CHEMICAL ENGINEERING TRANSACTIONS
VOL. 117, 2025
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Fabrizio Bezzo, Flavio Manenti, Gabriele Pannocchia, Almerinda di Benedetto

ISBN 979-12-81206-17-5; **ISSN** 2283-9216
DOI: 10.3303/CET25117186

Development of Machine Learning Models for Sandface Pressure Prediction in Oil Well

Lorraine P.Oliveira[a], Raul M.Foronda[a], Alexandre V.Grillo[b], Brunno F. dos Santos[a*]

[a]Department of Chemical and Material Engineering (DEQM). Pontifical Catholic University of Rio de Janeiro (PUC-Rio). Rua Marquês de São Vicente, 225 – Gávea, Rio de Janeiro – RJ, 22430-060, Brazil
[b]Federal Institute of Education, Science and Technology (IFRJ), Rua Lúcio Tavares, 1045 – Centro, Nilopolis – RJ, 26530-060, Brazil.
bsantos@puc-rio.br

The Oil & Gas (O&G) industry is increasingly leveraging Machine Learning (ML) techniques to predict well performance indicators, estimate missing operational metrics, and mitigate unexpected operational failures. However, the availability of extensive and high-quality datasets remains a major challenge due to the diversity of well characteristics and proprietary industry constraints. This study employs the Society of Petroleum Engineers' (SPE) Rate Transient Analysis (RTA) dataset to predict Sandface Pressure or Bottom Hole Pressure (BHP) using Decision Tree (DT) models. Initially, a comprehensive literature review on RTA was conducted, followed by an in-depth evaluation of the dataset and its variables. Feature selection was performed based on data availability, Spearman's correlation analysis, and Principal Component Analysis (PCA), leading to the exclusion of the "Oil Volume" variable from model training to improve predictive performance. The optimal DT model configuration was determined through cross-validation, utilizing Scikit-learn's GridSearchCV for hyperparameter optimization. The best-performing model achieved an $R2$ score of 0.982 and a Mean Squared Error (MSE) of 4.878×10^{-4}, demonstrating that RTA data effectively supports BHP prediction and that DT models are well-suited for this application. These findings highlight the potential of data-driven approaches in enhancing predictive analytics for well performance monitoring and optimization in the O&G sector.

1. Introduction

In 2023, fossil fuels accounted for 78 % of the global energy demand, and despite ongoing efforts for the current energy transition, they are estimated to remain responsible for at least 40 % by 2050 (MCKINSEY & COMPANY, 2024). Similarly, the oil and gas industry is expected to generate at least $ 2.5 trillion by 2030 (Deloitte, 2023). The application of Machine Learning in the Oil & Gas industry enables filling gaps in well logs and improving the quality of exploration data (BROWN; ROUBÍCKOVÁ; NORDLOH, 2020). To meet global economic needs while reducing operational costs and increasing the efficiency of O&G processes, new ML technologies have gained popularity in literature.Sandface pressure or Bottom Hole Pressure (BHP) is a crucial variable in O&G operations. It represents the pressure at the bottom of a well and can be used for dynamic monitoring of an operation, assisting in determining a well's production, or for more comprehensive evaluations such as Rate Transient Analysis (RTA) or Pressure Transient Analysis (PTA), which can determine productivity, extraction time, and geological characteristics of a well (Taghavinejad et al., 2022). BHP is essential for predicting reservoir performance, and optimizing recovery strategies, such as water or gas injection to maintain reservoir pressure. Factors influencing BHP include reservoir pressure, formation characteristics (permeability and porosity), production rates, fluid properties (viscosity and compressibility), and well conditions. BHP is measured using downhole instrumentation or estimated through numerical modeling and simulation, often supported by software such as IHS Harmony and Schlumberger's ECLIPSE (IHS Markit, 2020).
RTA has been widely used to interpret O&G data (such as phase flow rates, pressures, and production line pressures) and provide insights into reservoir behavior. RTA combines phenomenological models with production data to analyze how production rates vary over time, thereby inferring critical reservoir parameters such as sandface pressure (Clarkson et al., 2016). The quality and quantity of data available for training ML

