Scientific Research Publishing

Open Journal of Geology, 2026, 16(9), 577-590
https://www.scirp.org/journal/ojg
ISSN Online: 2161-7589
ISSN Print: 2161-7570

Explaining the Performance Hierarchy of Linear Regression, Support Vector Regression, and Random Forest on Physics-Generated PVT Data: A Real-Gas-Law and Gauss-Markov Perspective

Gnessoa Rene Hie*, Joseph Atubokiki Ajienka, Joseph Amieibibama

Department of Petroleum and Gas Engineering, University of Port Harcourt, Port Harcourt, Nigeria
Email: *gnessoa.hie@inphb.ci

How to cite this paper: Hie, G.R., Ajienka, J.A. and Amieibibama, J. (2026) Explaining the Performance Hierarchy of Linear Regression, Support Vector Regression, and Random Forest on Physics-Generated PVT Data: A Real-Gas-Law and Gauss-Markov Perspective. *Open Journal of Geology*, **16**, 577-590.
https://doi.org/10.4236/ojg.2026.169029

Received: August 16, 2026
Accepted: September 18, 2026
Published: September 21, 2026

Abstract

Machine learning (ML) regression models are increasingly proposed as fast surrogates for deterministic Pressure-Volume-Temperature (PVT) simulation in reservoir and production engineering, yet the choice among competing algorithms is rarely justified on physical or statistical grounds. This study trains, evaluates, and compares three regression algorithms, Linear Regression (LR), Support Vector Regression (SVR), and Random Forest Regression (RF), on a PVT dataset generated from an industry-standard Integrated Production Modelling (IPM) software platform (PROSPER) using the Glaso correlation for a representative mature onshore Niger Delta oil well. Gas density was predicted from pressure and associated fluid properties across a depletion range of 1000 - 3500 psig. Following a structured preprocessing pipeline (exploratory data analysis, Isolation Forest anomaly removal, Min-Max normalization, K-means regime clustering, and time-series smoothing), the three algorithms were trained on an 80:20 split and evaluated using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). Linear Regression achieved the highest accuracy (R^2 = 0.9954, RMSE = 0.2240 lb/ft^3), followed by SVR (R^2 = 0.9408, RMSE = 0.8042 lb/ft^3) and Random Forest (R^2 = 0.8389, RMSE = 1.4326 lb/ft^3). Physically consistent extrapolation beyond the training range (correctly predicting monotonically increasing gas density at 4000 and 5000 psig) confirmed that the Linear Regression model had captured the governing thermodynamic relationship rather than an artefact of the training data. We show that this performance hierarchy is not an empirical coincidence but a direct, predictable consequence of two established results: the real gas law,

which imposes a near-linear pressure-density relationship (Pearson r = 1.00) at approximately constant reservoir temperature, and the Gauss-Markov theorem, which guarantees that Ordinary Least Squares is the best linear unbiased estimator when the true relationship is linear. A bias-variance decomposition further explains why SVR's kernel mismatch and Random Forest's piecewise-constant approximation both add prediction variance without a compensating reduction in bias on this near-linear, physics-generated data. These findings provide petroleum engineers and applied ML practitioners with a principled, physics-grounded criterion for algorithm selection when training data are generated from simulator-based PVT correlations, rather than defaulting to the most complex available model.

Keywords

Machine Learning, PVT Modelling, Linear Regression, Support Vector Regression, Random Forest, Gauss-Markov Theorem, Real Gas Law, Physics-Informed Machine Learning, Reservoir Engineering

1. Introduction

Machine learning (ML) has been increasingly adopted in petroleum engineering as a computationally efficient alternative to physics-based simulation for tasks including production forecasting, well performance optimization, and fluid property prediction [1] [2]. A comprehensive review of 166 published studies by Tariq *et al.* [2] found that ensemble and deep-learning architectures dominate the recent literature but also identified the absence of physical consistency as the most frequently reported limitation of published ML models: fewer than one in five reviewed studies explicitly addressed the physical plausibility of model predictions. A model that fits historical data well but violates a governing physical law, for example, predicting that gas density decreases as pressure increases, cannot be trusted for engineering decision-making regardless of its statistical goodness of fit [3].

The physics-informed machine learning (PIML) paradigm has emerged as a principled response to this risk, integrating physical knowledge into the ML workflow either by embedding governing equations directly into the training objective, as in Physics-Informed Neural Networks [4], or by incorporating physical priors into model architecture and training data [3] [5]. A comparatively underexploited route to physical consistency is to generate the ML training data itself from a validated, physics-based simulator, so that the training distribution is thermodynamically and fluid-mechanically consistent by construction, and any model trained on it inherits this consistency [6]. This approach requires no modification to standard ML algorithms and is directly compatible with existing commercial production modelling software, making it readily transferable to industry practice.

