Journal of
Engineering Sciences
Faculty of Engineering
Assiut University
JES
مجلة
العلوم الهندسية
كلية الهندسة
جامعة أسيوط
journal homepage: http://jesaun.journals.ekb.eg

A Note on Predicting Rate of Penetration Using Machine Learning Models

Received 24 June 2025; Revised 25 September 2025; Accepted 25 September 2025

Ahmad Atef Husseiny[1]
Attia M. Attia[2]
Ahmed G. Hagag [3]

Keywords
Rate of Penetration, Machine learning, Directional Drilling, Dataiku, ROP prediction optimization

Abstract: Rate of Penetration (ROP) is a critical parameter influencing drilling efficiency and accelerating field development. Although conventional ROP models (e.g., Bourgoyne, Hareland) provided reasonable results, they often struggle with limited accuracy and adaptability across different well conditions. Recent advances in hybrid machine learning (ML)-physics or deep ML models improve ROP prediction; however, these methods typically require complex programming, limiting their practical adoption. This study addresses these gaps by introducing a prompt, simple, and strong ROP prediction for directional wells, eliminating the need for hybrid modelling through the platform Dataiku Data Science Studio (DSS). To evaluate the impact of domain-specific parameters, two calculated metrics, D-exponent and Mechanical Specific Energy (MSE) were integrated into the dataset. Three ML algorithms (Gradient Boosted Trees, XGBoost, and Support Vector Machines (SVM)) were trained and tested using R^2, Mean Absolute Error, and Root Mean Square Error (RMSE) across three directional offshore wells from the same field. XGBoost showed best performance and significant improvement R^2 scores for all wells after incorporating MSE and D-exponent: from 0.373 to 0.974 (Well-1), 0.216 to 0.945 (Well-2), and 0.862 to 0.983 (Well-3). Features importance and SHAP values analyses further quantified the contributions of MSE and D-exponent to all models' accuracy, demonstrating their role in enhancing predictions. This work provides a practical, programming-free solution for ROP optimization in directional drilling, achieving high performance without using advanced ML technologies.

1. Introduction

Drilling operations have critical importance within oil and gas industry as they are directly related to capital costs, production and field allocation. The efficiency of drilling operations requires minimization of drilling expenditures. Many efforts and approaches have been conducted to mitigate drilling issues to enhance operational performance by either minimum non-productive time (NPT) and/or minimum required drilling time. It is known that increasing the speed of drilling increases (ROP), the costs will be more efficient. This

[1] Ahmad.Atef@suezuniv.edu.eg - Dept. of Petroleum and Mining Engineering, Suez University, Egypt
[2] Attia.Attia@bue.edu.eg - Professor, Dept. of Petroleum and Gas Technology Engineering, the British University in Egypt.
[3] Ahmed.Hagag@suezuniv.edu.eg - Associate professor, of Petroleum and Mining Engineering, Suez University, Egypt

https://doi.org/10.21608/jesaun.2025.397652.1574

parameter (ROP) is one of the most focal parameters that influence both the expense and execution of drilling. In addition, it guides drilling engineers to select optimal variables for achieving the lowest cost per foot. To attain minimal costs and maximum efficiency, various parameters must be analysed to comprehend their impacts on ROP. Elkatatny et al. (2017) defined that Rate of Penetration (ROP) is the volume of rock fractured per foot per hour. Alternatively, it has been characterized as the velocity at which rock is drilled beneath the bit. ROP has been defined by Bourgoyne (1986) [5] as a metric quantifying the advancement of the bit through rock formations. Although the increase of ROP seems better for accelerating drilling process and reducing its time, the probability of operational issues such as stuck pipe and poor whole cleaning will be increased to be in critical condition. Therefore; this parameter must be optimized and monitored to avoid such these issues and to maintain cost efficiency (Akgun, (2002) [2]).

ROP is affected by several parameters while drilling such as the borehole dimensions, bit design, geological properties of the rock formation being drilled (including rock strength and drillability), and operational parameters such as WOB, rotational speed, torque, and hydraulic conditions. In addition to the drilling fluid parameters and BHA, factors. The prediction of ROP must be takin in considerations to enable precise calculation of drilling costs and timelines, thereby facilitating the design of drilling parameters, optimization of operational variables, and even support in refining wellbore trajectories and well structures. They will assist to guide field engineers to achieve strategic allocation of field production (Abdulmalek et al., (2018) [1]; Jahanbakhshi et al., (2012) [13]). In past decades, most approaches to predict ROP are basically relying on the interpretations of historical drilling data. However, most of these approaches are not effective for prediction as some of them are based on mathematical assumptions or based on specific field condition which is not applicable for another field. Consequently, the development of a reliable predictive model for drilling rates, which integrates empirical correlations with available data, has been identified as a pressing challenge within drilling engineering. The **Table-1**, represents common ROP correlations used for past decades.

Table 1: Previous researches on ROP

References	Input parameters	Output Results	Remarks
Speer, J.W. (1959) [25]	Rotation Rate, Bit Type, Properties of Circulating Mud, Weight on Bit (WOB), Hydraulic Horsepower	ROP optimization curves	Without statistical metrics
Cunningham R.A. (1960) [7]	Rotary Speed and WOB	Analytical ROP equations	Without validation metrics
Bingham, M.G. (1964) [4]	Rotary Speed and WOB, Drill Bit OD, formation type	Drillability index	Correlation with ROP
Bourgoyne A.T. et al. (1974) [6]	Rotary Speed, WOB, Drill Bit OD, Depth, Bit tooth wear, Compaction, Deferential Pressure, Pore pressure, Bit hydraulics and jet compact factor	ROP model with $R^2 = 0.80–0.95$	Calibrated to field data
Hareland, G. et al. (1994) [10].	Rotary Speed, WOB, Drill Bit OD, Rock compressive strength.	ROP prediction with $R^2 > 0.85$	

References	Input parameters	Output Results	Remarks
Maurer W.C. (1962) [19]	Weight on Bit (WOB), Rotary Speed, Rock Strength, Bit Geometry	Theoretical ROP	Without empirical validation metrics
Teale R. (1965) [30]	Mechanical Specific Energy (MSE), Weight on Bit (WOB), Drillability of the Formation	Rock strength	Correlation: MSE ∝ compressive strength; R implied
Warren T.M. (1987) [31]	Weight on Bit (WOB), Rotary Speed, Hydraulic impact factor, rock strength, differential pressure	ROP with $R^2 \approx 0.89$	ROP for roller cones
Detournay E. et al. (1992) [32]	Rock Cutting Mechanics, Bit Geometry, Intrinsic Specific Energy	ROP/Torque models	Analytical validation; no statistical metrics
Pessier R.C. et al. (1992) [33]	WOB, RPM, Stick-slip, lateral vibrations, Drill Bit size and cutter type, Formation Abrasiveness, Cuttings removal efficiency	ROP model with $R^2 = 0.92$	MSE-based dysfunction detection

As mentioned, the above that these listed methods are limited in obtaining precise ROP prediction. This limitation arises from the complex, indirect, and inherent relationships of drilling parameters, such as WOB, RPM, geomechanical rock properties, and drilling hydraulic efficiency, all of which control the behavior of ROP. These conventional equations are generally investigated and derived under laboratory conditions and/or mathematical conversions. These derivations do not reflect to onsite drilling environments.

With the evolution of technology, Artificial Intelligence (AI), Machine Learning (ML) and statistical analysis, drilling experts tried to advantage from AI into traditional ROP prediction models. This fusion has emphasized innovative approaches and insights in the field (Moran and Ibrahim 2010 [20]). Current approaches for predicting ROP are divided into two categories: those grounded in theoretical or empirical frameworks, and those leveraging statistical or machine learning (ML) techniques. Among ML algorithms applied to ROP forecasting, logistic regression, support vector machines (SVMs), Neural Networks, and Random Forest (RF) are prominent (Noshi et al. 2019 [15]). While Neural Networks and Random Forest can achieve prediction accuracies as high as 80%, their "black box" nature limits interpretability, leaving modelers unable to understand the internal mechanisms and reliant on trial-and-error adjustments. Shi X. et al. (2016) [17] stated that logistic regression which pairs with historical drilling data and real-time operational metrics can uncover underlying relationships with ROP. This approach highlights the use of these techniques in constructing predictive models for drilling efficiency. El-Sayed et al. (2023) [9] used unsupervised MLs coded by Python, K-Nearest Neighbours (KNN) and Multilayer Perceptron (MLP) in order to predict ROP in vertical offshore wells. The authors highlighted that greater accuracy of ROP can be achieved by removing outliers. Benzminabadi et al. (2017) [3] successfully predicted ROP using ANN and Multiple Nonlinear Regression (MNR) coded in Python. By combining operational drilling parameters and mechanical rock properties,

they achieved significant improvements in their results, highlighting the importance of nonlinear relationships in ROP prediction. Omogbolahan S. et al. (2019) [28] employed four models SVM, Least-Squared SVR, ANN, and Extreme Learning Machine (ELM) for ROP prediction. Their study observed that reducing the number of features yielded nearly the same performance for the models, which means that excessive features can impede or reduce model performance such as standpipe pressure or mud density which generally contributed less than 1% to model accuracy based on results. These features could to lead to data noise and overburden ML models without improving ROP prediction. This highlights the importance of effective feature selection and the ability of ML techniques to handle limited datasets. Mohamadian et al. (2021) [26], made statistical analysis of drilling parameters used for wellbore instability prediction such as stuck pipe and hole cleaning using AI. They found that most input parameters for wellbore problems are WOB which has been used in 70.97% of the reviewed papers), the flow rate (FR: 54.84%), the ROP (51.61%), the revolutions per minute (RPM: 38.71%) and the measured depth (MD: 35.48%) as these parameters are easy to obtain and not costly comparing to other parameters obtained by downhole tools which are expensive. Olukoga et al. (2021) [21] analysed 94 studies to evaluate prevalent Machine Learning algorithms in drilling applications. They reported that Artificial Neural Networks (18%), Support Vector Machines (17%), Regression (13%), Deep Learning (10%), Decision Trees (8%), and Random Forests (8%) dominated the field, collectively accounting for nearly 75% of the methodologies studied, **see the figure-1**. The reason dominance of using ANN and SVM due to their effectiveness in handling with complex features data as SVM is effective for smaller datasets, whereas ANN is effective for large datasets as highlighted by Hegde et al. (2020) [12].

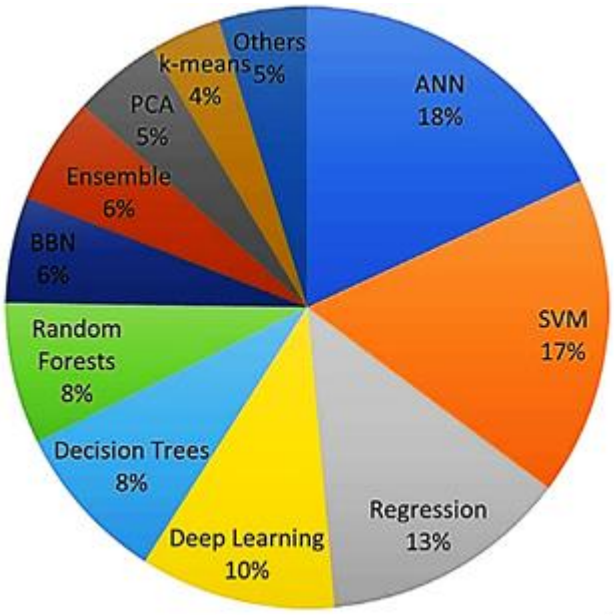

Figure 1: Most ML techniques used for well drilling (Olukoga et al. 2021 [21])

In short, it is clear that combination of AI or ML for petroleum industry will promote and enhance prediction of critical drilling parameters. This combination must conduct a clear comparative analysis of different ML techniques within real-time data. This comparison will

help to identify proper features selection and optimizing hyperparameters tuning to achieve best prediction possible.

