METHODOLOGICAL DEVELOPMENT FOR THE COMPARISON OF MULTIPLE MACHINE LEARNING ALGORITHMS IN THE RECONSTRUCTION OF WELL LOGGINGS IN THE MIDDLE MAGDALENA VALLEY BASIN IN COLOMBIA

Franco Bertaiola Ríos, Luis Fernando Duque and Andrés Mauricio Muñoz García
Grupo de Geofísica y Ciencias de la Computación (GGC3), Instituto Tecnológico Metropolitano de Medellín (ITM)

Abstract

In the Middle Magdalena Valley Basin (MMVB) in Colombia, there is the Mugrosa formation of high oil interest, since large hydrocarbon reserves accumulated in it. of the various wells that were drilled in the area, the three of main interest for this research are Tenerife 1 (T1), Tenerife 2 (T2) Of the various wells that were drilled in the area, the three of main interest for this research are Tenerife 1 (T1), Tenerife 2 (T2) where the dipolar P wave (DTCO), dipolar S wave (DTSM), and Gamma-Ray (GR) well logs were partially registered, and Tenerife 3 (T3) that does not have such logs. The lack of this information becomes an obstacle for present and future research and industrial projects in this and other areas of the country, since completing them by traditional methods is technically and economically unfeasible.

The present research proposes the implementation of a range of Machine Learning (ML) methods such as Random Forest, KNN, Gradient Boosting, AdaBoost, Multi-linear Regression, and Artificial Neural Networks, to provide an economically viable solution. to this problem, since the computational alternatives proposed in this work present a much lower cost than the entire technical and logistical process of obtaining records traditionally.

Introduction

The Magdalena Middle Valley Basin (MMVB) in Colombia is one of the most economically and historically important sedimentary basins in the country due to the exploitation of hydrocarbons (Velásquez-Espejo, 2011), in this, the Tenerife field is located, where three holes were made T1, T2 and T3 (Tenerife-1, Tenerife-2, and Tenerife-3 respectively), where the P-wave (DT) sonic records were taken, potential spontaneous (SP), density (RHO) and deep induction resistivity (ILD), in addition, in the first two there is partial information from the dipolar S-wave (DTSM), dipolar p-wave (DTCO) and lightning gamma (GR) and do not exist for the last well. Additionally, pseudo-records of porosity (PHIE), clay volume (Vclay), and water saturation (Sw) were calculated using rock physics.

The almost complete depletion of the conventional resources of the area today forces to explore other types of geological formations in which the data were not taken, however, repeating the drilling process would be technically and economically unfeasible. To solve this type of problem, many computational techniques have been developed over the years, such as parametric-based ones such as multi-component induction well-logging (MCIL) (Wang et al., 2008) and autoregressive models (Bianchin et al., 2019) and lately, those based on artificial intelligence (AI) techniques, where different Machine Learning (ML) algorithms are used, the objective of which is to make the program learn from known information (well logs) and be able to predict the missing information. Among them are multiple regression algorithms and neural networks (NN) that were used in South-West Iran (Eskandari et al., 2004) to reconstruct shear wave velocity from log data. Parapuram et al. (2018) uses ANNs to predict geomechanical properties of the Bakken formation in North Dakota (USA) and a special type of ANN called recurrent (RNN) are used by Zhang et al. (2018) to generate synthetic well data and compare the results with those obtained by a classic NN (Full Connected). These ML models have given promising results in the prediction of geophysical data, however, although ML has been showing good performance in geosciences over the years (Bergen et al., 2019), in Colombia the application of these tools Mainly in the field of Petrophysics , there has been little, even so, there are works such as the one developed by Iturrarán-Viveros et al., (2018) where the correct performance of the NN is satisfactorily validated to determine petrophysical properties in the same field Tenerife in Colombia, which motivates us to develop more studies on ML Applications in MMVB and Colombia.

The data from wells T1, T2 and T3 (Figure 1) are divided into two groups as follows: the input, which are present throughout the entire drilling as Depth, DT, SP, RHO, PHIE, Vclay, Sw and ILD and the objective data GR, DTCO and DTSM, which are partially along T1 and T2, and are unknown in T3.