This raises a question that has received little systematic attention in the petro-

leum engineering ML literature: when training data are generated from a physics-based correlation rather than sampled directly from noisy field measurements, does the statistical structure imposed by the underlying physics predictably favor one class of ML algorithm over another, and can that preference be explained rather than merely observed? Existing comparative studies typically report which algorithm performed best on a given dataset [7] without connecting the result to the governing physics of the data-generating process or to the statistical theory of the estimators being compared.

This study addresses that question directly. Using a PVT dataset generated from an industry-standard Integrated Production Modelling (IPM) platform (PROSPER) with the Glaso [8] correlation for a representative mature onshore Niger Delta well, we train, evaluate, and compare three regression algorithms representing three distinct algorithmic paradigms, Linear Regression (parametric linear modelling), Support Vector Regression (kernel-based regularized regression), and Random Forest Regression (ensemble decision-tree aggregation), on the task of predicting gas density from pressure and associated fluid properties. We then provide a physical and statistical explanation, grounded in the real gas law and the Gauss-Markov theorem, for the resulting performance hierarchy. The specific objective of this study is to train, evaluate, and compare the three algorithms using RMSE and R^2 as performance metrics, and to explain the physical and statistical basis of the observed performance hierarchy.

The remainder of this paper is organized as follows. Section 2 describes the physics-based data generation procedure, the preprocessing pipeline, and the model training and evaluation methodology. Section 3 presents the comparative performance results. Section 4 discusses the physical and statistical explanation for the observed hierarchy and its practical implications for algorithm selection in petroleum engineering ML applications. Section 5 concludes.

2. Materials and Methods

2.1. Study Context and Physics-Based Data Generation

The dataset used in this study was generated from a production system model of a representative mature onshore oil well in the Niger Delta petroleum province of Nigeria, constructed in PROSPER, an industry-standard Integrated Production Modelling software platform. Pressure-Volume-Temperature (PVT) fluid properties were computed using the Glaso [8] correlation, a generalized empirical PVT correlation widely applied for black-oil systems. Eight PVT parameters, pressure, Gas-Oil Ratio (GOR), oil density, oil viscosity, oil formation volume factor, oil compressibility, gas viscosity, and gas density, were extracted at 20 discrete pressure steps spanning the reservoir's depletion range, from the initial reservoir pressure of 3500 psig down to a late-decline pressure of 1000 psig, at an approximately constant reservoir temperature. Gas density, the fluid property governing the hydrostatic contribution of gas to wellbore mixture density in gas lift and vertical lift performance calculations, was selected as the prediction target. The reservoir's

bubble point pressure was approximately 1325 psig, so the dataset spans both the undersaturated (single-phase) and saturated (two-phase, solution-gas-liberating) flow regimes.

2.2. Data Preprocessing Pipeline

A five-step preprocessing pipeline was applied prior to model training. 1) Exploratory Data Analysis characterized the distributional properties (mean, standard deviation, skewness, kurtosis, range) and inter-variable correlation structure of all eight PVT variables. 2) Isolation Forest anomaly detection [9], an ensemble-based unsupervised method that isolates anomalous observations through recursive random partitioning of the feature space, identified and removed two observations that were statistically inconsistent with the thermodynamic trend of the surrounding data, likely reflecting transcription artefacts from manual entry of PROSPER outputs into the Python environment. Anomaly detection is applied as the first preprocessing step, before normalization and clustering, because both Min-Max scaling and K-means clustering are sensitive to extreme values, and removing anomalous records first prevents outliers from distorting the scaling range or cluster centroids. The five steps are applied in strict sequence for this same reason: Isolation Forest first; normalization second (because K-means uses Euclidean distance, which requires scaled features); K-means third; and smoothing last, so regime labels reflect instantaneous thermodynamic state before smoothing blurs the undersaturated-saturated boundary. Within the training partition, Isolation Forest was applied after the train-test split, so the four test observations were never exposed to or affected by anomaly-removal decisions. 3) Min-Max normalization [10] rescaled all input features to the range [0, 1] using $x' = \left(x - x_{\min}\right) / \left(x_{\max} - x_{\min}\right)$, applied after the train-test split and parameterized using training-set statistics only, to prevent test-set information leakage; this step is essential for SVR, whose objective function is sensitive to the relative magnitude of input features. 4) K-means clustering [11], specified with k = 2, partitioned the dataset into the undersaturated and saturated production regimes, providing an additional regime-indicator feature. 5) A simple moving-average smoothing step reduced high-frequency variation not attributable to the underlying pressure-dependent thermodynamic trend.

