PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
Ministry of Higher Education and Scientific Research

Serial N°: ……. / 2024

Kasdi Merbah Ouargla University

Faculty of Hydrocarbons, Renewable Energies and Sciences of Earth and the Universe

Department of Geology

Dissertation
To obtain the Master's degree
Option: Hydrocarbons Geology

Submitted by:

- **Khelifi Houari Boumedien**
- **Bousbia Brahim Naoufel**
- **Bouzine Mohieddine**

Supervisor: Ms. ROBEI SARRA

-TITLE-

Real-Time Prediction of Formation Types from Drilling Data Using Artificial Intelligence

Defended on: 08/06/2024

Jury:

President:Mr. BELKSIER MOHAMMED SALEH Professeu Univ. Ouargla

M.A.A Univ. Ouargla

Examiner: Mr. DRAOUI ABDELMALEK M.C.B Univ. Ouargla

Academic Year: 2023/2024

Dedication

To our dearest family and friends,

This thesis is dedicated to you.

Acknowledgements

First and foremost, we express our sincerest and deepest gratitude to Allah Almighty, who granted us the strength and patience to complete this work.

This thesis is the result of the collective efforts of many individuals whose support was instrumental in achieving this research's objectives.

We would like to extend our profound appreciation to all who contributed to its completion.

We begin with our supervisor, **Dr. Sarra Robei** , whose invaluable guidance, encouragement, and insightful advice were crucial throughout the research process.

Our gratitude also extends to **Mr. Said BENHAMOUCHE**, drilling engineer at **ENTP**, for his exceptional support during our training , and to all engineers for their unwavering support and collaboration.

We also thank Mr. **Mohammed Elmaki MOUHOUBI,** for his assistance in topic selection and developing the problematic.

We are deeply grateful to **Mr. Salah Eddine KAIFAS,** for his attentive support and invaluable development during the initial phases of our study.

Special thanks to **Mr. Sadjed BOUSBIA**, for his continuous help, and all the essential support he provided for the completion of this study.

Our sincere thanks go to **Mr. Hacene ABAYAHIA,** Mudlogger at Sonatrach, for his patience, cooperation, and for providing all the necessary information and explanations.

We acknowledge the professors who honored us by joining the review committee and agreeing to evaluate this work, lending their expertise to ensure the highest standards of scientific rigor.

Additionally, we extend our gratitude to our professors and administrative staff, for their continuous support and encouragement.

Finally, we express our appreciation to everyone who contributed, directly or indirectly, to the successful completion of this thesis.

Abstract

This study addresses the challenge of predicting formation types in petroleum field through a machine learning approach. Utilizing around 10,000 datasets from two wells in the Hassi Messaoud field, the model, named "Ama" (AI Mud Logger Assistant), demonstrates a mean accuracy of 70% in the training phase and high precision in predicting formation classes. Testing data confirms its effectiveness. The research overcomes limitations such as time lags and the lack of real-time data, employing seven drilling parameters rather than the traditional sole focus on the rate of penetration (ROP). The model was trained on a carefully preprocessed dataset from the Hassi Messaoud field to ensure unbiased and balanced samples. The primary purpose is to develop an machine learning model to accurately predict of geological formation types from drilling data. This leads to improve precision, reduce drilling time and costs, and enhance efficiency by minimizing uncertainties. This approach aspires to offer significant insights into geosciences, formation type detection and exploration practices.
Keywords: Machine learning, Artificial intelligence, drilling data, Formation type, Hassi Messaoud.

ملخص

تتناول هذه الأطروحة التحدي المتمثل في التنبؤ بأنواع الطبقات الارضية من خلال نهج التعلم الآلي.وباستخدام حوالي 10000 مجموعة بيانات من بئريين في حقل حاسي مسعود، حيث يُظهر نموذجنا "Ama" (AI Mud Logger Assistant)، دقة متوسطة تبلغ 70% في مرحلة التدريب ودقة عالية في التنبؤ بنوع الطبقات الارضية. ويستخدم نموذجنا سبعة معايير للحفر بدلاً من التركيز التقليدي الوحيد على معدل الاختراق (ROP). تم تدريب النموذج على مجموعة بيانات تمت معالجتها بعناية من حقل حاسي مسعود لضمان الحصول على عينات دقيقة ومتوازنة. الهدف الأساسي هو تطوير نموذج الذكاء الاصطناعي، للتنبؤ بدقة أنواع الطبقات الارضية من بيانات الحفر. ويهدف هذا إلى تحسين الدقة وتقليل وقت الحفر وتكاليفه وتعزيز الكفاءة من خلال تقليل حالات عدم اليقين. ويطمح هذا النهج إلى تقديم رؤى مهمة في علوم الأرض، مما قد يؤدي إلى إحداث تقدم في وصف الطبقات الارضية وممارسات الاستكشاف.

الكلمات المفتاحية: التعلم الآلي، الذكاء الاصطناعي، بيانات الحفر، نوع التكوين، حاسي مسعود.

