Ministry of Higher Education and Scientific Research

Kasdi Merbah University -Ouargla-

Institute of Technology

Department of Applied Engineering

Domaine : Science and Technology

Field: Optics and Precision Mechanics

Specialty: Measurements, Metrology, Quality

Thesis

LICENCE PROFESSIONNEL

Theme :

Lithology recognition during drilling through drilling parameters

Presented by :

Nour el Houda BENAISSA and Ayeterrahemene SIOUANE

Publicly defended on: June 04th 2024

Before the examination committee

Dr. MAHBOUB Abbdelbasset	**Président**
Pr. KADRI Mohemed el mehdi	**Supervisor**
Dr. HAMMAD Nabila	**Co-supervisor**
Dr. SETTOU Belkhir	**Examiner**

Année Universitaire : 2023/2024

Table of figures

Abbreviation List

ANN: artificial neural networks

BIT_RPM: Rotation per minute

CNN: Convolutional Neural Networks

DTs : Decision trees

ECDBIT: Effective Circulation Density on Bit

FLOWOUT: Outgoing mud flow

IADC: International Association of Drilling Contractors

LWD; Logging While Drilling

ML: Machine Learning

MWD: Measurement While Drilling

PUMP: Pump pressure

ROP: Rate Of Penetration

ROP_AVG: Average Rate Of Penetration

RPM: Rotation Per Minute

SPP: Stand Pipe Pressure

SVC: Support Vector Classification

SVM: Support Vector Machine

TG: Total Gas

TMD: True Measured Depth

TOTGAS: Total Gas

TVD: True Vertical Depth

WOB: Weight On Bit

WOB: Weight On Bit

Table of content

Thanks and appreciate

We would like to express our deepest appreciation to our thesis supervisor, Pr. KADRI Mohammed Mehdi, and co-supervisor, Dr. HAMMAD Nabila, for their invaluable guidance, support, and expertise throughout this research project. Their unwavering dedication, insightful feedback, and constant encouragement have played a pivotal role in shaping the success of this thesis.

Additionally, we would like to extend our sincere gratitude to the university and the institute family who have provided invaluable assistance and support throughout our studies. Your dedication and commitment to creating a conducive learning environment have been crucial to the completion of this thesis

Furthermore, we would also like to express our gratitude to our internships supervisors, Mr. YOUSFI Walid and Mr. BENACER Abdel Ali HSE Engineers, Mr. Mezzar Abdel Kamel and Mr. Achour Mohammed from the Mudlogging department at National Well Services Company (ENSP), and Mr. Mohamed KCHIDA and Mr. Ahmed LACHRAF Senior Instrumentation technicians at the National Algerian Drilling Company (ENAFOR), for the valuable experience and insights they have provided during our time with them.

We are also immensely thankful to our family and friends for their unconditional love and support. Their belief in us and their constant motivation have been a driving force that has kept us going, even during the most challenging times. We are particularly grateful to our parents, whose unwavering support and encouragement have been a source of strength, pushing us to persevere and achieve our goals.

Lastly, we want to convey our heartfelt appreciation to Dr. CHERGUI Abdelhakim, who has been a mentor, a second father, and a big brother to us. Your guidance, wisdom, and unwavering support have been instrumental in our personal and academic growth, and we are truly grateful for your presence in our lives.

To all of you, we extend our sincere and heartfelt gratitude. Your contributions have been invaluable, and we are forever indebted to you for your role in the success of this thesis

Abstract

This study addresses the problem of accurately predicting the tops of the formation and its lithology while drilling using a novel machine learning (ML) approach. Our model achieves 68% accuracy and 69% precision in real-time litho-facies identification using 11 drilling parameters. Extensive data preprocessing ensured unbiased and effective model performance. The model was carefully trained and evaluated, with extensive data preprocessing to reduce features, balance the sample distribution, and ensure an unbiased dataset. This attention to data quality and preprocessing is crucial for effective model performance.

A case study in the Hassi Messaoud oil field validated the model's practical applicability y. This ML-based methodology offers significant advantages for real-time geosteering, enhancing decision-making and efficiency in drilling operations. The model successfully predicts fine sand and fine sandstone lithologies, and for clay predictions are reasonably close to the real values. However, for salt, the model significantly overestimates the presence compared to the real data, suggesting challenges in predicting the “salt” lithology.

Résumé

Cette étude aborde le problème de la prévision précise des sommets des formations et de leur lithologie pendant le forage à l'aide d'une nouvelle approche de learning machine (ML). Notre modèle, atteint une précision de 68 % et une exactitude de 69 % dans la reconnaissance en temps réel des litho-faciès en utilisant 11 paramètres de forage. Un prétraitement des données exhaustif a assuré une performance du modèle non biaisée et efficace. Le modèle a été soigneusement entraîné et évalué, avec un prétraitement des données étendu pour réduire les caractéristiques, équilibrer la distribution des échantillons et garantir un ensemble de données non biaisé. Cette attention à la qualité des données et au prétraitement est cruciale pour une performance efficace du modèle.

Une étude de cas dans le champ pétrolifère de Hassi Messaoud a validé la mise en pratique du modèle. Cette méthodologie basée sur le ML offre des avantages significatifs pour la surveillance géologique en temps réel, améliorant la prise de décision et l'efficacité des opérations de forage. Le modèle prédit avec succès les lithologies de sable fin et de grès fin, et pour les argiles, les prédictions sont raisonnablement proches des valeurs réelles. Cependant, pour le sel, le modèle surestime considérablement sa présence par rapport aux données réelles, ce qui suggère des difficultés dans la prédiction de la lithologie "sel".

ملخص

هذا البحث يعالج تحدي التنبؤ الدقيق بالتكوينات الحجرية والحدود الطبقية في صناعة النفط باستخدام منهج جديد للتعلم الآلي " (ML)، يحقق دقة 68٪ و69٪ في تحديد ملامح الصخور في الوقت الفعلي باستخدام 11معلمة حفر .كفل التحضير الشامل للبيانات أداء نموذجي غير متحيز وفعال.

تم تدريب النموذج وتقييمه بعناية، مع التحضير الشامل للبيانات للحد من الميزات ، وموازنة توزيع العينة ، يُعد هذا الاهتمام بجودة البيانات والتحضير أمرًا حاسمًا لأداء النموذج .وضمان مجموعة بيانات غير متحيزة

الفعال
أثبتت دراسة الحالة في الحقل النفطي لحاسي مسعود صحة التطبيق العملي للنموذج. وتوفر هذه المنهجية القائمة على التعلم الآلي مزايا كبيرة للمراقبة الجيولوجية في الوقت الحقيقي، وتحسين عملية صنع القرار وكفاءة عمليات الحفر. يتنبأ النموذج بنجاح بالطبقات الصخرية الرملية الناعمة والحجر الرملي الناعم، وبالنسبة للصلصال، فإن التنبؤات قريبة بشكل معقول من القيم الفعلية. أما بالنسبة للطبقة الملحية، يبالغ النموذج في تقدير وجودها بشكل كبير مقارنةً بالبيانات الفعلية، مما يشير إلى وجود صعوبات في التنبؤ بالطبقات الصخرية "الملحية."

Introduction

Lithology prediction involves determining the types of rocks that will be encountered during drilling, which is essential for various considerations such as distinguishing rock types, preventing formation pressures, optimizing drilling parameters, and improving safety. The relationship between drilling parameters, lithology, and instrument parameters is crucial for optimizing drilling operations and improving the accuracy of subsurface characterization.

Historically, lithology identification relied on direct physical examination of drilling samples, which was time-consuming and limited in spatial coverage. The integration of geophysical logging data, such as gamma ray, resistivity, sonic, density, and neutron logs, emerged as a crucial technique for inferring lithological properties. However, this indirect data required expert interpretation and was not suitable for real-time, accurate, and cost-effective lithology recognition.

Recent advancements in artificial intelligence and machine learning have revolutionized lithology recognition, introducing innovative techniques that address the limitations of traditional methods. Intelligent algorithms, including artificial neural networks (ANN) and deep learning models like Convolutional Neural Networks (CNN), enable rapid and accurate identification of lithologies from rock images, making the process more accessible and less reliant on specialized petrology expertise.

The integration of machine learning tools and drilling parameters offers the potential to transform the way lithology is identified, providing instant insights and enabling more informed decision-making during exploration and production processes. This study aims to explore the frontiers of lithology recognition, leveraging the power of modern machine learning techniques to develop a robust and adaptive framework for accurate and timely lithology identification. By harnessing the wealth of information available from drilling parameters and other relevant data sources, this research seeks to establish a new paradigm in subsurface characterization, ultimately enhancing the efficiency and effectiveness of exploration, extraction, and resource management.

The dissertation is structured around four chapters:

Introduction to Lithology Recognition

- Importance of accurate lithology identification in various industries
- Historical progression of lithology recognition methods
- Limitations of traditional approaches
- Role of artificial intelligence and machine learning
- Emergence of techniques like ANN and CNN
- Benefits of rapid and accurate lithology identification from rock images

- Importance of real-time insights for time-sensitive operations

Study Area: Hassi Messaoud Region

An Overview of the Hassi Messaoud region and its relevance to the research

Materials and Methods

- Data sources: Drilling parameters (MWD) , Chromatographic data
- Data processing techniques
- Machine learning techniques and algorithms employed

Results and Interpretation

- Presentation of the findings from the proposed framework
- Analysis of the accuracy, efficiency, and adaptability of the lithology recognition approach
- Discussion of the implications and potential impact on exploration, extraction, and resource management

Chapter I : Introduction to the Lithology recognition

I. Overview of Drilling Operations

Drilling operations are a critical component of the oil and gas industry, involving the extraction of hydrocarbons from underground reservoirs. The process typically begins with exploration, where geologists and engineers identify potential drilling sites based on geological surveys and seismic data.

Once a site is selected, a drilling rig is set up (fig. 01) , and a well is drilled through the earth's crust to reach the targeted reservoir. The drilling process involves the use of specialized equipment, including drill bits, drill pipes, and drilling fluids, which are designed to manage the pressure and temperature conditions encountered during drilling.

In Algeria, various companies, including national and international oil and gas companies, carry out drilling operations. These operations are subject to regulations and guidelines set by the government and international organizations, such as the International Association of Drilling Contractors (IADC) and the International Labor Organization (ILO).

Conventional drilling is the most common method used in Algeria. It involves pumping a drilling fluid (Mud) , typically a mixture of oil and additives, down the drill pipe to cool and lubricate the drill bit and remove cuttings from the wellbore. This method is widely used due to its effectiveness in managing the pressure and temperature conditions encountered during drilling. (Solutions intégrées de sécurité Algeria)

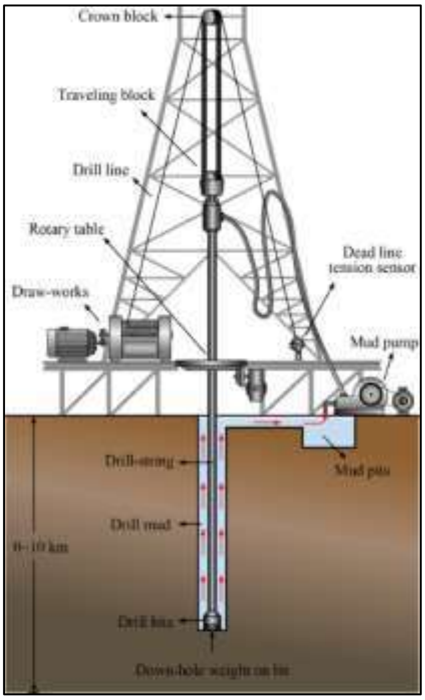


Figure 01 - Drilling rig

II. lithology recognition

Lithology identification is a critical aspect of drilling operations, as it provides crucial information about the geological composition of subsurface formations. Accurate lithology identification helps optimize drilling parameters, improve drilling performance, and reduce costs. Traditional methods of lithology identification rely on physical rock samples obtained during drilling (fig. 02-a,b) , by means of observing the color, structure, and mineral crystallinity, rocks can be classified into several general categories (Kathrada & Adillah , 2019). For further classification, thin sections are observed from such perspectives as the optical characteristics of minerals and crystal shapes (Hu et al., 2010; Chen et al., 2010). As well as geophysical log data or geochemical methods such as gamma ray, resistivity, sonic, density, and neutron logs. However, these methods can be time-consuming (fig. 03) , labor-intensive, and subjective, especially when dealing with large amounts of data (Chopra et al.,2002; Smith et al., 2002; Xu et al., 2021).

In Algeria, the National Wells Services Company has adopted a specialized model called GEOLOG to help with the visualization of lithology in a comprehensive master log. This model relies on the lithological information provided by the mud logger, who analyzes the rock samples obtained during the drilling process.