2.3. Machine Learning Model Selection and Training

Three regression algorithms, selected to represent three distinct algorithmic paradigms, were trained on a stratified 80:20 train-test split of the full 20-observation dataset (16 training observations, 4 held-out test observations), stratified to ensure representation of both undersaturated and saturated pressure conditions in both partitions; a simple random split would risk placing all four test observations within a single pressure regime and understate generalisation performance. The train-test split was performed before K-means clustering and moving-average smoothing, which were applied to the training partition only, ensuring that no

test-set information influenced these preprocessing parameters. Isolation Forest anomaly detection was subsequently applied within the 16-observation training partition, identifying and removing two anomalous records, so that the models were fitted on 14 effective training observations; the 4-observation test set was held entirely clean of anomaly-removal decisions. For each of the three models, the predictor set consisted of pressure as the primary input variable, with the regime-indicator feature (a binary label produced by K-means Step 4 encoding the undersaturated/saturated distinction) included as an additional predictor. For the multiple-predictor Linear Regression extension, pressure, GOR, and gas viscosity were used jointly. All predictor variables are available at deployment from standard PVT monitoring data without requiring gas density information.

Linear Regression was applied as the baseline parametric model, using Ordinary Least Squares (OLS) to fit gas density as a function of pressure. Under the Gauss-Markov theorem [12], OLS is the Best Linear Unbiased Estimator (BLUE) when the true relationship is linear and the classical regression assumptions hold. A multiple-predictor extension incorporating pressure, GOR, and gas viscosity jointly was also evaluated.

Support Vector Regression (SVR) was applied following Min-Max normalization, using a Radial Basis Function (RBF) kernel with scikit-learn's default hyperparameters (C = 1.0, $\varepsilon = 0.1$, γ = 'scale'). SVR was selected for its theoretical capacity to model nonlinear relationships through kernel transformation while controlling complexity via the epsilon-insensitive loss and the regularization parameter C [13] [14].

Random Forest Regression was applied with 100 trees, using bootstrap sampling and random feature selection at each split node, following Breiman [15]. Random Forest was selected for its established empirical robustness on complex, noisy engineering datasets and its capacity to generate feature importance rankings.

All computations were implemented in Python 3 using the scikit-learn library [10] within a Jupyter Notebook environment.

2.4. Performance Evaluation

Model performance was evaluated on the held-out test set using two complementary metrics: the Root Mean Squared Error (RMSE), which quantifies average prediction error in the target variable's original units (lb/ft^3) and penalizes large errors disproportionately, and the coefficient of determination (R^2), a scale-independent measure of the proportion of variance explained. The joint use of both metrics follows the recommendation of Chai and Draxler [16]. Physical consistency of the best-performing model was additionally assessed through an extrapolation test, evaluating predictions at pressures (4000 and 5000 psig) outside the training data range, on the principle that a model which has captured the true governing relationship, rather than overfit to the training sample, should extrapolate in a physically correct (monotonically increasing) direction [17].

3. Results

3.1. Dataset Characteristics

Gas density in the preprocessed dataset ranged from 3.95 lb/ft^3 at 1000 psig to 14.07 lb/ft^3 at 3500 psig, a factor of approximately 3.6 across the pressure depletion range, consistent with the real gas law's prediction that gas density scales approximately with pressure at constant temperature. As we can see in Figure 1, Pearson correlation analysis revealed a near-perfect linear relationship between gas density and pressure (r = 1.00), a very high correlation between gas density and gas viscosity (r = 0.99, reflecting their shared dependence on molecular packing density via Chapman-Enskog kinetic theory), and a moderate correlation between gas density and GOR (r = 0.58, arising because the dataset spans both undersaturated and saturated regimes). Figure 2 shows that Pairwise scatterplots across all eight PVT variables confirmed a predominantly linear inter-variable structure, with only mild, bubble-point-localized nonlinearity observed in oil viscosity and oil compressibility.