Paper Received: 18 December 2024; Revised: 1 March 2025; Accepted: 30 April 2025
Please cite this article as: Oliveira L.P., Foronda R.M., Grillo A.V., dos Santos B.F., 2025, Development of Machine Learning Models for Sandface Pressure Prediction in Oil Well, Chemical Engineering Transactions, 117, 1111-1116 DOI:10.3303/CET25117186

models are highly relevant factors in their predictive capacity (Géron, 2019). Some methods aim to improve data quality through variable selection or extraction techniques that reduce data dimensionality, such as Principal Component Analysis (PCA) (Kurita, 2020). PCA is widely used in the literature as an effective way to describe data variance or work with simpler components that accurately represent a dataset.

Obtaining real operational data for ML research in the O&G field remains a challenge, but some recent initiatives aim to change this scenario, such as Petrobras' 3W database, which gathers pressure and flow data from oil extraction wells (Vargas, R. E. V. et al., 2019). Another important initiative is the release of the public data repository by the Society of Petroleum Engineers (SPE), which consolidates data from three databases, including RTA analyses (Society of Petroleum Engineers, 2024). The SPE repository includes operational data from 53 wells, providing resources for researchers to develop ML models for reservoir analysis. Despite the amount of data, the RTA dataset, a key part of the SPE repository, has yet to be fully explored in terms of variable context and distribution, necessitating assumptions for ML development.

ML is a transformative approach within artificial intelligence, it enables systems to learn patterns from data and improve performance without explicit programming (Murphy, K. P., 2012). Supervised learning, particularly regression models, is fundamental for predicting continuous outcomes. Models such as Linear Regression, Decision Tree Regression, and Neural Networks address varying complexities and data types (Bishop, 2006; Quinlan, 1986; Goodfellow et al., 2016). These models have applications across industries, including forecasting market trends, optimizing production schedules, and predicting patient outcomes (Domingos, P., 2012). Advances in ML are driven by the increasing availability of large datasets and computational resources, enabling the development of more sophisticated and accurate models to tackle real-world problems (Smola, A. J. & Schölkopf, B., 2004).

This study used a ML model to determine BHP using RTA data provided by SPE. The dataset used contains detailed production information from 53 oil wells. Using ML for BHP determination can represent an operational gain in O&G activities, supporting platform operators in monitoring such a critical process parameter.

The aim of this work was to develop a Machine Learning model to determine BHP using RTA data provided by SPE, including evaluating the provided RTA data, identifying variable correlations through correlation matrices, applying PCA to identify the best variables that describe the dataset, and building a predictive ML model for BHP.

2. Methods

The workflow applied by this work follows three main steps: Data acquisition and evaluation, variables selection and model training; those steps will be better described by this section.

2.1 Data acquisition and cleaning

From SPE dataset, all data used in this work has been collected from dataset_1_all_wells's dataset "production_data.csv" file (SPE Data Repository: Data Set: 1, 2024). Each RTA analysis contains 13 variables: Lease, Time (days), Choke Size, Gas Volume (MMscf), Oil Volume (Stb), Water Volume (Stb), Gas Lift Inj Volume (MMscf), Casing Pressure (psi(a)), Tubing Pressure (psi(a)), Active Pressure (psi(a)), Line Pressure (psi(a)), Pressure Source, and Calculated Sandface Pressure (psi(a)).

Data processing, model training, and graph visualization were performed using Python version 3.10 (Python, 2019), through Google Colab. The source code for this work is available on the GitHub platform in a public repository (Freire, R. M., 2024). The main Python libraries used in this work were: sklearn, prince, pandas, numpy, seaborn, and matplotlib. Pandas dropna function was used to remove lines with only non-numerical values.

2.2 Variable selection

In order to better understand the statistical behaviour and relationship, spearman correlations and PCA methods were employed by this work.