2. Literature Review

2.1 D-exponent and Mechanical Specific Energy

One of the most important parameters that affects the performance of ROP in directional drilling is the rock brittleness. The quality of formation strength is intuitively expected to affect the well productivity. One of indirect parameter that represents the rock toughness is D-exponent. This exponent has a trend of drilling curves in overbalanced zones which helps to detect regimes of formation pressure from normal to abnormal pressure (Jorden and Shirley 1966 [29]). In addition, it represents the drillability of the formation and reflects the performance of ROP. This parameter as shown in **Equation-2** directly relates to the penetration rate and the bit size and inversely relates to the weight on the bit and the rotational speed.

$$\mathbf{D-exponent} = \frac{\mathrm{Log}\left(\frac{0.308*\mathrm{ROP}}{60*\mathrm{RPM}}\right)}{\mathrm{Log}\left(\frac{12*\mathrm{WOB}}{\mathrm{Bit\ size}}\right)} \quad \text{.....(1)}$$

- ROP in unit of m/hr.
- WOB in unit of lb.
- Bit size is drill bit diameter in unit of inches.

2.2 Mechanical Specific Energy (MSE)

It is used to evaluate drilling performance which means the energy required to cut and smash a specific volume of rock using a drill bit (Teale et al. 1965 [30]). It has quantitative assessment of how efficiently mechanical energy is converted into rock destruction. It is influenced by various drilling factors, including torque, rotary speed, weight on bit (WOB), and ROP, which are commonly used during present drilling operations. The efficiency of drilling process can be optimized through monitoring values of mechanical energy being put into the system during drilling and comparing that energy with in-situ rock strength (Majidi R. et al. 2017 [18]). The use of MSE for estimating pore pressure depends on the conditions of influenced subsurface rock's stresses, to some extent, by the fluid pressure in exerted in these pores. As a result, the pore pressure is needed for rock fracture energy while drilling (Majidi R. et al. 2017 [18]). In other words, MSE values shall be optimized through monitoring drilling parameters to achieve effective and high ROP performance and minimizing energy waste. The values of MSE have two indications, high MSE values indicate inefficient drilling (e.g., bit wear, excessive friction, or improper drilling parameters) whereas low MSE values indicate optimal drilling conditions, where energy is effectively used for rock breakage. This parameter can be calculated using the **Equation-2**.

$$MSE = \left(\frac{WOB}{A_b} + \frac{120.\pi.RPM.T}{A_b * 0.308 * ROP}\right) \quad \text{.....(2)}$$

- WOB is the weight on bit in unit of lb.
- Ab is drill bit diameter in unit of inches.
- ROP in unit of m/hr.
- RPM is in the revolutions per min.
- T is the torque in unit of ft.lb.

MSE can improve understanding the behaviour of ROP through pore pressure prediction. In other words, when MSE trendlines indicate abnormal pore pressures the ROP will tend to decrease. In underbalanced conditions or indicating a potential kick, the ROP will tend to increase. This will help to predict ROP and potentially gives an improved method of well control. Although we reviewed several studies related to ROP prediction using different parameters (Omogbolahan S. et al. (2019) [28], Hazbeh O. et al. (2021) [11], Shaygan K. et al. (2023) [16], Ehsan B. et al. (2021) [8], Li C. et al. (2020) [23], Noshi and Schubert (2019) [15], Li and Samuel (2019) [24], and Abdulmalek et al. (2018) [1]) none of them used D-exponent for during their studies except El-Sayed et al. (2023) [9] who used this parameter and they did not analysed the impact or the correlation of their features to ROP to optimize prediction performance. Therefore; it is decided to calculate D-exponent and MSE in our data for three directional wells in order to check their importance on predicting ROP.

2.3 Computational intelligence techniques

Machine Learning analysis can serve as a critical phase in knowledge discovery within databases, including the extraction of non-relationships between features in datasets. Drilling operations, for instance, generate vast and irregularly distributed data characterized by inherent relationships. This complexity necessitates advanced methods capable of interpreting such challenges. Machine Learning models can identify these hidden relationships and guide solutions for tackling real-world problems that defy conventional approaches (Siddique & Adeli, 2013 [27]). This section outlines previous studies of ML techniques for predicting ROP.

2.4 Machine Learning for ROP

Shaygan K. et al. (2023) [16] applied Random Forest (RF) and Multilayer Perceptron Neural Networks (MLPNN) to forecast the ROP in directional wells. They concluded that indirect parameters including Weight on Bit (WOB) and cutting transport efficiency are significantly influenced. The absence of these features in input datasets was noted to degrade model performance. Hazbeh O. et al. (2021) [11] used hybrid algorithms combining Multilayer Perceptron (MLP) with optimization techniques such as Artificial Bee Colony (MLP-ABC), Gravitational Search Algorithm (MLP-GSA), and Firefly Algorithm (MLP-FF). The MLP-ABC hybrid outperformed others and indicated that integrating MLP with optimizers enhances prediction the prediction. Ehsan B. et al. (2021) [8] optimized three neural network models, Multilayer Perceptron (MPNN), Cascade-Forward (CFNN), and Radial Basis Function (RBFNN) using backpropagation and biogeography-based algorithms. Their findings showed

that computational intelligence drastically improved ROP prediction compared to traditional neural networks. Li C. et al. (2020) [23] merged Artificial Neural Networks (ANN) with an Integrated Genetic Algorithm (IGA) to predict ROP and WOB in China's complex shale gas formations. After testing ANN and IGA separately, their hybrid approach achieved superior accuracy and real-time optimization during drilling. Kloucha C.K. et al. (2022) [14] used Dataiku Data Science Studio software which has machine learning techniques to recommend optimal drilling equipment for future scenarios. Their analysis showed that using ML techniques from this software outperformed conventional statistical approaches, streamlining equipment selection with greater precision.

These studies demonstrated the effectiveness of data-driven approaches for ROP modelling. However, they still challenge in programming advanced ML models which are relatively hard to be achieved from scratch (Naser M.Z. 2023) [38]. Our work will use Dataiku DSS platform as it is free and produces the most consistent performance models (Naser M.Z. 2023) [38]. The expected highest models performance in this study will be ensemble models and SVM due to their effective in both bagging and possessing boosting techniques which reduce training time, better generalization, and minimizing the multiclass error rate (Ganaie M.A. Et al. (2022) [39], Huang F. (2018) [40]), Tabik S. et al. (2020) [41]).

3. Methodology

In order to build ML models using Dataiku DSS, this flowchart shows steps of predicting ROP.

4. Data Collection

In this research, an actual data from an offshore three directional gas wells in sand reservoirs in the eastern portion of the West Delta Deep Marine concession, which lies offshore in the deep water of Nile Delta, Egypt. The exploratory well was targeting a thick Pliocene channel levee complex trends in a NNE-SSW orientation in which no wells have been penetrated the central Channel (Fahmy R. et al. 2025) [37]. Stratigraphically the central Channel can be correlated to the main channel, in which there are several development wells in the same field. Three deviated wells were being drilled with J-shape design with different depths based on reservoir intervals. Well-1 penetrating the previously undrilled central channel, while Wells-2 and -3 developed flanking intervals. All wells were drilled with similar BHAs and mud systems but varied in measured depth (MD) due to reservoir architecture. They also share common drilling parameters in addition to calculated D-exponent and Mechanical Specific Energy, see **table-2**. These data were available from the field, open-hole wireline logs data were unavailable for these wells. However, unavailable these data will not invalidate the models on ROP prediction for this study. Core predictors (WOB, RPM, MSE, D-exponent) were prioritized based on established physical relationships with ROP as per recommendations from Teale et al. (1965)

[30] and Jorden et al. (1966) [29]. The non-shared parameters (Pore pressure est. and pit volume) were being collected from available sensors data permitted cross-well consistency.

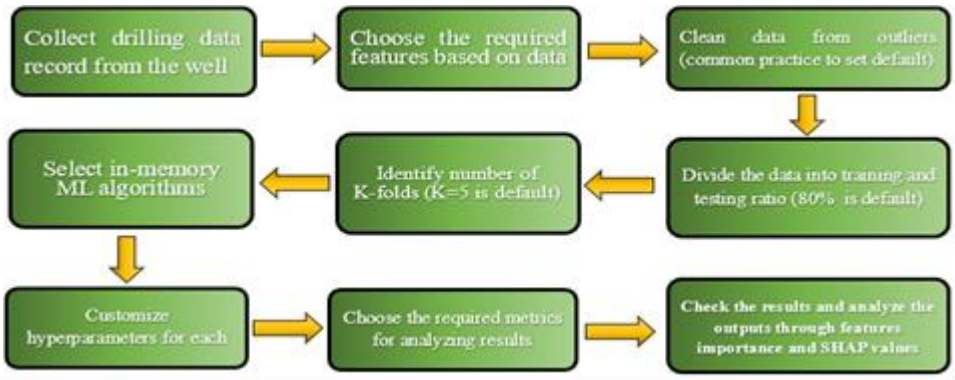

Figure 2: Data workflow procedures for predicting ROP.

Table 2: wells and their features used.

Well number	Target	Common shared Parameters	Non-shared parameters	Total number of parameters
Well-1	**ROP**	Measured depth (m), Inclination (degree), Surface WOB (klb), RPM, Surface Torque (lb.ft), Stand Pipe Pressure (psi), Mud Wieght (ppg), Flow in Pum (gpm), Hours On Bit (min), Downhole Revs On Bit (Krev), Bit Diameter (inch), D-Exponent, Mechanical Specific Energy (psi)	Pit volume (bbl), Pore Press Est. (ppg)	15
Well-2	**ROP**	Measured depth (m), Inclination (degree), Surface WOB (klb), RPM, Surface Torque (lb.ft), Stand Pipe Pressure (psi), Mud Wieght (ppg), Flow in Pum (gpm), Hours On Bit (min), Downhole Revs On Bit (Krev), Bit Diameter (inch), D-Exponent, Mechanical Specific Energy (psi)	Pore pressure Est. (ppg)	14
Well-3	**ROP**	Measured depth (m), Inclination (degree), Surface WOB (klb), RPM, Surface Torque (lb.ft), Stand Pipe Pressure (psi), Mud Wieght (ppg), Flow in Pum (gpm), Hours On Bit (min), Downhole Revs On Bit (Krev), Bit Diameter (inch), D-Exponent, Mechanical Specific Energy (psi)	Pit volume (bbl)	14

5. Data filtering and smoothing

5.1 Model Building

Dataiku DSS offers fourteen in-memory Machine Learning (ML) models which don't require python coding. Other advanced or customized ML techniques such as hybrid technology or Genetic Algorithms require special Python or Scala coding for complex data tasks to get better

results (Dataiku DSS 2025 [34]). To assess performance differences in this study, we calculated ROP using all in-memory ML models, then we filtered top three ML models based on R^2 score. These ML models are Extreme Gradient Boosted Machine (XGBoost), Gradient Boost Trees, and Support Vector Machine (SVM). The results will be shown as comparisons of predicted ROP versus actual ROP for both training and testing datasets, feature importance rankings, and their numerical effects on accuracy of ROP (SHAP values). Notably, the platform does not display regression plots for training dataset (Dataiku DSS 2025 [35]). The platform supports training and testing dataset through learning curves only. In order to validate normal data distribution and interpretability, we calculated Cumulative Distribution Functions (CDFs) through the software built in code, as illustrated in **table-6**. The **table-3, 4, and 5** show the list of calculated statistical parameters of the case study. These tables demonstrated that parameter characteristics of ROP, WOB, and MSE refer to tight clustering and confirm operational consistency. The Well-1 has ROP variability with 15% greater mean than in Wells-2/3, which attributes to central channel lithology. The shared 13 inputs will enable us to cross-well model comparisons. The skewness and kurtosis values in these tables refer to data distribution normality and quantitative indication for any deviation from symmetry. For the Well-1 ROP (skewness=4.58) means strong right-skewness which refer to intermittent high-ROP drilling phases. For the Well-2 RPM (skewness = -2.9) means left- skewness bias from frequent low-RPM operations. Kurtosis measures tailedness in normal distribution, with values greater than 3 (Leptokurtic) refer to outlier-prone distributions. The Kurtosis values with less than 3 (platykurtic) refer to light tails. For Wells-1/2 ROP (kurtosis = 67.04 and 88.7 respectively) are considered highly leptokurtic, which confirms outlier-prone events, whereas Well-3 ROP (kurtosis = -0.1) is Mesokurtic which indicates both limited outliers' existence and smooth drilling operations. These distributions will justify models' performance in this study.

