

## Performance analysis of random forest and support vector machine models in predicting pore pressure from well-log data

Gallileu Genesis<sup>1</sup>, Igor Fernandes Gomes<sup>1</sup>, José Antonio Barbosa<sup>2</sup>, Carlos Humberto Cuartas-Oquendo<sup>2</sup>

<sup>1</sup>*Dept. of Civil Engineering, UFPE.*

*Av da Arquitetura, s/n, Cidade Universitária, 50740540, Recife-PE, Brazil*

*gallileu.genesis@ufpe.br, igor.fernandes@ufpe.br*

<sup>2</sup>*GEOQUANTT Research in Geosciences, Dept. of Geology, UFPE*

*Av da Arquitetura, s/n, Cidade Universitária, 50740540, Recife-PE, Brazil*

*jose.antonio@ufpe.br, carlcartas@gmail.com*

**Abstract.** Pore pressure (PP) prediction is critical for well drilling operations and oil reservoir characterization and management. Recent advances in the development of Machine Learning (ML) models have led to a growing application of these methods for pore pressure prediction using well log records. In this work, we have evaluated the performance of two ML models for the task of PP prediction, one based on Random forest (RF) and another based on Support vector machine (SVM). The study used geophysical logs (Gamma-ray, Sonic, and Density) of stratigraphic wells drilled in the offshore Sergipe Basin, NE Brazil, to predict the PP in the regional sedimentary column of the basin. The values obtained by the ML models were compared with values of PP obtained by classic approaches used in the industry to establish the actual accuracy of the methods tested. We divided the data used in the study in training and testing into the proportion of 70% and 30%, respectively. We also used the metrics Mean square error – MSE and R-squared to evaluate the performance. The MSE of the SVM model was about one order of magnitude greater than that obtained by the RF in the training data. The validation data showed a similar result. This behavior appeared for different training data sizes, which shows the invariability of the relative performance of the models related to the amount of data used. Another aspect observed was the scalability of the models. The results show that the RF model presents a linear behavior concerning the model fitting time as a function of the amount of data, while the SVM model has an exponential behavior. Finally, in the test data, the RF model presented better results in all evaluated metrics, with an MSE of about 90%, which was lower than that obtained by the SVM model. By comparing the values predicted by the models and the actual values, the RF model has an r-squared of 0.99, while the SVM model has an r-squared of 0.96. Thus, the performance of the RF model was superior to that of the SVM in all treated aspects.

**Keywords:** pore pressure prediction, machine learning, random forest, SVM.

### 1 Introduction

Pore pressure (PP), also called formation pressure, is defined as the hydrostatic pressure exerted by the fluids (oil, water and gas) inside the pores of rocks in subsurface and can be expressed in the gradient form, as the change in pressure per unit depth (psi/ft or Pa/m) or, in field units, in (lbs/gal) Bruce and Bowers [1]. The formation of these pressures is associated with the compaction aspect of the different types of rocks and depends on depositional processes, sediment composition and types of the interpore fluids present in the early deposition.

Definition of PP during drilling operations is critical for the calculation of the weight of the drilling fluid, or drilling mud, which must be kept between the PP and the fracture pressure, which is defined as the pressure limit to create fractures in the formation. Thus, the prevention of uncontrolled artificial fracturing needs to be assured to avoid well damage and the consequent mud loss. On the other hand, if the weight of the mud is far below the pore pressure of geological formations, it can result in the catastrophic escape of fluids and trigger kicks and blowouts. The balance between these forces is particularly complex when the pore pressure is naturally close to the fracture gradient, a typical scenario found within overpressure intervals. Thus, understanding pore pressure behavior as a function of depth and identification of critical zones to help in the correct determination of

the drilling mud weight is a fundamental aspect of the oil industry operation.

Recent advances in the development of Machine Learning (ML) models have led to an increase in the application of these methods in the prediction of pore pressure, which aims more accurate models: Osarogiagbon [2], Abdelaal and Elkatatny [3], Yu [4], Ahmed [5], Hu [6], Keshavarzi and Jahanbakhshi [7].

In this work, we evaluated the performance of two ML models, one based on Random forest (RF) and another based on Support vector machine (SVM), through the prediction of pore pressures of stratigraphic wells drilled in the offshore Sergipe Basin, Northeastern Brazil. The PP estimation was achieved based on the relation of data from geophysical logs (gamma ray, sonic and density). The actual data used for comparison and training was provided by Cuartas-Oquendo [8], who calculated the PP gradients for the same wells through classical methods also used by the industry. This author calculated the PP gradient of wells drilled in the shallow and deep domains of the basin and also used seismic data as pseudo wells to establish a regional distribution of PP.