2.2.1 Spearman correlation

Reading the SPE dataset with pandas, spearman correlation was generated function "corr" using method as "spearman" (e.g. "DataFrame.corr(method='spearman')"), in order to improve data visualization, a heatmap was plotted using the library seaborn's heatmap function (e.g. "seaborn.heatmap") and spearman correlation's data.

2.2.2 PCA

Prince library was employed to perform PCA analysis in order to evaluate variables correlation and distribution. The non default parameters employed were: n_components equals to 6, n_iter was 10, rescale_with_mean used was True and random_state of 200. To plot the principal component tables,

their contributions to describing the dataset, and correlations, the parameters from the PCA class were called: .eigenvalues_summary, .column_contributions_, and .column_correlations. The sum of the product of each principal component's variance by its contribution to the variable was employed to rank the variables (HALFORD, 2023).

2.3 Model Training

Decision Tree (DT) regression was the chosen model to be applied for this work, employing the "DecisionTreeRegressor" function from sklearn. Cross validation technique was used with sklearn's GridSearchCV, to following topologies amongs the configuration of the used regressor: max_depth between 2, 4 and 8, min_samples_split of 2, 5 and 10, min_samples_leaf of 1, 4 and 8 and cv was default 5.
The whole dataset was splitted between training and test datasets by the proportion of 70 % of training data and 30 % being redirected for tests. The R^2 metric and Mean Squared Error (MSE) were adopted to compare the trained models, those metrics were also used to compare training and testing models to determine better fits.

3. Results and Discussion

3.1 Data evaluation

Initially, the SPE data was evaluated to gain a better understanding of its variables. The dataset used is a collection of time series related to reservoir testing analysis (RTA), encompassing 53 O&G wells, totaling 60,976 rows of information (SPE, 2024). Among RTA variables, "Lease" and "Pressure Source" represent, respectively, the identification of the evaluated well and the type of pressure assessed (casing or tubing). Similarly, Time corresponds to the day number in the time series. The distribution of null or zero data was analyzed by calculating their occurrence, as shown in Table 1.

Table 1: Proportion of null and zero values for each variable

Variables	Null Values Percentage (%)	Null and Zero Values (%)
Choke Size	96,2%	96,2%
Gas Volume (MMscf)	0,4%	5,1%
Oil Volume (stb)	13,8%	80,9%
Water Volume (stb)	0,0%	9,9%
Gas Lift Inj Volume (MMscf)	87,5%	91,3%
Casing Pressure (psi(a))	0,0%	1,4%
Tubing Pressure (psi(a))	0,0%	26,1%
Active Pressure (psi(a))	0,0%	1,0%
Line Pressure (psi(a))	0,0%	1,7%
Pressure Source	0,0%	0,0%
Calculated Sandface Pressure (psi(a))	1,4%	1,4%

It is noticeable that most values for "Choke Size" and "Gas Lift Inj Volume" are missing or zero. Therefore, the assumption of this study is that these columns are not important for predicting "Calculated Sandface Pressure" and can thus be disregarded. Leases that presented empty information were dropped.
"Tubing Pressure" also exhibits many zero vectors; however, it can be considered an operational data point since "Tubing Pressure" and "Casing Pressure" are indicated by the "Pressure Source". In other words, "Tubing Pressure" is zero when the "Pressure Source" is "Casing Pressure".
The "Oil Volume" column also raises questions about the dataset. Since crude oil is a three-phase compound of gas, oil, and water (Guan H.Y., Jiyuan T., 2009), more data in "Oil Volume" was expected, since "Gas Volume" and "Water Volume" didn't present the same absence in the dataset.
After proposed data cleaning and removal of lines with only non-numerical values, the dataset was left with 59,917 lines of information.

