosen to provide the ML algorithms with many samples for training. Python programming was utilized in this study, and multiple regression models were employed to determine the most effective model for predicting the recovery factor with minimal errors. The same out-of-sample data was used consistently throughout the study to ensure that all models were evaluated using similar data points.

Model-agnostic techniques were employed to analyze the relevance of input features to separate explanations from the machine learning models. To achieve this objective, both Permutation Feature Importance (PFI) and Shapley values were utilized for providing explanations. The desired properties were model representation and explanation flexibility, which are not specific to any model type. The inputs required for this analysis included the models, the target variable (RF), and the error metrics.

2.2.1 Permutation Importance Feature

The Permutation Feature Importance method is a model-agnostic technique used to assess the significance of features in a model, regardless of the model type. It involved randomly assigning a component with a nearly random value and permuting its relationship with the model output. If changing the feature value increases model error, then the feature is considered essential. The advantage of PFI is that it is straightforward to interpret and does not require retraining of the data. This work used a function based on a regression model to calculate the importance score for each input parameter. Features with high importance scores are considered to have more substantial predictive potential, as determined by criteria for importance score, as demonstrated in this study by Otchere et al. (2022).

2.2.2 Shapley Values

The Shapley value is a crucial concept in determining the significance of a feature and its contribution to a model's performance. It was developed by Lloyd Shapley from Cooperative Game Theory. Understanding how the model's parameters affect its output is vital in grasping the fundamental elements of creating its output. Unlike other techniques, the Shapley value can be used regardless of the

model type or structure. However, it has its limitations, the most significant being its high computational cost. As a result, an approximate solution may be the only feasible option for many real-world scenarios. Additionally, it can be easily misinterpreted. Fortunately, the Shap library, an open-source tool, is an excellent resource for working with Shap values and other metrics (Shapley, 1953).