## 2 Background

### 2.1 Random forest

Random Forests are a type of ML algorithm, called ensemble, originally proposed by Breiman [9], in which a set of decision trees are combined to produce a robust prediction. For regression problems, the final prediction is given by the average of the predictions of all the trees that produced the forest (Fig. 1).

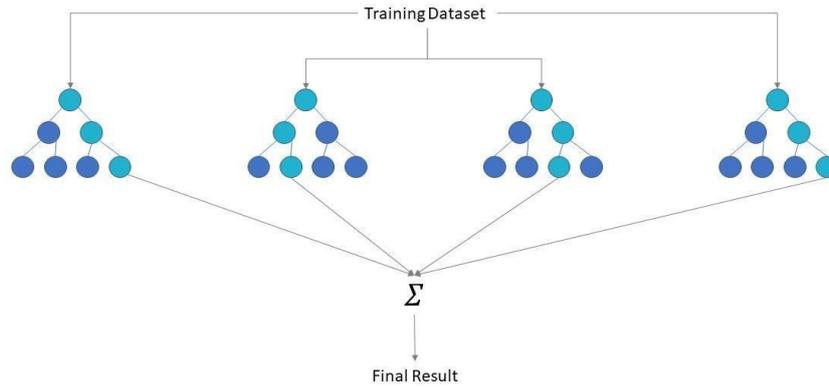


Figure 1. Random forest algorithm prediction process (IBM[10]).

In each tree, the regression process proceeds as follows: the predicted value at each node ( $\hat{y}_i$ ) will be the average of all instances ( $y_k$ ) belonging to the node (eq. (1)). Then, the Mean square error – MSE between the predicted value and the values of all instances present in the node is measured (eq. (2)).

$$\hat{y}_i = \frac{1}{n} \sum_{k=1, k \in i}^n y_k \quad (1)$$

$$MSE_i = \sum_{k=1, k \in i}^n (\hat{y}_i - y_k)^2 \quad (2)$$

The process of choosing the attributes that will compose the root node and the internal nodes is done recursively, over all training data features (or a subset of these). The best configuration will be the one that produces the smallest overall error. Thus, the split node will be considered the one that presents the minimum residual sum of squared errors or the mean squared error.

## 2.2 Support vector machine

Support vector machine is a discriminative machine learning model used for both classification and regression tasks, whose conception is based on statistical learning frameworks developed by Vapnik [11]. For linear regression problems an SVM model finds the hyperplane that best fits the data (solid line in Fig. 2), given a certain error tolerance, defined by an error margin, called maximum error ( $\varepsilon$ ). A higher value of  $\varepsilon$  leads to a higher error tolerance. Thus, the objective is to find a function  $h(x)$  (eq. (3)), with optimal parameters and whose errors with respect to the training points  $y$  do not exceed the value defined for  $\varepsilon$ .

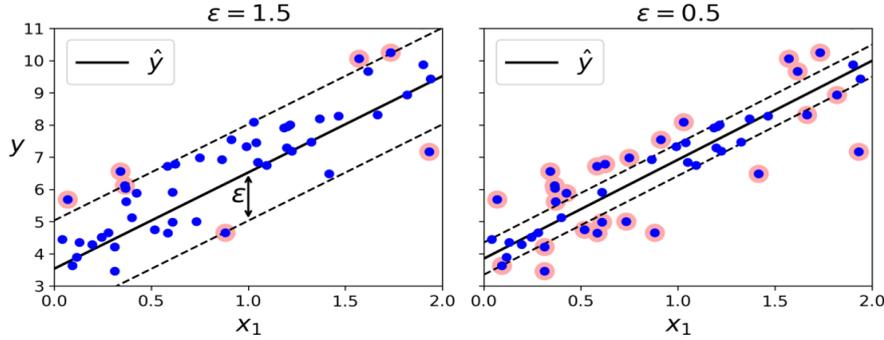


Figure 2: Regression scheme of a SVM model. Adapted from Gron [12].

In this way, let the hyperplane be defined by:

$$h(x) = W^T X + b = 0 \quad (3)$$

where  $X$  are the training features and  $W$  and  $b$  are the parameters to be optimized.