3.2 Variables correlation: Spearman correlation and PCA

The following spearman heatmap could be plotted from the data:

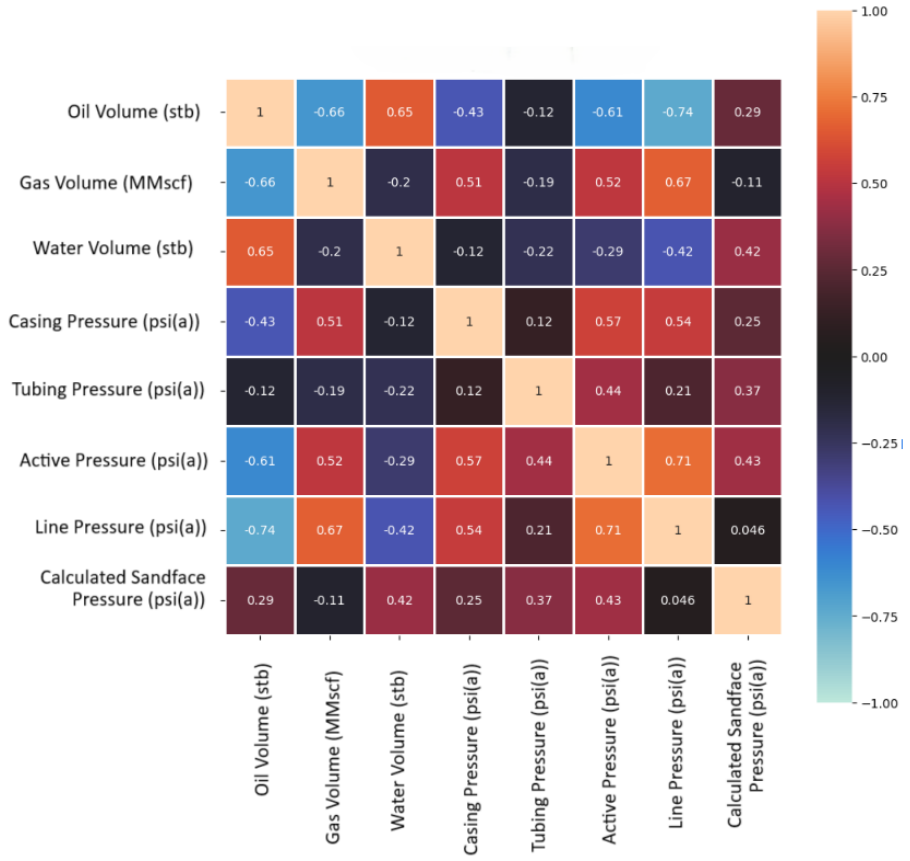

Figure 1: Spearman Correlation

Strong negative correlations were identified between “Line Pressure” and “Oil Volume” (-0.74) and strong positive correlations between “Line Pressure” and “Active Pressure” (0.71). Additionally, moderate correlations (greater than |0.6|) were observed between “Gas Volume” and “Oil Volume” (-0.66), “Water Volume” and “Oil Volume” (0.65), and “Line Pressure” and “Gas Volume” (0.67). None of the variables has shown significantly strong correlation (>0.8), therefore many of the variables are supposed to be important throughout data evaluation.

Through PCA analysis it’s possible to observe that at least 4 PC’s (principal components) are necessary to explain almost 90 % (89.15 %) of data variance (Table 2), i.e. it’s possible to create a model that explains almost the entire dataset within 4 PCs (Principal components).

Table 2: Eigenvalues, Variance, and Cumulative Variance for PCA

component	eigenvalue	% of variance	% of variance (cumulative)
1	3.138	44.83%	44.83%
2	1.532	21.88%	66.71%
3	1.103	15.75%	82.46%
4	0.468	6.69%	89.15%
5	0.353	5.04%	94.19%
6	0.304	4.35%	98.53%

It’s possible to rank out the variables of the dataset from columns contribution (Table 3) over each component’s % of variance (Table 2), concluding the rank from most important to less important: “Water Volume”>”Tubing Pressure”>”Gas Volume”>”Active Pressure”>”Casing Pressure”>”Oil Volume”> “Line Pressure”.