Figure 02,a - Obtained samples from down hole

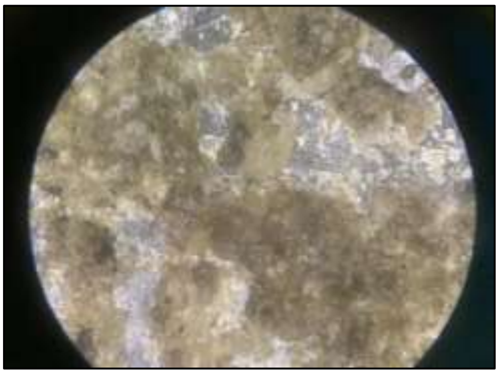


Figure 02 ,b - observation and analysis of the samples

III. Advanced Lithology identification researches

To address the mentioned challenges of time and others , researchers have explored intelligent methods for lithology identification using artificial intelligence and machine learning algorithms. These are some key advancements in lithology identification :

1. Improved Object Detection Algorithms

Single Shot Multi box Detector (SSD):

- Functioning Principle: SSD is an object detection algorithm that uses a single neural network to predict bounding boxes and class probabilities directly from full images. It is designed to handle a wide range of object sizes and aspect ratios.
- Improvements: The SSD algorithm was modified by adding residual networks and adaptive moment estimation to improve its performance. This modification enhanced the accuracy of the algorithm in identifying rock types.
- Results: The modified SSD algorithm achieved average accuracies of 89.4% and 98.4% in identifying 12 rock types in the Xingcheng area of China. (Hou, 2023)

Integration of Remote Sensing and Machine Learning

- Functioning Principle: This method combines remote sensing data from various sources (Landsat, Sentinel-2, ASTER, and Hyperion) with machine learning algorithms (SVM, RF, and ANN) to create detailed and accurate lithological maps.
- Principle: The remote sensing data provides detailed information about the geological composition of the area, which is then used to train machine learning models to identify lithological features (HAMMAD, 2016).
- Results: This method has been shown to be particularly effective in remote or inaccessible areas where traditional methods may be difficult to apply (Xu et al., 2021).

Application of Artificial Intelligence in Petroleum Reservoirs

- Functioning Principle: AI-based methods are used to analyze real-time drilling data and predict lithology types in low permeability reservoirs.
- Principle: The AI algorithms are trained on large datasets of drilling data and are able to identify patterns and relationships that are difficult for humans to recognize.
- Results: AI-based methods have been shown to improve the accuracy and speed of lithology identification compared to traditional techniques like cross plots and mathematical statistics.(Tiantai Li, 2022; Khalifa , 2023)

Deep Learning for Refined Lithology Identification

- Functioning Principle: Deep learning models such as Faster R-CNN and Region Vision Transformer are used to analyze rock images and identify lithology types.
- Principle: These models use convolutional neural networks to extract features from the rock images and then use these features to identify the lithology type.
- Results: These models have been shown to be highly accurate in identifying rock types, including sandstone lithofacies.(Xu et al., 2022)

These advancements in lithology identification using artificial intelligence and machine learning have significantly enhanced the efficiency, accuracy, and accessibility of the process, making it easier to study geological features and support applications in mineral exploration, geotechnical engineering, and petroleum geology.

Here's a detailed look at the relationship between lithology prediction and drilling tools calibration:

IV. Lithology Prediction for Drilling Tools Calibration

Lithology prediction involves determining the types and properties of rocks that will be encountered during drilling operations.

Calibration of drilling tools involves adjusting and fine-tuning the tools and equipment used in drilling operations to ensure they function correctly under the predicted conditions.

Lithology prediction and the calibration of drilling tools are closely connected in the context of oil and gas exploration, mining, and other subsurface investigations. Understanding this relationship is crucial for optimizing drilling operations, improving efficiency, and ensuring safety. Accurate lithology prediction helps in:

- **Drill Bit Selection and Optimization**: Accurate lithology prediction allows for the selection of the most appropriate drill bits and their calibration to minimize wear and optimize performance. Different lithologies cause different wear patterns, and understanding these can lead to better bit selection and longer bit life.
- **Torque and Drag Management:** Calibrating tools to handle expected resistances from different rock types.
- **Drilling Efficiency:** Knowing the lithology helps in calibrating the drilling parameters such as weight on bit, rotational speed, and drilling fluid properties. This ensures efficient penetration rates and reduces non-productive time.
- **Tool Life and Maintenance:** Predicting the types of lithologies and their abrasiveness can help in planning maintenance schedules and the calibration of

tools to prolong their life

- **Mud System Calibration:** Adjusting the properties of drilling mud to optimize hole cleaning and pressure control based on lithology.
- **Measurement While Drilling (MWD) and Logging While Drilling (LWD) Tools:** During drilling, real-time lithology prediction (through MWD/LWD tools) helps in making immediate adjustments to the drilling tools and parameters. This continuous calibration process enhances drilling efficiency and safety.
- **Data Integration:** Combining lithology prediction data with calibration parameters enhances the ability to predict tool performance and necessary adjustments, creating a feedback loop that improves future predictions and calibrations.
- **Formation Damage Prevention:** Proper calibration based on lithology can prevent issues such as formation damage or wellbore instability, which can arise from inappropriate drilling parameters or tools.

.

Chapter II : Study Area- Hassi Messaoud Region

V. Historical Background and geological situation of Hassi Massaoud rigion

Historical Background of the Hassi-Messaoud Field

Several wells have been drilled in the Hassi Messaoud region. The first well was drilled in 1956, to a depth of 3338 meters. A second well was drilled in 1957 to complete the first reservoir, which was put into production at the end of 1958. Subsequently, 20 more wells were drilled, allowing for a production of approximately 0.5 million tons. By the end of 1964, 153 wells were in production, with the start-up of the first two high-pressure gas re-injection stations. Seven injection wells were then equipped. Between 1963 and 1967, an average of 8 wells per year were drilled. By the end of 1975, 262 wells had been drilled, of which 222 were operational. In 2000, the number of wells drilled reached 1096, including 783 oil producers, 100 gas injection wells, and 37 water injection wells. The field also had 129 dry wells and 47 water-producing wells. Subsequently, several studies have been conducted to better understand the structure of the Hassi-Messaoud field (Djemili, & Boublal, 2016).

Geological structure of the Hassi-Messaoud Field

The Hassi-Messaoud field represents the largest oil deposit in Algeria. It is located in the northeastern part of the Saharan platform and corresponds to the northern extension of the Amghid el biod ridge (figure 03). It is a flattened, broad, oval anticline trending north- northeast to south- southwest (Djemili, & Boublal, 2016), parallel to the major fault zone (fig. 3b). It covers almost 2,000 Km^2 in the Oued Mya basin (Moussous, 2008; Loukil, 2016).

The field is bounded:

- To the north, by the Djamaa-Touggourt structure
- To the west, by the Oued Mia depression
- To the east, by the Dahar depression
- To the south, by the Amguid depression

In geographical coordinates, the field is situated between:

- 32°15' north latitude to the north
- 31°30' north latitude to the south
- 5°40' east longitude to the west

In Lambert coordinates, the field extends between:

- 790,000 and 840,000 m east for the X coordinate
- 110,000 and 150,000 m north for the Y coordinate

Figure 03 – a. geographical location of hassi messaoud region

Figure 03 – b. geological context of hassi messaoud region (Wec Algeria, 2007)

Zonatation of Hassi Massaoud field

The Hassi-Messaoud oil field is divided into numbered zones, which are naturally deduced from the production characteristics and the geology of the reservoir (Hamzioui, 2016). This subdivision is based on the evolution of well pressures as a function of production, allowing for the identification of 25 distinct producing zones (fig. 04)

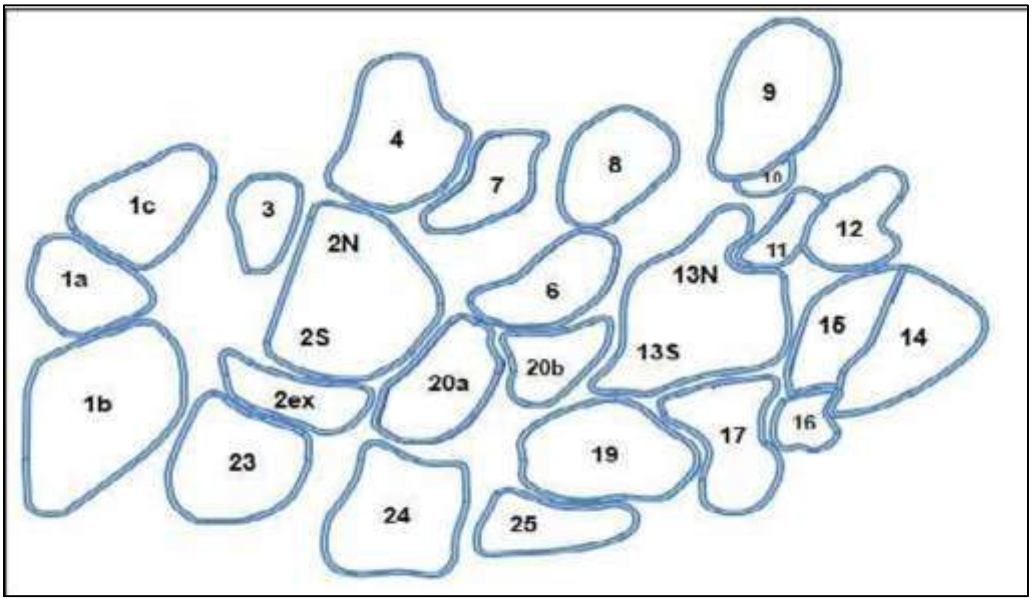


Figure 04- Zonatation of Hassi Massaoud field

VI. Internship at the National Company for Well Servicing (ENSP)- Insights into Drilling Operations and Lithology Recognition in the Hassi Messaoud Region

To gain a deeper understanding of the drilling operations and lithology recognition processes in this area, we conducted an internship at the National Company for Well Servicing (ENSP) from May 20 th, 2024, to May 31st, 2024.

During this internship, we had the opportunity to visit several key service areas within the ENSP, including the Mud Logging Cabin, the Control Quality Service, and the Well Site W1. This hands-on experience provided invaluable insights into the real-world practices and challenges associated with lithology recognition in the Hassi Messaoud region.

Well Site W1 :
At the Well Site W1 , we observed the drilling operations and the collection of drilling parameters (MWD) and chromatographic data. We witnessed the continuous monitoring of various parameters, such as mud flow rate, drill bit weight, and rotational speed, which are crucial for identifying changes in the subsurface lithology in real-time. Additionally, we learned about the process of lithology identification by collecting rock samples from the well and analyzing their physical and mineralogical properties. This firsthand experience highlighted the importance of integrating these drilling parameters with the chromatographic and rock sample data for effective lithology recognition.

Mud Logging Cabin:
The Mud Logging Cabin served as a central hub for data acquisition and analysis. Here, we gained insights into the processes of data collection, handling, and preliminary interpretation. We observed the real-time monitoring of drilling parameters, the integration of chromatographic data, and the use of specialized software for initial lithology identification (figure 05) . This experience highlighted the challenges and complexities involved in translating the raw data into meaningful geological insights.

Figure 05 - Drilling parameters

Control Quality Service:
The Control Quality Service department was responsible for ensuring the reliability and accuracy of the data collected during the drilling operations. We learned about the various quality control measures, such as instrument calibration, data validation, and error detection, that are essential for maintaining the integrity of the dataset used in the lithology recognition process.

The knowledge and practical experience gained through this internship at the ENSP have significantly enriched our understanding of the drilling operations and lithology recognition challenges in the Hassi Messaoud region. These insights have directly informed the design and development of the machine learning-based lithology recognition framework presented in this study.

Chapter III : Materials and methods

VII. Exploratory Data Analysis

Effective lithology recognition in the Hassi Messaoud region requires a thorough understanding of the available data sources and their characteristics. Before applying any machine learning techniques, it is essential to conduct a comprehensive exploratory data analysis (EDA) to gain insights into the data, identify potential patterns, and uncover any underlying relationships or anomalies. This preliminary step is crucial for informing the selection and optimization of the most appropriate modeling approaches.

In this section, we present the EDA conducted on the drilling parameters (MWD) and chromatographic data collected from the Hassi Messaoud region.