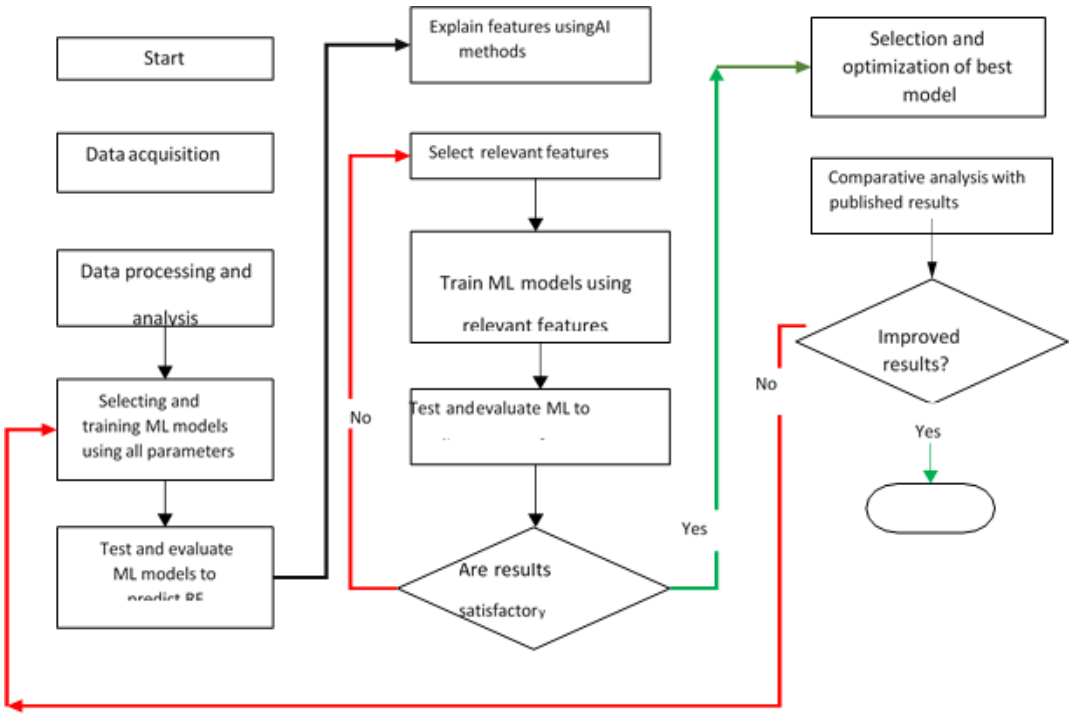

Fig1. Machine Learning Workflow

2.3 Model Evaluation Criteria

The prediction error was measured by the difference between the expected values and the best-fit line of the actual data to assess the accuracy of model predictions in this study. The appropriateness of the models used was determined by evaluating and rating their errors based on the following criteria:

1. The Akaike Information Criterion (AIC) assesses the precision and excellence of models by indicating a greater likelihood of the model is suitable for the data

$$AIC = 2K - 2(log - likelihood) \quad (1)$$

2. Mean Absolute Error (MAE) represents the average absolute difference between the actual and predicted values.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (2)$$

3. The Root Mean Squared Error (RMSE) is a commonly used metric for assessing model performance, as it provides an easily interpretable measure of the deviation of prediction errors, indicating the proximity of predicted values to their expected values. Specifically, the RMSE represents the standard deviation of the prediction errors.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (3)$$

4. The coefficient of determination (R2) is a commonly used criterion that measures how closely the dependent variable fits the regression line with the best fit.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \quad (4)$$

3. RESULTS

3.1 Error Metrics

The models were compared using AIC to determine the best fit (Figure 2). Lower AIC values indicate better predictions. AIC differences of less than 10 are insignificant, while those between 10 and 50 are moderate, and those between 50 and 100 are significant. Differences greater than 100 are extreme. The decision tree model was the most accurate for predicting recovery factors. Compared to the Boost model, which also had low AIC, the decision tree model had a difference of 260 AIC, indicating a significant improvement in fit. AIC can be used to determine the model's ability to fit the data accurately.

