*Original Research Article*

HYBRID PREDICTIVE MODELING OF WET GAS PRESSURE-VOLUME-TEMPERATURE PROPERTIES FOR NIGER DELTA RESERVOIRS

.

ABSTRACT

|  |
| --- |
| Pressure-Volume-Temperature (PVT) properties has always been a major interest to reservoir engineers for fluid characterization, reserves estimation and recovery. Existing PVT correlations often struggle to capture the intricate Niger Delta gas reservoirs. To address this gap, this study explores regression algorithms for the development of selected PVT properties of a wet gas reservoir in Niger Delta.  This work integrate both black box and white box modeling techniques to develop a hybrid predictive analytics models which formed the basis for optimal feature selection for the development of a simplified correlation from a blend of two linear models to predict PVT properties of Gas Compressibility factor (z-factor), Gas Formation Volume Factor (Bg) and Gas Viscosity (μg). The modelling architecture employs supervised machine learning regression algorithms that are developed on 1,111 wet gas data points with a 5-fold cross-validation technique. Statistical metrics of evaluation was used to validate performance.  Findings revealed that pseudopressure and gas viscosity are major determinants for gas z-factor, while gas density, gas viscosity, and pseudopressure are crucial for Bg. The hybrid models achieved AARE, R² and RMSE scores of 0.69%, 99.95% and 0.0067 for z-factor, 0.33%, 99.22% and 0.00023 for Bg, and 0.92%, 99.97% and 0.0056 for μg, respectively. Additionally, the developed mathematical correlations yielded R² of 99.68%, 94.67%, 99.45% for z-factor, Bg, and μg, respectively.  Hybrid wet gas Pressure-Volume-Temperature correlations from regression algorithms was developed for the Niger Delta region. The models ensures a contextualized and reliable representation of the region by reducing biases and improving correlation accuracy. |

*Keywords: Hybrid predictive model, PVT properties, Black box algorithm, Wet gas properties, Statistical metrics, White box algorithm*

1. INTRODUCTION

The understanding of hydrocarbon fluid PVT properties within the subsurface has played a significant role within the oil and gas industry towards the effective recovery of commercial hydrocarbon constituents in underlying reservoirs. In the realm of reservoir fluid behavior analysis, Pressure-Volume-Temperature (PVT) labs play a pivotal role (Adagunodo et al., 2022). Researchers (Ahmadi and Ebadi, 2014; Baniasadi et al., 2015; Dutta and Gupta, 2010; Elsharkawy, 1998; Freyss et al., 2013) deploy an array of instruments within PVT labs to decipher the behavior and properties of oil and gas samples extracted from reservoirs. Notably, Freyss et al. (2013) provide insights into PVT analysis for oil reservoirs, emphasizing the significance of tools such as high-pressure, high-temperature PVT cells for precise determination of fluid characteristics. In gas reservoirs, the PVT properties of the hydrocarbon fluids, including the compressibility, density, and phase behavior, dictate the reservoir's response to production activities. The complex interplay of these properties influences key factors such as fluid flow, phase transitions, and the ultimate recovery of hydrocarbons. Accurate characterization of PVT properties is, therefore, indispensable for reservoir engineers seeking to maximize production rates while ensuring sustainable reservoir performance.

**1.1 PVT System for Gas Reservoirs**

Gas reservoirs exhibit dynamic PVT behaviors due to factors such as temperature gradients, pressure differentials, and compositional variations. They play a pivotal role in meeting the world's increasing energy demands, and accurate modeling of their Pressure-Volume-Temperature (PVT) behavior is crucial for effective reservoir management. Several crucial gas PVT properties play a pivotal role in understanding and predicting the behavior of hydrocarbons under varying conditions. These properties are fundamental for characterizing gas reservoirs and formulating accurate models for reservoir management. Here are some key gas PVT properties: temperature, pressure, composition, formation volume factor, compressibility factor, viscosity, phase behavior, specific gravity and heat capacity.

Traditional PVT models, although valuable, often face challenges in accurately representing the complex behavior of gas reservoirs (Osmah et al., 2001). The conventional approaches rely on simplifications and assumptions that may not capture the intricacies of real-world reservoir conditions. As a result, discrepancies between model predictions and actual reservoir performance are common, leading to suboptimal production and recovery strategies. The challenges in traditional PVT modeling stem from the dynamic nature of gas reservoirs, which exhibit non-ideal behaviors such as phase transitions, compositional variations, and complex fluid interactions. Additionally, uncertainties in reservoir data further contribute to inaccuracies in predicting reservoir performance. These limitations highlight the need for advancements in PVT modeling techniques to enhance the precision of gas reservoir correlation.

This research seeks to address the existing gaps in gas reservoir PVT modeling by exploring the integration of advanced machine learning regression techniques. Through a comprehensive analysis of reservoir data and the application of sophisticated algorithms, the aim is to enhance the accuracy and reliability of PVT models. By doing so, this study contributes to the broader goal of optimizing gas reservoir management, improving production forecasts, and ultimately maximizing the recovery of valuable hydrocarbons. The dual challenge lies in the development of specialized PVT models tailored for gas reservoirs and the creation of a region-specific model based on the unique conditions of the Niger Delta reservoir. By addressing these issues, the research endeavors to contribute significantly to the advancement of PVT modeling techniques for gas reservoirs and improve the accuracy of reservoir correlation, particularly within the distinctive geological context of the Niger Delta.

