A stacked generalization ensemble method for rate of penetration prediction

Erasmo A. B. Silva[1], Antonio P. A. Ferro[1], Francisco A. V. Binas Júnior[1], Lucas P. Gouveia[1], Aline S. R. Barboza[1]

[1]*Laboratory of Scientific Computing and Visualization, Center of Technology, Federal University of Alagoas*
Av. Lourival Melo Mota, S/n, Tabuleiro do Martins, 57072-970, Maceió, Alagoas, Brazil
erasmo.bezerra@lccv.ufal.br, antonio.ferro@lccv.ufal.br, francisco.junior@lccv.ufal.br,
lucasgouveia@lccv.ufal.br, aline@lccv.ufal.br

Abstract. Efforts to reduce drilling costs and duration have made accurate predictive models for rate of penetration (ROP) essential in the drilling industry. These models assist decision-making concerning parameters that affect drill efficiency. Utilizing advanced machine learning algorithms, such as ensemble methods and artificial neural networks, has become a clear trend aimed at enhancing predictive precision. In this study, a staked generalized ensemble model is introduced with the objective of improve the performance of ROP prediction. This approach combines four base learners, namely Random Forest (RF), Gradient Boosting (GB), Multiple Linear Regression (LR), and Artificial Neural Networks (ANN). The resulting meta-data from these models are used to make final ROP prediction using Ridge Regression algorithm. Drilling data from two wells in the Volve Field are used for training, including various operational and formation related parameters, such as Weight on Bit (WOB), Average Rotary Speed (RPM), Mud Flow Rate (FR), and Delta-T Compressional (DTC). The performance of the model is evaluated on an unseen well using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The proposed approach has demonstrated superior performance compared to the base learners, as indicated by the comparative analysis. This suggests its potential to enable more accurate predictions, consequently improving the efficiency of the drilling process.

Keywords: rate of penetration; stacking model; machine learning.

1 Introduction

In the oil and gas industry, the Rate of Penetration (ROP) is a key measure of drilling efficiency. A higher ROP reduces operational time and associated costs. Improving drilling efficiency solving an optimization problem that aims to maximize ROP as a function of operational parameters and geological data. An accurate ROP model is essential to achieve this goal.

Various empirical models for ROP prediction have been proposed, including those by Bingham [1], Bourgoyne Jr. and Young Jr. [2], Warren [3, 4], Hareland and Rampersad [5], Hareland [6], and Motahhari [7]. More recent research has applied machine learning regression to predict ROP. This approach involves identifying complex patterns in training data to make predictions in new datasets. Machine learning models have been demonstrated higher accuracy on real-world field data compared to traditional models, as shown by Soares and Gray [8] and Ferro [9]. This superior performance is attributed to the ability to capture the complex, and nonlinear relationships among the various variables influencing the drilling rate.

Multiple architectures have been used in ROP prediction. Artificial Neural Networks (ANN) are the most used algorithms, according to Barbosa [10] and Li [11]. This category includes Multi-Layer Perceptron (MLP), Radial Basis Function Neural Networks (RBFNN), Extreme Learning Machines (ELM), Adaptive Neuro-Fuzzy Inference System (ANFIS), and other algorithms. Recent studies have achieved accurate results in ROP prediction,

such as those by Ashrafi [12], Abbas [13], Brenjkar [14], and Shi [15].

Ensemble methods represent another popular group of techniques used in ROP prediction. Li [11] notes that most applications in this category rely on Random Forest (RF), an ensemble based on decision trees. The interpretability of RF is a notable feature frequently used for feature selection. Research by Soares and Gray [8], Ferro [9] and Shaygan and Jamshidi [16] demonstrates that RF models can outperform ANN in terms of accuracy when predicting ROP.