Table 3: Statistical characterization summary of the Well-1

Statistical Parameter	Depth (m)	Incl (deg)	ROP Avg (m/hr)	Surface WOB Avg (klb)	RPM Surface Avg (rpm)	Torque Abs Avg (f-lb)	SPP Avg (psig)	Flow in Pum Avg (gpm)	Hours On Bit (min)	Revs On Bit (Krev)	Pit volume (bbl)	D-Exponent	Pore Press Est (ppg)	Mud Den (mg)	Bit Diam.	MSE (PSI)
Maximum	2251.5	٢٧,٣٧	١٥٨,٩٩	٣٢,٢٠	١٥٢,٠٠	٩٨٣٧,٩٠	٢٨٨٣,٠	١٠٩١,٤٠	٢٨٩٨,٠٠	٤٤٧,٧٠	٧٩٧,٠٧	١,٩٣	٩,٦٣	١٠,٧٥	١٧,٥٠	٨٤٥٤,١٧
Minimum	1330	٠,٤٤	١,٠٩	٠,٥٠	٣١,٠٠	١٣٥٥,٢٠	٨٣٥,٠٠	٢٣٣,٤٠	٠,٠٠	٠,٠٠	١٦٩,٣٠	٠,٤٨	٨,٦٨	٨,٦٠	١٢,٢٥	١٨,٦٣

Table 4: Statistical characterization summary of the Well-2

Statistical Parameter	Range	Mean	Median	Standard Deviation	Kurtosis	Skewness	Statistical Parameter	Maximum	Minimum
Depth (m)							Depth (m)	2255	1669
Incl (deg)	٢٦,٩٣	١٩,٣٢	٢٥,٥١	٩,٨٠	-0.90	-0.85	ROP Avg(m/hr)	١٥٨,٩٩	١,٦٧٨
ROP Avg (m/hr)	١٥٧,٩١	١٥,٣٥	١٥,١٦	٨,٠٤	٦٧,٠٤	٤,٥٨	Surface WOB Avg(klb)	٣١,١	٠,٥
Surface WOB Avg	٣١,٧٠	١١,٥٢	١٠,٥٠	٥,٥٩	٠,٣٥	٠,٧٧	RPM	١٥٢	٤٢
RPM Surface Avg (rpm)	١٢١,٠٠	١١٨,٥٠	١٢٦,٠٠	٢٢,٥٣	٣,٩٧	-2.15	Torque Abs Avg (f-b)	٩٨٣٧,٩	٢٨٥٣,٨
Torque Abs Avg (f-b)	٨٤٨٢,٧٠	٥٣٠٩,٣٤	٥٠٣٦,٤٥	١٦٥٢,٤٠	-١,١٦	٠,٢٧	SPP Avg (psig)	٢٨٥٤	١٤٧٥
SPP Avg (psig)	٢٠٤٨,٠	٢٢٣١,٥٦	٢٣٧٠,٠	٤٨٦,١٥	-٠,٤٥	-0.70	M.wt (ppg)	١٠,٧٥	١٠,٤
Flow in Pum Avg (gpm)	٨٥٨,٠٠	٨١٤,٠٩	٨١١,٥٠	٧٩,٦١	٩,٢٦	-0.64	Flow in Pum Avg(gpm)	٨٥٥,٦	٢٣٣,٤
Hours On Bit (min)	0٢٨٩٨,٠	١٠٢٥,٥٩	٨٤٣,٥٠	٧٧٤,٣٢	-0.71	٠,٦٩	Hours On Bit (min)	٢٩٢٠	٠
Revs On Bit (Krev)	٤٤٧,٧٠	١٢٠,٩١	٩٧,٣٥	١٠٥,٠٠	٠,٩٥	١,١٥	Revs On Bit Bit (Krev)	٣٤٥,٣	٠
Pit volume (bbl)	٦٢٧,٧٧	٢٥٩,٤٦	٢٣٥,٤٩	٩٥,٤٣	١٠,٤٠	٣,١٦	inclination (deg)	٢٧,٣٧	٢٣,٠3
D-Exponent	١,٤٥	١,٠٧	١,٠٦	٠,١٨	١,٦٥	٠,٥٣	Pore pressure (ppg)	٩,٦٣	٨,٩٩
Pore Press Est (ppg)	٠,٩٥	٩,٢٠	٩,١٩	٠,١٦	١,٦٠	-0.25	Bit Diameter	١٧,٥	١٢,٢٥
Mud Den (ppg)	٢,١٥	١٠,٢٩	١٠,٧٠	٠,٦٣	١,١٨	-1.56	D-exponent	٢,٢5	٠,٧3
Bit Diam.	٥,٢٥	١٥,١٨	١٧,٥٠	٢,٦١	-1.95	-0.23	MSE (PSI)	٨٤٥٤,١7	٣٨,١٦
MSE (PSI)	٨٤٣٥,٥٤	٧٥٢,٢٣	٤٩٣,٦٥	٧٨٦,٩٧	١٨,٤٣	٣,٣٤			

Statistical Parameter	Range	Mean	Median	Standard Deviation	Kurtosis	Skewness
Depth (m)	١٥٧,٣١٢	١٥,٦	١٥,٢٠٥	٨,٠	٨٨,٧	٥,٥
ROP Avg(m/hr)	٣٠,٦	١٣,٢	١٢,٣	٦,٢	-0.8	٠,٢
Surface WOB Avg(klb)	١١٠	١٢٠,٢	١٢٧	١٨,٦	٧,٨	-2.9
RPM Surface Avg (rpm)	٦٩٨٤,١	٦٦٦٩,٤	٦٩٣٢,٩	١١٧٣,٢	-0.8	-0.4
Torque Abs Avg (f-b)	١٣٧٩	٢٥٨٥,٨	٢٦٦٧	٢٢٢,٠	٥,٥	-2.1
SPP Avg (psig)	٠,٣٥	١٠,٧	١٠,٧	٠,٠	٢٥,٠	-4.5
M.wt (ppg)	٦٢٢,٢	٧٩٩,١	٨٠٨,٣	٤٤,٧	٥٣,٢	-6.6
Flow in Pum Avg(gpm)	٢٩٢٠	١٤٦٣,٣	١٥٥٦	٨٢٥,٤	-1.2	-0.1
Hours On Bit (min)	٣٤٥,٣	١٦٣,٧	١٧٤,٢	١٠١,٨	-1.2	٠,٠
Revs On Bit (Krev)	٤,٣٤	٢٦,٧	٢٧,٠٨	٠,٩	١,٩	-1.4
inclination (deg)	٠,٦٤	٩,٣	٩,٢٣	٠,٢	-0.6	٠,٥
Pore pressure (ppg)	٥,٢٥	١٦,٣	١٧,٥	٢,٢	-0.4	-1.2
Bit Diameter	١,٥٢	١,٣	١,٣٤	٠,٢	٠,٧	٠,٢
D-exponent	٨٤١٦	٧٥٩,٤	٤٩٤,٢٥	٧١٥,٧	١٦,٧	٣,١
MSE (PSI)						

Table 5: Statistical characterization summary of the Well-3

Statistical Parameter	Maximum	Minimum
Depth (m)	2092	1211.5
ROP Avg (m/hr)	٢١,٩	٥,٧
Surface WOB Avg(klb)	٢٢,٨	٠,٦
RPM Surface Avg (rpm)	١٢١	٥١
Torque Abs Avg (f-b)	٤٥٧٠,٣	١٧٤٩,٧
SPP Avg (psig)	٣٤٠٠	٢٣١٢
Flow in Pum Avg(gpm)	١٠٣٩	٨٨٨
M.wt (ppg)	١٠,٧٢	١٠,٢
HOB (hr)	٤٣,٨٥	٠,٤٨
Rev On Bit (krev)	٣٠٤,٠٠	١,٦٠
Bit Diameter	١٧,٥٠	١٢,٢٥
Pit volume (bbl)	٧٨٥,٢٦	٦٠٣,٦٨
D-exponent	١,٨٦	٠,٨٠
Inclination (deg)	٥٥,٠٥	٤٣,٨٩
MSE (PSI)	١٠٤٧,٣٦	٨٧,٧٠

Statistical Parameter	Skewness	Kurtosis	Standard Deviation	Median	Mean	Range
Depth (m)	٠,٧	-0.1	٣,٤	١١,٧	١٢,٤	١٦,٢
ROP Avg (m/hr)	٠,٨	٠,٢	٤,٣	٧,٥	٨,٢	٢٢,٢
Surface WOB Avg(klb)	-4.9	٢٣,٥	١١,٢	١١٩	١١٦,٧	٧٠
RPM Surface Avg (rpm)	٠,٥	٠,٤	٤٢٢,٠	٢٦٢٦,٥٥	٢٦٤٩,٤	٢٨٢٠,٦
Torque Abs Avg (f-b)	-0.4	-0.8	٢٧٦,٧	٢٩٥٠	٢٩١٣,٨	١٠٨٨
SPP Avg (psig)	-0.1	-1.9	٥١,٧	١٠٠٢	٩٦٦,٨	١٥١
Flow in Pum Avg(gpm)	-0.8	-0.5	٠,١	١٠,٥	١٠,٥	٠,٥٢
M.wt (ppg)	٠,٢٨	-0.94	١١,٤١	١٨,٨٨	١٩,٧٧	٤٣,٣٧
HOB (hr)	٠,٣٠	-0.94	٨١,٧٦	١٢٣,٦٥	١٣٠,٤٠	٣٠٢,٤٠
Rev On Bit (krev)	-0.12	-1.99	٢,٦٢	١٧,٥٠	١٥,٠٤	٥,٢٥
Bit Diameter	٠,٢٠	-0.67	٣٨,٣٤	٦٩٤,٨٠	٦٩٧,٩٦	١٨١,٥٨
Pit volume (bbl)	٠,١٤	٠,١٣	٠,١٦	١,٢٥	١,٢٥	١,٠٦
D-exponent	-2.21	٤,٣٦	٢,٥٥	٥٣,١١	٥٢,٧٩	١١,١٦
Inclination (deg)	٢,٤١	١٢,٧٨	١٤٢,٧٩	٢٩٧,٤٧	٣٢٨,٦٥	١٤٥٩,٦٧
MSE (PSI)						

To optimize ROP prediction, each ML technique has its own hyperparameters tuning or control parameters. These parameters were designed to reduce overfitting and avoiding excessive underfitting. This will help for a particular model to perform best results possible. To optimize tuning values, the platform offers “hyperparameters optimization” which makes several trials for values regularization parameter. Based on the tested results, the software recommended to keep control parameters for XGboost and Gradient Boost Trees with default values. These default values will help the model to adapt automatically with input data’s variance to reduce overfitting, bias, unnecessary computation, preventing excessive computation time, and avoiding severe imbalance regression problems. On the other hand, control parameters of SVM have been customized as seen in **table-7**. The **table- 8,** and **9** are control parameters for XGBoost and Gradient Boost Trees respectively.