1. Drilling parameters (MWD) :

Measurement While Drilling (MWD) sensors play a pivotal role in drilling operations by providing essential drilling parameters and formation evaluation data in real-time. Unlike Logging While Drilling (LWD) sensors, which primarily focus on geological formations, MWD sensors emphasize drilling dynamics and directional control. However, they also contribute significantly to lithology recognition by offering:

Formation Pressure Measurements: MWD sensors monitor formation pressures, enabling drillers to anticipate and mitigate potential well bore stability issues and fluid influxes.

Drill Bit Performance Data: By monitoring drill bit performance metrics such as weight-on-bit and torque, MWD sensors help assess drilling efficiency and identify drilling challenges in real-time.

Downhole Temperature and Pressure Measurements: These measurements provide insights into downhole conditions, including temperature gradients and pressure differentials, aiding in the interpretation of lithological boundaries and formation properties.

MWD sensors are integral to steering the drill bit accurately through the formation, optimizing drilling performance, and ensuring precise wellbore placement. Their contribution to lithology interpretation enhances overall drilling efficiency and facilitates informed decision-making during drilling operations.

a) - Weight on Bit (WOB):

Weight on Bit (WOB) is the weight applied by the bottom hole assembly (BHA) on the drill bit.

b) - Rotation Speed or RPM (Revolutions Per Minute):

Rotation speed is another variable that can influence the Rate of Penetration (ROP). The total tool rotation speed is equal to the surface rotation speed plus the downhole motor/turbine rotation speed. It is varied depending on the tool used and the geological formations encountered to reduce lateral vibrations and avoid triggering the drillstring resonance.

c) - Rate of Penetration (ROP):

Rate of Penetration (ROP) represents the speed at which the drilling tool advances into the rock. It is measured in meters per hour and is influenced by various factors.

d)- Torque:

Torque is an indicator of what is happening at the drilling tool level. In soft formations, torque can indicate that the tool is in contact with the bottom before the WOB. If downhole torque measurements are available, they can be used in combination with surface measurements for greater accuracy.

e) - Mud Flow in :

the flow rate of the drilling mud as it enters the wellbore during the drilling process.

f) - Stand pipe Pressure (SPP):

Standpipe pressure represents the sum of pressure losses generated by the mud flow through the drill string, passing through the tool, and returning through the annular space. It is measured at the surface through the standpipe and provides various information about well conditions, the resistance and hardness of the geological formation being drilled, as well as the quality of the drilling mud used.

G) - Drilling Depth:

Drilling depth is an important parameter that can influence the Rate of Penetration (ROP) during the drilling operation.

H)- Bit Drilling Time

The amount of time the drill bit is actively engaged in penetrating and drilling through the formation.

2. Chromatographic data (TOTAL Gas) :

The total gas detected during the drilling process refers to the measurement and analysis of the various gas components present in the drilling mud. This includes the detection and quantification of both hydrocarbon gases, such as methane, ethane, propane, and others, as well as non-hydrocarbon gases like carbon dioxide, hydrogen sulfide, nitrogen, and helium.

Several analytical techniques are employed to detect and characterize the total gas in the drilling fluid, including gas chromatography (GC), mass spectrometry (MS), Fourier transform infrared spectroscopy (FTIR), and laser absorption spectroscopy (LAS). These methods allow for the identification and precise measurement of the individual gas components, providing detailed information about the composition of the gas mixture in the drilling mud.

The analysis of the total gas detected during drilling is an important tool for understanding the subsurface conditions and potential hydrocarbon reservoirs encountered while drilling a well. The gas composition data can help geologists and engineers make informed decisions about the drilling operations and the potential for hydrocarbon production.[Mostafa Raouf ,Apr17,2023],[J. ERZINGER, T. WIERSBERG AND M. ZIMMER,June 2006],[Haibo Liang; Haifeng Chen; Jinhong Guo; Xing Zuo, 2019]

3. Data collection and description :

The analysis revolves around data from wells originating from the Hassi Massoud region, comprising 12 columns - 2 columns of depth, 8 columns of drilling parameters, Total Gas, and Lithology (the output variable) sourced from mud logging data. The dataset contains 4,492 rows without any missing values. The investigation of the Lithology column reveals a diverse array of 15 lithologies: Anhydrite, Argile, Argile silteuse, Calcaire, Calcaire dolomitique, Dolomite, Dolomite et calcaire crayeux, Dolomite et calcaire argileux, Dolomite et calcaire vacuolaire, Grès fins, Gypse, Sable fin, Sable moyen, Marne, and Sel. The study focuses on all these lithologies and aims to address the imbalanced class distribution using appropriate techniques, such as oversampling, to derive valuable insights from the drilling data.

4. Data Imputation (oversampling) :

Imputation of missing values is a procedure that replaces missing values with plausible estimates to enable accurate estimation of population parameters and maintain the power of data mining and analysis techniques. The optimal approach depends on the amount of missing data, and it is generally recommended to compare results before and after imputation if more than 25% of the data is missing .(Rubin ,1976)

In the context of this study, we encountered a significant amount of missing data in the total gas field. To address this challenge, we implemented an imputation procedure that increased the available data from 1,085 to 4,492 data points. This allowed us to better capture the underlying trends and patterns in the total gas variable, which is a crucial input for the lithology recognition framework developed in this study. By imputing the missing values, we were able to maintain the statistical power of our analysis and ensure more reliable estimates of the population parameters.

#	Column	Non-Null Count	Dtype
0	Depth: TMD[m]	4491 non-null	int64
1	Depth: TVD[m]	4491 non-null	float64
2	ROP[m/h]	4491 non-null	float64
3	ROP Inverse [min/m]	4491 non-null	float64
4	Bit Drilling Time[h]	4491 non-null	float64
5	RPM Avg[rpm]	4491 non-null	int64
6	WOB Avg[t]	4491 non-null	int64
7	SPP Avg[psi]	4491 non-null	int64
8	Torque Avg[lb*ft]	4491 non-null	int64
9	Flow IN Avg[l/min]	4491 non-null	int64
10	TGAS[%]	1085 non-null	float64
11	Lithology	4491 non-null	object

Figure 06,a - data infos before imputation

#	Column	Non-Null Count	Dtype
0	Depth: TMD[m]	4491 non-null	float64
1	Depth: TVD[m]	4491 non-null	float64
2	ROP[m/h]	4491 non-null	float64
3	ROP Inverse [min/m]	4491 non-null	float64
4	Bit Drilling Time[h]	4491 non-null	float64
5	RPM Avg[rpm]	4491 non-null	float64
6	WOB Avg[t]	4491 non-null	float64
7	SPP Avg[psi]	4491 non-null	float64
8	Torque Avg[lb*ft]	4491 non-null	float64
9	Flow IN Avg[l/min]	4491 non-null	float64
10	TGAS[%]	4491 non-null	float64
11	Lithology	4491 non-null	object

Figure 06 ,b- data infos after imputation

5. Data Normalization :

Data normalization is a crucial pre-processing step in machine learning that transforms data into a common format, ensuring that all input parameters are scaled to a common range. It is essential for datasets with different units or magnitudes across different features His function returns the standardized value of a distribution characterized by a mean value and a standard deviation [Goyal & Singh, 2019).

	Depth: TMD[m]	Depth: TVD[m]	ROP[m/h]	ROP Inverse [min/m]	Bit Drilling Time[h]	RPM Avg[rpm]	WOB Avg[t]	SPP Avg[psi]	Torque Avg[lb*ft]	Flow IN Avg[l/min]	TGAS[%]
0	13	13.00	2.38	25.53	0.51	31	1	5	1073	518	NaN
1	14	14.00	2.70	22.21	1.21	31	1	9	999	520	NaN
2	15	15.00	2.22	27.25	1.40	31	1	16	1340	518	NaN
3	16	16.00	0.96	62.67	1.65	35	1	18	1323	509	NaN
4	17	17.00	1.79	33.59	2.61	43	2	15	1845	488	NaN
...	...	...	...	...	...	...	...	...	...	...	...
4486	4499	4497.69	1.99	30.22	105.15	108	15	2229	9903	1849	0.56
4487	4500	4498.69	2.10	28.56	105.67	108	14	2248	10925	1856	0.60
4488	4501	4499.69	1.95	30.77	106.16	111	12	2242	9796	1855	0.61
4489	4502	4500.69	2.14	28.07	106.61	107	14	2241	11775	1858	0.62
4490	4503	4501.69	2.50	24.05	106.90	111	15	2219	4745	1850	0.64

Figure 07 ,a - Data before the normalization

	Depth: TMD[m]	Depth: TVD[m]	ROP[m/h]	ROP Inverse [min/m]	Bit Drilling Time[h]	RPM Avg[rpm]	WOB Avg[t]	SPP Avg[psi]	Torque Avg[lb*ft]	Flow IN Avg[l/min]	TGAS[%]
count	4491.000000	4491.000000	4491.000000	4491.000000	4491.000000	4491.000000	4491.000000	4491.000000	4491.000000	4491.000000	4491.000000
mean	0.500000	0.500039	0.149271	0.022982	0.221564	0.628772	0.392303	0.638915	0.456602	0.706956	0.011192
std	0.288772	0.288767	0.166136	0.042163	0.174916	0.259924	0.167922	0.176826	0.207966	0.201530	0.016442
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.250000	0.250035	0.022572	0.002362	0.098090	0.468208	0.250000	0.552392	0.299462	0.474069	0.011192
50%	0.500000	0.500070	0.077445	0.008039	0.183171	0.757225	0.416667	0.654328	0.475593	0.742778	0.011192
75%	0.750000	0.749636	0.223454	0.028640	0.291457	0.815029	0.541667	0.705011	0.615762	0.880613	0.011192
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 07 b - Data after the normalization

6. Data visualization :

Through series of statistical analyses, visualizations, and data transformations, we aim to:

- Describe the overall structure and distribution of the variables within the dataset.
- Identify any missing values, outliers, or data quality issues that may require attention.
- Explore the relationships between the drilling parameters, chromatographic data, and the known lithological classifications.
- Assess the suitability of the data for the subsequent machine learning-based lithology recognition tasks.

By thoroughly examining the data characteristics and establishing a solid understanding of the underlying patterns, we can lay the groundwork for the development of a robust and effective lithology identification framework tailored to the Hassi Messaoud region. The insights gained from this exploratory analysis will guide the selection of relevant features, the design of appropriate modeling architectures, and the interpretation of the final results (fig.08).

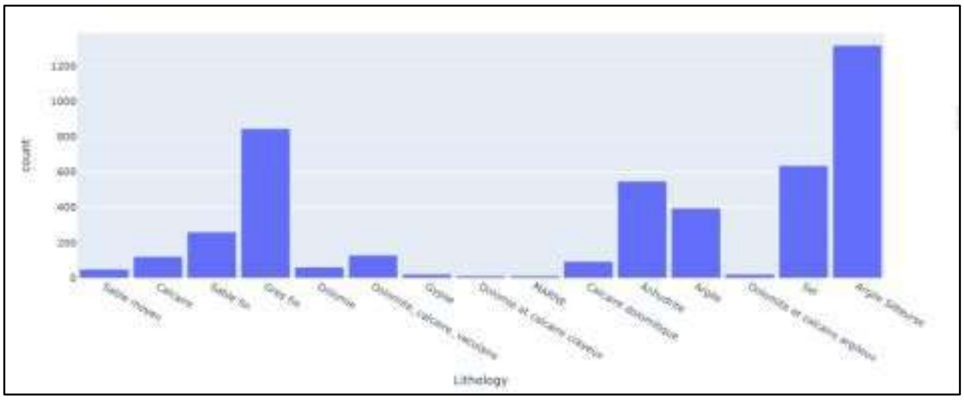


Figure 08 - Data Visualization

VIII. Machine learning tools

Machine learning is a field within artificial intelligence that focuses on developing and studying statistical algorithms capable of learning from data and making predictions without explicit programming instructions. It enables machines to learn from past experiences, identify patterns, and improve incrementally with minimal human intervention. Machine learning algorithms are trained on data sets to create models that can categorize images, analyze data, predict outcomes, and perform various tasks autonomously.

Machine learning approaches are typically categorized into three main paradigms: supervised learning, unsupervised learning, and reinforcement learning, each with its own characteristics and applications. Supervised learning involves learning from labeled data, unsupervised learning discovers patterns in unlabeled data, and reinforcement learning involves learning through interaction with an environment and rewards. These approaches cater to different learning scenarios and tasks, contributing to the versatility of machine learning in various fields such as natural language processing, computer vision, speech recognition, and more.

1. Supervised Learning Algorithms:

Supervised machine learning is a fundamental but strict technology. Operators provide the computer with input examples and the desired outputs, and the computer searches for solutions to obtain those outputs based on the inputs. The goal is for the computer to learn the general rule that maps inputs to outputs.

