ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.2591

New approach to Estimate Oil Recovery Factor for Water Drive Sandstones Reservoirs through Applications of Machine Learning

MSc. Ing. Víctor Alexei Huerta Quiñones, Universidad Nacional de Ingeniería, Lima-Perú, ahuerta@fip.uni.edu.pe
Bs. Leonardo Adriano, Universidad Nacional de Ingeniería, Lima-Perú ladrianos@fip.uni.edu.pe
Bs. Gregory Zuñiga, Universidad Nacional de Ingeniería, Lima-Perú, gzunigaa@fip.uni.edu.pe
MSc. Ing. Wladimir Chavez Ormaza, Escuela Politécnica Nacional de Quito, wladimir.chavez@epn.edu.ec

Abstract*– Mature heavy oilfields in the Northern Peruvian Jungle have produced oil for over 40 years under waterdrive mechanism, with a wide range of ultimate recovery factor in between 10% to 60% depending on some rock and fluid properties like permeability, API gravity and oil viscosity; a reasonable estimation of recovery factor at an early stage of development and/or production is quite critical for sizing and scheduling development plans, as well as CAPEX investments. This research article introduces a new approach that integrates empirical correlations, analytical methods and machine learning algorithms to estimate oil recovery factor in water-drive sandstone reservoirs at early development and production. Preliminary studies showed that the most representative reservoir and fluid parameters, such as reservoir size, porosity, permeability, net thickness, residual oil saturation, API gravity, oil viscosity and initial pressure, which are typically measured during the exploration, appraisal and early development stages, can be correlated to expected recovery factors obtained from mature fields with long production history. Literature empirical correlations will be initially tested with available information of oilfields of Marañón Basin to estimate correlation coefficient; in addition, material balance and analogy estimations will be integrated to this analysis, and a ranking of the best-performing methods will be prepared. Different regression ML learning algorithms will be compared using existing data to select the best one providing the most precise predictions. A new empirical correlation that significantly outperforms traditional industry equations will be adjusted with ML algorithms weights and biases. The results of this comprehensive study will contribute to a better understanding of the water drive mechanism in the oilfields of the Northern Peruvian Jungle, as well as a more reliable Recovery Factor and EUR estimations at early development stages, which is critical for investment decisions on development plans before extensive production data is available; also, some guidelines may be established to optimize development plans and improve operational practices to increase oil production and recovery in "green oilfields".

Keywords*—Recovery factor, waterdrive mechanism, machine learning, empirical correlations, oil viscosity, rock permeability, early development, Marañón basin.

I. INTRODUCTION

Oil recovery factor (RF) is a key factor for all exploration and development (E&P) companies mainly during the early reservoir life, because several investment decisions are based on the amount of recoverable hydrocarbons, which could be obtained from the target asset with the available techniques & technology, as well as operational practices, under current economic scenario.[1]

In early stages, Recovery factor is commonly estimated with field analogy or empirical correlations. Guthrie and Greenberger (1955)[2] related oil recovery factor in reservoirs under water drive empirically to reservoir rock and fluid properties. The authors considered 73 sandstone reservoirs, mostly under waterdrive mechanism. Oil recovery was expressed as a function of permeability, porosity, oil viscosity, formation thickness and connate water saturation, as indicated by (Eq. 1).

$$R_o = 0.114 + 0.272\log(k) + 0.256(S_w) - 0.136\log(\mu_o) - 1.538(\varnothing) - 0.00035(h) \qquad (1)$$

Later, an API statistical study was developed by Arps, using a database from 312 reservoirs (Arps et al. 1967)[3] under a waterdrive mechanism to develop an empirical equation to predict recovery efficiency. The correlation was expressed as a function of permeability, porosity, oil viscosity, water viscosity, oil formation volume factor, initial reservoir pressure, abandonment pressure and connate water saturation, as indicated by (Eq. 2).

$$R_o = 0.549\left[\varnothing\,\frac{(1-Swc)}{\beta oi}\right]^{0.0422}\left[k\,\frac{\mu_{wi}}{\mu_{oi}}\right]^{0.0770}(\mathrm{Swc})^{-0.1903}\left(\frac{Pi}{Pa}\right)^{-0.2159} \qquad (2)$$

In addition, Gulstad (1995)[4] used multiple linear regression techniques to study the determination of the oil recovery factor for water-drive reservoirs in both sandstone and carbonate formations. He observed that the original oil in place is strongly

[1] Mohamed Al-Jifri1(2021). New proxy models for predicting oil recovery factor in waterflooded heterogeneous reservoirs.