Table 6: CDFs data input variables for the three wells

Statistic	Well-1			Well-2			Well-3		
	XGBoost	Gradient Boosted Trees	SVM	XGBoost	Gradient Boosted Trees	SVM	XGBoost	Gradient Boosted Trees	SVM
Min. (raw)	-7.3413	-4.9380	-2.1960	-6.0974	-10.0190	-11.2970	-1.6548	-1.8241	-1.6756
Min.	- 2.1853	-3.0287	-	-2.6350	-2.4373	-1.2206	-1.1010	-1.1609	-
25th perc.	-0.5201	-0.5378	-0.2687	-0.4872	-0.4521	-0.1901	-0.2446	-0.2833	-0.0735
Median	0.0530	-0.0165	0.0046	0.0436	0.0839	-0.0300	0.0026	-0.0431	0.0165
75th perc.	0.4738	0.5128	0.3195	0.6947	0.5952	0.2721	0.2184	0.2224	0.0497
90th perc.	0.9325	1.1785	0.7836	1.3664	1.1010	1.2036	0.4291	0.5523	0.1467
Max.	2.7666	4.1201	3.0256	2.8054	3.3827	5.3283	0.8106	1.2578	0.4287
Max. (raw)	6.1368	19.3410	34.271	12.5510	10.7280	18.4570	2.3131	3.1107	1.0498
Average	0.0037	-0.0027	0.1305	0.0965	0.1142	0.2363	-0.0230	-0.0378	-0.0041
Standard Deviation	0.9384	1.1955	0.6960	1.0929	1.0625	1.0975	0.3901	0.4817	0.1307

Table 7: Control parameters for SVM model applied to the three wells

	Well-1	Well-2	Well-3
Kernel	rbf	rbf	rbf
Kernel coefficient (gamma)	scale	scale	scale
Regularization parameter C	15	15	15
Stopping tolerance	0.001	0.001	0.001
Max iterations	-1	-1	-1
Rows (before preprocessing)	1293	953	1300
Rows (after preprocessing)	1293	953	1300
Columns (before preprocessing)	15	15	15
Columns (after preprocessing)	14	14	14
Matrix type	dense	dense	dense
Policy	Split the dataset	Split the dataset	Split the dataset
Sampling method	First records	First records	First records
Partitions	All partitions	All partitions	All partitions
Record limit	100000	100000	100000
Split mode	Randomly	Randomly	Randomly
Train ratio	0.8	0.8	0.8
Number of Training	1293	953	1300
Number of Testing	315	220	320
Random seed	1337	1337	1337

Table 8: Control parameters for XGBoost model applied to the three wells.

Objective	**Reg_linear**
Time Method	Automatic-CPU only
Max number of trees	300
Early stopping rounds	4
Max depth of tree	3
Eta (learning rate)	0.2
Max delta step	0
Alpha (L1 regularization)	0
Lambda (L2 regularization)	1
Gamma (Min loss reduction to split a leaf	0
Min sum of instance weight in a child	1
Subsample ratio of the training instance	1
Columns subsample ratio for splits / levels	1
Columns subsample ratio for splits	1
Balancing of positive and negative weights	1
Value treated as missing	NaN
Matrix type	dense

Table 9: Control parameters for Gradient Boost Trees model for the three wells.

	Well-1	**Well-2**	**Well-3**
Number of Boost stages	100	100	100
Feature sampling strategy	Fixed Proportion	Fixed Proportion	Fixed Proportion
Proportion of features to sample	1	1	1
Learning rate	0.1	0.1	0.1
Loss	Try least square	Try least square	Try least square
Max depth of trees	3	3	3

All the above were processed and filtered data offline. However, real-time data processing will require addressing several operational challenges, including data latency from surface and downhole sensors, sensors' reliability (e.g. defected sensors, deviated data) and data filtration time. The platform can address these limitations after obtaining real-time data and processing them with low hardware feasibility.

6. Results and Discussion

Dataiku offers eight in-memory evaluation metrics, we selected three metrices, R^2 score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) as indicators of each model's performance. The **Table-10** shows R^2 score of the predicted ROP for both the training and testing datasets, under conditions before and after using the D-exponent and MSE, applied to three directional wells for three machine learning (ML) techniques. The Table-10 represents guidelines for R^2 scores (Sarjana K. et al. 2021) [36].

Table 10: Guidelines for R2 scores

R^2	**Evaluation**
0.00 – 0.199	Very Low
0.20 – 0.399	Low
0.40 – 0.599	Medium
0.60 – 0.799	Strong
0.80 – 1.00	Very Strong

The **table-12,** and **13** follow the same structure but present results using MAE and RMSE, respectively. Regression plots of the predicted ROP are provided for the testing data, while training results are accessible exclusively through learning curves. Based on these tables, it is observed that XGBoost and

Gradient Boosted Trees achieve strong overall performance for three wells for approximately all metrics used. In contrast, SVM underperforms compared to the ensemble models, particularly in Well-1, where its R^2 score and RMSE are poorer, and training scores are lower than testing scores (both before and after applying MSE and D-exponent). SVM also struggles in Well-2, exhibiting high RMSE values regardless of preprocessing (**Table-13**). Additionally, in the R^2 scores, SVM shows a pronounced gap between training and testing scores just after addition of MSE and D-exponent calculations (**Table-11)**. These discrepancies suggest issues such as suboptimal hyperparameter tuning, data quality limitations, dataset complexity, and insufficient training data volume, all of which constrain SVM's effectiveness relative to ensemble models. In this case, the performance of SVM can be improved for well-2 through filtering outliers by customizing numerical values and rescaling methods which are in Dataiku's features handling to remove automatically potential of outliers. In general, the improvement of ROP prediction across all metrics highlights the value of integrating MSE and D-exponent calculations with diverse ML models.

The Well-3 R^2 score (0.999) for SVM in training data seems overfitting due to kernel's excessive flexibility in modelling as the Well-3 has low-variance drilling parameters (see Table-5). In addition to its larger data inputs (rows=1300) which could increase the model overfitting. In order to limit overfitting, it is suggested to optimize cross validation approaches through tuning training/testing ratios, modifying number of K-folds based on number of intervals or formations (i.e. $K>5$) for better evaluation, and optimizing number of regularization parameter (i.e. $C<15$) to limit generalization and overfitting. However, the testing R^2 score (0.997) remains valid as it reflects performance on unseen data, and the near-identical training/testing scores which confirm this behaviour.

The **Figures-3-20** show regression plot of the predicted ROP vs actual ROP for testing dataset applied on the three wells using the three models. These plots were taken from the software directly and cannot be customized. Every figure and what's directly under it represent ROP prediction with inclusion and exclusion of D-exponent and Specific Energy respectively with same technique. By reviewing the mentioned tables and figures, it is clear that using both D-exponent and MSE have enhanced ROP prediction significantly for all three wells and that emphasise the importance of using calculated features which evaluate drilling performance and formation rock toughness. To understand the contribution of each feature and their quantity importance on ROP prediction, Dataiku offers this advantage in form of features importance percentage on its platform, so we calculated features importance percentage and summarized in **Table-14, 15, and 16** for well-1, 2, and 3 respectively. Feature importance means contribution percentage of each feature to predict the required target (ROP). In other words, the accuracy of the predicted target will be changed when adding or neglecting these features. Based on these tables, we can conclude that adding these features significantly enhanced performance across all models and wells as confirmed by all metrics. The dramatic increase of R^2 scores to exceed 90% indicate very strong alignment with actual ROP. On the other hand, reduction in RMSE and MAE (e.g., Well-1's XGBoost RMSE drops by 80%) as it confirms lower absolute error magnitudes, demonstrated robust performance post-feature inclusion and effectiveness of handling complex data.

Ensemble models (XGBoost and Gradient Boosted Trees) outperformed SVM, primarily because they inherently rank features based on their contribution to minimizing prediction error. This allows these models to prioritize the most informative features, such as MSE and put them to top ranks, then marginalize other features (e.g. D-exponent) to avoid under or overfitting predictions. In contrast, SVM has other mechanism, which it tries to classify data patterns differently and priorities features which has direct relationship with ROP. Thus, SVM considers D-exponent as top priority comparing to rest of features. The inclusion of MSE and D-Exponent strengthened their relationships with rock properties and drilling efficiency, and ensemble models will increase focusing on using these parameters and put

them with highest priority and rest of features will be less important and deprioritized. SVM showed also very good performance for all wells. However, due to be sensitivity to complex data and outliers, this model performed less precise than rest of techniques. In case exclusion of MSE and D-exponent, all models reconfigured features importance and their ranking have been changed. All features don't have same ranking except Surface torque which is the top importance compared to rest of features. The reasons behind non-uniform of features ranking for all wells are mainly due to quality of data, data complexity, presence of outliers, and/or lack of training data for the three wells. In general, it does not mean that weak features with importance below 4% have negligible effects or negative effects on ROP prediction. When utilizing these weak features, the accuracy of predicted ROP will be slightly increase.

However, the improvement of ROP is related to numerical values of each feature data and their contribution to the required target. In other words, some high numerical values for a particular feature could have either positively or negatively impact on ROP prediction and vice versa. In order to understand behaviors of features used for ROP, we extracted features effects (or SHAP values) which are automatically calculated through Dataiku's platform and put in this study, see **figures-21-38** were every figure and what's directly under it represent features effect after and before inclusion of D-exponent and MSE respectively with same technique. For each data point, a SHAP value for a particular feature quantitates how much that feature's actual value contributed to pushing the model's prediction from the average prediction to its final predicted value. For better illustration, a positive SHAP value means that feature pushed the prediction higher while negative SHAP value means that feature pushed the prediction lower. The **figures-21,22,23,27,28,29,33,34, and 35** demonstrated that all three wells share that MSE, D-exponent, Bit diameter, and flow in pump with low numerical values are positively impact ROP prediction. As an example for **figure-21**, low MSE values has span range 34 units that strongly contribute to pushing the predicted ROP higher. For D-exponent, it has span range greater than 10 units for SVM which also contribute better prediction for ROP unlike ensemble models which have lower span range due to their different prioritizing mechanism. For Surface Torque, surface RPM, surface WOB, and measured depth, all techniques showed that high numerical values have positively impact on ROP prediction and vice versa. Other features, such as mud weight, standpipe pressure (Omogbolahan S. et al. (2019) [28]), well inclination, bit revolutions, and pore pressure estimation, showed least influence on ROP prediction due to their low importance and span range. Thus, variations in their numerical values had no statistically significant effect on model outputs. The rest of **figures-24,25,26,30,31**,**32,36,37, and 38** are exclusion of Mechanical Specific Energy and D-exponent, the models reconfigured features effects and focused on weak features (e.g. Standpipe Pressure, Flow in Pump, well inclination etc.) and decreased degree of overfitting data. Additionally, it is observed that numerical values for the three wells are not uniformly distributed. In other words, some features with high numerical values such as surface WOB and surface RPM may either have positive or negative impact on ROP prediction due to lack of normalization. We can deduct that inclusion of MSE and D-exponent have significant effect on distributing numerical values of other features.