Supervised machine learning can be used to make predictions on unavailable or future data (referred to as "predictive modeling"). The algorithm tries to develop a function that accurately predicts the output based on the input variables. For example, predicting the value of a real estate property (output) based on inputs such as the number of rooms, year of construction, land area, location, etc. Supervised machine learning can be divided into two types:

- Classification: The output variable is a category.
- Regression: The output variable is a specific value.

The main algorithms used in supervised machine learning are as follows: random forests, decision trees, k-nearest neighbors (k-NN) algorithm, linear regression, support vector machines (SVM), logistic regression (fig. 09, a) and gradient boosting, naive Bayes classification.

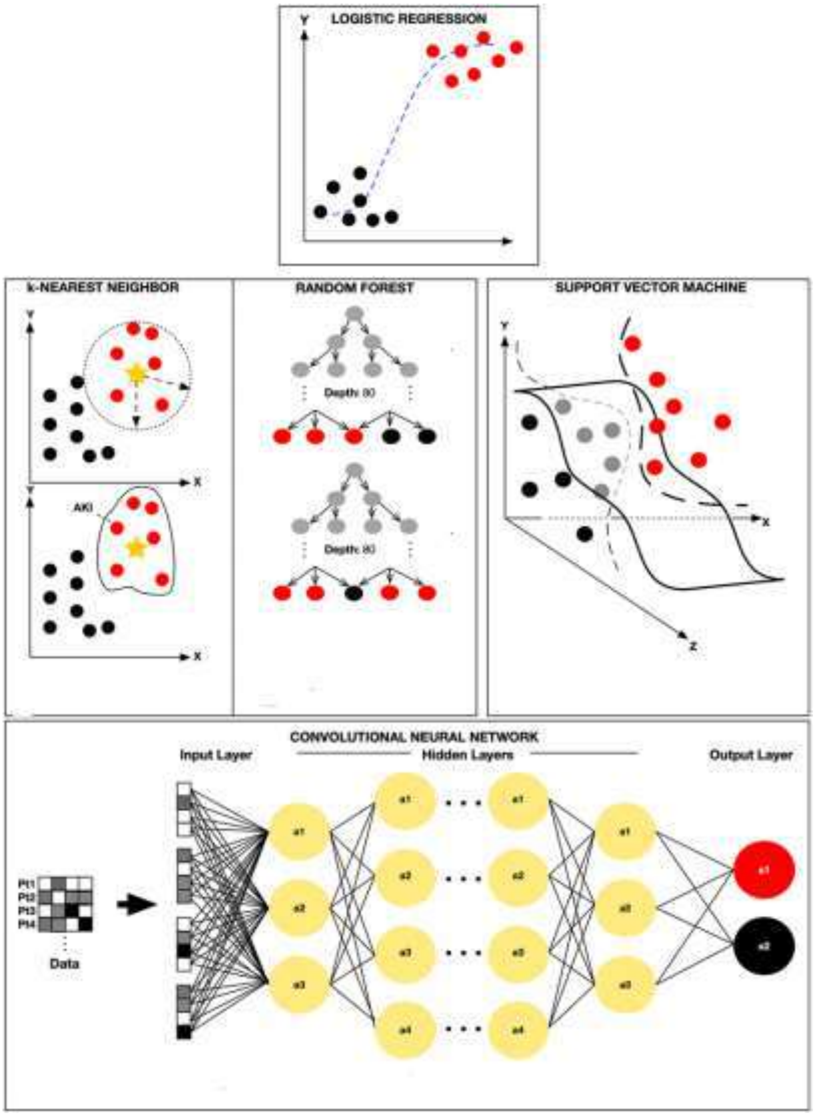


Figure 09,a. Machine Learning techniques (DNN, LR, K-NN, RF, and SVM) illustrated as conceptual drawings (Rashidi et al. 2020)

Supervised learning algorithms are trained using labeled data, where the input features are associated with corresponding target labels. In the context of lithology recognition, supervised learning algorithms can be trained to classify drilling data into different lithological formations based on known examples. Common supervised learning algorithms used for lithology recognition (Figure 09, b) include:

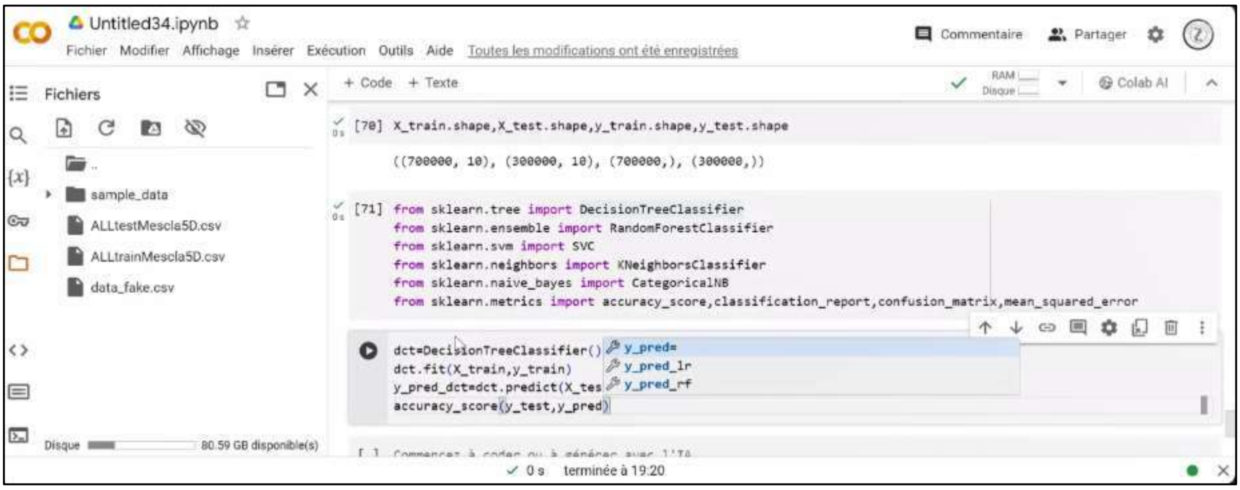


Figure09 b - Supervised learning algorithms

a)- Decision tree Classifier

Decision trees (DTs) are a powerful machine learning tool used for both classification and regression tasks. They construct a tree-like flowchart structure where each internal node represents a test on a feature, branches depict the outcomes of the test, and leaf nodes hold the class labels.

Compared to other algorithms, DTs provide a clear indication of the role and importance of each variable in making predictions (Higa, 2018; Zhao & Zhang, 2008). The DT algorithm works by recursively partitioning the training data into subsets based on feature values until a stopping criterion is met, such as maximum tree depth or minimum samples required to split a node. At each step, it selects the feature that provides the maximum information gain or minimizes impurity after the split, using metrics like entropy or Gini index (Criminisi et al., 2012; Gupta et al., 2017; Witten et al., 2016).

DTs are generally built using a top-down, recursive divide-and-conquer approach (Breiman et al., 2017; Witten et al., 2016). A standard decision tree consists of a root node, branching nodes, and leaf nodes. The algorithm starts by selecting the root node attribute, and then progressively splits the instances into sub samples based on the values of the chosen attributes at each branching node.

DTs offer several advantages, including generating interpretable rules, handling both continuous and categorical variables, providing clarity on important predictive features, ease of use, scalability, missing value tolerance, managing non-linear relationships, and handling imbalanced data. However, they can be prone to over fitting, especially with complex or deep trees, and may not perform as well as other algorithms for continuous target prediction.

b)- Random forest Classifier

The random forest (RF) classifier is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of classification tasks. It works by creating a set of decision trees from randomly selected subsets of the training data and features. Each tree in the forest votes for a class, and the class with the most votes becomes the predicted class for the input data.

Some key characteristics of the RF classifier include:

- It can handle both classification and regression problems.
- It is well-suited for large, complex data sets and high-dimensional feature spaces.
- It provides feature importance scores to understand the significance of different variables.
- It is resistant to over fitting compared to individual decision trees.

The RF algorithm has several advantages:

- It is versatile and can be used for many types of machine learning tasks.
- It is easy to use and requires little data preprocessing.
- It provides a good indicator of feature importance.
- It is hard to beat in terms of performance compared to other algorithms.

However, it also has some limitations:

- It can be computationally expensive due to the construction of multiple trees.
- It may be biased towards the majority class in imbalanced datasets.
- It can be challenging to interpret the reasoning behind individual predictions compared to a single decision tree.

Overall, the RF classifier is a powerful and widely-used machine learning algorithm that combines the strengths of multiple decision trees to deliver accurate and robust predictions. It has been successfully applied in a variety of remote sensing experiments, including farmland mapping, and has achieved efficient classification results in these domains (Chuvieco et al., 2002; Chen et al., 2018; Guru et al., 2017; Hentze et al., 2016; Karnieli et al., 2010).

c)- Support Vecto Machine /Support Vector Classifications :

Support Vector Machines (SVMs) and Support Vector Classifiers (SVCs) are powerful machine learning algorithms primarily used for classification tasks. SVM is a versatile method that can handle both classification and regression problems by constructing an optimal hyperplane to separate different classes in a multidimensional feature space.

On the other hand, SVC is a specific implementation of the SVM algorithm, designed exclusively for classification problems. The goal of SVC is to find the best hyperplane that can effectively classify data points into distinct classes.

Support Vector Machines (SVM)
SVMs are efficient in high-dimensional spaces and are widely used for classification tasks. They aim to find a maximum-margin hyperplane that separates the classes with the largest possible distance. The concepts of support vectors, hyperplanes, and margins are essential in understanding SVMs. SVMs can handle both linear and non-linear classification problems by using kernel functions. They are robust to outliers and can effectively handle datasets with a large number of features.

Support Vector Classification (SVC)
SVC is a classification-specific variant of the SVM algorithm, focused on categorizing data points into different classes. It is implemented in the sci kit-learn library as sklearn.svm.SVC and supports multiclass classification. SVC is mathematically grounded in linear algebra and optimization, with the goal of finding the optimal values for defining the separating hyper plane. Similar to the general SVM algorithm, SVC can handle both linear and non-linear classification problems by using kernel functions. SVC provides a flexible and powerful tool for solving a wide range of classification problems, especially when dealing with high-dimensional data.

d)- The K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a non-parametric, supervised learning technique used for both classification and regression problems in machine learning. It operates based on the principle of proximity, where the algorithm makes predictions by assessing the similarity between data points.

KNN is a straightforward yet powerful algorithm that classifies data points based on a majority vote of their nearest neighbors. For classification tasks, KNN uses this majority voting mechanism to assign class labels, with the choice of the k value determining the number of neighbors considered.

KNN is a lazy learning algorithm, meaning it stores the training data and performs the computational work during the prediction stage. This approach makes KNN memory-intensive, as it needs to retain the entire training dataset for making predictions.

The selection of the optimal k value is crucial, as it can impact the model's bias-variance trade-off. Higher k values tend to reduce the model's variance but increase the bias, while lower k values have the opposite effect. Cross-validation techniques are often employed to determine the best k value that provides the desired balance between performance and generalization.

2. Common Matrix

During the assessment of classification models, several common metrics play a vital role, including accuracy, precision, recall, and the F1 score. These metrics offer valuable insights into the model's performance and effectiveness. The equations and their respective

definitions are as follows:

- Accuracy: It is the ratio of correctly predicted observations to the total number of observations. This metric is useful when the classes of the target variable are nearly balanced.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{A1}$$

- Precision: It is the ratio of correctly predicted positive observations to the total number of predicted positive observations.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{A2}$$

- Recall: Recall is the ratio of correctly predicted positive observations to the total number of actual positive observations.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{A3}$$

- F1 Score: The F1 score is the harmonic mean of precision and recall, aiming to find a balance between precision and recall.

$$\text{F1Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{A4}$$

where

- TP: correctly predicted positive observations.
- TN: correctly predicted negative observations.
- FP: incorrectly predicted positive observations.
- FN: incorrectly predicted negative observations.

Chapter IV : Results and discussion

IX. Data correlation

Present the correlation matrix to show relationships between numerical features for our study this our correlation heat map (matrix) (fig. 10).