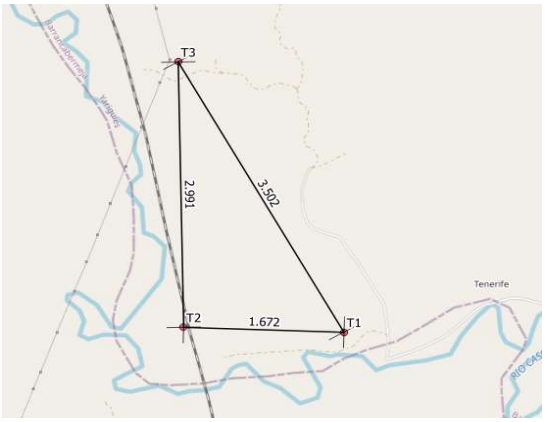

Figure 1. Distribution of Wells T1, T2 and T3, distance in meters

To do this and obtain the ML model with the best results, the task was approached from a generalized perspective that seeks to obtain, through a methodology, the best ML method for the area of interest, but which in turn, When using it in any other area, it is fully functional, that is, in this work we implement a methodology that autonomously compares six different ML methods, first by varying the hyperparameters of each of these, thus finding the optimal referent values. to the information of interest, using cross-validation as a selection parameter, later it creates the six best models and confronts them with each other so that in the end only the model with the best results remains.

Method

Currently, geoscientists face increasingly overwhelming amounts of data, they are in the task of extracting as much useful information as possible from these, for this they use tools such as Machine Learning (ML), which has demonstrated its efficiency when reconstructing geological information (Bergen et al., 2019). However, when addressing a data science problem, the distribution and correlation of information are one of the deterministic factors to obtain a good result (Provost & Fawcett, 2013), this is mainly due to the architecture of the different ML methods and their interaction with data; To achieve this good Data-Model combination, a methodology was developed, which is divided into three stages and aims to identify the ML algorithm that provides the best performance according to the distribution of the data in the target area; This methodology is summarized in Figure 2 and was used to successfully reconstruct some missing oil well logs in the MMVB in Colombia. Each stage is described below:

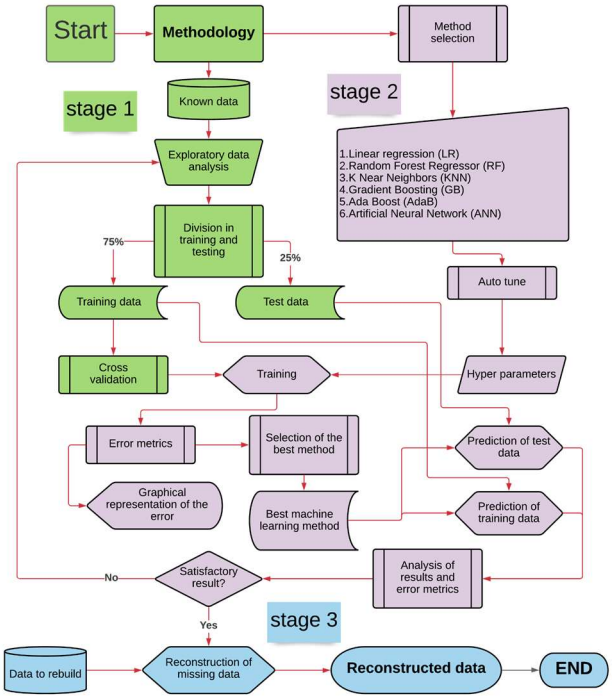

Figure 2. Applied methodology.

Stage 1: Exploratory Data Analysis (EDA): The EDA is one of the fundamental bases for obtaining new information from known data sets. It is based on the use of a series of tools that allow knowing the nature of the existing information to evaluate the most relevant parameters according to the analyst's experience (Milo & Somech, 2020). Among the various tools for exploratory analysis are filtering, aggregation, visualization and distribution techniques. In the analysis process, the two main points that were taken into account were the following:

I. Consistency of the data: through a statistical description, the distribution of each of the input columns was visualized, to find typing errors, anomalous data and behavior of the data by percentiles.

II. linear correlation and independence: Using a correlation matrix it was observed that the depth presents a great correlation with the output data, something to be expected due to the formations present in the well, but when predicting the records of another well where the formations are in different depths, this dependency increased the error, therefore this column was eliminated.