[2] Guthrie, R.K.; Greenberger, M.H. (1955). The Use of Multiple-correlation Analyses for Interpreting Petroleum-engineering Data. In Proceedings of the Drilling and Production Practice

[3] Arps, J.J.; Roberts, T.G. (1967). The effect of the relative permeability ratio, the oil gravity, and the solution gas-oil ratio on the primary recovery from a waterdrive type reservoir.

[4] Gulstad, R.L. (1995). The Determination of Hydrocarbon Reservoir Recovery Factors by Using Modern Multiple Linear Regression Techniques.

correlated to oil viscosity (original and at bubble point), original oil in place, reservoir temperature, initial reservoir pressure, abandonment pressure and connate water saturation (see Eq,3).

$$R_o = -274.94 + 0.44(STOIIP) - 56.70\ln(\mu_{ob}) - 119.45\ln(S_w) + 0.04(p_{ep}) \\ -4.73(\mu_{oi}) + 4.38(\mu_{oa}) + 0.24(STOIIP_{calc}) - 0.88(T) \quad (3)$$

It is observed than Recovery factor is usually affected by several engineering and geological factors that can be grouped into four groups (asset size, rock parameters, fluid properties, and reservoir energy). The authors noted that their developed model outperformed all the available empirical equations in the literature.

Many aspects make the estimation of the RF very complicated, since no clear approach that considers all these aspects is available. Understanding all the technical and non-technical parameters associated with the reservoir nature, technologies in use, economic conditions, and other factors is necessary for the reserve evaluation process.

The relevance of the Northern Peruvian Jungle as a strategic region for the recovery of the upstream industry in Peru is considerable, due to the number of fields and oil resources (**see figure 1**), An increase in the recovery factor in mature oilfields of north Peruvian jungle is strongly dependent on proper completion technology, well placement and trajectories, availability of of heavy oil diluent, economic feasibility of oil transportation, and effective reservoir management strategies[5]; many of these factors are strongly influenced by a sustainable management of field operations.

It should be noted that in many cases, the recovery factors for Vivian and Chonta formation in light oilfields are above 35% (**see figure 2**), and there is still left and by-passed oil that can be recovered, particularly in heavy oilfields. Based on that, we would like to propose a new approach to incorporating ***machine learning algorithms into the traditional RF empirical correlations*** to build strong correlations for fields in early development and production in Marañón field, based on data ***from 25 mature oilfields producers*** in Marañón Basin, as well as analogous field database worldwide.

Hence, in order to promote better reservoir characterization studies and practices, as well as the application of new subsurface and surface technologies during the planning and development of these "green fields", the Peruvian Government should play an active role to incorporate some changes to the hydrocarbon regulations, including fiscal incentives and modifications in contractual agreements (royalty release policies)[6].

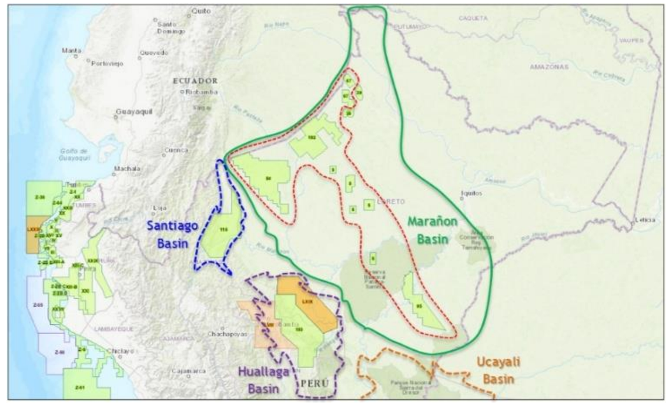

Figure 1: Marañón basin blocks 192, 64, 39, 67, 129, 123, 8 and 95 [7] have almost 1 billion barrels of oil resources.