Let  $T = (X, y)$  be the training set, then the following convex optimization problem is formulated:

$$\min_{(W,b)} \frac{1}{2} W^T W \quad \text{s.t.} \quad |y_i - (W^T x_i + b)| \leq \varepsilon, \forall (x_i, y_i) \in T \quad (4)$$

It is possible to adjust the value of  $\varepsilon$  to obtain the desired model accuracy, as shown in Fig. 2. However, these restrictions may be too strict for certain problems. To work around this problem, a pair of slack parameters ( $\varepsilon_i, \varepsilon'_i$ ) is introduced for each training data point, so that the optimization problem results in eq. (5):

$$\min_{(W,b)} \frac{1}{2} W^T W + C \sum_i^n (\varepsilon_i + \varepsilon'_i) \quad (5)$$

subject to,  $\forall (x_i, y_i) \in T$ :

- $|y_i - (W^T x_i + b)| \leq \varepsilon_i + \varepsilon'_i$
- $\varepsilon_i \geq 0$
- $\varepsilon'_i \geq 0$

The constant  $C > 0$  is a regularization term that controls the trade-off between the optimization of the hyperplane parameters and the tolerance for errors that the model can make and, therefore, helps to avoid overfitting.

## 3 Methodology

The dataset, composed of a total of 28,946 samples of profiles of Gamma ray, Sonic and Density, was divided into training data (70% of the total) and test data (remaining 30%). The model's hyperparameters tuning were obtained using the RandomizedSearchCV function of scikit-learn.

We used Mean Square Error (eq. (6)) and R-squared (eq. (7)) as the metrics to evaluate the models.

$$MSE = \frac{1}{m} \sum_{i=1}^m (\hat{y}^i - y^i)^2 \quad (6)$$

$$R^2 = 1 - \frac{(\hat{y}^i - y^i)^2}{(\bar{y} - y^i)^2} \quad (7)$$

Where  $\hat{y}^i$  is the  $i$ -th prediction of the model,  $y^i$  is the corresponding real value and  $\bar{y}$  is the mean of the values.

## 4 Results and discussion

Figure 3 shows the mean squared error of the models during the hyperparameters tuning stage. The model based on SVM presents a good consistency between the training and test values, which shows a good control of overfitting. However, the final MSE is much higher and with much more variability than that obtained by the RF model.

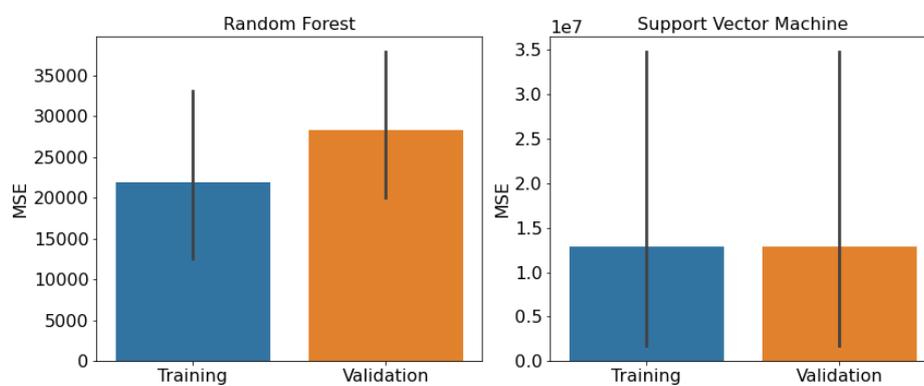


Figure 3. MSE scores obtained during the hyperparameter tuning process.

The learning curves in Fig. 4 show that both models have similar MSE scores in training and cross-validation. This procedure demonstrated a small generalization error. However, as in Fig. 3, the performance of the RF model is superior to the SVM. It is important to note that the SVM has consistently improved with the increases in training data, showing that, for a sufficiently large amount of data, its performance tends to approach the RF.

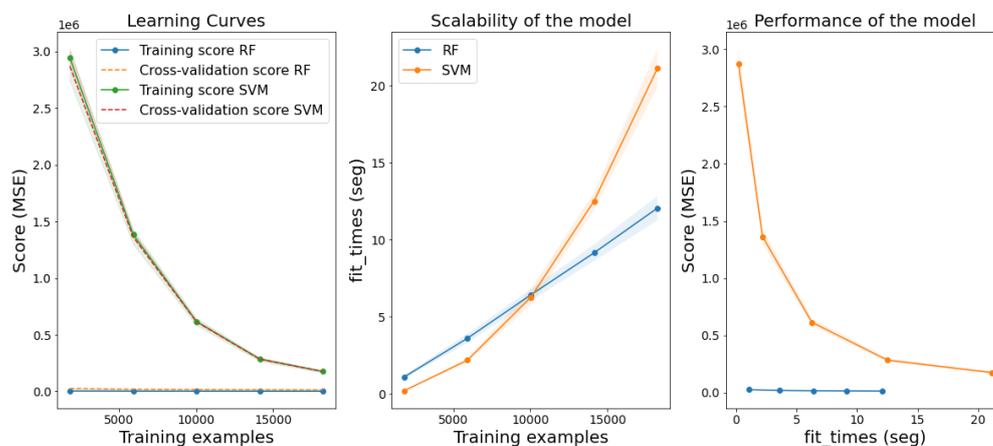


Figure 4. Learning curves (left), scalability (center) and performance (right) of the models.