Stage 2: Algorithm

a. Method selection: the algorithm consists of 6 ML methods, from which you can choose with which you want to carry out the entire selection, prediction and validation process, the ML models available for this work are:

I. Linear regression (LR)

II. Random Forest Regressor (RF)
III. K Near Neighbors Regressor (KNN)
IV. Gradient Boosting Regressor (GB)
V. AdaBoost (AdaB)
VI. Artificial neural networks (ANN)

b. Auto tune: The hyperparameters of ML methods, significantly the performance, whether in the computation time, overtraining, reliability of the result, among many other characteristics. The autonomy of the algorithm allows it to search, using tools from the Sklearn Python library, for the combination of hyperparameters that provide the best score in cross-validation with the data to be analyzed for each of the ML methods selected from the main algorithm, as in the case of Paper (2020), where they explain how to implement GridSearchCV to tune hyperparameters in ML.

c. Error metrics: Various statistical tools were used to determine how accurate each of the models is and then compare them with each other and choose the best one, the metrics used are:

I. MAE (Mean absolute error)
II. MSE (Mean square error)
III. RMSE (Root mean square deviation)
IV. Percentage absolute error
V. Normalized error

d. Best method comparison: After obtaining the evaluation metrics of each method, the algorithm analyzes which of these selected ML models gives the best result using the absolute error as a determining parameter of selection, then the algorithm makes a complete prediction of the data used for testing. and returns all the testing information in a DataFrame to thoroughly analyze its performance, in addition to returning the optimal model trained to make subsequent predictions.

e. Already having the best models from the process seen so far, we proceed to make a cross-prediction of the two databases and their combination, that is, use the model obtained from T1 to predict T2 and vice versa, then join T1 and T2 in the same TT (Tenerife Total) database to carry out the entire algorithm process and subsequently analyze its performance.

Stage 3: Reconstruction

a. Reconstruction of missing section of T1: the ML model obtained from the process in numbers 1 and 2 applied to T1 is used to predict the unknown zone of said well.

b. Reconstruction of missing section of T2: The ML model obtained from the process in numbers 1 and 2 applied to T2 is used to predict the unknown zone of said well.

c. Reconstruction of T3: The ML model obtained from the process in numbers 1 and 2 applied to TT is used to completely reconstruct the unknown records of well T3.

Results

Various experiments were carried out in wells T1 and T2, in these it was observed that when predicting T1 with the model obtained from T2 and vice versa the error increased, in the case of the DTSM prediction of T1 predicted with T2 an absolute error was reached approximately 20%, which is not a satisfactory error for this work, so a series of analyzes and tests were carried out with different combinations of input data finally, it was concluded that depth (Depth) is the parameter that generates this error, so by repeating the previous prediction, but without using Depth as an input parameter, it was possible to reduce the percentage of DTSM error of T1 predicted with T2 at 8%, a significant improvement (more than 10 percentage points), the DTCO also improved, it went from 10.6% to 5.7% absolute error, this is reflected in Tables 1 and 2 However, we could see that the GR went from a 13.2% error to 14.4%, which was assumed to be an acceptable cost due to the improvements obtained in the other predictions; It should be noted that as the various experiments were carried out, the algorithm developed seeks the ML method that will provide the best result for the Input-Output combination of the data, therefore, as seen in tables 1 and 2, the methods of ML used for the two experiments presented are not the same, but they are the ones with the best results for each case, which is precisely what is sought with this algorithm, adaptability to various scenarios.

Table 1

Prediction Metrics T1 with T2 (without depth)

Parameter	Method	% Mean error	MAE
GR	RF	14,41043	9,943142
DTCO	RF	5,691536	4,538581
DTSM	RF	8,006072	12,44406

Table 2

Prediction Metrics T1 with T2 (with depth)

Parameter	Method	% Mean error	MAE
GR	GB	13,157799	8,190888
DTCO	KNN	10,60402	8,668988
DTSM	KNN	19,424515	30,77529

Knowing the combination of the data that gives us the best results to predict records in distant wells, an analysis was carried out in the same well to obtain the information on the behavior of the ML methods for each one and thus measure their performance at the time to rebuild information about itself, this was done three times, once for each perforation (T1 and T2) and a third for the combination of these (Tenerife Total “TT”), as can be seen in Table 3, they were obtained More than satisfactory results, however, when analyzing the ranges in which the data varies, it was noted that the absolute error does not always reflect the reality of the data when the magnitudes present a very large offset, therefore a normalized error was added for taking into account the uniqueness of each well log.

Table 3				
Tenerife 2 metrics				
Parameter	Best method	% Mean error Test	Test MAE	% Normalized error
GR	RF	4,948	2,989	3,621
DTCO	RF	4,111	3,480	5,515
DTSM	RF	5,551	9,547	5,620

Now, seeing the information in Table 3, the normalized error at first glance does not seem to give new information, but when applied to the prediction made at T2 using T1 shown in Table 4, we see that it becomes relevant because shows us a uniformity in the predictions, which is not evidenced in the absolute error, this allows us to identify that in the fluctuation zone, the predicted and the real information differ homogeneously around 10%, which is a promising.