**1.2 Modeling PVT Gas Properties for Niger Delta Region**

The reliance on PVT correlations from non-Niger Delta reservoirs for PVT modeling of Niger Delta reservoirs introduces uncertainties and limitations. The Niger Delta, characterized by its diverse geological and fluid properties, warrants reservoir correlation or model to be derived directly from its own wells. Utilizing laboratory data from other regions may lead to inaccuracies in prediction, as the unique geological and fluid characteristics of the Niger Delta could significantly differ from those in other locations. To address this challenge, this work developed PVT models grounded in data obtained exclusively from the Niger Delta reservoir, providing a more contextually accurate representation, and enhancing the reliability of gas reservoir correlation within this specific region.

**1.3 Traditional and Mathematical Approaches to Gas PVT Prediction**

**Black Oil Models:** these simplified reservoir fluid models that categorize reservoir fluids into black oil, volatile oil, and gas based on API gravity.

Limitations: They are not suitable for highly compositional systems and may oversimplify fluid behavior, leading to inaccuracies in phase behavior predictions.

**Laboratory Experiments:** this is a direct measurement of PVT properties through laboratory experiments, such as constant composition expansion or differential liberation.

Limitations: These methods are time-consuming, expensive, and often impractical for obtaining real-time data from the reservoir. Additionally, the data obtained might not fully represent the reservoir conditions.

**Empirical Correlations:** this are correlations based on empirical relationships to estimate fluid properties from limited data.

Limitations: They are often specific to certain fluid types and may lack accuracy when applied outside their intended range. Limited adaptability to variations in reservoir conditions is a challenge.

**Standing's Correlation:** developed by Standing in 1947, this empirical correlation estimates gas formation volume factor (Bg) based on pressure, temperature, and gas gravity.

Limitations: It assumes constant gas gravity and is limited to specific pressure and temperature ranges. It may not accurately capture the behavior of gases with varying compositions and under extreme conditions.

**Thermodynamic Models:** such as Peng-Robinson, Soave-Redlich-Kwong (SRK), Redlich-Kwong-Soave (RKS) etc. These are equation of state used to predict phase behavior and PVT properties of hydrocarbon mixtures (Voutsas et al, 2018).

Limitations: It requires extensive input data, and the accuracy diminishes for highly non-ideal systems. There can be challenges in accurately characterizing complex fluid compositions.

## **1.4 Inclusion of Machine Learning in PVT Modelling**

Recent years have witnessed a surge in the application of Machine Learning (ML) techniques (Khoukhi and Albukhitan, 2011; Ali et al., 2013; Khosravi et al., 2018; Azizi et al., 2019; Deumah et al., 2021; Hamid et al., 2021; Rezaei et al., 2023) in the petroleum sector for the prediction of PVT fluid properties. Machine learning regression algorithms, offers a promising avenue for improving the accuracy of PVT models (Mohamadi-Baghmolaei et al., 2015; Oloso et al., 2016; Ramirez et al., 2017; Sola-Aremu, 2019; Uzogor and Akinsete, 2020; Xi et al., 2020; Rezaei et al., 2022). The ability of these ML techniques to identify complex patterns within large datasets (Sola-Aremu, 2019; Ikpabi and Akinsete, 2022) aligns well with the intricate nature of gas reservoir behavior. By leveraging advanced regression algorithms, there is a potential to develop more robust and reliable PVT models that can better represent the diverse conditions encountered in gas reservoirs.

Machine learning (ML), a transformative branch of artificial intelligence, revolutionizes the prediction and modeling of complex systems, offering unprecedented adaptability to the challenges in PVT modeling for gas reservoirs. Unlike traditional rule-based programming, ML enables algorithms to autonomously learn patterns from data, making it highly effective for capturing the nuanced relationships inherent in gas reservoir systems.

Supervised ML, a cornerstone of this research, involves training a model on a labeled dataset, where the model learns to map input features to corresponding output values. In the realm of PVT modeling, supervised ML brings unparalleled efficiency and accuracy. By harnessing a diverse array of features encompassing PVT properties, compositional details, and petrophysical attributes, the model becomes adept at discerning intricate correlations.

2. methodology

The objective of this study is to develop a comprehensive model and correlation that can be used in the determination of selected key PVT gas properties that can be implemented in diverse gas applications like reservoir simulation, material balance calculations, and effective recovery techniques. Supervised machine learning algorithms was used to develop a correlation for the prediction of Gas Formation Volume factor, Gas Compressibility Factor (z-factor) and Gas Viscosity.

### **2.1 Research Design**

This study adopts a multifaceted approach, integrating both quantitative and qualitative methodologies to comprehensively address the assigned objectives concerning gas reservoir PVT properties. This dual approach ensures that conclusions drawn from the study are well-justified and align with fundamental principles governing gas reservoir systems. Recognizing the intricate nature of these properties, characterized by their complexity and multifactorial determinants, a dual approach is essential to ensure robust conclusions aligned with fundamental principles governing gas reservoir systems.

Quantitative methods, including mathematical modeling and data analysis, are pivotal in providing precise numerical estimations of gas reservoir properties. These calculations serve as the backbone for validating the study's findings, offering verifiable and reproducible results that bolster the credibility of the research outcomes. Furthermore, within the quantitative domain, the study delves into the development of mathematical correlations, leveraging both linear and non-linear models. While linear models offer foundational insights, the incorporation of more advanced predictive techniques, such as tree-based and ensemble models, enriches the predictive capabilities, providing nuanced insights into feature importance and non-linear relationships within the data.