Stacked generalization is an ensemble learning technique that combines predictions from multiple individual algorithms using a meta-model to achieve higher prediction accuracy. Although less popular than other methods, this approach has the potential to improve ROP predictions by leveraging the strengths of the base learners. Promising results were achieved by Liu [17], who proposed a stacking model for ROP prediction using Extreme Gradient Boosting (XGB) as a meta-model and five base learners: Support Vector Regression (SVR), Extremely Randomized Trees (ET), Random Forest (RF), Gradient Boosting (GB), and Light Gradient Boosting (LGB). A similar approach was adopted by Alsaihati [18], using RF as a meta-model, and ANN and ANFIS as base learners. In both cases, the highest accuracy was achieved by the stacking model.

This study proposes an ROP model based on stacked ensemble learning, using four base learners, namely RF, GB, Multiple Linear Regression (MLR), and ANN. The stacked model combines meta-data from these models to calculate ROP predictions using Ridge Regression. A case study is presented using data from the Volve oil field. Two wells provide data for training, and the performance of the model's generalization is assessed using data from an unseen well. The stacking ensemble model was seen to outperform the base learners in terms of prediction accuracy.

2 Machine learning algorithms

Machine learning algorithms play an essential role in improving the accuracy of ROP prediction. In this section, some of the most popular approaches will be explored.

Random Forest (RF) is an ensemble machine learning algorithm that combines the predictions of multiple decision tree structures. A decision tree is built on a series of decisions, where each decision is based on variable values to choose one path or another [19]. In RF, each tree is created using the bagging method, which involves generating different random subsets of the original data by sampling with replacement. Moreover, during the construction of each tree, only a sample of features is randomly selected to consider for splitting at each decision node [20]. Each tree is trained on a different subset of training data. In regression problems, final predictions are obtained by averaging the predictions from all the trees. This process helps to reduce the variance and improve the overall model performance compared to a single decision tree.

Similarly, Gradient Boosting (GB) is another ensemble learning algorithm based on multiple decision trees. However, GB constructs trees sequentially, using information from previously grown trees [20]. Starting with an initial solution, the error between the predicted and observed values is evaluated using a loss function, typically the Mean Squared Error (MSE) in regression. A new tree is built to minimize this residual. This involves finding the best combination of features and split points for the decision tree. The process is repeated iteratively, with each tree aiming to reduce the accumulated residuals from previous trees. Final predictions are obtained by summing the predictions from all the trees [21].

Moving on to a linear technique, Multiple Linear Regression (MLR) assumes that there is approximately a linear relationship between the set of features and the target variable. The MLR model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{1}$$

where X_j represents the jth predictor, β_j is the coefficient that relate this feature and the response Y, and p is the number of predictors. The coefficients are obtained by minimizing the sum of squared residuals

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip} \right)^2 \tag{2}$$

where y_i is the ith target value observed, and $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_p$ are the values that minimize Eq. 1.

In problems with many predictors or when the predictors are highly correlated, there can be a lot of variability in the least squares fit, leading to overfitting and consequently poor predictions for new data. This can be handled by constraining the estimated coefficients, which can reduce variance with a small increase in bias [20]. One such approach is Ridge Regression, which imposes a penalty on the sum of squared, so that the function to be minimized becomes

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

where $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage applied to the coefficients.

In contrast to linear models, Artificial Neural Networks (ANN) are nonlinear models composed of layers of artificial neurons, each connected to the next through adjustable weights. Each neuron in a layer receives inputs, performs a linear combination of these inputs (by multiplying them with weights and adding a bias), and applies an activation function to produce an output. These outputs are then passed to the next layer, and the process repeats until the output layer is reached. The output of the neural network has the form

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j \right) \tag{4}$$

where X is an input vector of p variables, K is the number of hidden units, g is a nonlinear activation function, and the parameters $\beta_0, \dots, \beta_k$ and $w_{10}, \dots, w_{Kp}$ need to be estimated from data [20]. The activation function may be considered in many forms such as linear, sigmoidal and ReLU. Fitting ANN involves estimating the unknown parameters in Eq. 4. Popular strategies for training include gradient descent and regularization algorithms.