Regarding scalability, results show that despite having an approximately linear behavior, the computational cost growth rate of the RF model is lower than that of the SVM for training sample sizes greater than 10,000. This fact can also be noticed by analyzing the performance of the models regarding the adjustment time.

The histograms in Fig. 5 show the distribution of the test data and of the values predicted by the RF and SVM models. A visual analysis shows that the RF model performed better than SVM.

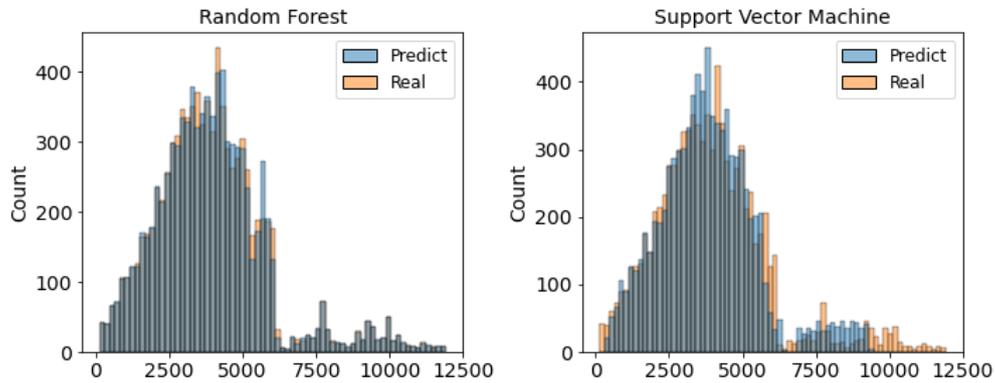


Figure 5. Histograms showing the distributions of predictions obtained by the RF (right) and SVM (left) models compared with actual values.

The qualitative interpretation of results shown in Fig. 5 is confirmed by the MSE scores of the models in the test data, which are shown in Fig. 6. The SVM score is 7 times higher than the RF score.

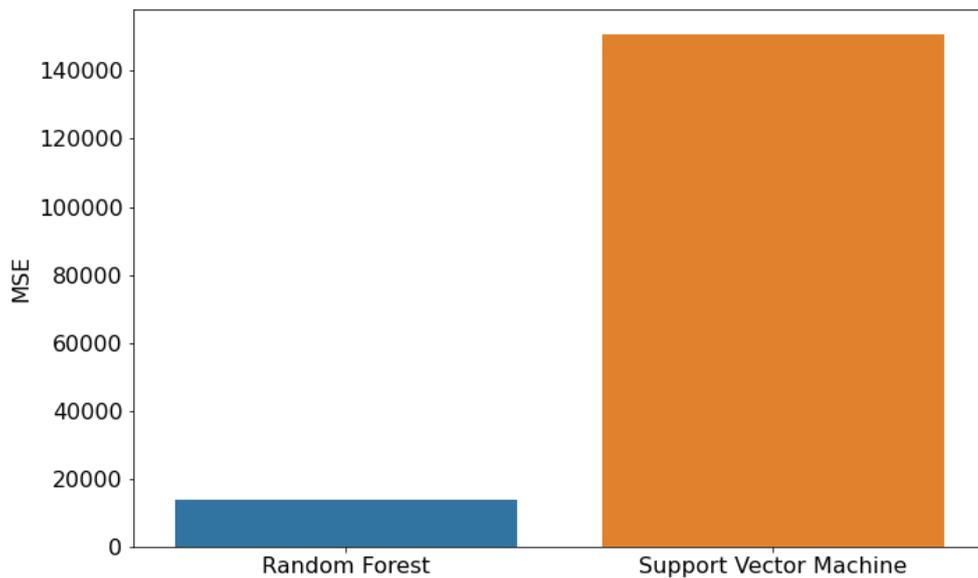


Figure 6. MSE scores of the models in test data.

Furthermore, the R-squared between the test data and the predictions of each model is shown in Fig. 7. One can note that the RF performs better, with more stable forecasts and an r-square of 0.996, compared to the SVM performance, which exhibits more dispersed forecasts and an r-square of 0.961.