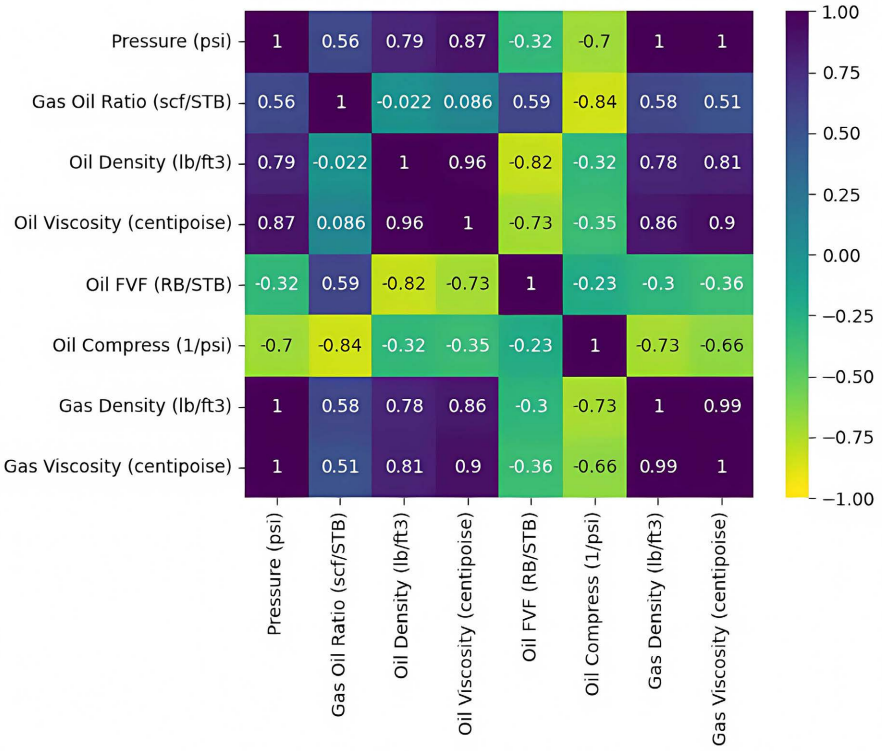

Figure 1. Heatmap of Pearson correlations between PVT dataset variables.

3.2. Comparative Model Performance

Table 1 presents the consolidated test-set performance of the three algorithms. Linear Regression achieved the highest accuracy (R^2 = 0.9954, RMSE = 0.2240 lb/ft^3), followed by SVR (R^2 = 0.9408, RMSE = 0.8042 lb/ft^3) and Random Forest (R^2 = 0.8389, RMSE = 1.4326 lb/ft^3). The Random Forest RMSE was approximately

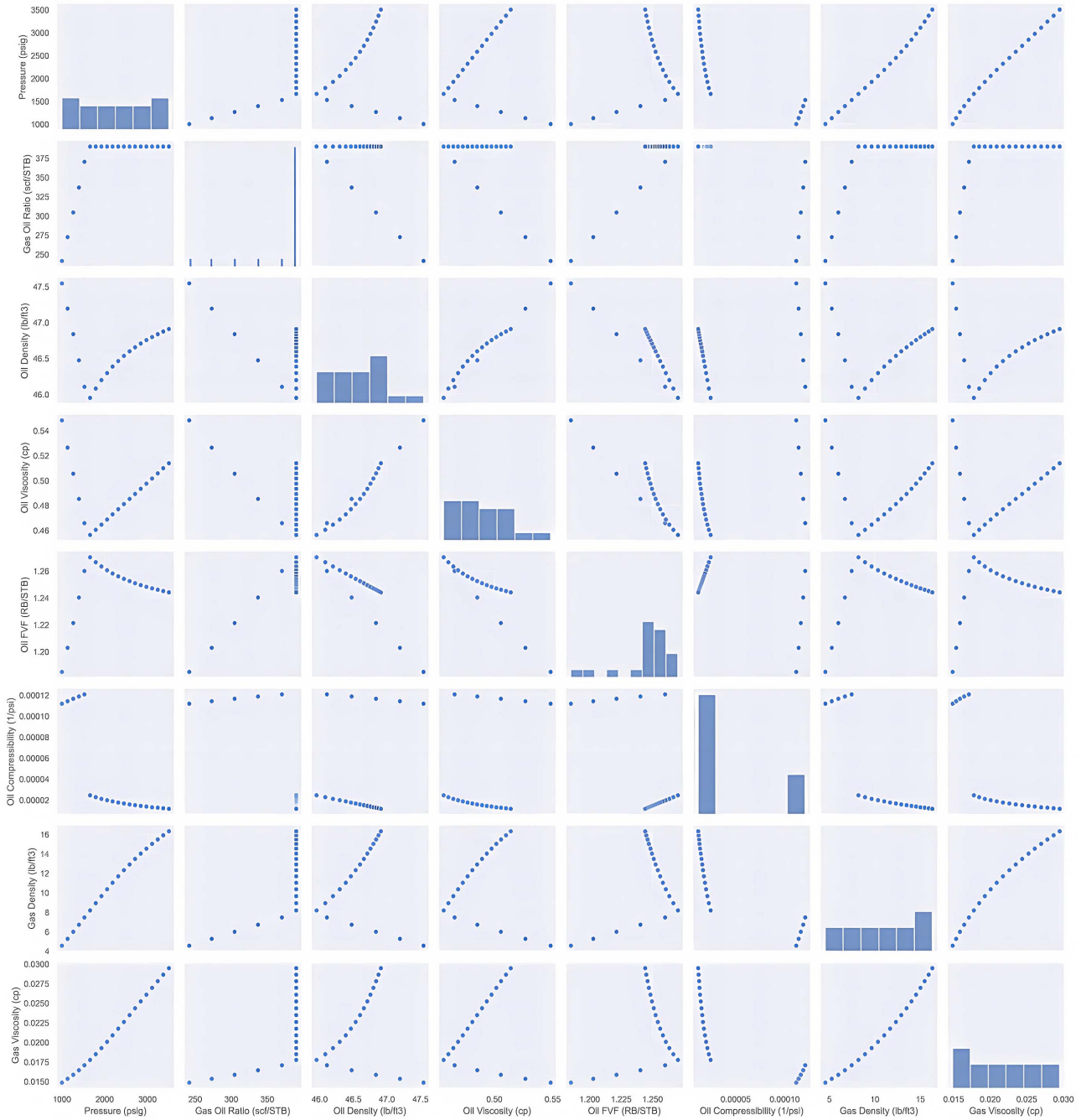

Figure 2. Pairplot of PVT dataset variable relationships.