Finally, Stacked Generalization, also known as Stacking, is an ensemble learning technique introduced by Wolpert [22] that combines the predictions of multiple models to achieve a more accurate final prediction. This approach can be seen as a more sophisticated version of cross-validation, as it combines the strengths of each model, instead of simply select the single best-performing model.

The method consists of using predictions from base learners as inputs to train a meta-model, which is responsible for the final predictions. According to Breiman [23], Stacking is likely to be more beneficial when combining dissimilar models. Similar regressors tends to capture the same information from the data, leading to very similar predictions.

3 Methodology

Data from two wells in the North Sea Volve oil field are used to train the base learners and subsequently the meta-model. The stacking proposed structure is presented in Fig. 1. A total of 20038 data points are utilized to train, covering the lithologies of claystone, limestone, sandstone, siltstone, marl, and coal. The set of features includes mud logging and Logging While Drilling (LWD) information, such as Weight on Bit (WOB), Average Rotary Speed (RPM), Mud Flow Rate (FR), and Delta-T Compressional (DTC). The selected features include three operational parameters of interest for optimization, WOB, RPM and FR; and DTC, which provides geophysical information about the drilled formation. The model's performance is evaluated using five-fold cross validation. Additionally, 5457 data points from an offset well not used during training is also assessed.

The data were pre-processed with noise reduction using the Savitzky-Golay smoothing filter and data normalization [11]. Table 1 provides a data description of the data. Wells F15S and F10 were used for training and cross-validation, while F14 was used for testing.

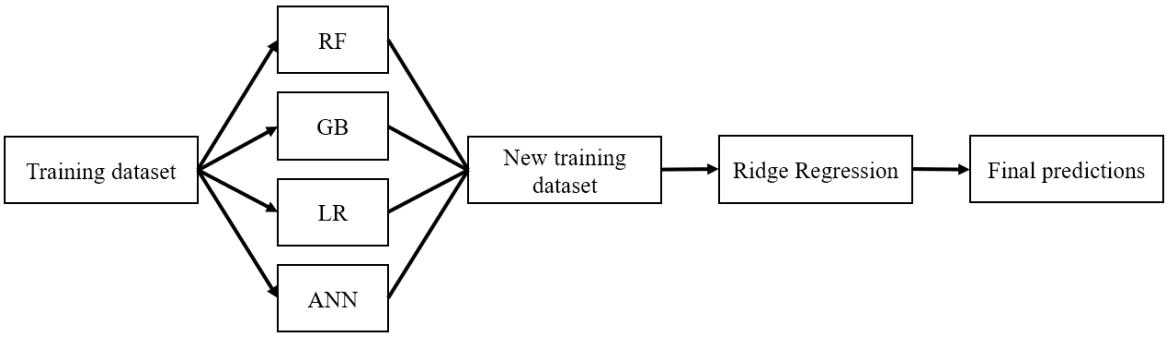

Figure 1. Stacked model.

Table 1. Data description per well.

Well	Number of points	Stat.	Measured depth [m]	ROP [m/h]	WOB [tf]	RPM [rev/min]	FR [L/min]	DTC [µs/ft]
F15S	15220	Min.	2554.83	0.666	0.000	0.00	791.376	52.922
		Max.	4062.98	55.836	22.060	290.856	2382.388	116.549
		Mean	3439.907	14.365	6.987	200.484	1825.933	73.215
F10	4818	Min.	3800.128	2.585	7.851	119.538	1929.294	53.103
		Max.	4299.725	34.649	17.569	204.885	2528.527	76.735
		Mean	4038.582	21.253	9.653	179.618	2324.392	61.023
F14	5457	Min.	2780.282	1.484	0.453	67.769	1650.853	58.825
		Max.	3465.850	39.274	15.187	181.856	2081.097	106.672
		Mean	3083.112	18.020	7.602	153.79	1973.95	80.262

The performance of cross-validation is assessed using the Mean Squared Error (MSE), calculated as

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

where n is the number of points, y_i is the ith observed target value, and $\hat{y}_i$ is the corresponding predicted value. In the test well, the performance is further evaluated using the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE), calculated as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

4 Results and Discussion

The base learners and meta-model were trained using five-fold cross validation. In this process, the training set is divided into five folds. In each iteration, one of the folds is used as the validation set, while the remaining four are used for training. The final validation metric was obtained by averaging the MSE from all five iterations, providing a more robust estimate of the model's performance. Hyperparameter tuning was performed using grid search, with the search space and selected parameters presented in Tab. 2.