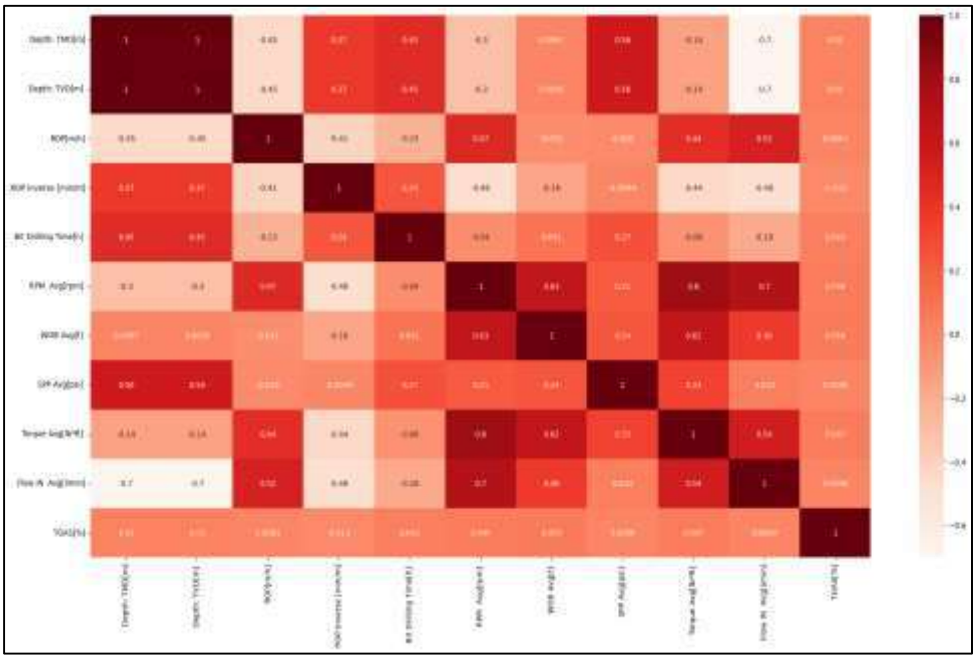


Figure 10- correlation heat map

- The diagonal elements represent the perfect correlation (value of 1) between each parameter and itself.
- The color scheme indicates the strength and direction of the correlations, with red representing strong positive correlations, blue representing strong negative correlations, and lighter colors indicating weaker correlations.
- The parameters that appear to have the strongest positive correlations are Depth: TMD and Depth: TVD, indicating they are closely related.
- There are also strong positive correlations between ROP Inverse and Bit Drilling Time, as well as between RPM Avg and WOB Avg, suggesting these parameters are likely interdependent.
- Some parameters, such as TGAS and the various drilling performance metrics, show relatively weak correlations with the other variables, indicating they may be more independent.

Based on the analysis of the correlation heat map, we can draw the following key results:

Interdependent parameters:

- Depth: TMD and Depth: TVD are highly correlated, indicating they are closely linked and interdependent.

- ROP Inverse and Bit Drilling Time also show a strong positive correlation, suggesting these parameters are interconnected.
- RPM Avg and WOB Avg exhibit a strong positive correlation, implying they are likely correlated in their influence on the drilling process.

Independent parameters:

- TGAS appears to have relatively weak correlations with the other variables, indicating it may be more independent and influenced by different factors.
- Some of the other drilling performance metrics, such as SPP Avg and Torque Avg, also display relatively weaker correlations, suggesting they may be more independent.

X. Models performances

The selection of appropriate modeling techniques was guided by the unique characteristics of the dataset, including the nature and distribution of the input variables, as well as the presence of missing data. We carefully considered the trade-offs between model complexity, interpret ability, and predictive accuracy to identify the most suitable approaches for this application.

In the following sections, we present a detailed analysis of the models' performances, highlighting their strengths, weaknesses, and the specific scenarios in which they excel.

1. Random forest classifier

```
The accuracy classification of random forest classifier 0.680756399
The confusion matrix classification of random forest classifier
 [[ 68  15   8   3   1   2   0   0   0   0   0   0   0   0   6]
 [ 10  56   0   2   0   1   0   1   0   2   0   0   0   0   6]
 [ 12   0 179   1   0   0   0   0   0  48   0   0   5   0  13]
 [  2   4   5   8   0   0   0   0   0   0   0   0   1   1   2]
 [  4   0   0   0  17   0   0   0   0   0   0   0   0   0   0]
 [  0   2   0   0   0   7   0   0   0   0   0   0   0   0   0]
 [  1   0   0   0   0   0   0   0   0   0   0   3   0   0   0]
 [  0   0   2   0   0   0   0   2   0   1   0   0   0   0   0]
 [  5   2   0   0   0   0   0   0  21   0   1   2   0   0   0]
 [  1  11  60   0   0   0   0   0   0  93   0   0   2   0   0]
 [  0   0   0   0   1   0   0   0   3   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   1   0   0   0   0   0   1   6   0   0  48   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   9   0]
 [ 14   6   6   0   1   0   0   0   0   0   0   0   0   0 104]]
```

```
The classification report classification of random forest classifier
                                 precision    recall  f1-score   support

                      Anhydrite       0.58      0.66      0.62       103
                         Argile       0.58      0.72      0.64        78
              Argile Silteurse       0.69      0.69      0.69       258
                       Calcaire       0.57      0.35      0.43        23
          Calcaire dolomitique       0.85      0.81      0.83        21
                        Dolomie       0.70      0.78      0.74         9
   Dolomie et calcaire crayeux       0.00      0.00      0.00         4
Dolomite et calcaire argileux       0.67      0.40      0.50         5
  Dolomite_calcaire_vaculaire       0.84      0.68      0.75        31
                       Gres fin       0.62      0.56      0.59       167
                          Gypse       0.00      0.00      0.00         4
                          MARNE       0.00      0.00      0.00         8
                     Sable fin       0.86      0.86      0.86        56
                   Sable moyen       0.90      1.00      0.95         9
                            Sel       0.79      0.79      0.79       131

                       accuracy                           0.68       899
                      macro avg       0.58      0.55      0.56       899
                   weighted avg       0.68      0.68      0.68       899
```

2. SVM Classifier :

```
The accuracy classification of SVM classifier 0.6318131256952169
The confusion matrix classification of SVM classifier
 [[ 62  12  12   0   0   0   0   0   3   0   0   0   0  14]
 [ 12  48   0   0   0   0   0   0   1   6   0   0   0  11]
 [  7   0 211   0   0   1   0   0   0  27   0   0   0  12]
 [  8   4   5   0   0   0   0   0   0   0   0   1   1   4]
 [  3   0   4   0   0   0   0   0  11   0   0   0   0   3]
 [  1   1   0   0   0   7   0   0   0   0   0   0   0   0]
 [  1   0   0   0   0   0   0   0   3   0   0   0   0   0]
 [  2   0   3   0   0   0   0   0   0   0   0   0   0   0]
 [  5   3   0   0   0   0   0   0  23   0   0   0   0   0]
 [  0  16  86   0   0   1   0   0   0  64   0   0   0   0]
 [  0   0   0   0   0   0   0   0   4   0   0   0   0   0]
 [  0   1   1   0   0   0   0   0   1  20   0  33   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   1   8   0]
 [ 16   3   0   0   0   0   0   0   0   0   0   0   0 112]]
```

3. Decision tree Classifier :

```
The accuracy classification of decision tree classifier 0.6162402669632926
The confusion matrix classification of decision tree classifier
 [[ 63  10  12   5   0   2   0   0   3   0   0   0   0   0   8]
 [ 15  50   1   2   0   0   0   1   0   5   0   0   1   0   3]
 [ 18   0 162   2   2   1   0   1   0  62   0   0   4   0   6]
 [  4   3   2   9   0   0   0   0   0   1   0   0   1   1   2]
 [  4   0   0   0  15   0   0   0   1   0   0   1   0   0   0]
 [  0   1   0   0   0   7   0   0   0   1   0   0   0   0   0]
 [  1   0   0   0   0   0   1   0   0   0   1   1   0   0   0]
 [  0   0   3   0   0   0   0   2   0   0   0   0   0   0   0]
 [  1   2   0   0   0   0   0   0  22   0   4   2   0   0   0]
 [  1   8  65   2   0   3   0   0   0  87   0   0   1   0   0]
 [  0   0   0   0   1   0   0   0   3   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   1   5   0   0   0   0   0   1   5   0   0  44   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   9   0]
 [ 16  11  20   1   0   0   0   0   0   0   0   0   0   0  83]]
```

```
The classification report classification of decision tree classifier
                              precision    recall  f1-score   support

                   Anhydrite       0.51      0.61      0.56       103
                      Argile       0.58      0.64      0.61        78
            Argile Silteurse       0.60      0.63      0.61       258
                    Calcaire       0.43      0.39      0.41        23
        Calcaire dolomitique       0.83      0.71      0.77        21
                     Dolomie       0.54      0.78      0.64         9
  Dolomie et calcaire crayeux       1.00      0.25      0.40         4
Dolomite et calcaire argileux       0.50      0.40      0.44         5
  Dolomite_calcaire_vaculaire       0.73      0.71      0.72        31
                    Gres fin       0.54      0.52      0.53       167
                       Gypse       0.00      0.00      0.00         4
                       MARNE       0.00      0.00      0.00         0
                   Sable fin       0.86      0.79      0.82        56
                 Sable moyen       0.90      1.00      0.95         9
                         Sel       0.81      0.63      0.71       131

                    accuracy                           0.62       899
                   macro avg       0.59      0.54      0.54       899
                weighted avg       0.63      0.62      0.62       899
```

```
The classification report classification of SVM classifier
                              precision    recall  f1-score   support

                   Anhydrite       0.53      0.60      0.56       103
                      Argile       0.55      0.62      0.58        78
            Argile Silteurse       0.66      0.82      0.73       258
                    Calcaire       0.00      0.00      0.00        23
        Calcaire dolomitique       0.00      0.00      0.00        21
                     Dolomie       0.78      0.78      0.78         9
  Dolomie et calcaire crayeux       0.00      0.00      0.00         4
Dolomite et calcaire argileux       0.00      0.00      0.00         5
  Dolomite_calcaire_vaculaire       0.50      0.74      0.60        31
                    Gres fin       0.55      0.38      0.45       167
                       Gypse       0.00      0.00      0.00         4
                   Sable fin       0.94      0.59      0.73        56
                 Sable moyen       0.89      0.89      0.89         9
                         Sel       0.72      0.85      0.78       131

                    accuracy                           0.63       899
                   macro avg       0.44      0.45      0.44       899
                weighted avg       0.59      0.63      0.60       899
```

4. K Nearest Neignbors Classifier :

```
The accuracy classification of KNeighbors classifier 0.6440489432703004
The confusion matrix classification of KNeighbors classifier
[[ 64  13  10   2   2   2   0   0   0   1   0   0   0   0   9]
 [ 15  52   0   1   0   0   0   0   0   3   0   0   1   0   6]
 [ 14   0 176   0   1   1   0   0   0  49   0   0   5   0  12]
 [  3   4   5   8   0   0   0   0   0   0   0   0   1   1   1]
 [  2   0   2   0  13   0   0   0   2   0   0   0   0   0   2]
 [  0   1   0   0   0   7   0   0   0   1   0   0   0   0   0]
 [  1   0   0   0   3   0   0   0   0   0   0   0   0   0   0]
 [  1   0   3   0   0   0   0   1   0   0   0   0   0   0   0]
 [  7   2   0   0   2   0   0   0  19   0   0   1   0   0   0]
 [  1  10  67   0   0   0   0   0   0  84   0   0   5   0   0]
 [  0   0   0   0   2   0   0   0   2   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   1   5   0   0   0   0   0   1   6   0   0  43   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   1   8   0]
 [ 16   6   5   0   0   0   0   0   0   0   0   0   0   0 104]]
```

```
The classification report classification of KNeighbors classifier
                             precision    recall  f1-score   support

                  Anhydrite       0.52      0.62      0.56       103
                     Argile       0.58      0.67      0.62        78
           Argile Silteurse       0.64      0.68      0.66       258
                   Calcaire       0.73      0.35      0.47        23
       Calcaire dolomitique       0.57      0.62      0.59        21
                    Dolomie       0.70      0.78      0.74         9
  Dolomie et calcaire crayeux     0.00      0.00      0.00         4
Dolomite et calcaire argileux     1.00      0.20      0.33         5
  Dolomite_calcaire_vaculaire     0.79      0.61      0.69        31
                   Gres fin       0.58      0.50      0.54       167
                      Gypse       0.00      0.00      0.00         4
                      MARNE       0.00      0.00      0.00         0
                  Sable fin       0.77      0.77      0.77        56
                Sable moyen       0.89      0.89      0.89         9
                        Sel       0.78      0.79      0.78       131

                   accuracy                           0.64       899
                  macro avg       0.57      0.50      0.51       899
               weighted avg       0.64      0.64      0.64       899
```