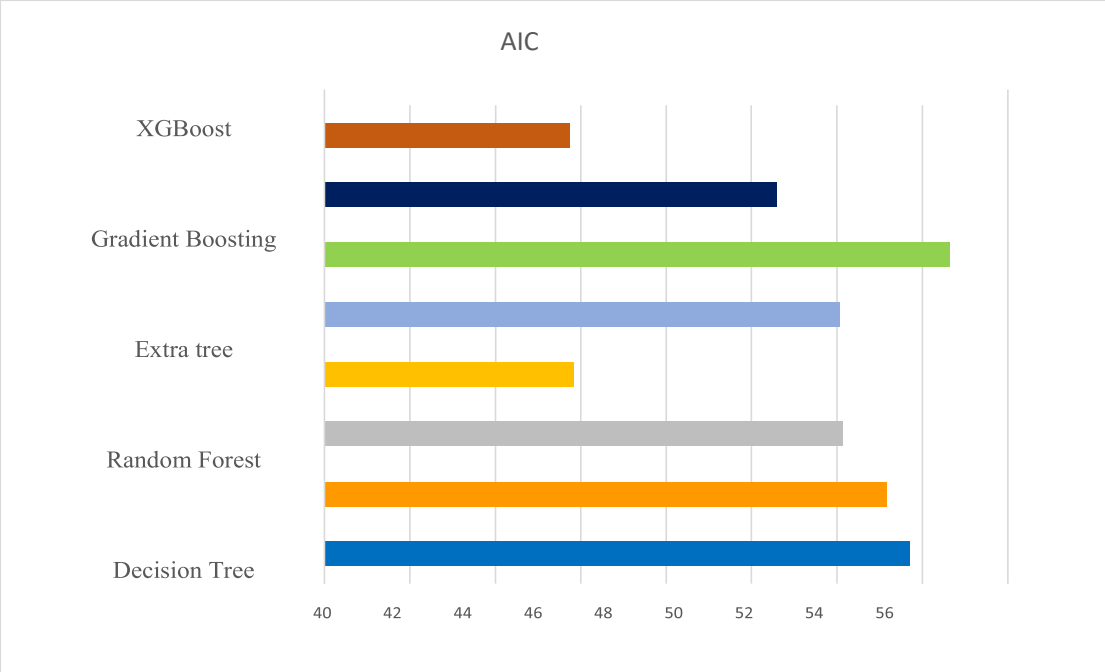

Fig2. *AIC Results of Models on Test Data*

The models' robustness is evaluated using unseen test data. While high training accuracy is expected since the models have already seen the data, it can lead to overfitting and a non-generalized model that include noise. Out-of-sample accuracy, which measures correct predictions on unseen data, is more crucial as models need to perform well on unknown data. In this study, R^2 values exceeding 0.85 were deemed favorable for both training and out-of-sample accuracy.

Table 3. *Correlation Coefficient Score of all Models for Train and Test Data.*

SN	Models	Train score	Test score
1.	Ridge Regression	0.372394	0.381738
2.	Lasso Regression	0.367822	0.37421
3.	Support Vector Machine	0.994892	-0.027344
4.	Decision Tree	1	0.90989
5.	Random Forest	0.984328	0.855492
6.	Extra tree Regression	1	0.861617
7.	Gradient Boosting Regression	0.868513	0.801631
8.	XGBoost	0.99873	0.839402

Additionally, Figure 3 illustrates the precision, accuracy, and consistency of the models' outcomes from a cross-validation analysis. Among all the supervised machine learning models, the decision tree model yielded the most consistent results compared to the actual RF values, as indicated by the RMSE results. Furthermore, the decision tree model was the most accurate for RF forecasts based on the MAE metric. The success of the decision tree model can be attributed to the utilization of the bias-variance concept in its constructionconcept

Fig3. *Comparison of Prediction Error of all Models based on MAE and RMSE*

3.2 Model Agnostic Analysis

Permutation importance using XGBoost was used to analyze feature importance and evaluate the impact of multicollinearity. Figure 4 highlights the strong correlation between pay zone thickness (h) and permeability(K) in predicting RF. Permuting input features decreased model accuracy, indicating the significance of these features.