Table 2. Hyperparameter grid search.

	Hyperparameters	Search space	Selected
RF	*n_estimators*	50, 100, 200	200
	min_samples_leaf	8, 10, 12	8
	max_features	*"sqrt"*	*"sqrt"*
	max_depth	8	8
GB	*n_estimators*	50, 100, 200	200
	min_samples_leaf	8, 10, 12	8
	max_features	*"sqrt"*	*"sqrt"*
	max_depth	8	8
	learning_rate	0.001, 0.005, 0.01	0.001
ANN	*hidden_layer_sizes*	5, 10, 20, 30, 40, (5, 5), (10, 10), (15, 15), (20, 20)	(15, 15)
	learning_rate	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}	10^{-3}
	activation	ReLU	ReLU
	max_iter	200	200
	early_stopping	True	True
	n_iter_no_change	15	15
	validated_fraction	15%	15%
	solver	SGD	SGD
Ridge	*alpha*	0.05, 1.0, 5.0	1.0

The validation metrics for each fold, as well as the mean and standard deviation, are presented in Tab. 3. The stacking model shows the lowest mean error, demonstrating an improved accuracy of 24.63% compared to the best performing base learner, GB. This suggests that the proposed model may be well-suited for the problem. In contrast, LR performs significantly worse, with high validation error, indicating it may not effectively capture the complexity of the data. RF, GB, and ANN models show intermediate performance, outperforming LR.

Table 3. MSE five-fold cross-validation.

	RF	GB	LR	ANN	Stacking
Split 1	9.744	9.045	284.839	18.297	7.067
Split 2	10.180	9.553	289.722	20.138	7.302
Split 3	9.416	8.685	293.743	18.254	6.441
Split 4	10.071	8.607	292.023	20.548	7.069
Split 5	10.076	9.317	289.479	18.308	6.801
Mean	9.897	9.207	289.961	19.109	6.936
Std.	0.281	0.310	3.002	1.016	0.294

Li [11] emphasizes the importance of testing and validating data-driven models under conditions that truly reflect the complexity and variability of the data they will encounter in practice. According to the same author, most of the ROP prediction studies use a random percentage of the original data for testing. Since this data is part of a continuous sequence, some of the testing data may have very little difference from those used for training, which limits the assessment of the model's generalization ability. True generalization capability is better evaluated by testing the model with data from a completely new well.

To assess the generalization capability of the models, an adjacent well was selected for testing. Figure 2 shows the observed and predicted ROP values for this well, along with the associated generalization errors. The proposed Stacking model achieved the lowest error compared to the base learners. With MAE values of 6.72 m/h and RMSE of 7.91 m/h; it demonstrated an improvement in accuracy ranging from 8.44% to 33.33% over the MAE, and from 6.05% to 33.27% over the RMSE.