6.4 times higher, and the SVR RMSE approximately 3.6 times higher, than the Linear Regression RMSE. Figures 3-5 show five predicted-versus-actual pairs; the additional point reflects a supplementary prediction at the boundary pressure included for visual completeness and was not used in computing the tabulated test-set metrics. The multiple-predictor Linear Regression model (pressure, GOR, and gas viscosity jointly) achieved a marginally improved R^2 of 0.9975, confirming that pressure alone is already the dominant predictor of gas density.

Table 1. Comparative test-set performance of the three regression algorithms.

Model	RMSE (lb/ft³)	R^2	Interpretation
Linear Regression	0.2240	0.9954	Highest accuracy; best fit for a near-linear, physics-generated dataset
Support Vector Regression	0.8042	0.9408	High accuracy; RBF kernel introduces mild curvature bias at the range extremes
Random Forest Regression	1.4326	0.8389	Lowest accuracy; piecewise-constant partitioning penalized on near-linear data

3.3. Error Structure and Physical Consistency

Examination of the predicted-versus-actual plots revealed distinct error structures for each algorithm. SVR exhibited systematic positive bias at low gas density (low pressure) and systematic negative bias at high gas density (high pressure), indicating that the RBF kernel introduced curvature where the true relationship is effectively linear (Figure 4). Random Forest exhibited the widest scatter of the three models (Figure 5), with the largest deviations concentrated at the extremes of the pressure range, consistent with its piecewise-constant approximation strategy: each decision tree partitions the input space into rectangular regions and predicts a constant value within each region, so approximation error is largest where the underlying gradient is steep relative to the local region width, that is, at the pressure extremes.

The physical consistency of the Linear Regression model was directly tested by extrapolating beyond the training pressure range. At 4,000 psig, the model predicted a gas density of 16.95 lb/ft³; at 5,000 psig, 24.10 lb/ft³. Both predictions are physically correct, monotonically increasing with pressure, as required by the real gas law, despite lying outside the range of conditions represented in the training data. An overfitted model would typically fail to generalize correctly beyond its training range [17]; the correct extrapolation behavior therefore indicates that the Linear Regression model captured the governing thermodynamic relationship rather than a spurious pattern specific to the training sample.

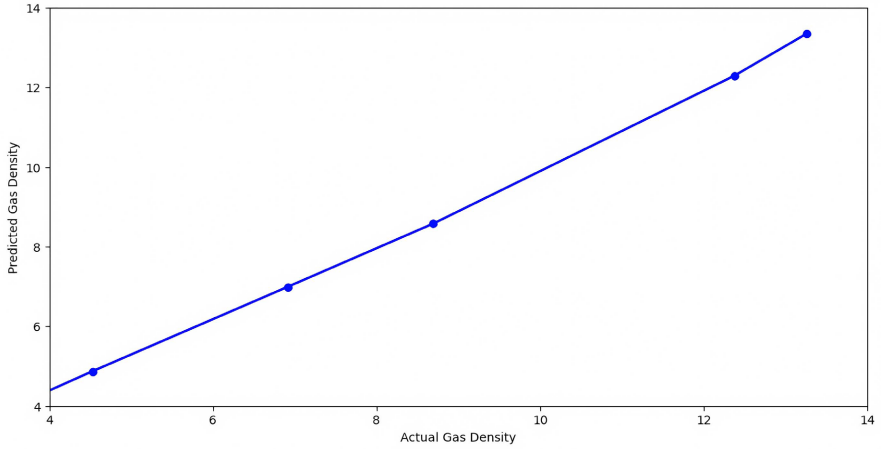

Figure 3. Predicted versus actual gas density—linear regression.

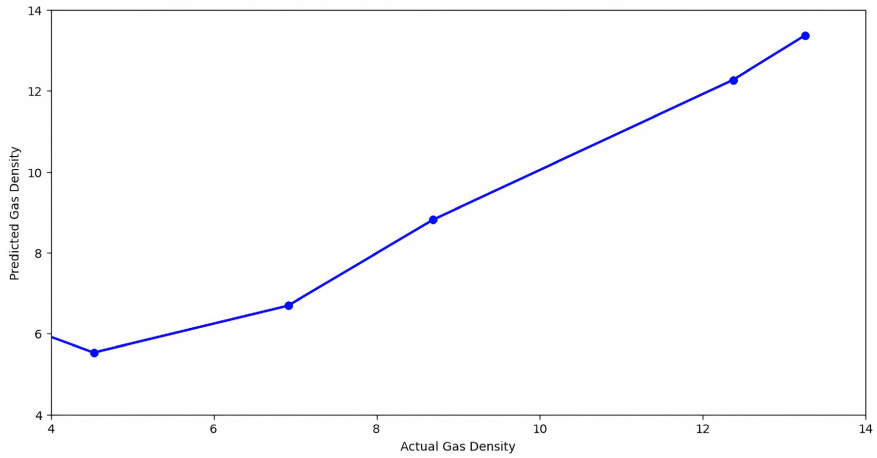

Figure 4. Predicted versus actual gas density—SVR.