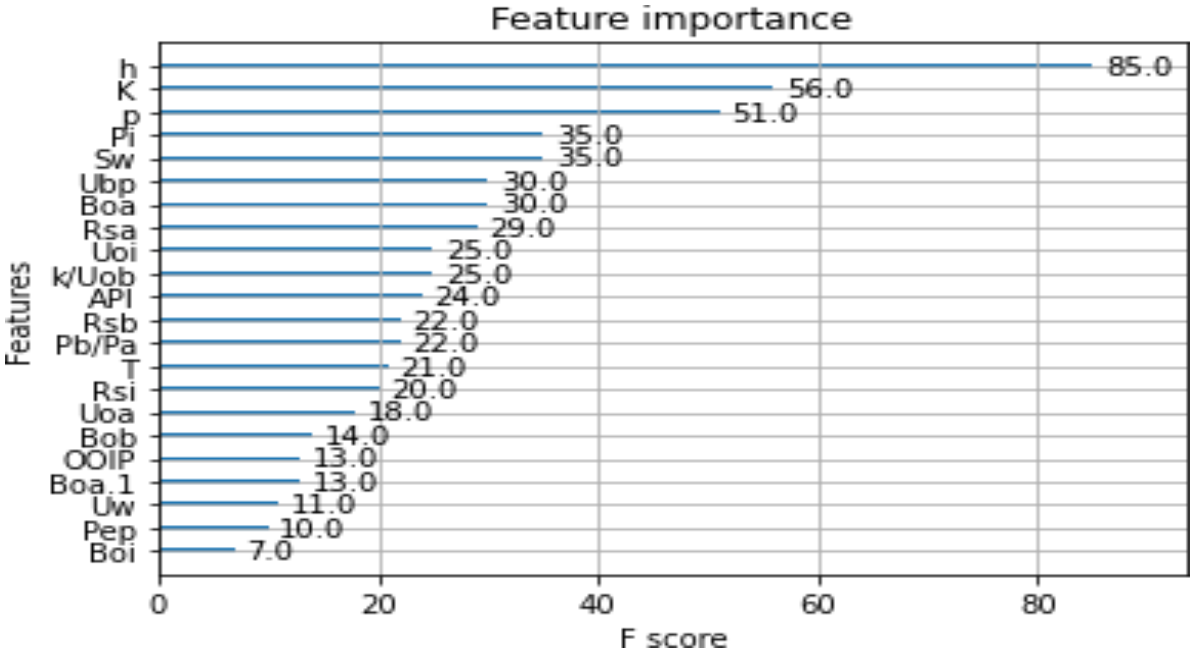

Fig4. *XGBoost Feature Importance of the Input Features to the Target*

3.2.1 Shapley Value Analysis

The decision tree model predicted the recovery factor using all features and the same datasets for training, testing, and feature significance analysis. Model agnostic metrics, like Shapley values, help explain how the model made predictions. Shapley values, originating from game theory, represent the impact of features on predictions. The feature significance plot in Figure 6 displays the absolute values of Shapley values, indicating the most relevant features with high absolute values. Permeability (K) was the most significant feature, with an average effect of 4% (0.04) on the target. API also had an impact. The plot revealed little connection between PFI results and the relationship of Boa, OOIP, and Pep to the target, requiring further research on characteristic differences. The permutation feature importance is defined by the decline in model performance, while Shapley values depend on feature magnitudes. The bee swarm plot provides more informative details than the feature importance plot.