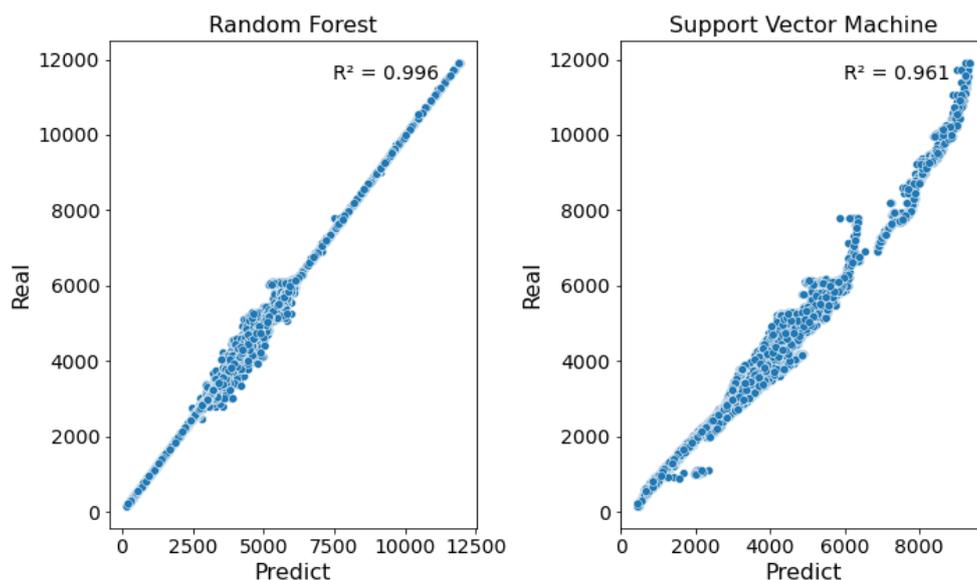


Figure 7. R-squared scores of the models in test data

## 5 Conclusion

The results showed that the RF model was superior to the SVM model performance in the pore pressure prediction task for the case study presented here. This superiority of the SVM model was observed in the stages of hyperparameter tuning, training, and testing in all analyzed metrics. In addition, the RF model proved superior in terms of scalability, with a lower computational cost than the SVM.

## References

- [1] B. Bruce and G. Bowers. Pore pressure terminology. *The Leading Edge*, 21(2), 170–173, 2002.
- [2] A. Osarogiagbon, O. Oloruntobi, F. Khan, R. Venkatesan, P. Gillard. Combining porosity and resistivity logs for pore pressure prediction. *Journal of Petroleum Science and Engineering*, 205(04), 108819, 2021.
- [3] A. Abdelaal and S. Elkatatny. Data-Driven Modeling Approach for Pore Pressure Gradient Prediction while Drilling from Drilling Parameters. *ACS Omega*, (05), 2021.
- [4] H. Yu, G. Chen and H. Gu. A machine learning methodology for multivariate pore-pressure prediction. *Computers Geosciences*, 143(07), 104548, 2020.
- [5] A. Ahmed, S. Elkatatny, A. Ali, M. Mahmoud and A. Abdurraheem. New Model for Pore Pressure Prediction While Drilling Using Artificial Neural Networks. *Arabian Journal for Science and Engineering*, 44(10), 2018.
- [6] L. Hu, J. Deng, H. Zhu, H. Lin, Z. Chen, F. Deng and C. Yan, Chuanliang. A new pore pressure prediction method-back propagation artificial neural network. *Electronic Journal of Geotechnical Engineering*, 18(01), 4093–4107, 2013.
- [7] R. Keshavarzi and R. Jahanbakhshi. Real-time prediction of pore pressure gradient through an artificial intelligence approach: A case study from one of middle east oil fields. *European Journal of Environmental and Civil Engineering*, 17(09), 2013.
- [8] C. H. Cuartas-Oquendo. *Regimes de pressão de poro na porção Sul offshore da sub-bacia de Sergipe, NE do Brasil*. Ph.D. thesis, Universidade de Brasília, 2020.
- [9] L. Breiman. Random Forests. *Machine Learning* 45 (1) pp. 5–32, 2001.
- [10] IBM. *Random Forest*. *IBM Cloud Learn Hub*, accessed 09 June 2022, <<https://www.ibm.com/cloud/learn/random-forest>>.
- [11] V. Vapnik and A. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theoretical Probability and its Applications*. 17(01), 264–280, 1971.
- [12] A. GRON. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2ed edn. O’Reilly Media, Inc, 2019.