Conversely, the qualitative aspect of the research focuses on interpreting the results in the context of fundamental principles underlying gas reservoir systems. By examining the underlying mechanisms and processes at play, the study aims to provide deeper insights into the behavior of gas reservoirs beyond numerical predictions alone. This qualitative analysis helps to validate the quantitative findings and ensures that they are grounded in sound scientific reasoning. In a bid to create a representative formula or model, most researchers are leaned to either one of these and are limited to the evaluation metrics as basis of study emphasis overlooking the potential synergies of a blend of both linear and non-linear model for my insights into feature importance of the data, by exploring the feature importance of the non-linear model a better interpreted view is portrayed of how the model is created, that is, what features played the most important roles towards the prediction of either the z factor, gas viscosity or the gas formation volume factor. These insights can include identifying influential factors, understanding non-linear relationships, and discovering previously unrecognized correlations within the data. Such qualitative insights can be valuable for refining reservoir models, improving understanding of reservoir behavior, and guiding decision-making processes.

### **2.1.1 Black Box Model**

In the context of predictive modeling, a black box model refers to a computational model or algorithm whose internal workings are opaque or not easily interpretable. Black box models focus solely on optimizing predictive performance, often at the expense of interpretability. The purpose of utilizing this model in gas PVT properties prediction is that the relationships between input parameters and target variables are often intricate and non-linear. Traditional linear models may struggle to capture these complex data patterns adequately. By contrast, black box models excel at modeling non-linear relationships, enabling them to capture subtle interactions and dependencies among variables. This capability is particularly advantageous in gas reservoir engineering, where understanding the nuanced interplay of various factors is crucial for accurate predictions. The black box algorithms used in this work are random forest, and lightGBM algorithm (ensemble-based algorithms).

### **2.1.2 White Box Model**

A white box model, also known as a transparent or interpretable model, refers to a computational model or algorithm whose internal workings are easily understandable and interpretable by humans. White box models prioritize transparency and explicability, providing clear insights into the relationships between input variables and model predictions (Ikpabi and Akinsete, 2022). White box models, such as linear regression, decision trees, or generalized linear models, offer straightforward interpretations of their predictions. By explicitly representing the relationship between input variables and model outputs, these models provide insights into the factors influencing gas PVT properties. This interpretability is invaluable in gas reservoir engineering, where stakeholders require clear explanations of model predictions to inform decision-making processes. Also, white box models facilitate the identification of causal relationships between input variables and model predictions. By examining the coefficients, feature importance of the model, researchers can discern the relative importance of different factors in influencing gas PVT properties. This causal understanding is crucial for elucidating the underlying mechanisms driving reservoir behavior and guiding effective reservoir management strategies.

**2.2 Dataset Description and Summarization**

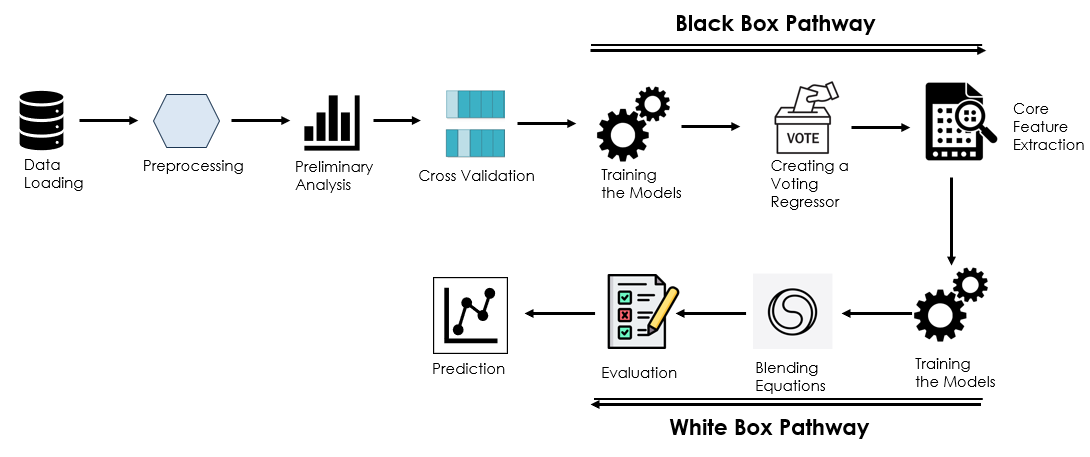
The dataset (Tables 1 and 2) utilized in this study was a combination of PVT properties from a Niger Delta wet gas reservoir. The Count in Tables 1 and 2 indicates the number of non-null values for each variable. In this dataset, there are 1,111 non-null values for each variable, suggesting there are no duplicates in the dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1** Summarization 1 | | | | | | |
|  | **Temperature** | **Pressure** | **z Factor** | **Gas FVF** | **Gas Viscosity** | **Gas Density** |
| **Count** | 1111 | 1111 | 1111 | 1111 | 1111 | 1111 |
| **Mean** | 120 | 2750 | 0.8447 | 0.0057 | 0.5928 | 45.7866 |
| **Std** | 18.9822 | 1312.5550 | 0.2962 | 0.0025 | 0.3284 | 9.2366 |
| **Min** | 90 | 500 | 0.3116 | 0.0043 | 0.0132 | 10.1354 |
| **25%** | 102 | 1625 | 0.5768 | 0.0046 | 0.3322 | 43.3023 |
| **50%** | 120 | 2750 | 0.832 | 0.0049 | 0.6041 | 48.8061 |
| **75%** | 138 | 3875 | 1.1044 | 0.0056 | 0.8518 | 51.9178 |
| **Max** | 150 | 5000 | 1.3959 | 0.0238 | 1.328 | 55.6477 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 2** Summarization 2 | | | | | |
|  | **Pseudopressure** | **Water FVF** | **Water Viscosity** | **Water**  **Density** | **Water**  **Compressibility** |
| **Count** | 1111 | 1111 | 1111 | 1111 | 1111 |
| **Mean** | 94666930.78 | 1.00257 | 0.64008 | 62.27225 | 2.92E-06 |
| **Std** | 24338986.64 | 0.00607 | 0.11781 | 0.37663 | 1.85E-08 |
| **Min** | 33780500 | 0.98909 | 0.47685 | 61.3762 | 2.89E-06 |
| **25%** | 78468700 | 0.99812 | 0.529 | 62.0043 | 2.91E-06 |
| **50%** | 94600800 | 1.00243 | 0.62448 | 62.2789 | 2.92E-06 |
| **75%** | 112382500 | 1.00687 | 0.74772 | 62.5475 | 2.93E-06 |
| **Max** | 147556000 | 1.01717 | 0.85087 | 63.1186 | 2.97E-06 |