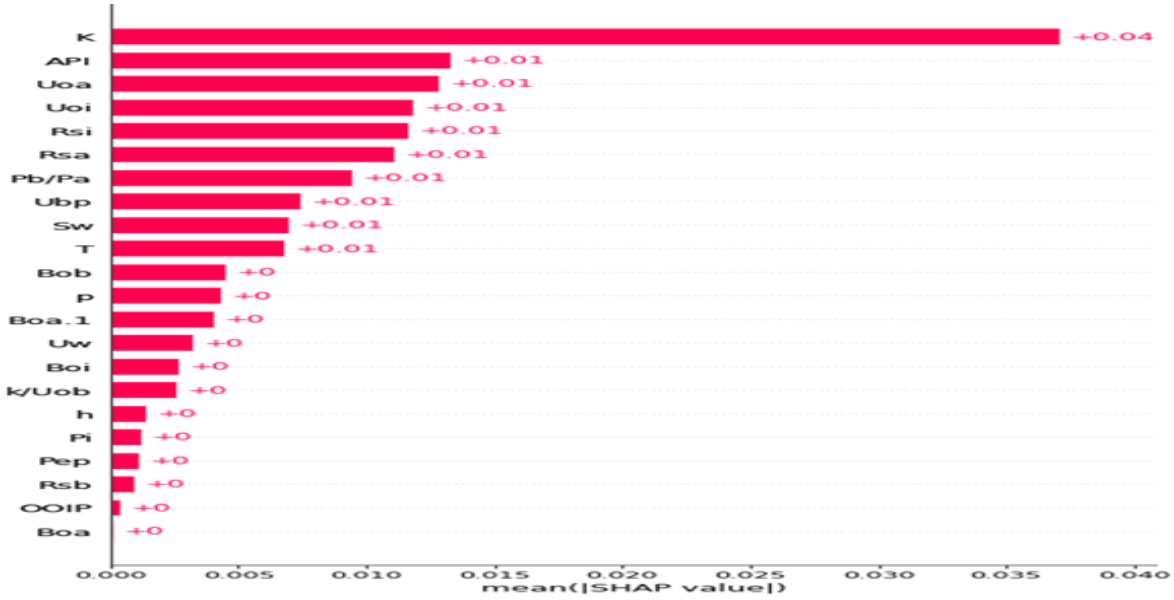

Fig5. *PFI Analysis of Data Input Parameters*

The bee swarm summary plot in Figure 6 displays the absolute values of the Shapley values for both the train and test datasets. The y-axis represents the features, while the x-axis represents the Shapley values. This summary figure provides the first cues to the positive and negative correlation between a feature's value and its effect on the target. The distribution of Shapley values for each feature is shown in this type of visualization, listed in descending order of significance. Blue denotes low feature values, while red denotes high feature values. Examining the impact of permeability reveals that low values predict high recovery factor values, while high values predict low recovery factor values. This research emphasizes the significance and necessity of model agnostic to comprehend the impact of other factors on recovery factor prediction. The bee swarm plot demonstrates that approximately 15 features can globally explain how the predictions were made.

The results of the bee swarm plot confirmed that features that have high importance in boththe train and test results exhibit their importance in the global explanation of the target. From the demonstrated results, the following analysis was derived.

1. Feature importance: The features are ranked in descending order, and from the train and test data plot, permeability is ranked first. The Boa feature has close to zero importance because it does not have any causal effect in predicting recovery factor; and
2. Impact: The horizontal location of the data points shows that permeability has a negative correlation and a high prediction effect in general.

3.3 Evaluation of Top Features

The top 5 performing models for the top 15 features based on PFI and Shapley values underwent additional analysis to see if they will outperform the initial model forecasts. PFI had the effect of eliminating OOIP, Rsb, Boa, K/Uob, Pi, and Pep. When Shapley values were applied, OOIP, Rsb,Boa, K/Uob, and Pep were eliminated.

When the features based on PFI and Shapley values were applied, all of the top-performing models performed better, as shown in Table 3. However, it was clear that Shapley values chose the most important features, which is mostly seen in the enormous gain in model accuracy.

Table 3. *Computed Accuracy on Test Data Using Top 5 Models*

SN	Models	Accuracy basedon all features	Accuracy basedon PFI's top 13 features	Accuracy based on Shap values top 13 features
1.	Gradient Boosting	0.8016	0.8057	0.8573
2.	Decision Tree	0.9127	0.9510	0.9971
3.	Random Forest	0.9000	0.9290	0.9634
4.	Extra Tree	0.9095	0.9453	0.9811
5.	XGBoost	0.9162	0.9276	0.9795

3.4 Model Optimization

The model parameters were optimized using Bayesian Optimization (BO) to improve performance. Figure 7. demonstrates the performance of the Decision tree model using the top 15 characteristics determined by Shapley values and the BO-DT model. Results were compared with findings from Gulstad et al. (1995) and Noureldine et al. (2016), who used different methods and models for recovery factor prediction with the same data. The selected features based on Shapley values significantly reduced the model error. MAE decreased by 91%, from 1.01 to 0.023, when using Shapley-selected features compared to the initial extra tree model. The RMSE for the decision tree model with all 15 features was 1.28, while the Shapley-selected features had an RMSE of 0.032, representing a 97% decrease. Hyperparameter adjustment with BO-DT further improved the decision tree, resulting in an MAE of 0.02 and an RMSE of 0.022. Overall, the Shapley selected features proved highly relevant, providing global generalization and improving model efficiency due to the bias-variance concept used in building the decision tree model.