5. Gaussian NB Classifier :

```
The accuracy classification of GaussianNB classifier 0.3259176863181313
The confusion matrix classification of GaussianNB classifier
[[  1   0   0   1   0  19   0   0   2   0   2   2  17   0  59]
 [  0   3   8   0   0  10   0   2   2   0   0   0  42   0  11]
 [  0   0  83   0   0  60   0  21   0   9   0   0  10   0  75]
 [  0   0   0   0   0  14   0   0   0   0   0   0   3   1   5]
 [  0   0   0   0   0   0   0   0   0   0   0  11   0   0  10]
 [  0   0   0   0   0   8   0   0   0   0   0   0   1   0   0]
 [  0   0   0   0   0   0   0   0   1   0   3   0   0   0   0]
 [  0   0   1   0   0   0   0   2   2   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0  13   0   9   1   8   0   0]
 [  0   2  59   0   0  38   0   2   0  18   0   0  43   0   5]
 [  0   0   0   0   0   0   0   0   1   0   2   1   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   2   0   0   8   0   0   0  46   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   1   8   0]
 [  0   0   0   0   1  15   0   0   0   0   0   0   6   0 109]]
```

```
The classification report classification of GaussianNB classifier
                             precision    recall  f1-score   support

                  Anhydrite       1.00      0.01      0.02       103
                     Argile       0.60      0.04      0.07        78
           Argile Silteurse       0.55      0.32      0.41       258
                   Calcaire       0.00      0.00      0.00        23
       Calcaire dolomitique       0.00      0.00      0.00        21
                    Dolomie       0.05      0.89      0.09         9
  Dolomie et calcaire crayeux     0.00      0.00      0.00         4
Dolomite et calcaire argileux     0.07      0.40      0.12         5
  Dolomite_calcaire_vaculaire     0.45      0.42      0.43        31
                   Gres fin       0.67      0.11      0.19       167
                      Gypse       0.12      0.50      0.20         4
                      MARNE       0.00      0.00      0.00         0
                  Sable fin       0.26      0.82      0.39        56
                Sable moyen       0.89      0.89      0.89         9
                        Sel       0.40      0.83      0.54       131

                   accuracy                           0.33       899
                  macro avg       0.34      0.35      0.22       899
               weighted avg       0.55      0.33      0.29       899
```

XI. Models performance comparison

The evaluation of the machine-learning models was performed, and the confusion matrices for each model were presented in the previous figures. These confusion matrices highlight the superior predictive capability of the random forest model (tab. 01, Fig. 21).

All the models were able to perfectly identify the 'sable moyen' class. However, the random forest model excelled in accurately predicting both the 'sable moyen' and 'sable fin' classes.

In contrast, the Gaussian NB model faced challenges in accurately classifying the 'sable fin' class, as evidenced by the confusion matrix (Fig. 20). This suggests that the random forest model outperformed the Gaussian NB model in discriminating between the different sand types.

Table 01 - Models Performance comparison

	The Model	Its accuracy
0	Random Forest	0.680756
1	Decision Tree	0.616240
2	SVM	0.631813
3	K Neighbors	0.644049
4	Gaussian NB	0.325918

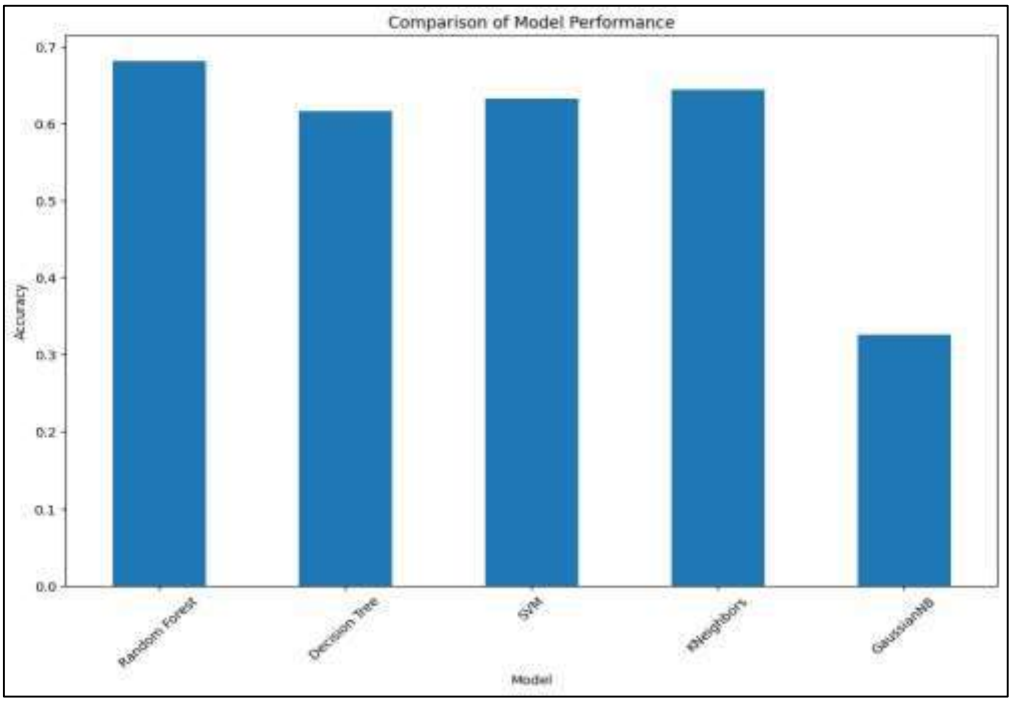


Figure 21 - Histogram of Models Performance comparison

The random forest classifier emerged as the top-performing model with anaccuracy of 68,07%, whereas the Gaussian NB model registered the lowest accuracy of 32% .

Figure 22 shows the precision of models in order by Lithology Classification Probability Distributions.

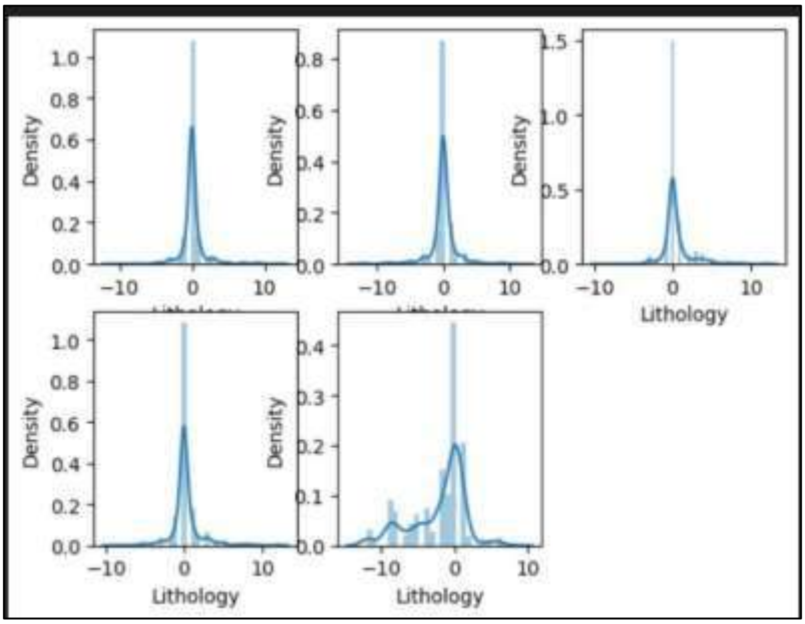


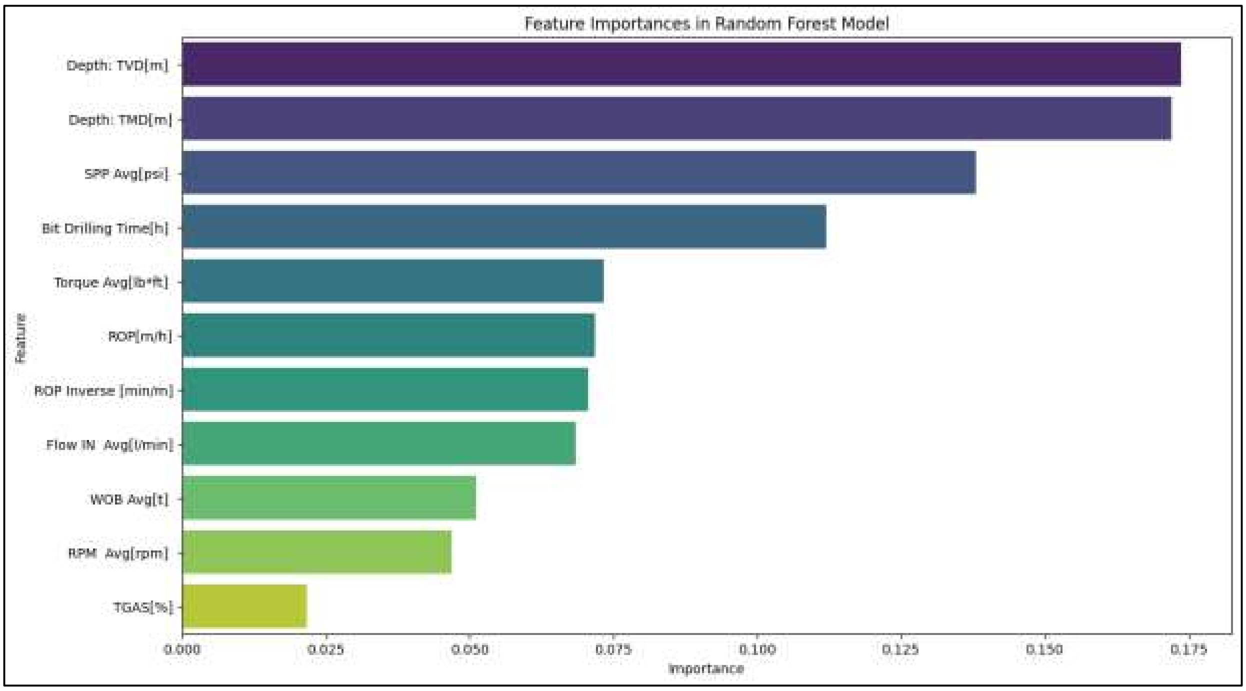
Figure 22 - Lithology Classification Probability Distributions (by order)

XII. Features importance in the model

The feature importance analysis of the Random Forest model reveals the key factors that drive its predictive capabilities (Fig. 23). At the top of the list is the true vertical depth (TVD) of the well, indicating that this depth-related parameter is the most significant input for the model's predictions. Closely following is the measured depth (TMD), further emphasizing the critical role of depth characteristics in the model's decision-making.

The third most important feature is the average standpipe pressure (SPP Avg), suggesting that this drilling parameter is also a crucial input for the model. The next two features of importance are the bit drilling time and the average torque, highlighting the relevance of drilling dynamics and operational factors in the model's performance.

The remaining features, such as rate of penetration (ROP), inverse ROP, flow rate, weight on bit, and rotational speed, have progressively lower importance scores. This implies that while these parameters do contribute to the model's predictions, they play a relatively less significant role compared to the top-ranked depth, pressure, and drilling-related inputs



Figue 23 - Histogram of Features importance

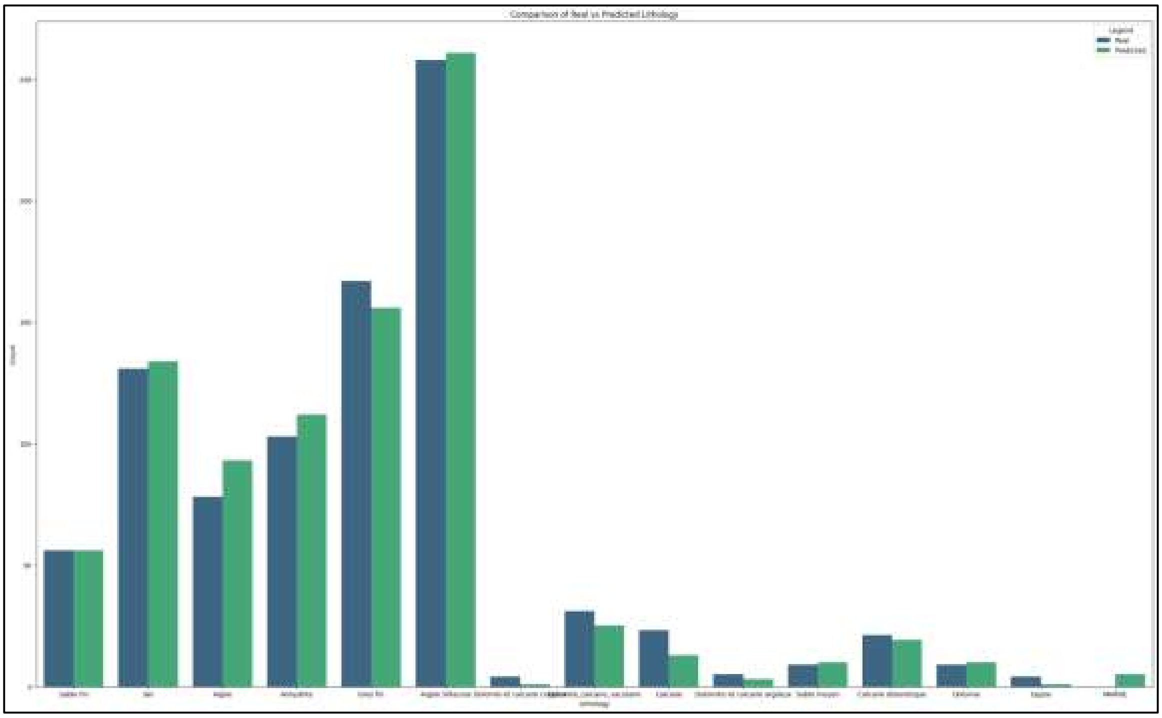
XIII. Results Sammury

The comparison between the real and predicted lithologies (Fig. 24, 25 and 26) reveals several key findings about the model's performance. For the "sable fin" (fine sand) and "gres fin" (fine sandstone) lithologies, the real and predicted values closely align, indicating the model accurately captures these rock types. However, for "sel" (salt), the model significantly overestimates the presence compared to the real data, suggesting challenges in predicting the salt lithology.