**2.3 Overall Model Framework**

The architecture design (Figure 1) for the prediction of gas PVT properties (z-factor, gas formation volume factor, and gas viscosity) in this work employs a comprehensive approach integrating both black box and white box modeling techniques. This dual-pathway framework ensures robustness and interpretability in the predictive models developed. Each component of the methodology is structured to refine and validate the predictive capability at various stages.



**Figure 1: Model Architecture**

**2.3.1 Data Loading**

The initial phase in the predictive modeling process involves loading the requisite data set. This dataset is critical as it contains the empirical data necessary to train and validate the predictive models. The data encompasses various properties and conditions under which the gas PVT properties have been previously observed and recorded. A Niger Delta gas reservoir dataset of 1,111 x 11 dimension is utilized in this research.

**2.3.2 Preprocessing**

Once the data is loaded, it undergoes preprocessing to ensure quality and consistency. This step typically involves cleaning the data, dropping duplicates (to reduce redundancy), handling missing values (for which none was found during the course of preprocessing) and selecting features that are most relevant to the predictive tasks at hand (further defined in later steps). The goal is to prepare a refined dataset that will facilitate effective model training.

**2.3.3 Preliminary Analysis**

A preliminary analysis follows, where exploratory data analysis techniques are employed to uncover trends, patterns, and distributions within the data. This stage is crucial for understanding the underlying structure of the data and for providing the appropriate theoretical justification to the subsequent modeling steps.

**2.3.4 Cross Validation**

To ascertain the robustness and generalizability of the model, cross-validation (Isemin and Akinsete, 2024) is conducted. This process involves partitioning the data into multiple subsets, allowing the model to be trained on one subset and validated on another, then the metrics from each subset is averaged. This technique helps in mitigating overfitting and ensures that the model performs well on unseen data. The K-Fold technique is used in this research. It is a technique used in regression problems to assess the performance of a machine learning model, especially when data is limited. It involves partitioning the dataset into k subsets of equal size. The model is trained k times, each time using k-1 folds as training data and one fold as validation data. Performance is evaluated using metrics like Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) on the validation set. Results from each iteration are averaged to provide an overall assessment.

**2.3.5 Training the Models (Black Box Pathway)**

In the black box pathway, multiple models are trained. This ensemble approach leverages different algorithms to capture various aspects and relationships within the data. Each model may focus on different features or relationships, providing a comprehensive understanding that enhances the overall predictive performance. A voting regressor is created for post-training. This regressor aggregates the predictions from decision tree, random forest and lightgbm models, employing a voting mechanism to finalize the prediction. This step enhances the prediction accuracy by reducing the variance of individual model predictions.

Simultaneously, core feature extraction is carried out to identify and utilize crucial features that significantly impact the PVT properties. This is done by using feature importance of each model and combining the report. This step is pivotal for the white box pathway, where interpretability and direct understanding of model decisions are essential. Within the context of core feature extraction, feature importance facilitates interpretability, offering insights into the model's decision-making process. By discerning which features carry more weight in predictions, domain experts can gain a deeper understanding of the factors guiding the model's outputs. Moreover, feature importance guides feature selection and engineering endeavors, aiding in the prioritization of pertinent features for further analysis or enhancement. Through visualization techniques such as bar plots feature importance is effectively communicated.

**2.3.6 Training the Models (White Box Pathway)**

Following the features identified by the hybrid ensemble models the white box or linear models, linear regression and elastic net regression are developed to have simple representative equations of the linear relationship with the corresponding target class.

**2.3.7 Blending Equations**

To integrate the outputs from both black box and white box pathways, blending equations are formulated. These equations strategically by averaging coefficients from both linear regression and elastic net regression to balance accuracy and interpretability. Before deployment, the combined model undergoes rigorous evaluation to assess its performance using various metrics tailored to the specific needs of gas PVT property prediction. This evaluation helps in fine-tuning the model and adjusting blending equations if necessary.

Finally, the refined model is used for making predictions on the z factor, Gas FVF and Gas Viscosity based on a set aside proportion of the data which is 30% of the initial data used.

This comprehensive methodology not only ensures the accuracy and robustness of the predictions but also maintains a balance between black box and white box approaches, catering to both performance and transparency needs in predictive modeling.