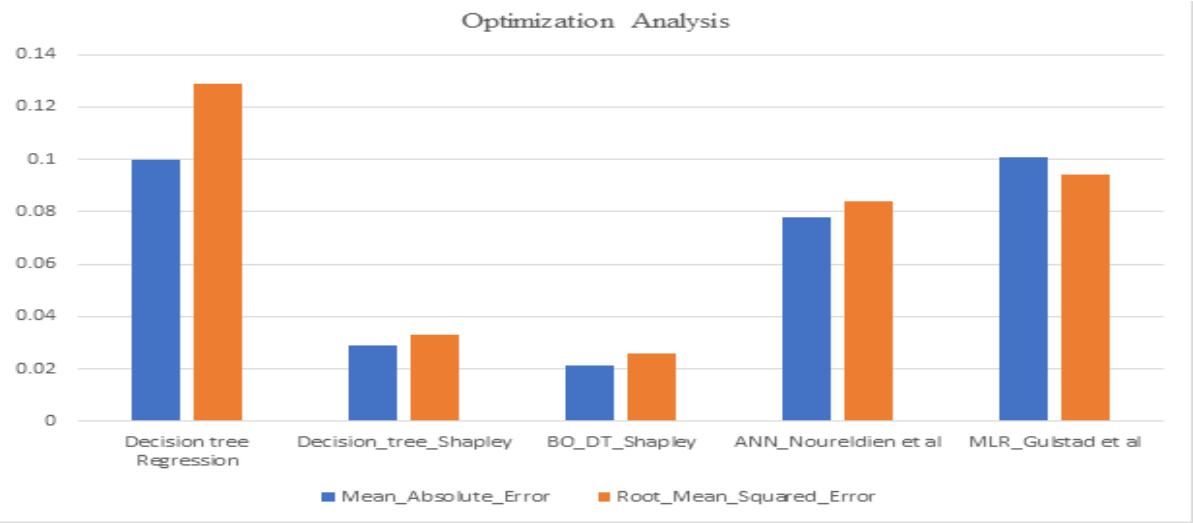

Fig7. *Comparison of RMSE and MAE for Top-Performing Models and Models by Previous Works on-Test Data*

3.5 Sensitivity Analysis

Figure 8. displays the kernel density estimation of expected and actual recovery factor data, with blue indicating expected values and yellow for actual values. The results indicate that the BO-DT model is considerably more precise than the actual data, as revealed by the test data. An analysis of errors reveals that the BO-DT model has a superior capability to capture a broad range of values, making it suitable for estimating the recovery factor in solution gas drive reservoirs for other wells. Consequently, the sensitivity analysis reinforces the previously established evaluation metrics.

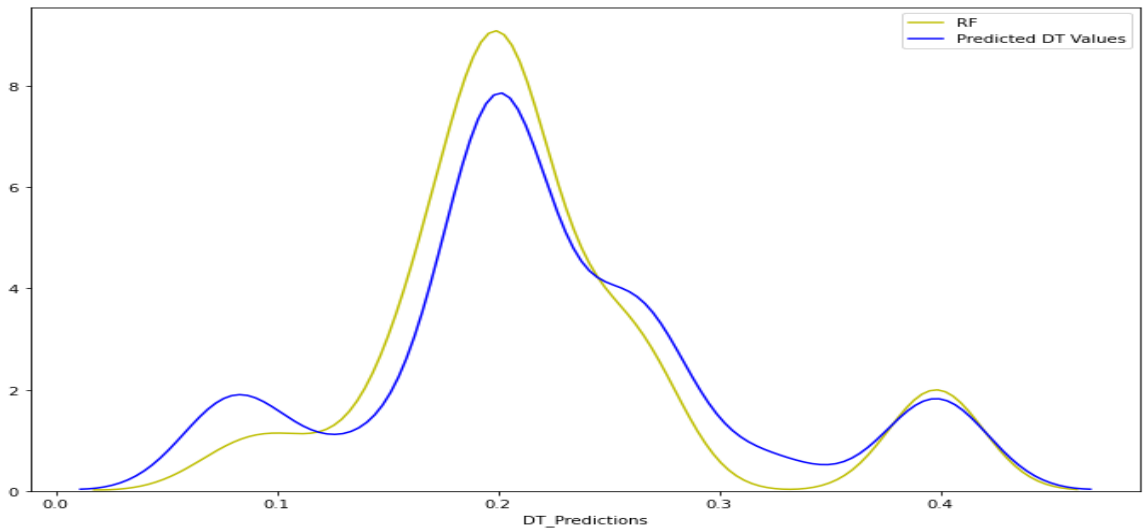

Fig8. *Kernel Density Estimation Showing Proximity of BO-DT Recovery Factor Prediction.*