While all models showed robust ROP prediction using our data from three wells, it is believed that this study will warrant further investigation for other fields with different formations. However, the variation effect of drilling parameters, geological properties, and data scalability on ROP prediction will be mitigated through augmenting data training data by simulating varied drilling conditions, while preserving physical relationships encoded in MSE and D-exponent. Future work should validate these models against larger datasets and other formations to assess broader applicability.

Table 11: Results of predicted ROP using coefficient R^2 score

	Before using Specific Energy and D-Exponent						After using Specific Energy and D-Exponent					
	XGBoost		Gradient Boosted Trees		SVM		XGBoost		Gradient Boosted Trees		SVM	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Well-1	0.570	0.373	0.745	0.520	0.342	0.383	0.995	0.974	0.989	0.946	0.791	0.903
Well-2	0.589	0.216	0.752	0.311	0.240	0.123	0.995	0.945	0.994	0.956	0.752	0.896
Well-3	0.944	0.862	0.914	0.846	0.868	0.815	0.997	0.983	0.990	0.978	0.999	0.997

Table 12: Results of predicted ROP using Mean Absolute Error (MAE)

	Before using Specific Energy and D-Exponent						After using Specific Energy and D-Exponent					
	XGBoost		Gradient Boosted Trees		SVM		XGBoost		Gradient Boosted Trees		SVM	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Well-1	3.501	3.776	2.822	3.143	3.220	3.336	0.413	0.748	0.600	0.898	0.557	0.614
Well-2	3.81	4.499	2.920	4.122	3.720	4.508	0.441	0.959	0.459	0.867	0.514	0.798
Well-3	0.595	0.947	0.753	1.002	0.873	1.036	0.132	0.311	0.260	0.371	0.082	0.104

Table 13: Results of predicted ROP using Root Mean Square Error (RMSE)

	Before using Specific Energy and D-Exponent						After using Specific Energy and D-Exponent					
	XGBoost		Gradient Boosted Trees		SVM		XGBoost		Gradient Boosted Trees		SVM	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Well-1	5.600	5.801	4.137	5.077	6.656	5.758	0.550	1.172	0.800	1.710	3.75	2.287
Well-2	5.23	6.524	4.04	6.117	7.066	6.880	0.593	1.727	0.626	1.541	4.05	2.381
Well-3	0.810	1.269	1.00	1.341	1.240	1.467	0.210	0.439	0.350	0.529	0.130	0.186

Table 14: Feature importance percentages for predicting ROP in Well-1

Feature Name	Well-1					
	After Using D-exponent and Specific Energy			Before using D-exponent and Specific Energy		
	XGBoost	Gradient Boost Trees	SVM	XGBoost	Gradient Boost Trees	SVM
MSE	39 %	40 %	8 %	-	-	-
D-exponent	10 %	9 %	34 %	-	-	-
Surface Torque	22 %	24 %	5 %	32 %	33 %	35 %
MD (Measured Depth)	12 %	10 %	7 %	36 %	32 %	5 %
Surface RPM	6 %	6 %	8 %	2 %	4 %	5 %
Surface WOB	6 %	6 %	27 %	6 %	5 %	6 %
Hours on Bit	2 %	1 %	1 %	8 %	9 %	8 %
Pit volume	1 %	<0.5%	1 %	3 %	5 %	2 %
Inclination	1 %	1 %	2 %	1 %	3 %	5 %
SPP (Stand Pipe P.)	<0.5%	<0.5%	2 %	5 %	2 %	9 %
Flow in Pump	<0.5%	<0.5%	1 %	<0.5%	2 %	4 %
Pore Pressure Est.	<0.5%	<0.5%	1 %	4 %	2 %	4 %
Revs. on Bit	<0.5%	1 %	2 %	4 %	2 %	10 %
Mud density	<0.5%	2 %	2 %	<0.5%	<0.5%	8 %

Table 15: Feature importance percentages for predicting ROP in Well-2

Feature Name	Well-2					
	After Using D-exponent and Specific Energy			Before using D-exponent and Specific Energy		
	XGBoost	Gradient Boost Trees	SVM	XGBoost	Gradient Boost Trees	SVM
MSE	49 %	50 %	7 %	-	-	-
D-exponent	1 %	1 %	38 %	-	-	-
Surface Torque	13 %	1 %	2 %	25 %	26 %	25 %
MD (Measured Depth)	<0.5%	3 %	1 %	<0.5%	6 %	13 %
Surface RPM	10 %	14 %	7 %	19 %	14 %	9 %
Surface WOB	1 %	1 %	32 %	12 %	14 %	9 %
Hours on Bit	7 %	1 %	1 %	23 %	14 %	6 %
Bit Diameter	17 %	17 %	6 %	<0.5%	<0.5%	8 %
Inclination	<0.5%	1 %	3 %	6 %	5 %	4 %
SPP (Stand Pipe P.)	<0.5%	<0.5%	1 %	4 %	8 %	13 %
Flow in Pump	<0.5%	<0.5%	1 %	4 %	4 %	3 %
Pore Pressure Est.	<0.5%	<0.5%	1 %	7 %	3 %	5 %
Revs. on Bit	<0.5%	3 %	1 %	<0.5%	8 %	5 %
Mud density	<0.5%	<0.5%	1 %	<0.5%	<0.5%	1 %

Table 16: Feature importance percentages for predicting ROP in Well-3

Feature Name	Well-3					
	After Using D-exponent and Specific Energy			Before using D-exponent and Specific Energy		
	XGBoost	Gradient Boost Trees	SVM	XGBoost	Gradient Boost Trees	SVM
MSE	31 %	30 %	21 %	-	-	-
D-exponent	2 %	1 %	15 %	-	-	-
Surface Torque	3 %	23 %	11 %	40 %	41 %	33 %
MD (Measured Depth)	18 %	23 %	8 %	7 %	8 %	6 %
Surface RPM	<0.5%	4 %	3 %	1 %	4 %	2 %
Surface WOB	3 %	2 %	14 %	4 %	4 %	6 %
Hours on Bit	5 %	3 %	3 %	14 %	12 %	9 %
Pit volume	3 %	4 %	1 %	14 %	12 %	8 %
Inclination	3 %	3 %	1 %	9 %	9 %	6 %
SPP (Stand Pipe P.)	1 %	<0.5%	2 %	7 %	5 %	7 %
Flow in Pump	10 %	7 %	5 %	2 %	2 %	9 %
Bit Diameter	1 %	<0.5%	12 %	<0.5%	<0.5%	6
Revs. on Bit	1 %	4 %	3 %	3 %	6 %	9 %
Mud density	1 %	<0.5%	2 %	<0.5%	<0.5%	<0.5%

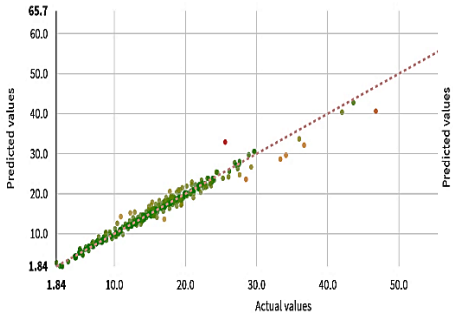

Figure 3: Predicted ROP vs actual ROP for XGBoost after using D-exponent & MSE (Well-1 R^2 = 0.974).

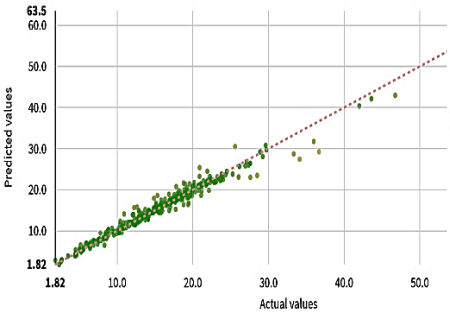

Figure 4: Predicted ROP vs actual ROP for Gradient Boosted Trees after using D-exponent & MSE (Well-1 R^2= 0.946).

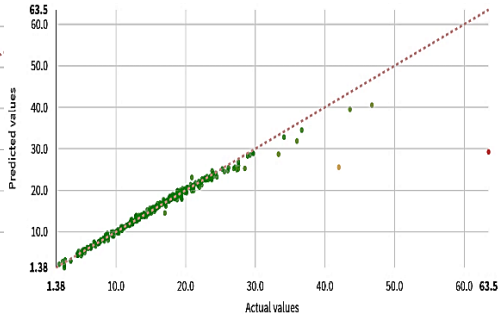

Figure 5: Predicted ROP vs actual ROP for SVM after using D-exponent & MSE (Well-1 R^2= 0.903).

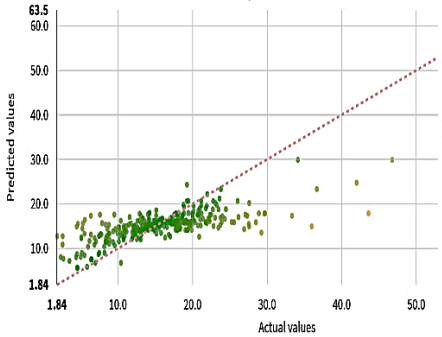

Figure 6: Predicted ROP vs actual ROP for XGBoost before using D-exponent & MSE (Well-1 R^2 = 0.373).

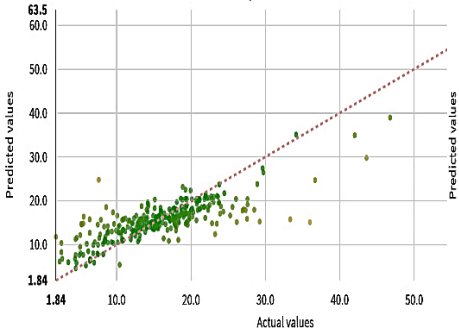

Figure 7: Predicted ROP vs actual ROP for Gradient Boosted Trees before using D-exponent & MSE (Well-1 R^2 = 0.52).

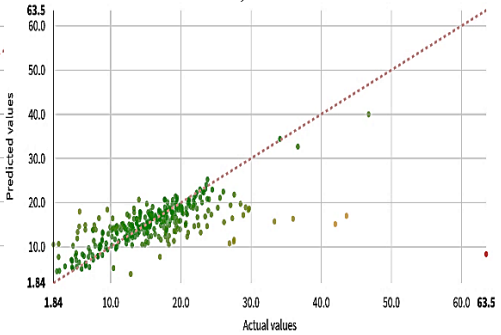

Figure 8: Predicted ROP vs actual ROP for SVM before using D-exponent & MSE (Well-1 R^2 = 0.383).

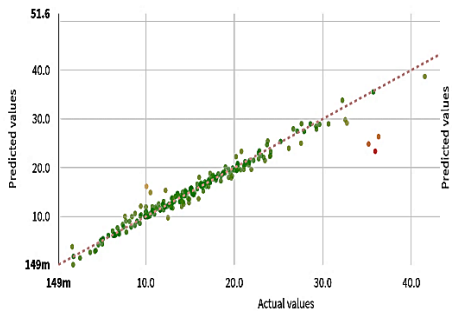

Figure 9: Predicted ROP vs actual ROP for XGBoost after using D-exponent & MSE (Well-2 R^2= 0.945).

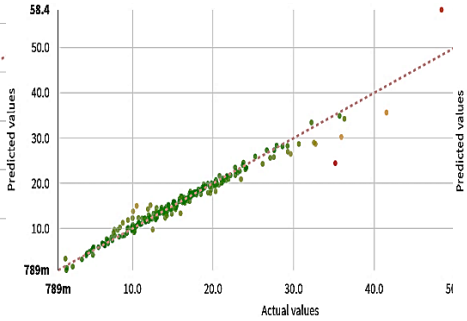

Figure 10: Predicted ROP vs actual ROP for Gradient Boosted Trees after using D-exponent & MSE (Well-2 R^2 = 0.956).