### 

### **2.4 Research Algorithms**

**2.4.1 Linear Regression**

Linear regression is a statistical method used to estimate the linear relationship between a single outcome variable (also known as the dependent variable) and one or more predictor variables (also known as independent variables). When there's only one predictor variable, it's called simple linear regression, while with multiple predictors, it's termed multiple linear regression. This concept differs from multivariate linear regression, which predicts multiple correlated outcome variables simultaneously. In linear regression, we model relationships using linear predictor functions, with the parameters estimated from the data. These models are referred to as linear models. Typically, we assume that the conditional mean of the outcome variable, given the predictor variables, follows an affine (straight line) function of these values. Linear regression focuses on the conditional probability distribution of the outcome variable given the predictor variables, rather than the joint probability distribution of all variables, which is the concern of multivariate analysis. The most common method to fit linear regression models is the least squares approach, where we minimize the sum of squared differences between observed and predicted values.

**2.4.2 ElasticNet Regression**

ElasticNet regression is a regularization technique that combines the penalties of both Lasso (L1-norm penalty) and Ridge (L2-norm penalty) regression methods. It was introduced as a solution to the limitations of these individual techniques, aiming to strike a balance between feature selection and model flexibility. The fundamental idea behind ElasticNet regression lies in its objective function, which incorporates both L1 and L2 penalties. ElasticNet combines the benefits of Lasso and Ridge regularization. The L1 penalty encourages sparsity in the coefficient estimates, leading to automatic feature selection by driving some coefficients to zero. On the other hand, the L2 penalty helps address multicollinearity and stabilizes the coefficient estimates. The choice of *λ*1​​ and *λ*2​ balances the trade-off between the two penalties, controlling the amount of shrinkage applied to the coefficients. Larger values of *λ*1​ tend to result in more coefficients being set to zero, leading to a sparser model, while larger values of *λ*2​ increase the overall shrinkage of the coefficients.

**2.4.3 Decision Tree Regressor**

Decision tree learning is a supervised learning technique utilized in statistics, data mining, and machine learning. It involves using a classification or regression decision tree as a predictive model to derive insights from a dataset. Classification trees are employed when the target variable has a discrete set of values, with leaves representing class labels and branches representing combinations of features leading to those labels. On the other hand, regression trees are used when the target variable can take continuous values, such as real numbers. Moreover, the notion of regression trees can be extended to any object with pairwise dissimilarities, like categorical sequences.

Decision trees are highly favored in machine learning due to their interpretability and straightforwardness. They are widely employed in decision analysis to visually and explicitly depict decisions and decision-making processes. Furthermore, in data mining, decision trees serve to describe data, with resulting classification trees serving as inputs for decision-making procedures.

**2.4.4 Random Forest Regressor**

Random Forest Regression is an ensemble learning technique used for regression analysis, which constructs a multitude of decision trees during the training phase and outputs the average prediction of the individual trees. It combines the concept of bootstrap aggregating (bagging) with random feature selection to create a diverse set of decision trees, thereby reducing overfitting and enhancing predictive accuracy.

**2.4.5 LightGBM Regressor**

The LightGBM framework supports various algorithms, including GBT, GBDT, GBRT, GBM, MART, and RF. It shares many advantages with XGBoost, such as sparse optimization, parallel training, multiple loss functions, regularization, bagging, and early stopping. However, a significant distinction lies in its tree construction method. Unlike most implementations that grow trees level-wise (row by row), LightGBM adopts a leaf-wise approach, selecting the leaf expected to yield the largest decrease in loss. Moreover, LightGBM deviates from the sorted-based decision tree learning algorithm commonly used by XGBoost and others. Instead, it employs a highly optimized histogram-based decision tree learning algorithm, offering efficiency and memory consumption benefits. The LightGBM algorithm incorporates two innovative techniques, Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), enhancing speed while preserving accuracy.

### **2.5 Model Hybridization**

An array containing the coefficients of the linear regression model and the elastic net models are blended together using a statistical measure that gives an accurate coefficient value to each parameter for accurate formula based prediction. Voting Classifier is employed to blend the predictions of individual models, such as Decision Tree, Random Forest, and LightGBM, into a single final model. The Voting Classifier combines the outputs of these base models by aggregating their predictions through a majority voting mechanism or by taking a weighted average.

The potential effectiveness of a Voting Classifier lies in its ability to leverage the diverse strengths of the individual base models. Each base model may excel in capturing different aspects of the data or exhibit varying degrees of bias and variance. By combining multiple models, the Voting Classifier aims to mitigate the weaknesses of individual models and produce a more robust and accurate prediction. Moreover, the ensemble nature of the Voting Classifier tends to reduce over-fitting, as it aggregates predictions from multiple models trained on different subsets of the data. This ensemble approach helps generalize better to unseen data and improves the overall performance of the final model. Another advantage of the Voting Classifier is its flexibility in model selection. It allows for the inclusion of heterogeneous base models, including classifiers with different underlying algorithms or hyperparameters. This versatility enables practitioners to leverage the strengths of various modeling techniques and tailor the ensemble to the specific characteristics of the dataset.

Furthermore, the Voting Classifier can be customized to suit different requirements and preferences. For instance, practitioners can choose between hard voting, where predictions are based on majority voting, or soft voting, where predictions are weighted based on the confidence scores of the base models. Additionally, the weights assigned to individual models can be adjusted to prioritize the contributions of more reliable or high-performing models.

### 

### **2.6 Evaluation Metrics**

The evaluation metrics used in this work are:

**2.6.1 Average Absolute Relative Error (AARE)**

This measures the average absolute percentage difference between predicted and actual values. It is particularly useful for assessing the accuracy of predictions relative to the magnitude of the actual values. A lower AARE indicates better prediction accuracy, with values closer to zero implying more accurate predictions.