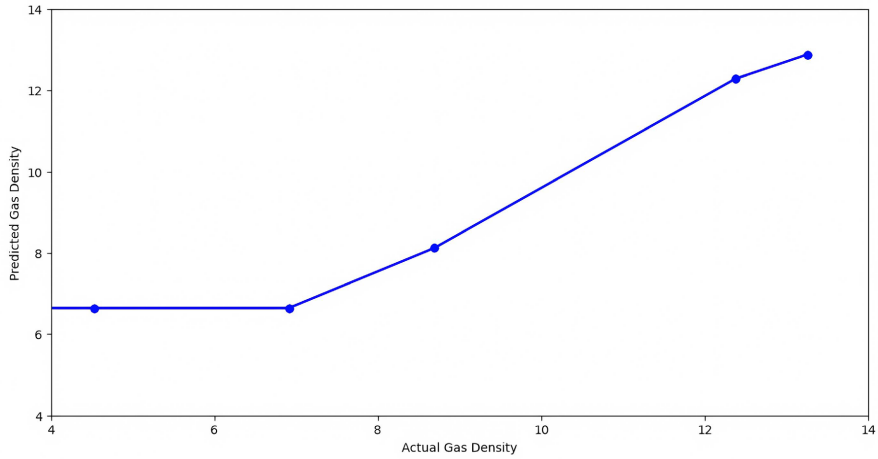

Figure 5. Predicted versus actual gas density—random forest.

4. Discussion

4.1. Physical and Statistical Explanation of the Performance Hierarchy

The observed performance hierarchy, Linear Regression > SVR > Random Forest, comparing the OLS Linear Regression implementation with the scikit-learn default SVR (RBF kernel, C = 1.0, $\varepsilon = 0.1$, $\gamma =$ 'scale') and the default Random Forest (100 trees, bootstrap sampling) implementations, on this single held-out test partition, is not an arbitrary empirical outcome but is fully explicable from the physics of the data-generating process and the statistical theory governing the three estimators. The explanation proceeds in three steps.

First, the real gas law prescribes the underlying physical relationship: $\rho_{gas} = \mathrm{PM}/\mathrm{ZRT}$, where P is pressure, M is the gas molecular weight, Z is the gas compressibility factor, R is the universal gas constant, and T is the (approximately constant) reservoir temperature. Over the pressure range studied (1000 - 3500 psig), the Z-factor for a typical Niger Delta gas composition varies only modestly,

from approximately 0.82 to 0.95 [18], so gas density varies approximately linearly with pressure. This near-linearity is directly confirmed by the correlation coefficient r = 1.00 observed between gas density and pressure in the dataset.

Second, for a dataset in which the true relationship between the target variable and the primary predictor is approximately linear, the linear hypothesis class is the physically correct functional form, and the Gauss-Markov theorem guarantees that the OLS estimator is the Best Linear Unbiased Estimator, the minimum-variance unbiased estimator within that hypothesis class [12]. Any departure from the linear functional form, whether the nonlinear kernel transformation applied by SVR or the piecewise-constant partitioning applied by Random Forest, adds model complexity, and therefore variance, without a compensating reduction in bias, because the residual bias of Linear Regression on this dataset is already negligible.

Third, the bias-variance decomposition of prediction error [17] [19] explains the magnitude of the observed hierarchy. Linear Regression achieves near-zero bias, because the linear model correctly represents the true relationship, combined with minimum variance, guaranteed by the Gauss-Markov theorem, yielding the lowest expected prediction error. Random Forest introduces a piecewise-constant approximation error that is only partly offset by bootstrap aggregation, producing the largest deviation from optimal performance among the three models (R^2 = 0.8389). SVR introduces variance through the mismatch between its RBF kernel's nonlinear hypothesis class and the effectively linear target function, together with the equal weighting of all within-margin predictions imposed by the epsilon-insensitive loss, without any corresponding reduction in bias, resulting in intermediate performance (R^2 = 0.9408).