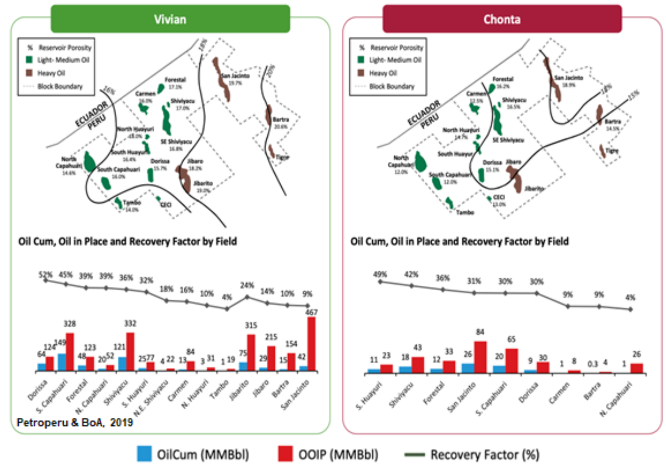

Figure 2: Current recovery factors by field in Block 192 oilfields for Vivian and Chonta reservoirs

II. Methodology

The methodological framework proposed in this study follows a structured workflow that integrates classical reservoir engineering concepts with modern data-driven techniques. The input database was compiled from 25 mature oilfields producing under water-drive mechanism, mainly located in the Marañón Basin, complemented with analogous worldwide fields. The selected variables correspond to parameters typically available during exploration, appraisal, and early development stages, including reservoir size, porosity, permeability, net pay thickness, residual oil saturation, oil viscosity, API gravity, and initial reservoir pressure. These variables were selected to ensure practical applicability of the methodology in early decision-making scenarios where production history is limited.

[5] *Babadagli, T. (2005). Mature Field Development – A Review. SPE 93884.*
[6] *Voskanian, M. (2009). Incentives to Revitalize Mature Fields in an Environmentally Safe Manner. SPE 121819.*
[7] *Perupetro website: www.perupetro.com.pe, North Peruvian Blocks Map.*

Prior to model development, an extensive data preprocessing stage was conducted to improve robustness and prediction reliability. This stage included data cleaning, detection and removal of outliers, consistency checks, and normalization of continuous variables to ensure comparable scales among predictors. The dataset was then randomly divided into training and testing subsets to avoid information leakage and to guarantee unbiased evaluation of model performance. This step is critical in machine learning applications, as the quality of input data strongly controls the generalization capability of the predictive models.

A statistical and correlation-based feature analysis was performed to identify the most influential variables affecting the recovery factor under waterdrive mechanism. Pearson correlation coefficients and exploratory data analysis were applied to detect redundant or weakly informative variables. This process ensured that the final model retained physically meaningful predictors while reducing dimensionality and overfitting risks. The selected features reflect fundamental reservoir engineering principles, particularly the impact of rock quality and fluid mobility on displacement efficiency.

Several supervised regression algorithms were implemented and compared, including linear regression-based methods and nonlinear ensemble techniques. Each model was trained using the same training dataset to ensure consistency in performance comparison. Hyperparameters were optimized through iterative testing and cross-validation. Among all evaluated algorithms, the Random Forest regressor demonstrated superior capability in capturing nonlinear interactions between reservoir and fluid properties, which are commonly observed in water-drive sandstone reservoirs.

Model performance was evaluated using standard statistical metrics, including the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). Cross-validation was applied to assess model stability and generalization across different data subsets. Classical empirical correlations, such as Guthrie, API, and Gulstad equations, were also tested using the same dataset to establish a direct benchmark against machine learning predictions. This comparative framework allowed a transparent evaluation of the added value introduced by data-driven models.

Beyond predictive accuracy, model interpretability was addressed using feature importance analysis and SHAP (Shapley Additive Explanations). These tools were applied to the best-performing algorithm to verify that the predictions were consistent with reservoir engineering physics. The Random Forest model not only achieved the highest accuracy but also provided physically coherent explanations of variable influence, leading to its selection as the optimal predictive model for recovery factor estimation in water-drive reservoirs.