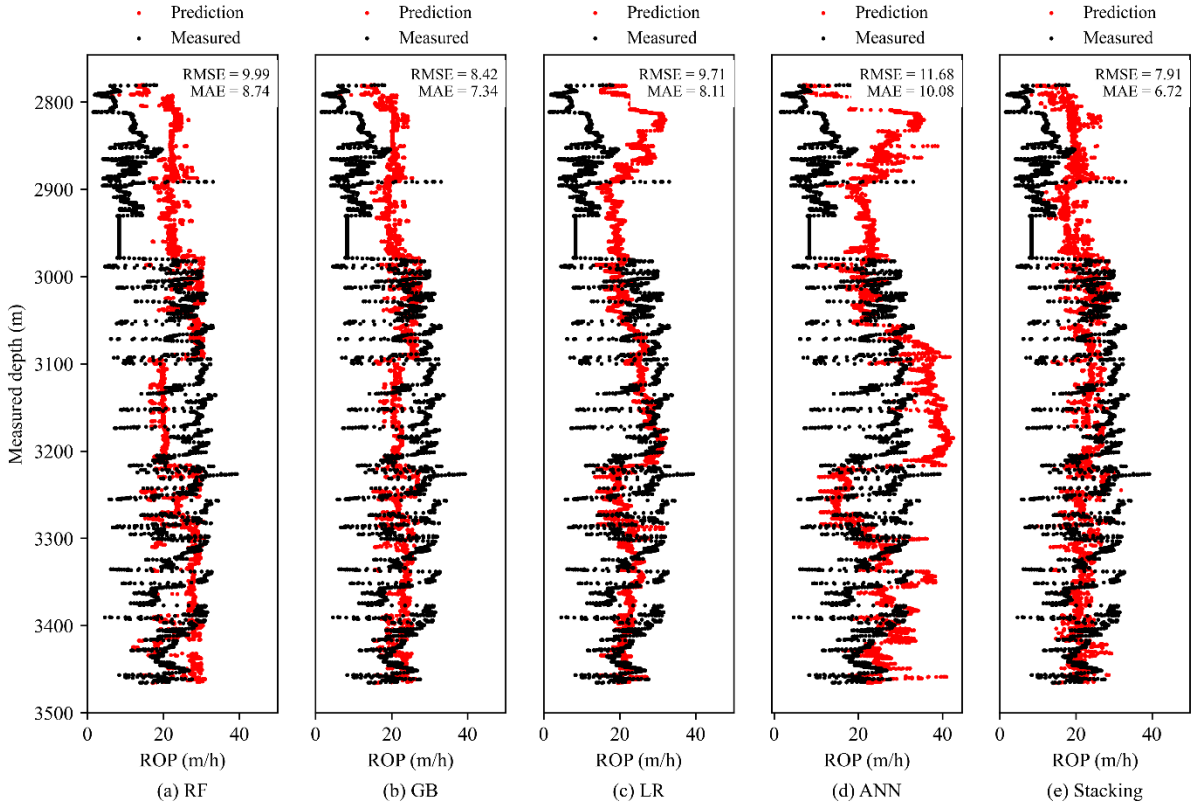

Figure 2. ROP predictions.

Although cross-validation showed that LR had the worst performance, the ANN stood out negatively by having the poorest performance on unseen data. This may suggest that the ANN is overfitting, meaning it is too closely tuned to the training data and struggles to generalize to new data. In contrast, the Stacking model demonstrated robustness against base learners' errors. By combining predictions from multiple models, the proposed model smooths out errors and showed superior generalization both in cross-validation and with new data for the study case.

5 Conclusions

The proposed Stacked Generalization Ensemble method for ROP prediction combines four base learners: RF, GB, LR and ANN. A study case using wells from North Sea Volve oil field showed superior generalization performance of the proposed model, outperforming the base learners both cross-validation and with data from an offset well not used during training.

For optimal results, it is important that the base learners are diverse, and that overfitting is properly managed.

By integrating the strengths of different models and reducing their errors, stacking models has the potential to enable more accurate ROP prediction. This could be valuable for recommendation systems of operational limits for drilling oil and gas wells, where precise ROP predictions are essential.