Table 3: PC's columns contribution

Component	1	2	3	4
Variable				
Oil Volume (stb)	11%	23%	6%	12%
Gas Volume (MMscf)	11%	4%	46%	1%
Water Volume (psi(a))	4%	35%	15%	25%
Casing Pressure (psi(a))	18%	5%	4%	49%
Tubing Pressure (psi(a))	9%	22%	29%	1%
Active Pressure(psi(a))	24%	10%	0%	2%
Line pressure(psi(a))	23%	1%	0%	9%

3.3 Models training and testing

Decision Tree regression models present a strong capability of identifying nonlinear correlation between variables. Since we observed small impact of Oil Volume variable among the dataset, a small contribution of this variable this variable's principal components and this is one of the variables with large amount of 0 vectores, one of the main propositions of this work is to ignore "Oil Volume", therefore the variables used for model training were: "Gas Volume", "Water Volume", "Casing Pressure", "Tubing Pressure", "Active Pressure" and "Line Pressure". The results indicated that the best predictive model for training and testing data used the combination of max_depth equals to 8, min_samples_leaf equals to 4 and min_samples_split equals to 10, from a combination of 135 trained models.By comparing testing results against Real calculated sandface pressure, it was possible to reach a 4,878E-04 Mean Squared Error (MSE) with R^2 of 0,982. The error distribution along real and predicted values can be observed below:

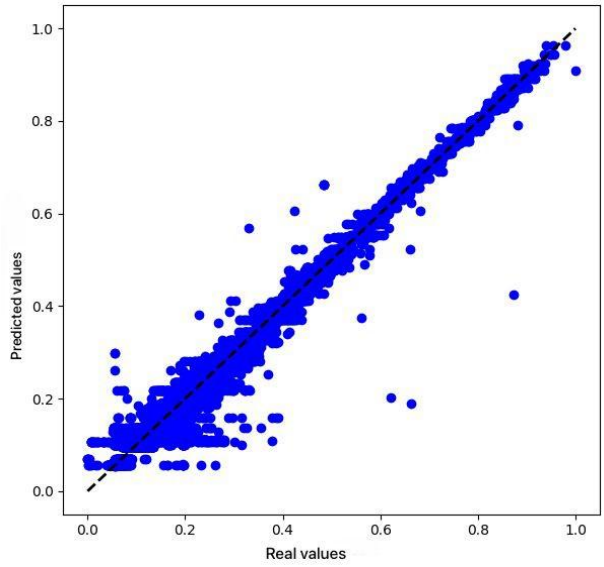

Figure 2: Error distribution. Real values vs Predicted values

4. Conclusion

This survey proposed investigating the SPE RTA's dataset and using it to forecast BHP with DT regression models. Data analysis and consecutive cleaning was performed under the dataset, with consecutive analysis of variables and model training and testing. After evaluating the dataset, it was observed that, despite the amount of data, the dataset lacks specific documentation that could prevent some assumptions when working with it. Using Spearman's correlation, the dataset showed stronger monotonic correlations between Line Pressure and Oil Volume (-0.74), and Line Pressure and Active Pressure (0.71), as well as moderate correlations for Gas Volume and Oil Volume (-0.66), Water Volume and Oil Volume (0.65), and Line Pressure and Gas Volume (0.67). Through PCA, it was possible to determine the variables rank as "Water Volume">"Tubing Pressure">"Gas Volume">"Active Pressure">"Casing Pressure">"Oil Volume"> "Line Pressure".The use of machine learning has proven to be an easier way to determine this important parameter. It was possible to

generate a machine learning model capable of predicting the calculated BHP from the SPE dataset with an R^2 of 0.982 and MSE of 4,878E-04.For future jobs, it would be interesting to evaluate different combinations of variables, by dropping variables with less significance on describing data variance, or including Oil volume, which was dropped for this work. Applying different combinations for the DT regressor, considering better timing performance for on-line practices, would also be of great value for a more advanced model.

Acknowledgments

The authors would like to express their gratitude to CNPq, CAPES, FAPERJ (Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro, E-26/200.282/2023-283570), and the Department of Chemical Engineering at PUC-Rio for their institutional support. Additionally, they acknowledge the Programa de Recursos Humanos da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (PRH-ANP) for the financial support that made this study possible.

References

Bishop, C. M., 2006, Pattern Recognition and Machine Learning, Springer, New York, NY, 778.