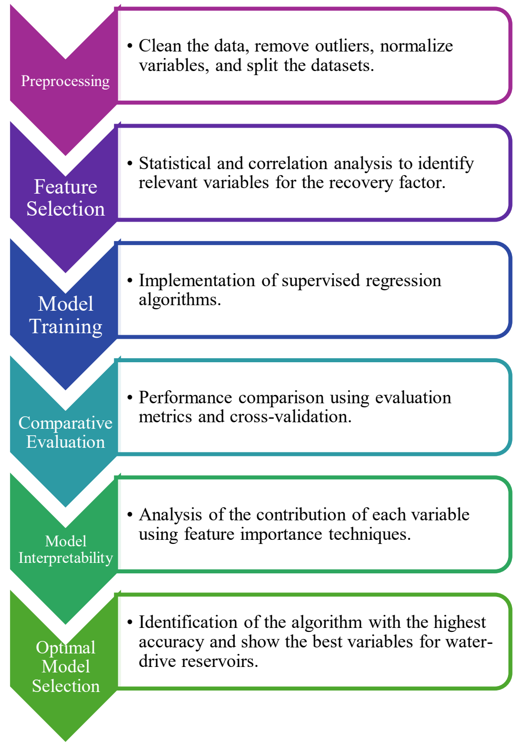

Figure 3: Methodological workflow for recovery factor estimation.

III. RESULTS

The results show a significant improvement in prediction accuracy when machine learning models are compared to traditional empirical correlations. Classical methods yielded R^2 values close to 0.40, indicating limited capability to generalize across heterogeneous reservoirs. In contrast, supervised machine learning algorithms substantially increased predictive performance, with the Random Forest model achieving R^2 values exceeding 0.94, demonstrating excellent agreement between predicted and observed recovery factors.

Graphical comparison of model predictions reveals that Random Forest consistently outperforms other regression techniques across the entire range of recovery factors. While linear and polynomial-based models tend to underestimate recovery at higher values, Random Forest accurately captures nonlinear trends and complex variable interactions. This behavior confirms the suitability of ensemble tree-based models for reservoirs characterized by strong geological and fluid heterogeneity.

Scatter plots of predicted versus observed recovery factors show that Random Forest predictions closely follow the 45-degree reference line, indicating minimal bias and dispersion. Other algorithms exhibit larger scatter and systematic deviations, particularly at low and high recovery extremes. This result highlights the robustness of Random Forest in handling nonlinearity and variable interactions without requiring explicit functional assumptions.

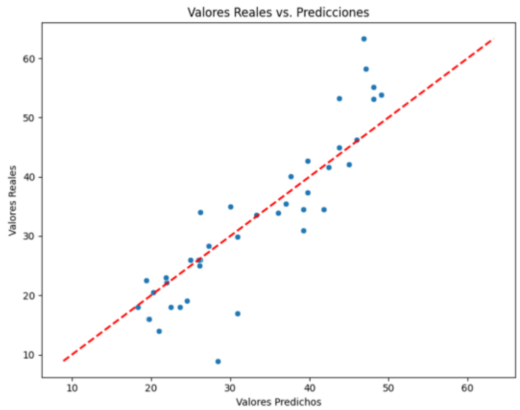

Figure 4: Comparison of recovery factor prediction accuracy.

Feature importance analysis indicates that permeability and oil viscosity are the most influential variables controlling recovery factors, followed by porosity and net thickness. Table 1 shows the statistics of the dataset.

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
Permeability (md)	1,342.7	1,000.0	1,080.9	40.0	4,000.0
Oil Viscosity (cp)	20.4	4.0	40.5	0.4	224.0
OOIP (MM STB)	407.5	95.0	1,491.6	9.4	9,626.0
Net pay (ft)	53.8	38.0	50.6	7.9	231.0
Initial Water Saturation (%)	28.7	30.8	9.0	16.0	56.1
Initial Oil Saturation (%)	47.9	61.3	23.7	16.0	84.0
Porosity (%)	22.1	20.0	7.1	12.0	34.0
Initial Pressure (psia)	3,214.2	4,030.0	1,733.0	486.0	5,719.0
Abandonment Pressure (psia)	2,475.9	3,143.3	1,318.8	376.7	4,117.7
Relative Permeability Oil-Water	297.5	143.0	420.7	4.0	2,000.0
$\sqrt{\text{Permeability/Porosity}}$	71.4	66.9	33.0	17.7	148.3
Real Recovery Factor (%)	32.3	30.9	13.6	8.9	63.3

Table 1: Dataset preparation.