**2.6.2 Mean Absolute Error (MAE)**

The MAE calculates the average absolute difference between predicted and actual values. It provides a measure of the average magnitude of errors without considering their direction. MAE is robust to outliers and provides a straightforward interpretation of prediction accuracy. Lower MAE values indicate better model performance, with zero representing perfect predictions.

**2.6.3 Root Mean Squared Error (RMSE)**

This is the square root of the average squared differences between the predicted and actual values. Converts error units back to the original units of the measured data, enhancing interpretability. It penalizes larger errors more heavily due to squaring of the errors, making it sensitive to outliers. Commonly used in regression tasks to measure the accuracy of predictions. Lower RMSE values indicate better model performance, with zero representing perfect predictions.

**2.6.3 Coefficient of Determination (R2)**

The R2 Score measures the proportion of variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, where 1 indicates perfect predictions and 0 indicates that the model performs no better than a baseline model. The R2 Score provides an intuitive measure of how well the model fits the data relative to a simple mean model. Higher R2 Score values signify better model performance, with values closer to 1 representing stronger relationships between predictors and the target variable.

3. results and discussion

### **3.1 Compressibility Factor (z-factor) Model Performance**

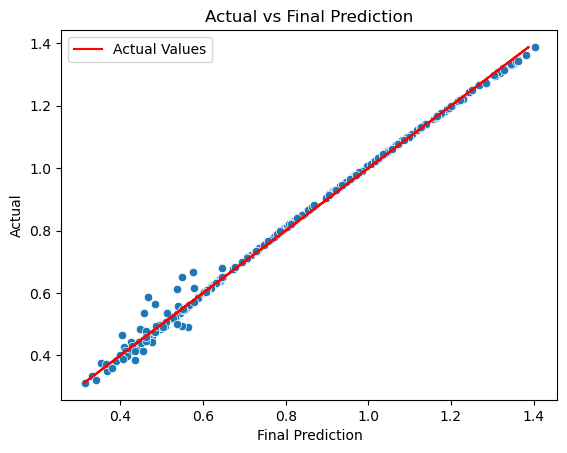
After training the dataset with the selected black box models decision tree, random forest, and lightgbm regressors with a 5 K-fold cross validation technique, the following scores for the AARE, MAE, R2 and RMSE were obtained (Table 3)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3a:** Black Box z-factor Model Performance | | | | | | | | |
|  | **AARE (%)** | | **MAE** | | | **R2 (%)** | | **RMSE** | |
| **Decision Tree** | 1.2647 | | 0.00834 | | | 0.9985 | | 0.01167 | |
| **Random Forest** | 0.6380 | | 0.00411 | | | 0.9995 | | 0.00679 | |
| **LightGBM** | 0.7531 | | 0.00463 | | | 0.9989 | | 0.01001 | |
| **Table 3b:** White Box z-factor Model Performance | | | | | | | | | |
|  | | **AARE (%)** | | **MAE** | **R2 (%)** | | **RMSE** | | |
| **ElasticNet** | | 1.818537 | | 0.011285 | 0.996165 | | 0.018613 | | |
| **Linear Regression** | | 1.818552 | | 0.011285 | 0.996165 | | 0.018614 | | |

Then proceeding to create a hybrid model of these models to account for their errors and inadequacies whilst raising the accuracy, a voting regressor of these models yielded AARE (0.61%), MAE (0.0043), R2 (99.95%) and RMSE (0.0067). Combining these complex models not only gives us a high predictive performance but also serves as an interpreting mechanism for the central properties that are required for adequate prediction of z-factor. Results revealed that the major parameters for z-factor prediction for this wet gas reservoirs are pressure, pseudo pressure, viscosity, density and temperature. Modeling the linear models with the selected parameters from the hybrid black box models with the evaluation of the white box model of Elastic Net and Linear Regression (Table 3b), by blending the coefficients of these models, a singular representative correlation is shown in equation 1.

 1

A final dataset was then tested with the modelled z-factor correlation when compared with the actual z-factor value the scatterplot is obtained below (Figure 2) displaying a strong correlation (R2=99.68%) between actual and final prediction.



**R2 = 99.68%**

**Figure 2: z-Factor Actual versus Modeled Samples**

### 

### **3.2 Gas Formation Volume Factor (Bg) Model Performance**

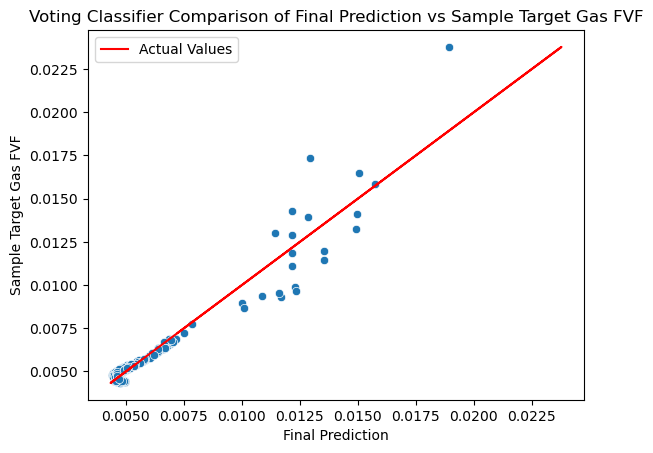
The training of the black box models prediction of gas formation volume factor yielded the following scores for decision tree, random forest and lightgbm (Table 4a).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 4a:** Black Box Bg Model Performance | | | | | | | | |
|  | **AARE (%)** | | **MAE** | | | **R2 (%)** | | **RMSE** |
| **Decision Tree** | 0.2566 | | 2.9394E-05 | | | 0.9972 | | 0.000129 |
| **Random Forest** | 0.1789 | | 2.1376E-05 | | | 0.9974 | | 0.000129 |
| **LightGBM** | 0.8545 | | 8.8742E-05 | | | 0.9719 | | 0.000404 |
| **Table 4b:** White Box Bg Model Performance | | | | | | | | |
|  | | **AARE (%)** | | **MAE** | **R2 (%)** | | **RMSE** | |
| **ElasticNet** | | 4.4619 | | 0.000313 | 0.9168 | | 0.000659 | |
| **Linear Regression** | | 4.4651 | | 0.000313 | 0.9168 | | 0.000659 | |