Table 4				
Metrics Prediction T2 with T1				
Parameter	Method	% Mean error	MAE	% Normalized error
GR	RF	12,34551	7,479539	9,059519
DTCO	RF	6,594747	5,967743	9,457595
DTSM	RF	9,420959	17,48694	10,29492

After obtaining these results, we advanced to the analysis of the TT model, which was used to predict T3, since it is the one that presents more generalized information; As can be seen in Table 5, which shows the result of the TT test, this model presents better results when predicting DTCO and DTSM than the T2 model for itself as can be seen in Table 3, this gives us to understand, that by joining information from several wells to develop a generalized model, it is possible to have an even greater precision than that of a model created only with information from itself , however to be able to ensure this, it is necessary to carry out more studies and determine what other parameters influence this. Even so, having satisfactory results such as those obtained in the ML Models developed, the reconstruction of the missing information of T1, T2 and the total reconstruction of T3 continued.

Table 5				
Tenerife Total Metrics				
Parameter	Best method	% Mean error Test	Test MAE	% Normalized error
GR	RF	5,269	3,248	3,93
DTCO	RF	3,677	3,070	4,866
DTSM	RF	5,144	8,558	5,038

In Figure 3 we can see the behavior of the prediction in T1 with TT and the union with the reconstruction of the missing data, it can be seen that the model presents a continuity with the already known data and that its fluctuation range is consistent with that of the existing information, it should be noted that as seen in the area where the predicted and real information coexist, the prediction model faithfully follows the movement of the signal but does not follow the points where peaks occur, this is a good thing, because if it did, it would show an over-fit of the model, which would make the reconstruction of the missing information unfeasible, Figure 4 shows the correlation between the real data and the predicted GR in T1 Using the TT model, this shows the high density in the central area and as it decreases at the ends of the scatter diagram, this reflects the peaks that were discussed previously, which the algorithm is not able to follow.

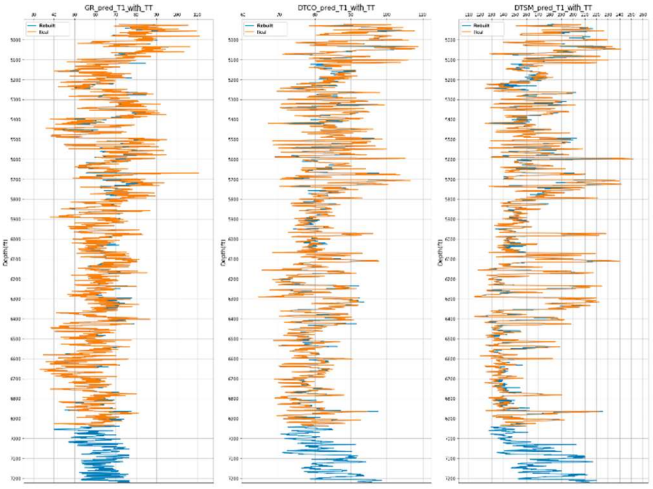

Figure 3. Reconstruction of T1 with TT, GR, DTCO and DTSM respectively, Orange real value, blue reconstructed value

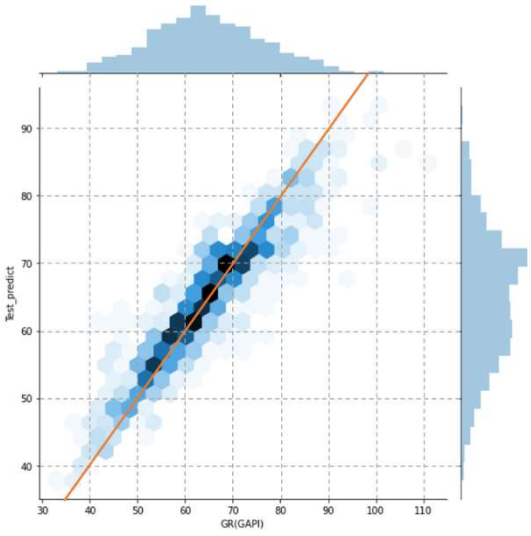

Figure 4. Scatter plot and histogram of the prediction of the GR test in T1 with TT

Finally, the T3 prediction was carried out using the TT model. As it does not have information on the predicted data, the metrics that were used previously cannot be

presented, however, an indication that the model is correct is the fluctuation ranges of the estimated records. In Figure 5 we can see that T3 reconstructed GR, DTCO and DTSM vary in a similar range as in the case of T1 reconstruction (Figure 3)