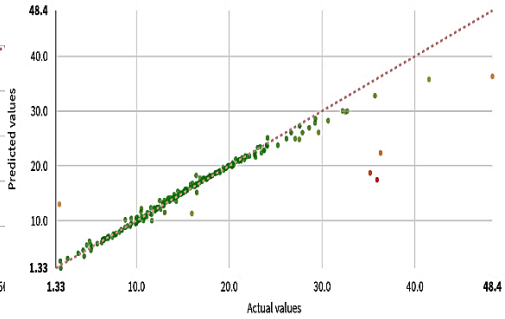

Figure 11: Predicted ROP vs actual ROP for SVM after using D-exponent & MSE (Well2 R^2 = 0.896).

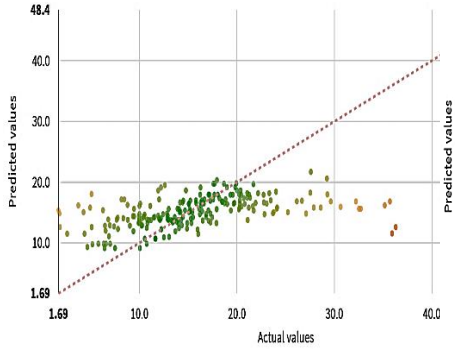

Figure 12: Predicted ROP vs actual ROP for XGBoost before using D-exponent & MSE (Well-2 R^2 = 0.216).

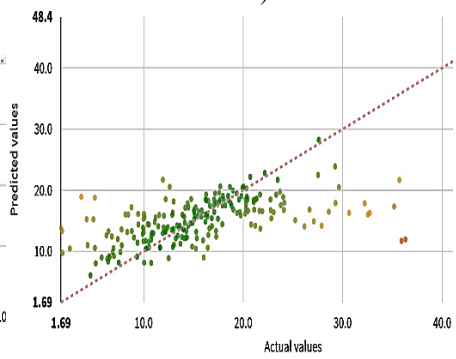

Figure 13: Predicted ROP vs actual ROP for Gradient Boosted Trees before using D-exponent & MSE (Well-2 R^2 = 0.311).

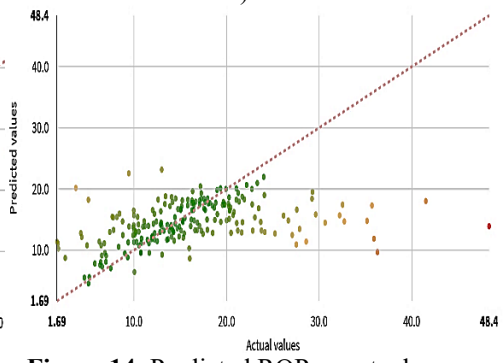

Figure 14: Predicted ROP vs actual ROP for SVM before using D-exponent & MSE (Well-2 R^2=0.128).

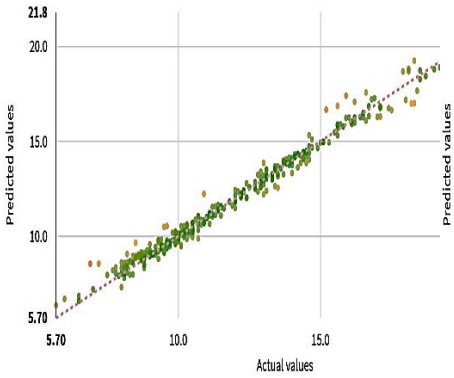

Figure 15: Predicted ROP vs actual ROP for XGBoost after using D-exponent & MSE (Well-3 R^2=0.983).

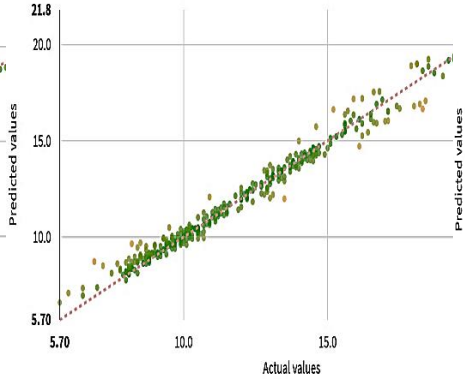

Figure 16: Predicted ROP vs actual ROP for Gradient Boosted Trees after using D-exponent & MSE (Well-3 R^2= 0.976).

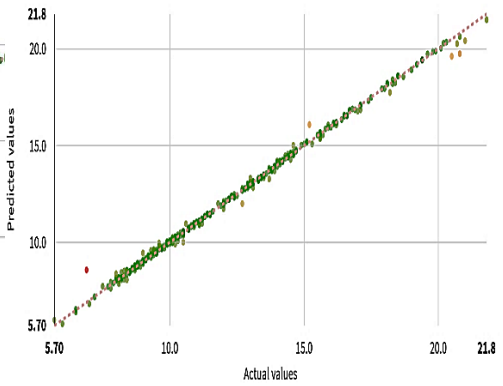

Figure 17: Predicted ROP vs actual ROP for SVM Boosted Trees after using D-exponent & MSE (Well-3 R^2= 0.997).

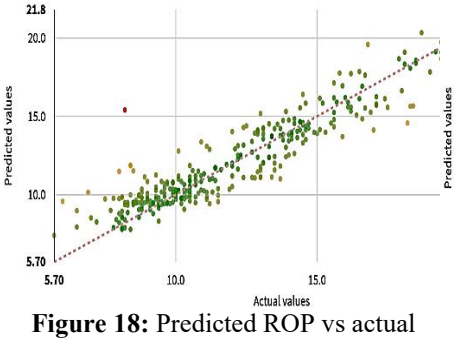

Figure 18: Predicted ROP vs actual ROP for XGBoost before using D-exponent & MSE (Well-3 R^2=0.862).

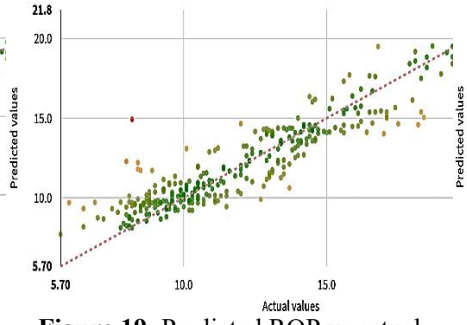

Figure 19: Predicted ROP vs actual ROP for Gradient Boosted trees before using D-exponent & MSE (Well-3 R^2 = 0.846)

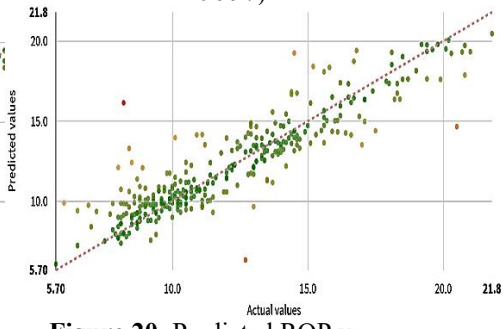

Figure 20: Predicted ROP vs actual ROP for SVM before using D-exponent & MSE (Well-3 R^2 = 0.815)

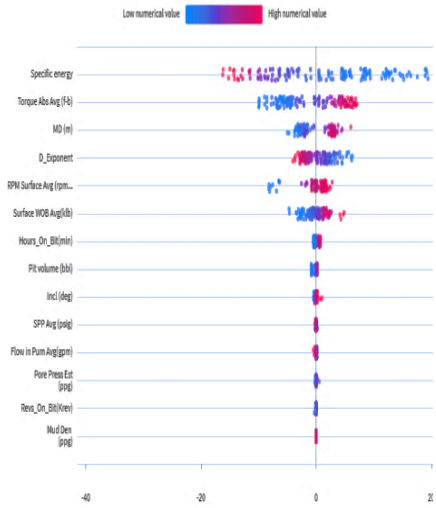

Figure 21: Features Effect for ROP prediction using XGBoost technique after using D-exponent & MSE (Well-1)

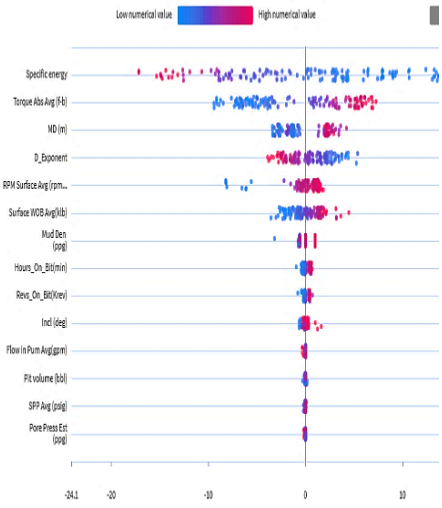

Figure 22: Features Effect for ROP prediction using Gradient Boosted Trees technique after using D-exponent & MSE (Well-1)

Figure 23: Features Effect for ROP prediction using SVM technique after using D-exponent & MSE (Well-1)

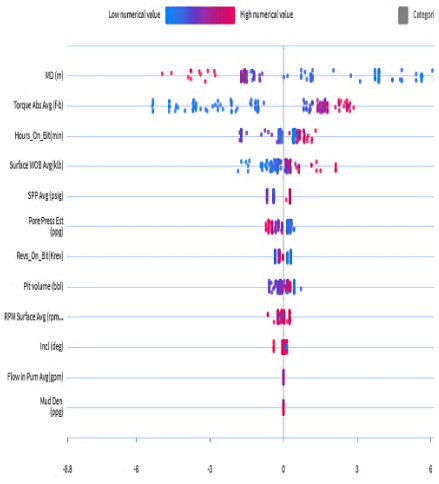

Figure 24: Features Effect for ROP prediction using XGBoost technique before using D-exponent & MSE (Well-1)

Figure 25: Features Effect for ROP prediction using Gradient Boosted Trees technique before using D-exponent & MSE (Well-1)

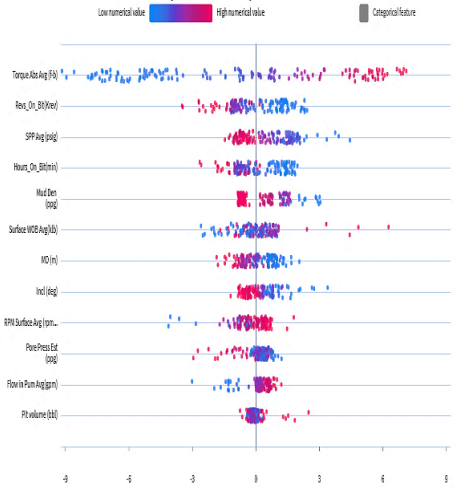

Figure 26: Features Effect for ROP prediction using SVM technique before using D-exponent & MSE (Well-1)

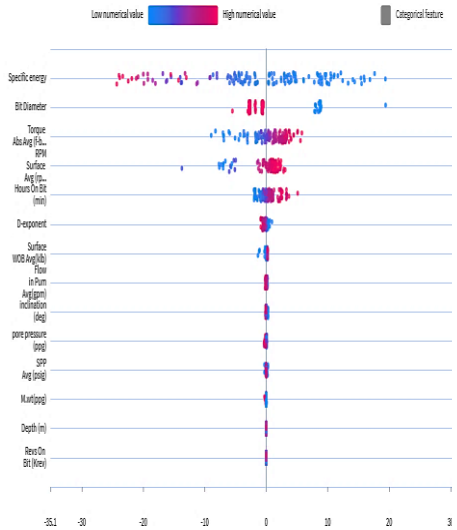

Figure 27: Features Effect for ROP prediction using XGBoost technique after using D-exponent & MSE (Well-2)