4.2. Practical Implications for Algorithm Selection

These findings carry a specific and generalizable implication for ML practitioners in petroleum engineering: the optimal algorithm for a given prediction task is determined, in part, by the statistical structure that the underlying physics imposes on the training data, and this structure depends on how the training data were generated. For datasets generated from simulator-based PVT correlations, such as PROSPER's Glaso-correlation workflow, which produce approximately linear inter-variable relationships as a direct consequence of the real gas law and related thermodynamic constraints, Linear Regression is both the theoretically justified and empirically best-performing choice, and its use avoids the unnecessary variance cost of more complex algorithms. This is a case for principled, physics-informed model selection rather than a default preference for algorithmic complexity.

The converse also holds. Field-measured PVT and production datasets incorporate additional noise sources, sensor error, sampling and operational variability, and potentially genuine nonlinear contributions from geological heterogeneity and multiphase interaction effects that are absent from a simulator-generated

dataset. Under these conditions, the linearity assumption that favors Linear Regression in this study is weaker, and the added flexibility of Random Forest or SVR may become an asset rather than a source of unnecessary variance [2] [7]. The performance hierarchy identified here should therefore be interpreted as specific to physics-generated training data rather than as a general recommendation against ensemble or kernel-based methods in petroleum engineering ML applications.

More broadly, this study demonstrates a practical and readily transferable route to physically consistent machine learning: generating ML training data from an industry-standard, physics-based simulator rather than embedding physical constraints directly into the model architecture or loss function, as in Physics-Informed Neural Networks [4]. This route requires no modification to standard ML algorithms or software, is compatible with existing commercial IPM platforms, and, as shown by the correct extrapolation behavior of the trained Linear Regression model, produces models whose predictions remain physically consistent even outside the training data range.

4.3. Limitations

Several limitations qualify these findings. First, the dataset was generated from a single-well production system model using the Glaso [8] PVT correlation, which was originally calibrated on North Sea crude oils and may not perfectly represent Niger Delta fluid compositions; any systematic bias in the correlation propagates into the ML training data and, in turn, into the model predictions. Second, the training dataset, drawn from 20 simulated pressure steps, is small by conventional ML standards, reflecting the discrete nature of simulator-generated PVT data rather than continuously sampled field data; the resulting train-test split (16/4 observations) limits the statistical power of the test-set evaluation, and the reported metrics should be interpreted as indicative rather than as precise population estimates. Furthermore, the performance hierarchy reported here is based on a single stratified 80:20 train-test split. The dataset's origin as physics-simulation output at discrete pressure steps, rather than a randomly sampled field time-series, constrains the resampling options available: standard k-fold cross-validation applied to an ordered pressure series would break temporal ordering and produce splits in which training observations are pressure-adjacent to test observations, inflating apparent generalisation. Time-series cross-validation requires minimum fold sizes that this small dataset cannot support. The stratified split used here was the methodologically appropriate choice given these constraints, ensuring both pressure regimes were represented in training and testing, but the four-observation test set limits the statistical precision of the reported metrics. Future work with larger simulator-generated datasets, spanning a wider pressure range at finer discretization, should validate the hierarchy using repeated resampling. Third, because the models were trained to reproduce the Glaso correlation's predictions rather than direct field measurements of gas density, the reported accuracy reflects

fidelity to the correlation, not independently validated accuracy against measured field gas density. Fourth, the SVR and Random Forest models were evaluated at scikit-learn default hyperparameter values rather than values tuned within the training data; the conclusions therefore compare these specific default implementations rather than the globally optimised algorithm classes. Hyperparameter tuning within the training partition may improve SVR and Random Forest accuracy and should be evaluated in future work. Importantly, however, the underperformance of SVR relative to Linear Regression on this dataset cannot be attributed to SVR as an algorithm class in general, but specifically to the mismatch between a flexible nonlinear learner and a near-linear data-generating process: the Gauss-Markov theorem and the bias-variance decomposition both predict that any expansion of the hypothesis class beyond the linear functional form, whether through kernel choice or hyperparameter tuning, adds variance without reducing the near-zero bias that OLS already achieves on this data structure. This is a structural prediction, not a contingent one, and it reinforces rather than undermines the algorithm-selection argument. Fifth, a minor numerical inconsistency was identified between the reported Random Forest R^2 (0.8389) and the value implied by the RMSE (1.4326 lb/ft^3) relative to the test-set variance implied by the Linear Regression and SVR metrics; the reported values are those produced directly by the scikit-learn evaluation, and any rounding or precision difference should be interpreted in that context. Extension of this comparison to field-measured, multi-well datasets is an important direction for future work.