Acknowledgements. The authors acknowledge the financial and technical support provided by Petróleo Brasileiro S.A. – PETROBRÁS.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] M. G. Bingham, *A new approach to interpreting rock drillability.* Petroleum Publishing Company, 1965.
[2] A. T. Bourgoyne Jr. and F. S. Young Jr., "A multiple regression approach to optimal drilling and abnormal pressure detection". *SPE J.*, vol. 14, n. 4, pp. 371–384, 1974.
[3] T. M. Warren, "Drilling model for soft-formation bits". *J. Pet. Technol.*, vol. 33, n. 6, pp. 963–970, 1981.
[4] T. M. Warren, "Penetration rate performance of roller-cone bits". *SPE Drill. Eng.*, vol. 2, n. 1, pp. 9–19, 1987.
[5] G. Hareland and P. R. Rampersad, "Drag-bit model including wear". In: *SPE Latin America/Caribbean Petroleum Engineering Conference*, 1994.
[6] G. Hareland, A. Wu, B. Rashidi, et al., "A new drilling rate model for tricone bits and its application to predict rock compressive strength". In: *44th U.S. Rock Mechanics Symposium and 5th U.S.-Canada Rock Mechanics Symposium*, 2010.
[7] H. R. Motahhari, G. Hareland, and J. A. James, "Improved drilling efficiency technique using PDM and PDC bit parameters". *J. Can. Petrol. Tecnol.*, vol. 48, n. 10, pp. 45–52, 2010.
[8] C. Soares and K. Gray, "Real-time predictive capabilities of analytical and machine learning rate of penetration (ROP) models". *Journal of Petroleum Science and Engineering*, vol. 172, pp. 934–959, 2018.
[9] A. P. A. Ferro, Modelos preditivos para ROP como suporte à otimização em tempo real de parâmetros operacionais na perfuração de poços de petróleo. MSc Dissertation, Federal University of Alagoas, 2024.
[10] L. F. F. M. Barbosa, A. Nascimento, M. H, Mathias, et al., "Machine learning methods applied to drilling rate of penetration prediction and optimization – a review". *Journal of Petroleum Science and Engineering*, vol. 183, 2019.
[11] Q. Li, J. P. Li, L. L. Xie, "A systematic review of machine learning modeling processes and application in ROP prediction in the past decade". *Petroleum Science*, in press.
[12] S. B. Ashrafi, M. Anemangely, M. Sabah, et al., "Application of hybrid neural networks for predicting rate of penetration (ROP): a case study from Marun oil field". *Journal of Petroleum Science and Engineering*, vol. 175, pp. 604–623, 2019.
[13] A. K. Abbas, S. Rushdi, M. Alsaba, et al., "Drilling rate of penetration prediction of high-angled wells using artificial neural networks", *J. Energy Resour. Tecnol.*, vol. 141, n. 11, 2019.
[14] E. Brenjkar, E. B. Delijani, and K. Karroubi, "Prediction of penetration rate in drilling operations: a comparative study of three neural network forecast methods". *Journal of Petroleum Exploration and Production*, vol. 11, pp. 805–818, 2021.
[15] X. Shi, G. Liu, X. Gong, et al., "An efficient approach for real-time prediction of rate of penetration in offshore drilling". *Mathematical Problems in Engineering*, vol. 2016, 2016.
[16] K. Shaygan, and S. Jamshidi, "Prediction of rate of prediction in directional drilling using data mining techniques". *Journal of Petroleum Science and Engineering*, vol. 221, 2023.
[17] N. Liu, H. Gao, Z. Zhao, "A stacked generalization ensemble method for optimization and prediction of the gas well rate of penetration: a case study in Xinjiang". *Journal of Petroleum Exploration and Production Technology*, vol. 12, pp. 1595–1608, 2022.
[18] A. Alsaihati, S. Elkatatny, and H. Gamal, "Rate of penetration prediction while drilling vertical complex lithology using an ensemble learning model". *Journal of Petroleum Science and Engineering*, vol. 208, 2022.
[19] G. Rebala, A. Ravi, A., and S. Churiwala. *An introduction to machine learning.* Springer Cham, 2019.
[20] G. James, D. Witten, T. Hastie, et al. *An introduction to statistical learning with applications in Python.* Springer, 2023.
[21] J. H. Friedman, "Greedy function approximation: a gradient boosting machine". *Ann. Statist.*, vol. 29, n. 5, pp. 1189–1232, 2001.
[22] D. H. Wolpert, "Stacked generalization". *Neural Networks*, vol. 5, pp. 241–259, 1992.
[23] L. Breiman, "Stacked regressions". *Machine Learning*, vol. 24, pp. 49–64, 1996.