The model's predictions for "argile" (clay) are reasonably close to the real values, though still somewhat higher, pointing to room for improvement in this area. Conversely, larger discrepancies exist between real and predicted values for other lithologies, such as "Angle Structure calcaire" and "Calcaire oolitique", highlighting the need to enhance the model's predictive capabilities for those specific rock types.

Overall, the variability in the model's performance across the different lithologies suggests that further refinement and optimization of the model would be beneficial to improve its consistency and accuracy in predicting a wider range of geological characteristics. This analysis provides valuable insights that can guide future model development efforts and help target areas for model enhancement to better align the predicted lithologies with the observed real-world data.

Figure 24- Comparison of Real vs Predicted Lithology (lithology in function of counts)



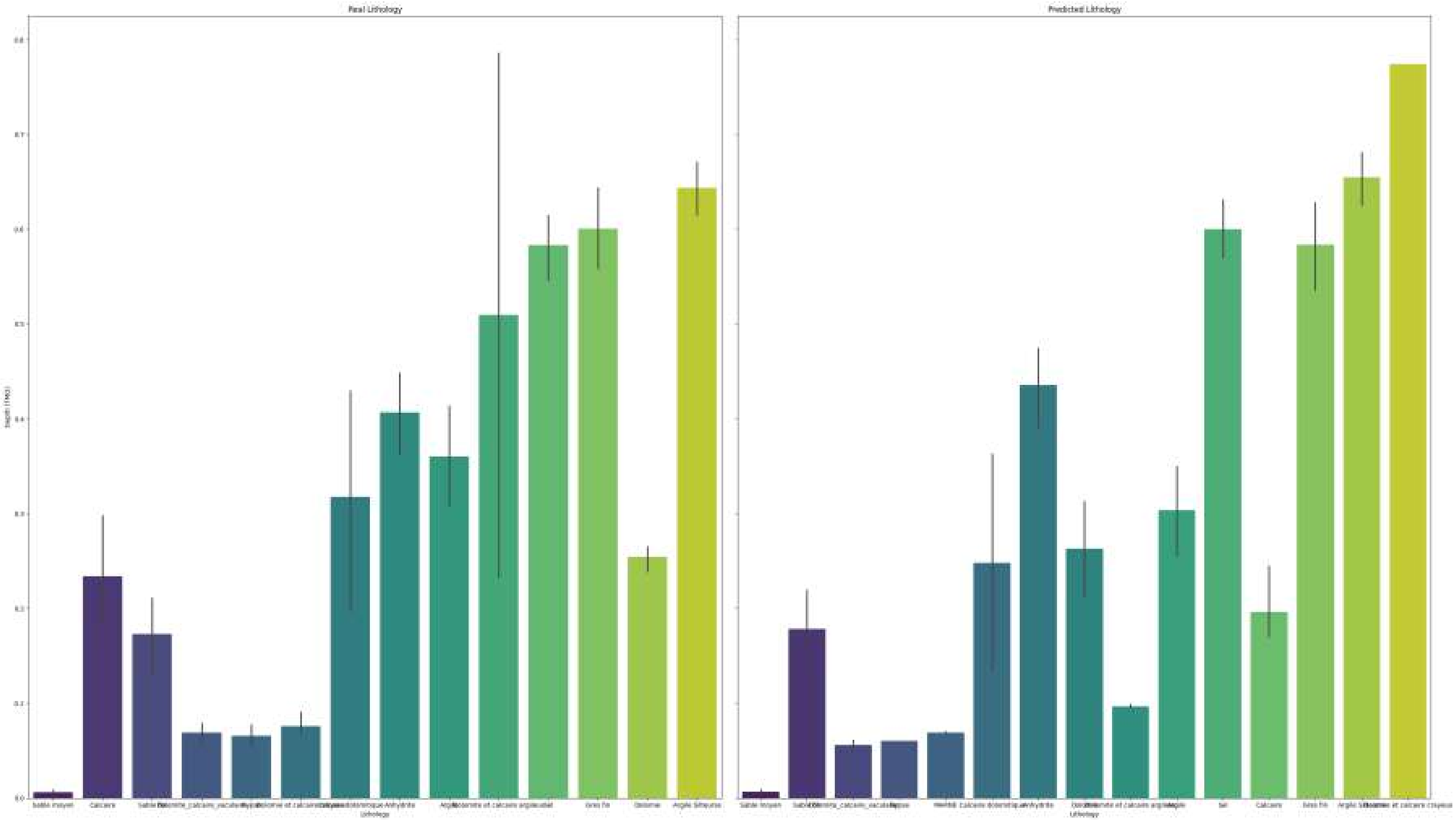


Figure 25 - Comparison of Real vs Predicted Lithology (lithology in function of depth)

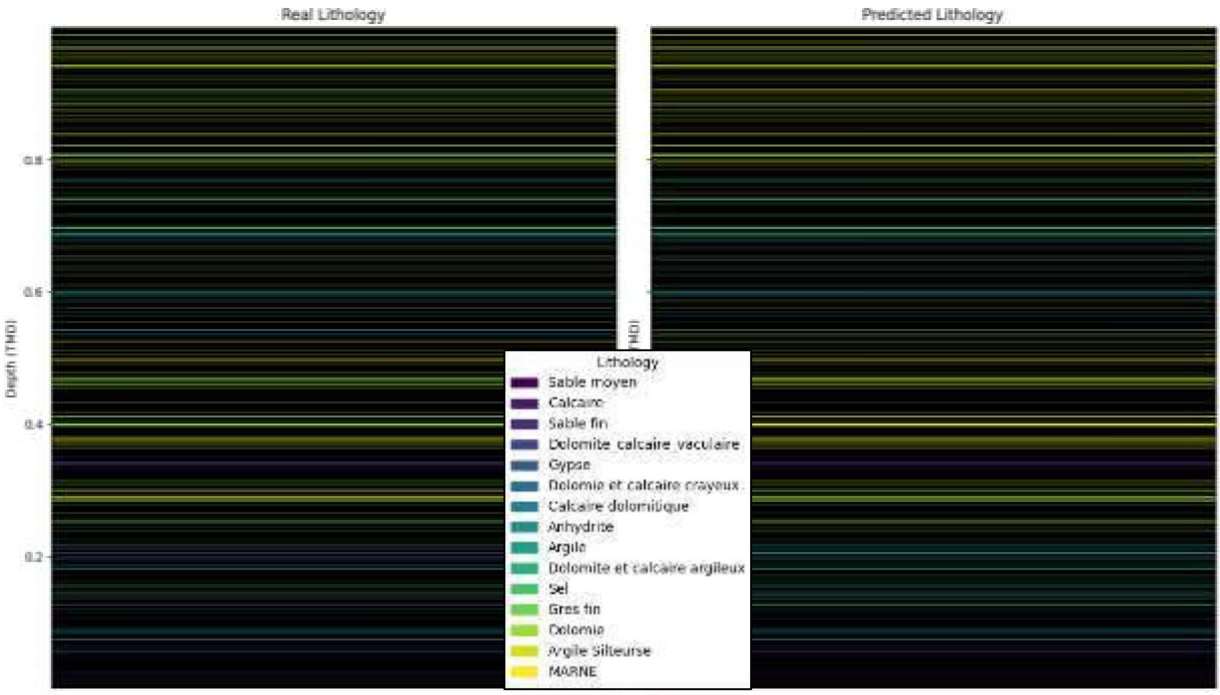


Figure 26- Comparison of Real vs Predicted Lithology (all data)

XIV. Limitations and Challenges

1. Complexity of Drill Parameters:
The model relies on analyzing a vast array of drill parameters, such as weight-on-bit, rotational speed, rate of penetration, and torque, among others. The sheer volume and complexity of these parameters can make it challenging to extract meaningful patterns and correlations with the underlying lithology. Effectively processing and interpreting this data in real-time is a significant challenge.

2. Lithology Variability:
Geological formations can exhibit a high degree of heterogeneity, with abrupt changes in lithology types, mineral compositions, and physical properties. The model needs to be able to quickly adapt to these rapid changes in lithology and update its recognition models accordingly. Accommodating this variability in a robust and accurate manner is a key limitation.

3. Execution Time Constraints:
Real-time lithology recognition based on drill parameters requires the system to process the data and provide accurate results within a very short time frame, often in the order of seconds or less. This places significant constraints on the computational

algorithms and hardware used, as they need to be highly efficient and optimized to meet the stringent timing requirements.

4. Dynamic Data Collection and Sensor Degradation:
The drill parameters used by the model may be affected by various factors, such as sensor degradation, drilling conditions, and changes in the sampling methods. Maintaining the system's performance and accuracy in the face of these dynamic factors can be a significant challenge, requiring robust data preprocessing and quality control techniques.

5. Validation and Uncertainty Quantification:
Ensuring the reliability and trustworthiness of the model's output is crucial, as the results may be used for critical decision-making in exploration, development, and production activities. Developing robust validation and uncertainty quantification techniques is essential to address this limitation, particularly when dealing with the inherent uncertainties associated with drill parameter data.

XV. Futur work

As the field of lithology progresses with the integration of advanced machine learning models, the next significant step is to seamlessly incorporate these models into real-time applications. One such potential application is LithoVision, a comprehensive tool designed for geologists and engineers to visualize and analyze lithology data efficiently. The integration of our Random Forest model for lithology recognition into LithoVision can revolutionize the way geological data is interpreted and utilized in the field. Here are some key points for future work in this area:

1. Seamless Integration with LithoVision:

- **API Development:** Develop a robust API that allows LithoVision to communicate with the machine learning model. This API will handle data transfer, model inference, and result retrieval in real-time.
- **User Interface (UI) Enhancements:** Update the LithoVision UI to include features that allow users to input new data and visualize predictions. The UI should be intuitive and provide clear visual cues about lithology classifications.
- **Batch and Real-Time Processing:** Ensure that the model can handle both batch processing of historical data and real-time processing of streaming data from drilling operations.

2. Model Optimization and Validation:

- **Model Retraining:** Periodically retrain the model with new data to improve accuracy and adapt to changes in geological formations. This ensures that the model remains up-to-date and reliable.
- **Validation and Testing:** Rigorously test the model within LithoVision under various scenarios to validate its performance. This includes edge cases and unusual lithology types to ensure robustness.

3. Enhanced Visualization Tools:

- **Interactive Logs:** Implement interactive lithology logs that allow users to drill down into specific depths and see detailed predictions and feature importances.
- **3D Visualization:** Integrate 3D visualization capabilities that provide a comprehensive view of the geological formations, helping users better understand the spatial relationships and lithology distributions.
- **Heatmaps and Annotations:** Include heatmaps to highlight areas with high uncertainty in predictions, and allow users to annotate these areas for further investigation.

4. User Feedback Loop:

- **Feedback Mechanism:** Implement a feedback mechanism within LithoVision where users can provide insights and corrections to model predictions. This user-generated data can be invaluable for retraining and improving the model.
- **Continuous Improvement:** Use the feedback loop to continuously refine the model. Incorporate active learning techniques where the model prioritizes learning from user-identified errors.

5. Integration with Other Tools:

- **Data Sources:** Integrate LithoVision with other data sources such as seismic data, core samples, and well logs to provide a holistic view of the geological environment. This can enhance the model's predictive capabilities by providing more context.
- **Collaboration Tools:** Incorporate collaboration tools that allow multiple users to work on the same dataset simultaneously, facilitating better decision-making and knowledge sharing.