Figure 28: Features Effect for ROP prediction using Gradient Boosted Trees technique after using D-exponent & MSE (Well-2)

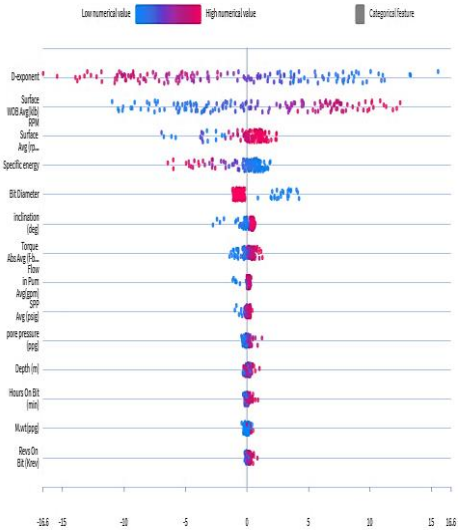

Figure 29: Features Effect for ROP prediction using SVM technique after using D-exponent & MSE (Well-2)

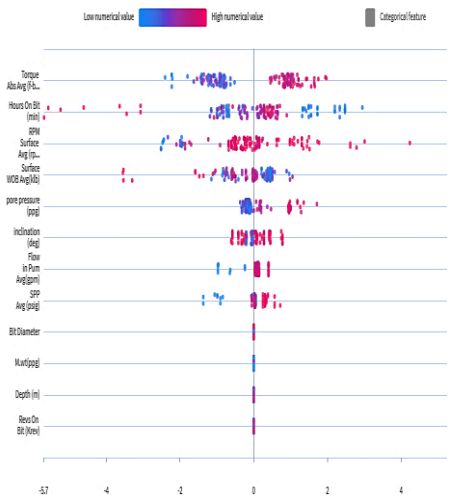

Figure 30: Features Effect for ROP prediction using XGBoost technique before using D-exponent & MSE (Well-2)

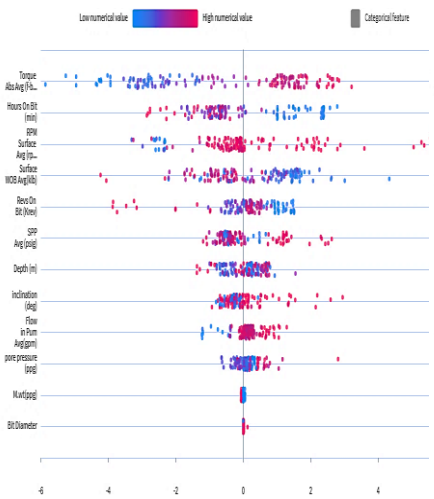

Figure 31: Features Effect for ROP prediction using Gradient Boosted Trees technique before using D-exponent & MSE (Well-2)

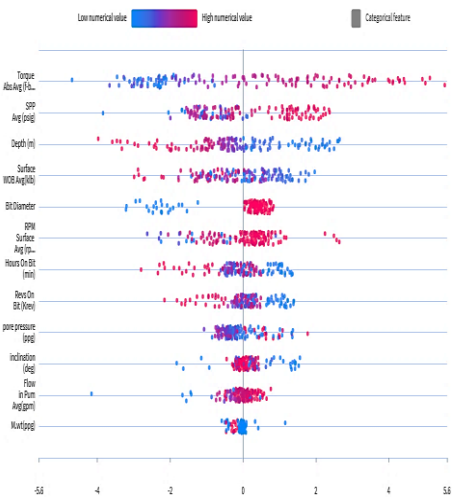

Figure 32: Features Effect for ROP prediction using SVM technique before using D-exponent & MSE (Well-2)

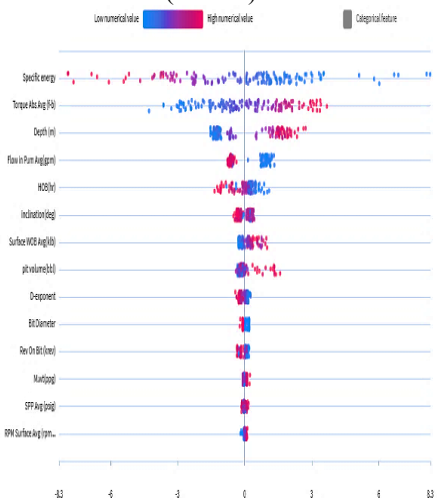

Figure 33: Features Effect for ROP prediction using XGBoost technique after using D-exponent & MSE (Well-3)

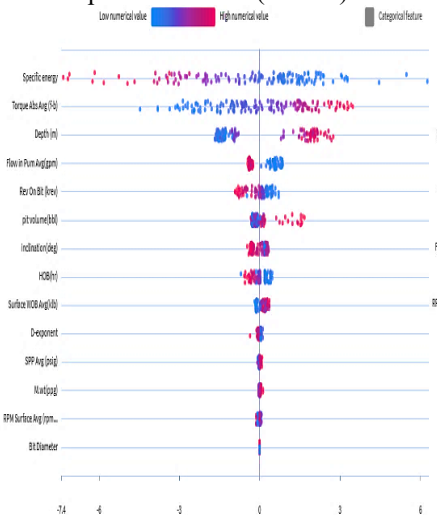

Figure 34: Features Effect for ROP prediction using Gradient Boosted Trees technique after using D-exponent & MSE (Well-3)

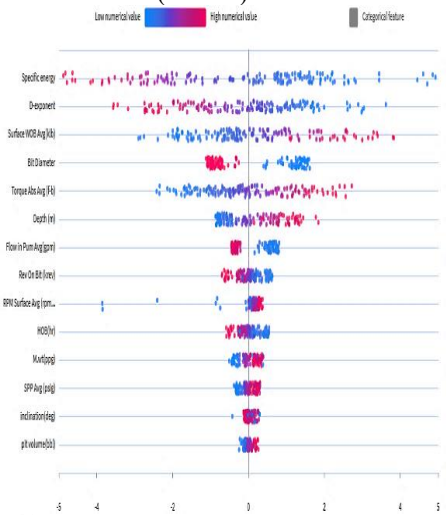

Figure 35: Features Effect for ROP prediction using SVM technique after using D-exponent & MSE (Well-3)

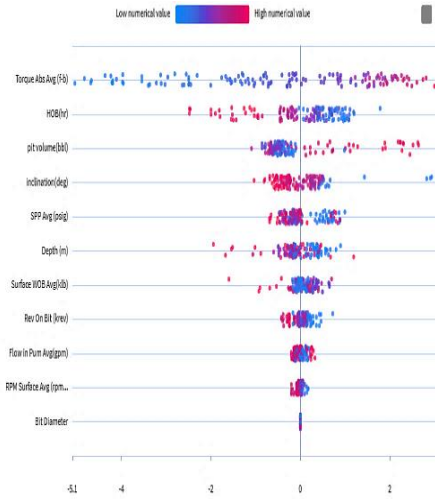

Figure 36: Features Effect for ROP prediction using XGBoost technique before using D-exponent & MSE (Well-3)

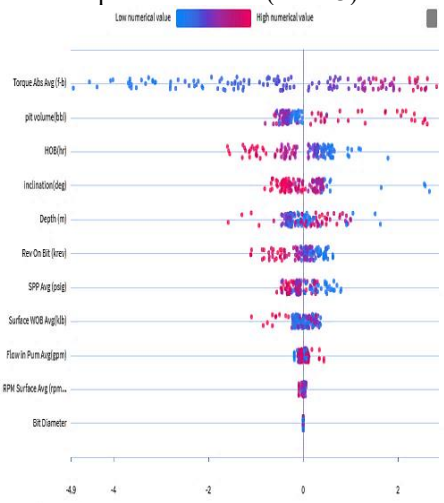

Figure 37: Features Effect for ROP prediction using Gradient Boosted trees technique before using D-exponent & MSE (Well-3)

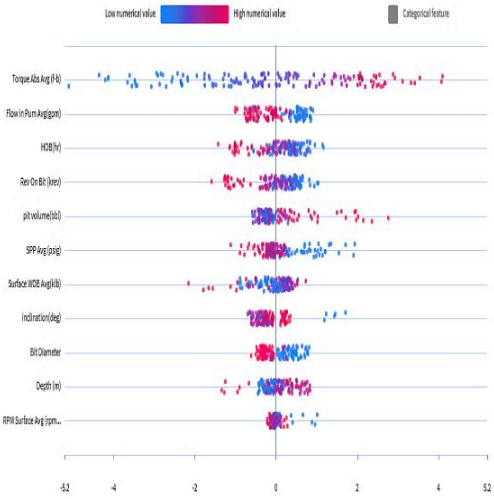

Figure 38: Features Effect for ROP prediction using SVM technique before using D-exponent & MSE (Well-3)

7.Conclusions

- The ROP was predicted using Dataiku DSS with built-in ML tools. This approach minimizes reliance on Python scripting, streamlining the workflow for users without advanced programming skills. While domain expertise remains critical for tasks such as model selection, hyperparameter tuning, and interpretation, the software reduces dependency on external consultants for routine coding tasks. However, the built-in ML tools in the software are constrained with flexibility such as algorithm modifications for advanced or hybrid technology so, the alternatives will require to add custom python scripting in the software for creating new model.
- All models showed results with very strong performance after incorporating Mechanical Specific Energy and D-exponent, where significant improvement of all metrics indicates a strong performance across the three wells. An example quantitative improvement to support this study that XGBoost showed increase R^2 from 0.373 to 0.974 for Well-1, from 0.216 to 0.945 for Well-2, and from 0.862 to 0.983 for Well-3.
- The significant improvement of ROP prediction for all models is due to using calculated parameters Mechanical Specific Energy and D-exponent which represent formation rock roughness and energy required for enhancing drilling performance.
- The use of features importance percentage helped us to identify and highlight the most significant features that dominate ROP prediction. All three wells have varied in features percentages and their sorting due to data quality and complexity. However, they share that Mechanical Specific Energy and D-exponent are the most important feature importance.
- The use of features effect or SHAP values highlighted that using calculated parameters (MSE and D-exponent) which evaluate drilling performance indicated that weak features (e.g., mud density, SPP) don't have influence on predicted ROP regardless of their numerical values either high or low. In contrast, strong features e.g. MSE and D-exponent have positively impact on ROP prediction when having low numerical values across all models specially SVM and vice versa. This analysis will help engineers to enhance data quality and to identify relevant relationship between drilling parameters used and ROP in order to reach the optimum performance.
- Future works will focus on using new approaches (e.g. advanced and/or hybrid models, modified MSE equations) for predicting ROP with comparative analysis to enhance transparency and interpretability of model predictions. In addition to their implications on other fields.

List of abbreviations

ROP	Rate of Penetration	**MLP-NN**	Multilayer Perceptron-Neural Network
ML	Machine learning	**MLP-FF**	Multilayer Perceptron-Firefly algorithm
WOB	Weight on bit	**CFNN**	Cascade-Forward Neural Network
RPM	Revolution per minute	**RBFNN**	Radial Basis Function Neural Network
T	Surface Torque	**IGA**	Integrated Genetic Algorithm
SPP	Standpipe pressure	**MNR**	Multiple Nonlinear Regression
Q	Pumping rate	**ELM**	Extreme Learning Machine
HOB	Hours on bit	**MD**	Measured depth
TFA	Total flow area	**MD**	Well measured depth
KNN	K-nearest neighbors	**D**	Drill Bit diameter
R^2	Coefficient of determination	**PP**	Pump pressure

MAE	Mean Absolute Error	**BFR**	Bit flow rate
RMSE	Root mean squared error	**M.wt**	Mud weight
MSE	Mechanical Specific Energy	**PV**	Plastic viscosity
AAPRE	Average Absolute Percentage Relative Error	**D_{bit}**	Bit diameter
ANN	Artificial neural networks	**PP**	Pore pressure estimation
MLP	Multilayer Perception	**OVB**	Over-burden pressure
MPNN	Multilayer Perceptron Neural Network	**Inc.**	Inclination
MLP-GSA	Multilayer Perceptron-Gravitational Search Algorithm	**CDF**	Cumulative Distribution Functions

8. Statements and Declarations

8.1 Availability of data and materials

Data will be available on request.