These findings are fully consistent with displacement efficiency theory in water-drive reservoirs, where high permeability enhances sweep efficiency and oil viscosity controls mobility ratio. The Random Forest model successfully quantifies these effects while accounting for their nonlinear interactions.

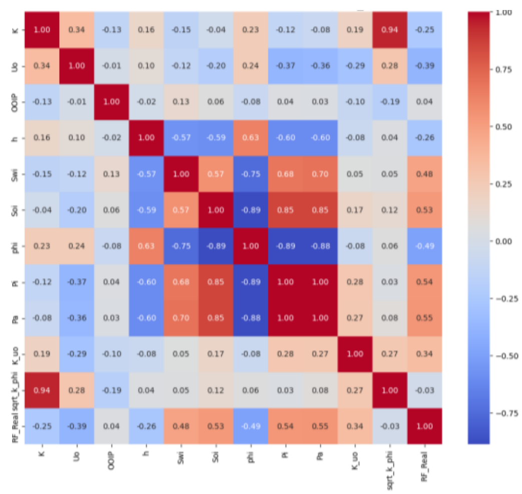

Figure 5: Heatmap of the variables.

When applied to datasets from different geological provinces, the model identifies distinct regional patterns. For instance, viscosity-dominated behavior is observed in heavy oilfields, while permeability-driven recovery dominates in clean sandstone reservoirs. These results demonstrate the model's capability to capture both global trends and regional specificities, reinforcing its applicability as a screening tool in early field development.

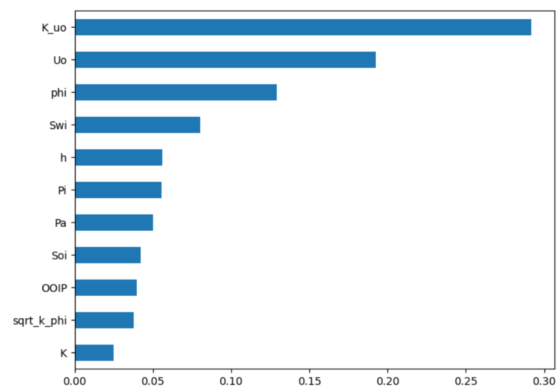

Figure 6: Feature importance derived from Random Forest model.

IV. Discussions

The success of machine learning models lies not only in their mathematical capabilities but also in their alignment with the fundamental principles of reservoir engineering. Variable importance analysis using SHAP (Shapley Additive Explanations) and feature permutation validated the physical logic of the predictions.

A. Non-Linearity in Reservoir Behavior

The comparative analysis reveals a critical insight: traditional empirical correlations, often based on linear or log-linear regressions, are insufficient for the heterogeneous reservoirs of the Northern Peruvian Jungle. The transition from Linear Regression (RMSE 12.16) to Random Forest (RMSE 6.77) represents a 44% reduction in prediction error. This improvement confirms that the relationship between rock/fluid properties (e.g., k, ϕ, μ_o) and the Recovery Factor is inherently non-linear and multi-dimensional.

B. Non-Linearity in Reservoir Behavior

The results from the SHAP analysis align perfectly with fundamental reservoir engineering principles, specifically Darcy's Law and Fractional Flow Theory:

- The dominance of the k_uo ratio (Permeability / Viscosity) as the primary driver is physically consistent for water-drive reservoirs. This ratio serves as a proxy for mobility, determining how easily oil moves relative to the displacing water.
- The model autonomously "learned" that dynamic flow parameters (k, μ_o) are more critical for determining the ultimate recovery efficiency than purely static volumetric parameters like Initial Oil in Place (OOIP) or Net Pay (h).

V. CONCLUSIONS

The integration of supervised machine learning models represents a significant advance in predicting the recovery factor for reservoirs with hydraulic drive. It was demonstrated that:

- Machine learning models, specifically Random Forest, consistently outperform traditional empirical correlations (Guthrie, API, Gulstad) in terms of accuracy and generalizability. While classical correlations exhibit an R^2 value close to 0.40, ML models achieve values exceeding 0.86.