4. CONCLUSION

a workflow for choosing relevant features for recovery factor predictions has been described in this study. The approaches assessed in this study showed that explainable AI can provide insight into the input features by using model-agnostic metrics. Simple statistical techniques might not be able to identify the causal effects of each input on the target because correlation does not imply causation. The models' likelihood of fitting the test data was assessed using the AIC result. On the test data, the R2, MAE, and RMSE were also compared for accuracy, consistency, and precision.

The input attributes are now explainable thanks to the models and how the target is anticipated using the PFI and Shapley values. When considering pertinent causal features, model agnostic has offered some reliable and practicalsolutions for recovery factor prediction.

5. ACKNOWLEDGMENT

The authors express their sincere appreciation to the University of Mines and Technology and the Department of Petroleum and Natural Gas Engineering for supporting this work.

REFERENCES

[1] Aliyuda, M., Kachalla, V., and Howell, J. (2019). “Machine-learning algorithm for estimating oil-recovery factor using a combination of engineering and stratigraphic dependent parameters”. Interpretation, 7(3), SE151–SE159.
[2] Arps, J. J., Aime, Member. (1955). “Reasons for Differences in Recovery Efficiency”.
[3] Breiman L. Random forests. Mach Learn 2001;45:5–32.https://doi.org/10.1023/A:1010933404324
[4] Chen, T., and Guestrin, C. (2016). “XGBoost: A scalable tree boosting system”. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13- 17-Augu, pp 785–794.
[5] Gordon, A.D, Breiman, L., Friedman, J.H, Olshen, R.A, Stone, C.J. “Classification and Regression Trees. Biometrics” 1984;40:874. https://doi.org/10.2307/2530946.
[6] Gulstad, L. R., (1995), “The Determination of Hydrocarbon Reservoir Recovery Factors by Using Modern Multiple Linear Regression”.
[7] Hoerl, A.E, Kennard, R.W, (1970), “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. Technometrics 1970;12:55–67. https://doi.org/10.1080/00401706.1970.10488634
[8] Lee, B. B., and Lake, L. W. (2015). Using data analytics to analyze reservoir databases. Proceedings - SPE Annual Technical Conference and Exhibition, 2015-January (September), pp 2481–2491.
[9] Noureldien, D. M, Ahmed, H. E, (2015), “Using Artificial Intelligence in Estimating Oil Recovery Factor”. Proceedings- SPE North Africa Technical Conference and Exhibition, 2015- Seprember, SPE-175867-MS.
[10] Otchere DA, Arbi Ganat TO, Gholami R, Ridha S. Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANNand SVM models. J Pet Sci Eng 2021;200:108–82. https://doi.org/10.1016/j.petrol.2020.108182.
[11] Shapley, L.S, (1970), “A Value for n-Person Games Contribution to Theory Games” (AM-28), Vol. II, vol. 2, Princeton University Press; 1953, p. 307–18. https://doi.org/10.1515/9781400881970-018
[12] Tibshirani, R.,(1996), “Regression Shrinkage and Selection Via the Lasso”. J R Stat Soc Ser B 1996;58:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[13] Vapnik, V., Lerner, A. (1963), “Pattern Recognition using Generalized Portrait Method”. Autom Remote Control 1963;24:774–80.

Citation: *Eric M. Amarfio., et..al.(2025).“Enhancing Hydrocarbon Recovery Factor Prediction Using Ensemble Machine Learning Workflow”, International Journal of Petroleum and Petrochemical Engineering (IJPPE), 10(1), pp.1-9, DOI: https://doi.org/10.20431/2454-7980.1001001.*