5. Conclusion

This study trained, evaluated, and compared Linear Regression, Support Vector Regression, and Random Forest Regression on a physics-generated PVT dataset for gas density prediction, derived from an industry-standard Integrated Production Modelling platform using the Glaso [8] correlation for a mature Niger Delta oil well. Linear Regression achieved the highest predictive accuracy (R^2 = 0.9954, RMSE = 0.2240 lb/ft^3), outperforming SVR (R^2 = 0.9408) and Random Forest (R^2 = 0.8389), and produced physically consistent extrapolations beyond the training pressure range. This performance hierarchy was shown to be a direct and predictable consequence of the real gas law, which imposes a near-linear pressure-density relationship at approximately constant reservoir temperature, combined with the Gauss-Markov theorem, which guarantees the optimality of Ordinary Least Squares under linear data conditions. The bias-variance decomposition further clarified why the added flexibility of SVR's kernel transformation and Random Forest's ensemble partitioning each introduced variance without a compensating reduction in bias on this near-linear dataset. These findings offer petroleum engineers and applied ML practitioners a principled, physics-grounded basis for machine learning algorithm selection, specifically comparing OLS Linear Regression with SVR and Random Forest at their scikit-learn default configurations, when training data are generated from simulator-based PVT correlations, and they il-

lustrate a practical, readily transferable route to physically consistent machine learning that requires no modification to standard algorithms or existing commercial modelling software.

Acknowledgements

The author gratefully acknowledges the guidance and supervision of Prof. Joseph A. Ajienka and Dr. Joseph Amieibibama, Department of Petroleum and Gas Engineering, University of Port Harcourt.

Author Contributions

Conceptualization, Gnessoa Rene Hie and Joseph A. Ajienka; methodology, Gnessoa Rene Hie; software, Gnessoa Rene Hie; validation, Gnessoa Rene Hie, Joseph A. Ajienka, and Joseph Amieibibama; formal analysis, Gnessoa Rene Hie; investigation, Gnessoa Rene Hie; resources, Gnessoa Rene Hie; data curation, Gnessoa Rene Hie; writing—original draft preparation, Gnessoa Rene Hie; writing—review and editing, Gnessoa Rene Hie; visualization, Gnessoa Rene Hie; supervision, Gnessoa Rene Hie; project administration, Gnessoa Rene Hie. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[2] Tariq, Z., Aljawad, M.S., Hasan, A., Murtaza, M., Mohammed, E., El-Husseiny, A., *et al.* (2021) A Systematic Review of Data Science and Machine Learning Applications to the Oil and Gas Industry. *Journal of Petroleum Exploration and Production Technology*, **11**, 4339-4374. https://doi.org/10.1007/s13202-021-01302-2

[3] Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., *et al.* (2017) Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, **29**, 2318-2331. https://doi.org/10.1109/tkde.2017.2720168

[4] Raissi, M., Perdikaris, P. and Karniadakis, G.E. (2019) Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *Journal of Computational Physics*, **378**, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045

[5] Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S. and Yang, L. (2021) Physics-Informed Machine Learning. *Nature Reviews Physics*, **3**, 422-440. https://doi.org/10.1038/s42254-021-00314-5

[6] Nwachukwu, A., Jeong, H., Pyrcz, M. and Lake, L.W. (2018) Fast Evaluation of Well Placements in Heterogeneous Reservoir Models Using Machine Learning. *Journal of Petroleum Science and Engineering*, **163**, 463-475. https://doi.org/10.1016/j.petrol.2018.01.019

[7] Temizel, C., Canbaz, C.H., Palabiyik, Y. and Ozcan, G. (2014) Machine Learning Ap-

plications in Unconventional Resources: An Analysis of Production Data. *SPE Annual Technical Conference and Exhibition*, Amsterdam, 27-29 October 2014.

[8] Glaso, O. (1980) Generalized Pressure-Volume-Temperature Correlations. *Journal of Petroleum Technology*, **32**, 785-795. https://doi.org/10.2118/8016-pa

[9] Liu, F.T., Ting, K.M. and Zhou, Z. (2008) Isolation Forest. 2008 *Eighth IEEE International Conference on Data Mining*, Pisa, 15-19 December 2008, 413-422. https://doi.org/10.1109/icdm.2008.17

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.

[11] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 21 June-18 July 1965 and 27 December 1965-7 January 1966, 281-297.

[12] Greene, W.H. (2018) Econometric Analysis. 8th Edition, Pearson.

[13] Vapnik, V.N. (1995) The Nature of Statistical Learning Theory. Springer. https://doi.org/10.1007/978-1-4757-2440-0

[14] Smola, A.J. and Schölkopf, B. (2004) A Tutorial on Support Vector Regression. *Statistics and Computing*, **14**, 199-222. https://doi.org/10.1023/b:stco.0000035301.49549.88

[15] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[16] Chai, T. and Draxler, R.R. (2014) Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Arguments against Avoiding RMSE in the Literature. *Geoscientific Model Development*, **7**, 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

[17] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition, Springer. https://doi.org/10.1007/978-0-387-84858-7

[18] Standing, M.B. and Katz, D.L. (1942) Density of Natural Gases. *Transactions of the AIME*, **146**, 140-149. https://doi.org/10.2118/942140-g

[19] Geman, S., Bienenstock, E. and Doursat, R. (1992) Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, **4**, 1-58. https://doi.org/10.1162/neco.1992.4.1.1