The hybrid of these three models gave AARE (0.42%), MAE (0.000054), R2 (99.26%) and RMSE (0.00023). Results revealed that the major parameters for gas formation volume factor prediction for this wet gas reservoirs are density, viscosity, pseudo pressure, pressure and temperature. Modeling the linear models with the selected parameters from the hybrid black box models with the evaluation of the white box model of Elastic Net and Linear Regression (Table 4b), blending these models resulted into correlation as shown in equation 2.

 2

A voting classifier comparison of final prediction with the sample target gave the scatterplot (Figure 3) displaying a strong correlation (R2=94.67%) between actual and final prediction.



**R2 = 94.67%**

**Figure 3: *Bg* Factor Actual versus Modeled Samples**

### 

### **3.3 Gas Viscosity Modelling**

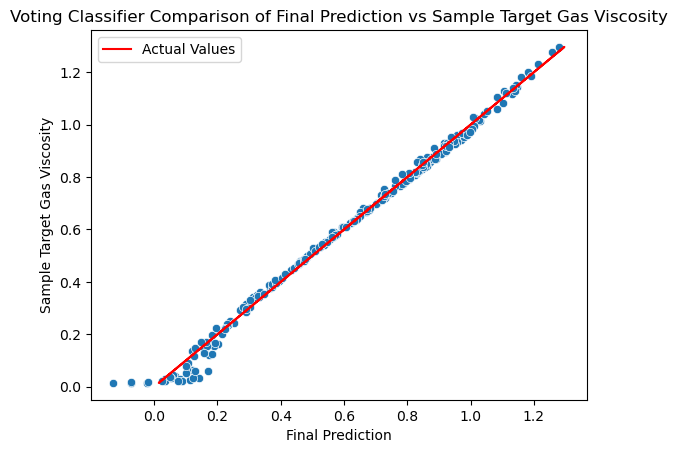
Evaluating the performance of gas viscosity prediction with the black box models (decision tree, random forest and lighgbm) produced the following metrics score as shown in the table below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 5a:** Black Box Gas Viscosity Model Performance | | | | | | | | |
|  | **AARE (%)** | | **MAE** | | | **R2 (%)** | | **RMSE** |
| **Decision Tree** | 1.2754 | | 0.00593 | | | 0.9995 | | 0.00766 |
| **Random Forest** | 0.9267 | | 0.00434 | | | 0.9997 | | 0.00586 |
| **LightGBM** | 1.1799 | | 0.00413 | | | 0.9995 | | 0.00732 |
| **Table 5b:** White Box Gas Viscosity Model Performance | | | | | | | | |
|  | | **AARE (%)** | | **MAE** | **R2 (%)** | | **RMSE** | |
| **ElasticNet** | | 8.9444 | | 0.01647 | 0.9949 | | 0.02358 | |
| **Linear Regression** | | 13.0173 | | 0.01623 | 0.9950 | | 0.02339 | |

The hybrid model developed from these three models gave AARE (0.01%), MAE (0.004), R2 (99.96%) and RMSE (0.0056). Results revealed that the major parameters for the gas viscosity prediction for this wet gas reservoirs are gas formation volume factor, pseudo pressure, z-factor, temperature, and gas density. Modeling the linear models with the selected parameters from the hybrid black box models with the evaluation of the white box model of Elastic Net and Linear Regression (Table 5b), blending these models resulted into correlation as shown in equation 3.

 3

The figure 4 below shows the correlation between the actual gas viscosity and predicted gas viscosity on data newly observed by the gas viscosity formula.



**R2 = 99.45%**

**Figure 4: Gas Formation Volume Factor (*μ*g) Actual versus Final**

4. Conclusion

This research adopts a comprehensive and systematic approach to predict PVT properties such as Gas Compressibility Factor, Gas Formation Volume Factor and Gas Viscosity in wet gas reservoirs, leveraging both black box and white box models. The black box model lays the groundwork, upon which the white box model builds, enhancing interpretability through ensemble methods that pinpoint critical features for developing simplified linear models with manageable errors and high generalizability.

References

Adagunodo, T. A., Bayowa, O. G., Alatise, O. E., Oshonaiye, A. O., Adewoyin, O. O., and Opadele, V. O. (2022). Characterization of reservoirs and depositional study of JP Field, shallow offshore of Niger Delta Basin, Nigeria. Scientific African, 15, e01064.

Ahmadi, M. A. and Ebadi M. (2014). Evolving smart approach for determination dew point pressure through condensate gas reservoirs, Fuel, 117(B), 1074-1084.

Ali Chamkalani, Sohrab Zendehboudi, Reza Chamkalani, Ali Lohi, Ali Elkamel, Ioannis Chatzis (2013). Utilization of support vector machine to calculate gas compressibility factor, Fluid Phase Equilibria, 358, 189-202.