BROWN, N.; ROUBÍCKOVÁ, A.; NORDLOH, V. A. Machine Learning for Gas and Oil Exploration. arXiv preprint arXiv:2010.04186, 2020. <https://arxiv.org/abs/2010.04186> accessed: 15 dez. 2024.

Clarkson, C. R., Nobakht, M., Kaviani, D., Ertekin, T., 2016, Production Analysis of Tight-Gas and Shale-Gas Reservoirs Using the Rate-Transient Analysis Technique, SPE Reservoir Evaluation & Engineering, 19(3), 312-329, DOI: 10.2118/144317-PA.

Deloitte, 2023, Managing and aligning expectations of the energy transition: energy and chemicals signature issue FY24, Deloitte. <deloitte.com/content/dam/insights/articles/us176669_e-i_fy24-energy---chemicals-signature-issue_managing-and-aligning-expectations-of-the-energy-transition-soft-launch/DI_EI-FY24-Energy-Chemicals-Signature-Issue.pdf> accessed 15.12.2024.

Domingos, P., 2012, A few useful things to know about machine learning, Communications of the ACM, 55(10), 78-87, DOI: 10.1145/2347736.2347755.

Freire, R. M., 2024, SPA RTA data survey, GitHub. From URL: https://github.com/raul-macedo-freire/spe-rta-data-survey.

Géron, A., 2019, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.), O'Reilly Media, 848.

Goodfellow, I., Bengio, Y., Courville, A., 2016, Deep Learning, The MIT Press, 800.

Guan H.Y., Jiyuan T., 2016, Introduction, in: Guan Heng Yeoh, Jiyuan Tu (Eds.), Computational Techniques for Multiphase Flows (Second Edition), Butterworth-Heinemann, Pages 1-18, DOI: 10.1016/B978-0-08-102453-9.00001-5.

Halford, M. (2023). Prince: A Python library for multivariate exploratory data analysis. <https://github.com/MaxHalford/prince> accessed: 15 dez. 2024.

IHS Markit, 2020, IHS Harmony Enterprise, IHS Markit. <ihsenergy.ca/support/documentation_ca/Harmony_Enterprise/2020_2/content/html_files/start_here.htm> accessed 15.12.2024.

Kurita, T., 2020, Principal Component Analysis (PCA), Computer Vision, p. 1–4.

MCKINSEY & COMPANY, 2024, Global energy perspective 2024, McKinsey & Company. <mckinsey.com/~/media/mckinsey/industries/energy%20and%20materials/our%20insights/global%20energy%20perspective%202024/global-energy-perspective-2024.pdf> accessed 15.12.2024.

Murphy, K. P., 2012, Machine Learning: A Probabilistic Perspective, MIT Press, Cambridge, MA, USA.

Python. Python Org. <www.python.org/> accessed 15.12.2024.

Quinlan, J. R., 1986, Induction of Decision Trees, Machine Learning, 1, 81-106, DOI: 10.1007/BF00116251.

Smola, A. J., Schölkopf, B., 2004, A tutorial on support vector regression, Statistics and Computing, 14, 199–222. DOI: 10.1023/B:STCO.0000035301.49549.88.

Society of Petroleum Engineers, 2024, Industry data repository, Society of Petroleum Engineers. <spe.org/en/industry/data-repository/> accessed 15.12.2024.

SPE Data Repository, 2024, Data Set: 1, Well Number: all, From URL: <spe.org/datasets/dataset_1/csv_files/dataset_1_all_wells/> accessed 15.12.2024.

Taghavinejad, A., Ostadhassan, M., Daneshfar, R., 2022, Unconventional reservoirs: rate and pressure transient analysis techniques: a reservoir engineering approach, Springer, Cham, 111.

Vargas, R. E. V., Munaro, C. J., Ciarelli, P. M., Medeiros, A. G., Amaral, B. G., Barrionuevo, D. C., Araújo, J. C. D., Ribeiro, J. L., Magalhães, L. P., 2019, A realistic and public dataset with rare undesirable real events in oil wells, Journal of Petroleum Science and Engineering, v. 181, p. 106223.