8.2 Competing interests

The authors declare that they have no competing interests.

8.3 Funding

The authors did not receive support from any organization for the submitted work, and no funds, grants, or other support were received.

8.4 Author Contribution declaration

There is no declaration of interest in this research.

8.5 Authors' Contributions

- Ahmad Atef, methodology, software preparation, validation of software, investigation, and original draft writing.
- Ahmed G. Hagag, Methodology, Supervision, and editing.
- Attia M. Attia, Technical consultation, and draft supervision. The authors read and approved the final manuscript.

References

[1] Abdulmalek A. S., Elkatatny, S., Abdulraheem, A. Mohammed M., Abdelwahab Z.A. and Mohamed I.M. "Prediction of Rate of Penetration of Deep and Tight Formation Using Support Vector Machine." Paper SPE-192316-MS, 2018.

[2] Akgun F. "How to Estimate the Maximum Achievable Drilling Rate without Jeopardizing Safety." Paper SPE 78567 presented at Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, United Arab Emirates, 13-16 October, 2002.

[3] Benzminabadi S.N, Ahmad R., Seyed M.E.J. Behzad T., and Abbas R. "Effect of rock properties on ROP modeling using statistical and intelligent methods: a case study of an oil well in southwest of Iran." PAN Journals. DOI 10.1515/amsc-2017-0010, 2017.

[4] Bingham, M.G., "A New Approach to Interpreting Rock Drillability." The Petroleum Publishing Co., 1969.

[5] Bourgoyne A.T., Millheim K.K., Chenevert M.E., and Young F.S. "Applied Drilling Engineering." Society of Petroleum Engineers Text Book Series, Vol. 1, Richardson, TX. 1986.

[6] Bourgoyne, A.T. and Young, F.S. "A Multiple Regression Approach to Optimal Drilling and Abnormal Pressure Detection." Society of Petroleum Engineers Journal. 14 (4): 371–384., 1974.

[7] Cunningham, R.A. "Analytical Determination of Optimum Bit Weight and Rotary Speed Combinations." Presented at the Fall Meeting of the Society of Petroleum Engineers of AIME, Dallas, Texas, 1960.

[8] Ehsan B., Ebrahim B.D., and Kasra K. "Prediction of penetration rate in drilling operations: a comparative study of three neural network forecast methods". Journal of Petroleum Exploration and Production Technology. DOI: 10.1007/s13202-020-01066-1, 2021.

[9] El-Sayed Y., Salem A.S., and El-Rammah S. "Rate of Penetration Prediction in Drilling Operation in Oil and Gas Wells by K-nearest Neighbors and Multi-layer Perceptron Algorithms." Journal of Mining and Environment (JME), Vol. 14, No. 3, 2023, 755-770. DOI:10.22044/jme.2023.12694.2306, 2023.

[10] Hareland, G. and Rampersad, P.R. "Drilling Optimization Using Drilling Data and Available Technology." Presented at the SPE Latin America/Caribbean Petroleum Engineering Conference, Buenos Aires, Argentina. pp. 657–667. https://doi.org/10.2118/26957-ms, 1994.

[11] Hazbeh, O., Aghdam, S.K. ye, Ghorbani, H., Mohamadian, N., Ahmadi Alvar, M., and Moghadasi, J. "Comparison of accuracy and computational performance between the machine learning algorithms for Rate of Penetration in directional drilling well." Petroleum Research. 6 (3): 271–282. Research DOI.org/10.1016/j.ptlrs.2021.02.004, 2021.

[12] Hegde, C., Pyrcz, M., Millwater, H. Daigle H., and Gray K. "Fully Coupled End-to-End Drilling Optimization Model Using Machine Learning." J Pet Sci Eng 186 (March): 106681. https://doi.org/10.1016/j.petrol.2019.106681. 2020.

[13] Jahanbakhshi, R.K.R., "Real-time prediction of Rate of Penetration during drilling operation in oil and gas wells. Am." Rock Mech. Assoc. 53, 127. ARMA 12-244, 2012.

[14] Kloucha C.K., El-Yossef B.S. Al-Hamlawi I., Salim M., Pausin W. Pal A., Mustapha H. Shah S., and Hussein A.N. "Machine Learning Model for Drilling Equipment Recommender System forImproved Decision Making and Optimum Performance." Paper SPE-211731-MS, DOI 10.2118/211731-MS, 2022.

[15] Noshi C. I. and Schubert, J. J. "Application of Data Science and Machine Learning Algorithms for ROP Prediction: Turning Data into Knowledge." Paper presented at the Offshore Technology Conference, Houston, Texas, USA, and 6–9 May. OTC-29288-MS. https://doi.org/10.4043/29288-MS, 2019.

[16] Shaygan K., and Jamshidi S. "Prediction of Rate of Penetration in directional drilling using data mining techniques." Journal of Petroleum Science and Engineering 221, 2023. 111293. https://doi.org/10.1016/j.petrol.2022.111293

[17] Shi X., Liu, G., Gong, X., Zhang, J., Wang J., and Zhang H., "An efficient approach for real-time prediction of Rate of Penetration in offshore drilling." Math. Probl Eng. 13, 2016. https://doi.org/10.1155/2016/3575380

[18] Majidi R. Albertin M., and Last N. "Pore-Pressure Estimation by Use of Mechanical Specific Energy and Drilling Efficiency." Paper SPE 178842 Drilling & Completion, June 2017.

[19] Maurer W.C. "The perfect-cleaning theory of rotary drilling." J. Petrol. Technol. 14 (11), 1270–1274, 1962. https://doi.org/10.2118/408-pa.

[20] Moran D., Ibrahim H., Purwanto, A., and Osmond, J. "Sophisticated ROP prediction technologies based on neural network delivers accurate drill time results." Proc, IADC/SPE Asia Pacific Drill. Technol. Conf. Exhib. Held Ho Chi Minh City, Vietnamm, 1–3 November, 100–108, 2010.

[21] Olukoga T.A. and Feng Y. "Practical Machine-Learning Applications in Well-Drilling Operations." Paper SPE-205480-PA Drilling & Completion, 2021. https://doi.org/10.2118/205480-PA

[22] Lenwoue A., Li Z., Tang C., Zhang W., Ding S., Hu P., and Sun W. "Recent Advances and Challenges of the Application of Artificial Intelligence to Predict Wellbore Instabilities during Drilling Operations." Paper SPE 215830-PA., 2023.

[23] Li C. and Cheng C. "Prediction and Optimization of Rate of Penetration using a Hybrid Artificial Intelligence Method based on an Improved Genetic Algorithm and Artificial Neural Network." Paper SPE-203229-MS presented at the Abu Dhabi International Petroleum Exhibition & Conference to be held in Abu Dhabi, 2020.

[24] Li Y. and Samuel, R. "Prediction of Penetration Rate Ahead of the Bit through Real-Time Updated Machine Learning Models." Paper presented at the SPE/IADC International Drilling Conference and Exhibition, The Hague, The Netherlands, 5–7 March. SPE-194105-MS, 2019. https://doi.org/10.2118/194105-MS.

[25] Speer, J.W. "A Method for Determining Optimum Drilling Techniques." Paper SPE-1242-G presented at the Gulf Coast Drilling and Production Meeting, Lafayette, Louisiana, 1959.

[26] Mohamadian, N., Ghorbani, H., Wood D.A. Merhad M., Davoodi S., Rashidi S., Soleimanian A., and Shahvad A.K. "A Geomechanical Approach to Casing Collapse Prediction in Oil and Gas Wells Aided by Machine Learning." *J Pet Sci Eng* 196 (107811), 2021. 107811. https://doi.org/10.1016/j.petrol.2020.107811.

[27] Siddique, N., and Adeli, H., "Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing. Wiley, 2013.
[28] Omogbolahan S. A, Ahmed A. A., and Ariffin S. "Computational intelligence-based prediction of drilling Rate of Penetration: A comparative study." Journal of Petroleum Science and Engineering. 2019. 172. 1-12. DOI.org/10.1016/j.petrol.2018.09.027
[29] Jorden, J. R. and Shirley, O.J. "Application of Drilling Performance Data to ovepressure Detection." J. Pet Technol 18 (11): 1387-1394. SPE-1407-PA., 1966. DOI.org/10.2118/1407-PA
[30] Teale R. "The Concept of Specific Energy in Rock Drilling." Int. J. Rock Mech. Mining Sci. 2 (1): 57–73, 1965. https://doi.org/10.1016/0148-9062(65)90022-7.
[31] Warren, T. M. "Penetration Rate Performance of Roller Cone Bits." Paper SPE-13259-PA., 1987. DOI: 10.2118/13259-PA.
[32] Detournay, E., & Defourny, P. "A phenomenological model for the drilling action of drag bits." International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts. 1992. DOI: 10.1016/0148-9062(92)91041-3.
[33] Pessier & Fear "Quantifying Common Drilling Problems With Mechanical Specific Energy and a Bit-Specific Coefficient of Sliding Friction." Paper SPE-24584-MS. presented at the SPE Annual Technical Conference and Exhibition, Washington, D.C., 1992. https://doi.org/10.2118/24584-MS**.**
[34] Dataiku. "In-memory Python". 2025. Retrieved from In-memory Python — Dataiku DSS 14 documentation.
[35] Dataiku. "Settings: Train/Test set." 2025. Retrieved from Prediction settings — Dataiku DSS 14 documentation.
[36] Sarjana K., Hayati L., and Wahidaturrahmi W. "Mathematical modelling and verbal abilities: How they determine students' ability to solve mathematical word problems?" Beta Jurnal Tadris Matematika, 2021. 13(2):117-129. DOI:10.20414/betajtm.v13i2.390.
[37] Ramy F., El Kammar M., Dahroug A., & Abd-Elfattah N. "AVO class II advanced prospectively workflow for deep Miocene in west delta deep marine, offshore Nile Delta, Egypt." Journal of African Earth Sciences. Vol. 227, 2025. 105634. https://doi.org/10.1016/j.jafrearsci.2025.105634
[38] Naser M.Z. "Machine learning for all! Benchmarking automated, explainable, and coding-free platforms on civil and environmental engineering problems." Journal of Infrastructure Intelligence and Resilience, 2023. https://doi.org/10.1016/j.iintel.2023.100028.
[39] Ganaie M.A., Hu M., Malik A.K., Tanveer M., and Suganthan P.N. "Ensemble deep learning: A review." Engineering Applications of Artificial Intelligence 115, 2022. 105151. https://doi.org/10.1016/j.engappai.2022.105151.
[40] Huang, F., Ash, J., Langford, J., Schapire, R., "Learning deep resnet blocks sequentially using boosting theory." International Conference on Machine Learning. PMLR, pp. 2058–2067, 2018.
[41] Tabik, S., Alvear-Sandoval, R.F., Ruiz, M.M., Sancho-Gómez, J.-L., Figueiras-Vidal, A.R., Herrera, F., "MNISTNET10: A heterogeneous deep networks fusion based on the degree of certainty to reach 0.1% error rate. Ensembles overview and proposal." Inf. Fusion 62, 73–80, 2020. http://dx.doi.org/10.1016/j.inffus.2020.04.002.