- Permeability (K) and oil viscosity (μo) are the physical parameters with the greatest impact on displacement efficiency. Machine learning models successfully identify and quantify the nonlinear interaction between these parameters.

- Applying a common predictive framework to geologically diverse basins, such as the Maranon (foreland basin) and the Albertine (rift basin), allows for the identification of patterns in hydraulic drive behavior, while also highlighting regional particularities, such as the impact of crude oil viscosity in Uganda versus the high permeability of quartz sandstones in Peru.

- The computational efficiency of machine learning models allows their use as rapid assessment tools in exploration and early development stages, providing a low-cost alternative with comparable accuracy to complex numerical simulations when data is limited.

VI. LIMITATIONS AND FUTURE WORK

Despite the accuracy achieved, machine learning models exhibit a critical dependence on the quality and representativeness of the training dataset. Therefore, applying a model trained in a specific region to a new geological province requires learning transfer processes to adjust the weights and biases to local conditions. In addition, the current model relies on initial reservoir conditions (P_i, S_{oi}). It does not account for dynamic operational changes such as production rates, well spacing, or artificial lift mechanisms implemented over the field's life.

Future work should focus on integrating hybrid models that combine physical laws (Physics-Informed Neural Networks) with data algorithms, ensuring that predictions always respect the principles of mass and energy conservation. Furthermore, incorporating dynamic, real-time production data and 4D seismic characterization parameters could enable adaptive prediction of the recovery factor throughout the field's productive life.

To enhance the robustness and applicability of this research, the following steps are proposed:

1. **Probabilistic Assessment**: Implement Quantile Regression Forests to generate prediction intervals (P10, P50, P90) rather than point estimates. This would allow engineers to quantify the uncertainty and risk associated with the reserve booking.

2. **Physics-Informed Machine Learning (PIML)**: Develop hybrid models that constrain the Machine Learning predictions using material balance equations, ensuring that predicted Recovery Factors do not violate physical limits (e.g., $\mathrm{RF} < 1 - \mathrm{S_{wi}} - \mathrm{S_{or}}$).

3. **Expansion of Feature Space**: Incorporate geological attributes (e.g., depositional environment labels) and operational parameters (e.g., well density) to capture the impact of reservoir management on the ultimate recovery.

ACKNOWLEDGMENT

We express our sincere gratitude to the Faculty of Petroleum, Natural Gas and Petrochemical Engineering of the National University of Engineering (Universidad Nacional de Ingeniería – UNI) for its continuous academic support and for providing an environment that promotes research and innovation in reservoir engineering. Special recognition is extended to the TRM Research Group for their technical guidance, constructive discussions, and valuable contributions throughout the development of this study, which were fundamental to strengthening both the methodological framework and the interpretation of results.

REFERENCES

[1] Mohamed Al-Jifri1(2020). New proxy models for predicting oil recovery factor in waterflooded heterogeneous reservoirs. Journal of Petroleum Exploration and Production Technology (2021). https://doi.org/10.1007/s13202-021-01095-4.

[2] Guthrie, R.K.; Greenberger, M.H. (1955). The Use of Multiple-correlation Analyses for Interpreting Petroleum-engineering Data. In Proceedings of the Drilling and Production Practice; American Petroleum Institute: Washington, DC, USA, 1955; pp. 130–137.

[3] Arps, J.J.; Roberts, T.G. (1967). The effect of the relative permeability ratio, the oil gravity, and the solution gas-oil ratio on the primary recovery from a waterdrive type reservoir.

[4] Gulstad, R.L. (1995). The Determination of Hydrocarbon Reservoir Recovery Factors by Using Modern Multiple Linear Regression Techniques. Master's Thesis, Texas A&M University, College Station, TX, USA, 1995.

[5] Babadagli, T. (2005). Mature Field Development – A Review. SPE 93884.

[6] Voskanian, M. (2009). Incentives to Revitalize Mature Fields in an Environmentally Safe Manner. SPE 121819.

[7] Perupetro website: www.perupetro.com.pe, North Peruvian Blocks Map.