6. Scalability and Performance:

- **Scalable Infrastructure:** Deploy the model on a scalable cloud infrastructure to handle large volumes of data and multiple simultaneous users. Ensure that the system can scale up during peak usage times.

- **Performance Optimization:** Optimize the model and API for low latency to ensure that predictions are provided in real-time, which is crucial during drilling operations.

7. Regulatory and Compliance:

- **Data Security:** Ensure that all data handling complies with relevant data protection regulations. Implement strong security measures to protect sensitive geological data.
- **Compliance:** Ensure that the application and its usage comply with industry standards and regulations.

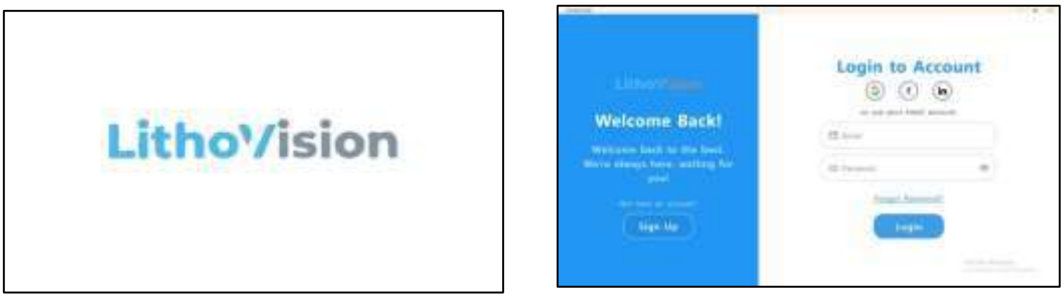


Figure 27 - LithoVision logo and interface

Conclusion

Conclusion

In this research, we present a machine learning-based approach to real-time lithology prediction using drilling data, focusing on the application of a Random Forest classifier. Our proposed methodology, termed LITHOVISION, successfully differentiates between various lithologies, achieving an accuracy of 68% in our study. This represents a significant improvement over traditional methods that primarily rely on the rate of penetration (ROP) for lithology prediction.

The relationship between lithology prediction and drilling tools calibration is reciprocal and critical for successful drilling operations. Accurate lithology predictions inform the calibration of drilling tools, leading to optimized performance, enhanced safety, and reduced operational costs. Conversely, well-calibrated tools provide more accurate data, improving future lithology predictions. This interaction is essential for advancing drilling technologies and achieving efficient resource extraction.

The LITHOVISION model uses a comprehensive set of drilling data features, including ROP, total gas content, and torque. This diverse feature set enables the model to capture the intricate relationships between different drilling parameters and their influence on lithology identification. By doing so, our approach provides a more nuanced and accurate prediction of subsurface lithologies, enhancing the understanding of geological formations encountered during drilling operations.

Despite the challenges in identifying rare lithologies, our research demonstrates significant progress in immediate subsurface analysis. The robustness of our results is evidenced by consistent precision and recall values around 69%, underscoring the efficacy of our methodology. These real-time lithology insights are particularly valuable for geosteering, a critical aspect of maintaining the optimal well trajectory within the pay zone, thereby improving drilling efficiency and reducing operational risks.

We are committed to the continuous enhancement of our LITHOVISION web application. Future improvements will focus on incorporating a dedicated section for diverse data visualization features, making the app more user-friendly and providing deeper insights. These enhancements will empower users with real-time, data-driven decision-making capabilities during drilling operations.

In conclusion, our study highlights the transformative potential of machine learning in real-time lithology prediction. While the current accuracy of 68% indicates room for further refinement, our research lays a solid foundation for revolutionizing subsurface analysis and geosteering in the oil and gas industry. By continually refining our models and integrating them into user-friendly applications like LITHOVISION, we aim to significantly enhance the efficiency and effectiveness of drilling operations, driving innovation and progress in the field.

References

- Aristodemou E., Pain C., De Oliveira C., Goddard T., Harris C. (2004). Inversion of nuclear well-logging data using neural networks. Geophysical Prospecting: Volume53, Issue1 January 2005 Pages 103-120. https://doi.org/10.1111/j.1365-2478.2005.00432.x
- Bobbitt, Z. (2022, April 19). How to Use SMOTE for Imbalanced Data in R (With Example).
- Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (2017). Classification and regression trees. In *Classification and Regression Trees*. 10.1201/9781315139470.
- Cassin Carli, M. (2014). Horizontal Directional Drilling Construction Methods in Mainland China.
- Chopra P., Papp É. and Gibson (2002) GEOPHYSICAL WELL LOGGING.
- Criminisi A., Shotton J., Konukoglu E. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning Found. Trends® Comput. Graphic. Vis., 7 (2–3) (2012), pp. 81-227
- Dey, R., & Mathur, R. (2022). Frequent Relational Smote For Class Imbalance An Experimental Approach.
- Djemili, B., & Boublal, M. S. (2016). Validation des donnés structurales de la région de Hassi Messaoud par la comparaison entre les données d'imagerie de fond et les cartes des linéaments gravimétriques.
- Donges, N. (2024, Mar 08). Random Forest: A Complete Guide for Machine Learning.
- Drilling Contractor. (n.d.). Algerian underground blowout - Drilling Contractor.
- Ellis D.V., Singer J.M. (2007) Well Logging for Earth Scientists (2nd Edition), Springer, Dordrecht, The Netherlands
- Encyclopaedia Britannica. (2024, Apr 9). Methane chemical compound.
- Erzinger, T. Wiersberg and M. Zimmer. (2006, June). Real-time mud gas logging and sampling during drilling.
- Garcia, D., Martinez, F., & Rodriguez, M. (n.d.). Application of Machine Learning Algorithms for Lithology Recognition While Drilling. SEG Annual Meeting, Society of Exploration Geophysicists.
- Goyal, S. K., & Singh, A. K. (2019). Data Normalization Techniques for Machine Learning. International Journal of Advanced Research in Computer Science and Software Engineering, 8(3), 1-8. DOI: 10.23956/ijarcse.2019.8.03.001
- Group News. (2015, December). Total Gas & H2S Gas Detection System in Mud: Forties Delta Platform.
- Gu, Y., Bao, Z., Song, X., Patil, S., & Ling, K. (2019). Complex lithology prediction using probabilistic neural network improved by continuous restricted Boltzmann machine and particle swarm optimization. Journal of Petroleum Science and Engineering, 179, 966-978.
- Hamzioui, M. L. (2016). Etude de la précipitation des asphaltènes dans la zone1 de Hassi Messaoud.
- Ho, T. K. (2016, April 17). Random Decision Forests.

- Hou, Z. (2023). Intelligent Lithology Identification Methods for Rock Images Based on Object Detection. Computers & Geosciences.
- Inglis, T. A. (1987). Measurement While Drilling (MWD). In: Directional Drilling. Petroleum Engineering and Development Studies, vol 2. Springer, Dordrecht.
- Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset.
- Jian H., Chenghui L., Zhimin C., Haiwei M. (2020) Integration of deep neural networks and ensemble learning machines for missing well logs estimation, Flow Measurement and Instrumentation, Volume 73,2020, 101748, ISSN 0955-5986, https://doi.org/10.1016/j.flowmeasinst.2020.101748.
- Jin Lai, Yang Su, Lu Xiao, Fei Zhao, Tianyu Bai, Yuhang Li, Hongbin Li, Yuyue Huang, Guiwen Wang, Ziqiang Qin, (2024). Application of geophysical well logs in solving geologic issues: Past, present and future prospect, Geoscience Frontiers, Volume 15, Issue 3, 2024,101779, ISSN 1674-9871, https://doi.org/10.1016/j.gsf.2024.101779.
- Jost, D. (2019, Jul 9). What is a proximity sensor?
- Kathrada M. , Adillah B. J.(2019). Visual Recognition of Drill Cuttings Lithologies Using Convolutional Neural Networks to Aid Reservoir Characterisation . SPE Reservoir Characterisation and Simulation Conference and Exhibition, Abu Dhabi, UAE, SPE-196675-MS https://doi.org/10.2118/196675-MS
- Khalifa, H., Tomomewo, O. S., Ndulue, U. F., & Berrehal, B. E. (2023). Machine Learning-Based Real-Time Prediction of Formation Lithology and Tops Using Drilling Parameters with a Web App Integration. Computers & Geosciences.
- Kheladi, D. M. (2020/2021). CARACTERISATION PETROLIERE DE RESERVOIR SIEGENIEN A L'AIDE DES METHODES MACHINE LEARNING, APPLICATION AU CHAMP DE - WADI EL TEH - BASSIN DE BERKINE OUEST.
- Kinney, T. A. (2024, May 21). An Overview of Proximity Sensors.
- Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., Ding, Y., Li, Y., Jian, L., Cheng, J., & Guo, H. (2024, February 16). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets.
- Li, T., Xiang, R., Jiao, T., & Yan, X. (2022). Research on Intelligent Lithology Identification Method Based on Real-Time Data of Drilling Wells. Computers & Geosciences.
- Liang, H., Chen, H., Guo, J., & Zuo, X. (2022, February 11). Detection and Identification Method of Drilling Total Hydrocarbon Gas Based on Infrared Spectroscopy and KL+BP-RBF Algorithm.
- Loukil, L., (2016) Énergie, environnement et développement durable en Algérie étude de cas de la région de Hassi-Messaoud , THESE , Université d'Oran.
- Moussous, Y. (2008). Applicability of HAZOP to Under-Balanced Drilling Surface Facilities in HMD Field. Volume 02, N° 02, Juillet 2008.
- Nabawy B.S. , Rochette P., Géraud Y. (2010) Electric pore fabric of the Nubia sandstones in south Egypt: Characterization and modelling, J. Afr. Earth Sci., 183 (2) (2010), pp. 681-694
- Qi, L., & Carr, T. R. (2006). Neural network prediction of carbonate lithofacies from well logs, Big Bow and Sand Arroyo Creek fields, Southwest Kansas. Computers & Geosciences, 32(7), 947-964.

- Raouf, M. (2023, Apr 17). Mud Gas While Drilling (GWD) Detection and Analysis.
- Rashidi, H.H., Sen, S., Palmieri, T.L. et al. Early Recognition of Burn- and Trauma-Related Acute Kidney Injury: A Pilot Comparison of Machine Learning Techniques. Sci Rep 10, 205 (2020). https://doi.org/10.1038/s41598-019-57083-6
- RS. (2023, Jan 24). A Complete Guide to Proximity Sensors.
- Salunke, D. (2023, Aug 28). SVC (Support Vector Classifier).
- Santos, P., Oliveira, R., & Silva, M. (2020). Recent Advances in Downhole Sensing Technologies for Lithology Recognition in Drilling Operations. Journal of Petroleum Science and Engineering, 170, 206-218.
- SLB Fully Integrated Algerian Exploration Wells Project Integrated Well Construction (01/01/2005)
- Smith, L., Jones, R., & Johnson, T. (2002). Lithology Recognition Techniques in Drilling Operations. Journal of Petroleum Geology, 25(3), 123-140.
- Tutorials Point. (2019, Dec). Support Vector Machines with Scikit-learn Tutorial.
- Tutorials Point. (2024, Jan 31). Random Forest Classifier using Scikit-learn.
- Wang, K., & Zhang, L. (2008). Predicting formation lithology from log data by using a neural network. Petroleum Science, 5(3), 242-246.
- William. (2023, March 6). Different Types of Proximity Sensors & Their Applications.
- Witten, I.H., Frank, E., Hall, M.A. & Pal, C.J. (2016). Data mining: practical machine learning tools and techniques. In *Data Mining: Practical Machine Learning Tools and Techniques*. 10.1016/c2009-0-19715-5.
- Wu, X., Zhao, Y., & Liu, D. (n.d.). Supervised Learning Algorithms for Lithology Recognition Using Drilling Data. SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers.
- Xu Z. , Shi H., Lin P., Liu T. (2021) Integrated lithology identification based on images and elemental data from rocks, Journal of Petroleum Science and Engineering; Volume 205, 108853, ISSN 0920-4105, https://doi.org/10.1016/j.petrol.2021.108853.
- Xu Z., Shi H., Lin P. and Ma W., (2022). "Intelligent On-Site Lithology Identification Based on Deep Learning of Rock Images and Elemental Data," in *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022, Art no. 6511205, DOI: 10.1109/LGRS.2022.3179623.
- Xu, Z., & Lin, P. (2021). Deep learning of rock images for intelligent lithology identification. Computers & Geosciences, 154, 104799.