Azizi, N., Rezakazemi, M. and Zarei, M.M. (2019). An intelligent approach to predict gas compressibility factor using neural network model. *Neural Comput & Applic* 31, 55–64

Baniasadi, H., Kamari, A., Heidararabi, S., Mohammadi, A. H., Hemmati-Sarapardeh, A. (2015). Rapid method for the determination of solution gas-oil ratios of petroleum reservoir fluids, Journal of Natural Gas Science and Engineering, 24, 500-509,

Deumah, S, S., Yahya, W. A., Al-khudafi, A. M., Ba-Jaalah, K. S. and Waleed, T. A. (2021). "Prediction of Gas Viscosity of Yemeni Gas Fields Using Machine Learning Techniques." Paper presented at the SPE Symposium: Artificial Intelligence - Towards a Resilient and Efficient Energy Industry, Virtual, October 2021. <https://doi.org/10.2118/208667-MS>

Dutta, S. and Gupta, J. P. (2010). PVT correlations for Indian crude using artificial neural networks. Journal of Petroleum Science and Engineering, Volume 72, Issues 1–2, Pages 93-109.

Elsharkawy, A. M. (1998). Modeling the Properties of Crude Oil and Gas Systems Using RBF Network. Paper presented at the SPE Asia Pacific Oil and Gas Conference and Exhibition, Perth, Australia, October 1998. <https://doi.org/10.2118/49961-MS>

Freyss, H., Guieze, P., Varotsis, N., Lestelle, K., and Simper, D. (2013). PVT Analysis for Oil Reservoirs. Schlumberger.

Hamid, R., Paiman, M., Hamzeh, A., Tatyana, P. and Ian, G. (2021). The application of supervised machine learning techniques for multivariate modelling of gas component viscosity: A comparative study, Fuel, 285, 119-146. <https://doi.org/10.1016/j.fuel.2020.119146>.

Isemin, I. A. and Akinsete, O. O. (2024) “Predicting Bubble Point Pressure and Gas-Oil Ratio PVT Properties Using Bagging Ensemble Techniques”, Journal of Engineering Research and Reports, 26(12), 368–377.

Khosravi, A., Machado, L. and Nunes, R.O. (2018). Estimation of density and compressibility factor of natural gas using artificial intelligence approach. *Journal of Petroleum Science and Engineering*, 168, 201-216. https://doi.org/10.1016/j.petrol.2018.05.023.

Khoukhi, A. and Albukhitan, S. (2011). PVT properties prediction using hybrid genetic-neuro-fuzzy systems. International Journal of Oil, Gas and Coal Technology, 4(1), 47. doi:10.1504/ijogct.2011.037744

Mohamadi-Baghmolaei, M., Azin, R., Osfouri, S., Mohamadi-Baghmolaei, R. and Zarei, Z., 2015. Prediction of gas compressibility factor using intelligent models. Natural Gas Industry B, 2(4), 283-294. https://doi.org/10.1016/j.ngib.2015.09.001.

Oloso, M. A., Hassan, M. G., Buick, J., & Bader-El-Den, M. (2016). Oil PVT characterisation using ensemble systems (2016). International Conference on Machine Learning and Cybernetics (ICMLC), Jeju, Korea (South), 61-68, doi: 10.1109/ICMLC.2016.7860878.

Osmah, E. A., Abdel-Wahhab, O. A. and Al-Marhoun, M. A. (2001). Prediction of Oil PVT Properties using Neural Networks. In SPE Middle East Oil Show, Society of Petroleum Engineers.

Ikpabi, P. B. and Akinsete, O. O. (2022). Correlation for Predicting Bubble Point Pressure for 22.3≤°API≥45 Crude Oils: A White-Box Machine Learning Approach. *International Journal of Frontiers in Engineering and Technology Research*, 3(2), 15-27.

Ramirez, A. M., Valle, G. A., Romero, F. and Jaimes, M. (2017). Prediction of PVT Properties in Crude Oil Using Machine Learning Techniques MLT. SPE Latin America and Caribbean Petroleum Engineering Conference. doi:10.2118/185536-ms

Rezaei, F., Jafari, S., Hemmati-Sarapardeh, A. and Mohammadi, A.H. (2022). Modeling of gas viscosity at high pressure-high temperature conditions: Integrating radial basis function neural network with evolutionary algorithms. *Journal of Petroleum Science and Engineering*, *208*, 109328.

Rezaei, F., Akbari, M., Rafiei, Y., Hemmati-Sarapardeh, A. (2023)*.* Compositional modeling of gas-condensate viscosity using ensemble approach. *Sci Rep* **13**, 9659. https://doi.org/10.1038/s41598-023-36122-3

Sola-Aremu, Oluwapelumi (2019). An Inferable Machine Learning Approach to Predicting PVT Properties of Niger Delta Crude Oil using Compositional Data. SPE Annual Technical Conference and Exhibition, Calgary, Alberta, Canada. doi: <https://doi.org/10.2118/199783-STU>

Uzogor, K. and Akinsete, O. (2020). Improved correlations and predictive models for Nigerian crude oil PVT properties using advanced regression and intelligent techniques. In SPE Nigeria Annual International Conference and Exhibition (p. D013S004R006). SPE.

Voutsas, E., Novak, N., Louli, V., Pappa, G., Petropoulou, E., Boukouvalas, C., Panteli, E., Skouras, S. (2018). Thermodynamic Modeling of Natural Gas and Gas Condensate Mixtures. 10.1002/9781119269618.ch3.

Xi Yang, Birol Dindoruk, Ligang Lu (2020). A comparative analysis of bubble point pressure prediction using advanced machine learning algorithms and classical correlations. Journal of Petroleum Science and Engineering, 185, 106598.