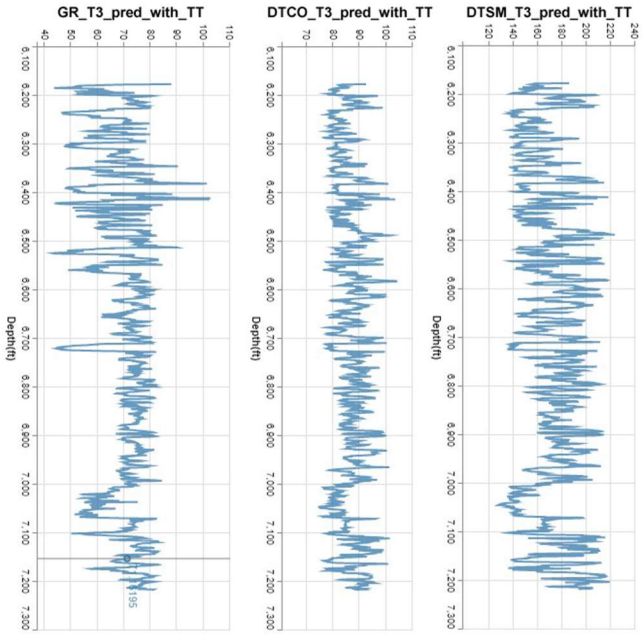

Figure 5. Reconstruction of T3 with TT, GR, DTCO and DTSM respectively.

Conclusions

The "classic" ML models presented better results than ANN in all cases, this agrees with Bergen et al. (2019), where it is said that this happens due to the reduced volume of data with which it works, in turn, it was evidenced that a generalized model (TT) when predicting the records with which the error was trained is similar to the average of the errors of the unique models of each well, but the generalized model, allows to predict information of unknown wells with greater precision.

For future work, it is expected to be able to validate the reconstructed information from Geophysics, analyzing its consistency with respect to local geology and also to be able to test these tools in other geoscience environments such as in the mining sector.

Acknowledgments

We thank the Instituto Tecnológico Metropolitano of Medellín (ITM), project code P20248, Ecopetrol S.A. and Colciencias for the information and data provided under the 0266-2013 Colciencias-Ecopetrol program.

References

Bergen, K. J., Johnson, P. A., Hoop, M. V. de, & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*(6433). https://doi.org/10.1126/science.aau0323

Bianchin, L., Forte, E., & Pipan, M. (2019). Acoustic impedance estimation from combined harmonic reconstruction and interval velocity. *GEOPHYSICS*, *84*(3), R385-R400. https://doi.org/10.1190/geo2018-0296.1

Eskandari, H., Rezaee, M., & Mohammadnia, M. (2004). Application of multiple regression and artificial neural network techniques to predict shear wave velocity from wireline log data for a carbonate reservoir South-West Iran. *CSEG recorder*, *42*, 48.

Iturrarán-Viveros, U., Muñoz-García, A. M., Parra, J. O., & Tago, J. (2018). Validated artificial neural networks in determining petrophysical properties: A case study from Colombia. *Interpretation*, *6*(4), T1067-T1080. https://doi.org/10.1190/INT-2018-0011.1

Milo, T., & Somech, A. (2020). Automating Exploratory Data Analysis via Machine Learning: An Overview. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2617-2622. https://doi.org/10.1145/3318464.3383126

Paper, D. (2020). Scikit-Learn Classifier Tuning from Complex Training Sets. En D. Paper (Ed.), *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python* (pp. 165-188). Apress. https://doi.org/10.1007/978-1-4842-5373-1_6

Parapuram, G., Mokhtari, M., & Ben Hmida, J. (2018). An Artificially Intelligent Technique to Generate Synthetic Geomechanical Well Logs for the Bakken Formation. *Energies*, *11*(3), 680. https://doi.org/10.3390/en11030680

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51-59. https://doi.org/10.1089/big.2013.1508

Velásquez-Espejo, A. J. (2011). *3D multicomponent seismic characterization of a clastic reservoir in the Middle Magdalena Valley Basin, Colombia* [Text, Colorado School of Mines]. https://mountainscholar.org/handle/11124/78119

Wang, H., Tao, H., Yao, J., & Chen, G. (2008). Fast Multiparameter Reconstruction of Multicomponent Induction Well-Logging Datum in a Deviated Well in a Horizontally Stratified Anisotropic Formation. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(5), 1525-1534. https://doi.org/10.1109/TGRS.2008.916080

Zhang, D., Chen, Y., & Meng, J. (2018). Synthetic well logs generation via Recurrent Neural Networks. *Petroleum Exploration and Development*, *45*(4), 629-639. https://doi.org/10.1016/S1876-3804(18)30068-5