Résumé

Cette étude aborde le défi de la prédiction des types de formations souterraines dans l'industrie pétrolière grâce à une approche d'apprentissage automatique. Utilisant environ 10 000 jeux de données provenant de deux puits de forage du champ de Hassi Messaoud, le modèle, baptisé « Ama » (AI Mud Logger Assistant), démontre une précision moyenne de 70 % dans la phase d'entraînement et une haute précision dans la prévision des types de formation. Les tests sur de nouvelles données confirment son efficacité. La recherche surmonte les limitations telles que les décalages temporels et le manque de données en temps réel, en utilisant sept paramètres de forage plutôt que de se concentrer uniquement sur le taux de pénétration (ROP). Le modèle a été formé sur un ensemble de données soigneusement prétraitées du champ de Hassi Messaoud pour garantir des échantillons impartiaux et équilibrés. L'objectif principal est de développer un modèle d'IA, en particulier un réseau neuronal, pour prédire avec précision les types de formations géologiques à partir des données de forage. Cela vise à améliorer la précision, à réduire le temps et les coûts de forage et à améliorer l'efficacité en minimisant les incertitudes. Cette approche aspire à offrir des informations significatives sur les géosciences, détection de type de formation et d'exploration.

Mots-clés : Machine learning, Intelligence artificielle, données de forage, Type de formation, Hassi Messaoud.

List of Figures

List of Tables

Table of contents

List of Abbreviations

AI	Artificial Intelligence
Ama	AI Mud Logger Assistant
ANNs	Artificial Neural Networks
BPNNs	Backpropagation Neural Networks
C.F.P.A	Compagnie Française des Petroles Algeria
KNN	K-Nearest Neighbors
ML	Machine Learning
MSE	Mean Squared Error
NaN	Not a Number
ROP	Rate Of Penetration
SPM	Strokes Per Minute
SPP	Stand Pipe Pressure
SVR	Support Vector Regression
W	Weights
WITS	Wellsite Information Transfer Specification
WOB	Weight on bit

Nomenclatures

xj	Input nodes
wij	Weights from the input layer
W	Weight matrix
uj	Biases

General Introduction

General Introduction

Drilling operations are fundamental to the exploration and production of oil and gas. The ability to predict formation types in real-time during drilling can significantly enhance the efficiency and safety of these operations. Traditional methods of identifying formation types involve post-drilling logging and analysis, which, although accurate, are time-consuming and can lead to delays in decision-making.

This approach involves using various drilling parameters such as revolutions per minute (RPM), rate of penetration (ROP), torque, weight on bit (WOB), standpipe pressure (SPP), depth, and flow pump rate. These parameters provide valuable insights into the behavior of the drilling process and the characteristics of the formations being penetrated. By analyzing these parameters, we can predict the type of formation encountered, allowing for immediate adjustments to the drilling process, which can improve efficiency, reduce costs, and enhance safety.
Machine learning and data-driven models have shown great promise in this regard. These models can learn from historical drilling data to identify patterns and correlations between drilling parameters and formation types. Once trained, these models can be deployed in real-time to predict formation types as drilling progresses, providing drilling engineers with critical information to prompt and well-informed decisions.

This study outlines the process of developing a predictive model for formation type identification using real-time drilling data. It covers data collection and preprocessing, exploratory data analysis, model selection and training, and finally, deployment and real-time prediction. By following this approach, drilling operations can be optimized to achieve better accuracy in formation identification, leading to more efficient and cost-effective exploration and production activities. Chapter 1 delves into a comprehensive analysis of the Hassi Messaoud reservoir, a massive oil field with unique geological characteristics and production challenges

Chapter 2 In this chapter we have provided an overview about mudloging, its objective and the mud logging sensors witch collecting the data

Chapter 3 Examine machine learning and its various algorithms principles, and its potential for formation type prediction. The chapter emphasizes the ability of machine learning to handle datasets and find patterns.

Chapter 4 translates theory into practice by applying machine learning techniques to the Hassi Messaoud reservoir data. This section details the data preparation steps, the selection of suitable algorithms, and the training and evaluation of the prediction models. By analyzing the models' performance in predicting formation types .

Chapter I: Presentation of Hassi-Messaoud Field

Introduction

Hassi-Messaoud field represents one of the most complex fields in the world. During geological history, this field has undergone, on the one hand, an intense tectonic evolution characterized by distinctive compressive phases. On the other hand, by the diagenetic transformation in the reservoir, during its burial over geological time, until it took the current shape or configuration. These events can sometimes improve the petrophysical parameters (natural hydraulic fracturing, dissolution, etc.) as well as reduce them (reduction of porosity, cementation due to solution pressure phenomena, creation of matrices of small grains, etc…).

I.1 Geographic location

Hassi Messaoud field is located 850 km southeast of Algiers (650 km as the crow flies) and 350 km from the Tunisian border. The dimensions of the field reach 2500 km2 with an oil-impregnated surface of approximately 1600 km2 (KECHAR, 2020). Its location in Lambert South Algeria coordinates is as follows:

- 790,000 to 840,000 East,
- 110,000 to 150,000 North.

In geographic coordinates:

- To the north by latitude 32° 15′,
- To the west by longitude 5° 40′,
- In the South by latitude 31° 30′,
- To the East by longitude 6° 35′.

I.2 Geological setting:

The Hassi Messaoud field occupies the central part of the Triassic province, east of the Oued Mya depression in district IV which, by its surface area and its reserves, is the largest oil deposit in Algeria which covers an area of almost 2500 km² (KECHAR, 2020). It is limited:

- In the North-West by the Ouargla deposits (Gellala, Ben Kahla and Haoud Berkaoui) .
- To the South-West by the deposits of El Gassi, Zotti and El Agreb.
- To the South-East by the deposits; Rhourde El Baguel and Mesdar.

Geologically, as shown in figure I-1 it is limited:

- To the West by the Oued M'ya,
- To the South by the Amguid El Biod mole,
- To the North by the Djammâa-Touggourt structure,
- To the east by the Dahar shoals, Rhourde El Baguel and the Ghadames.

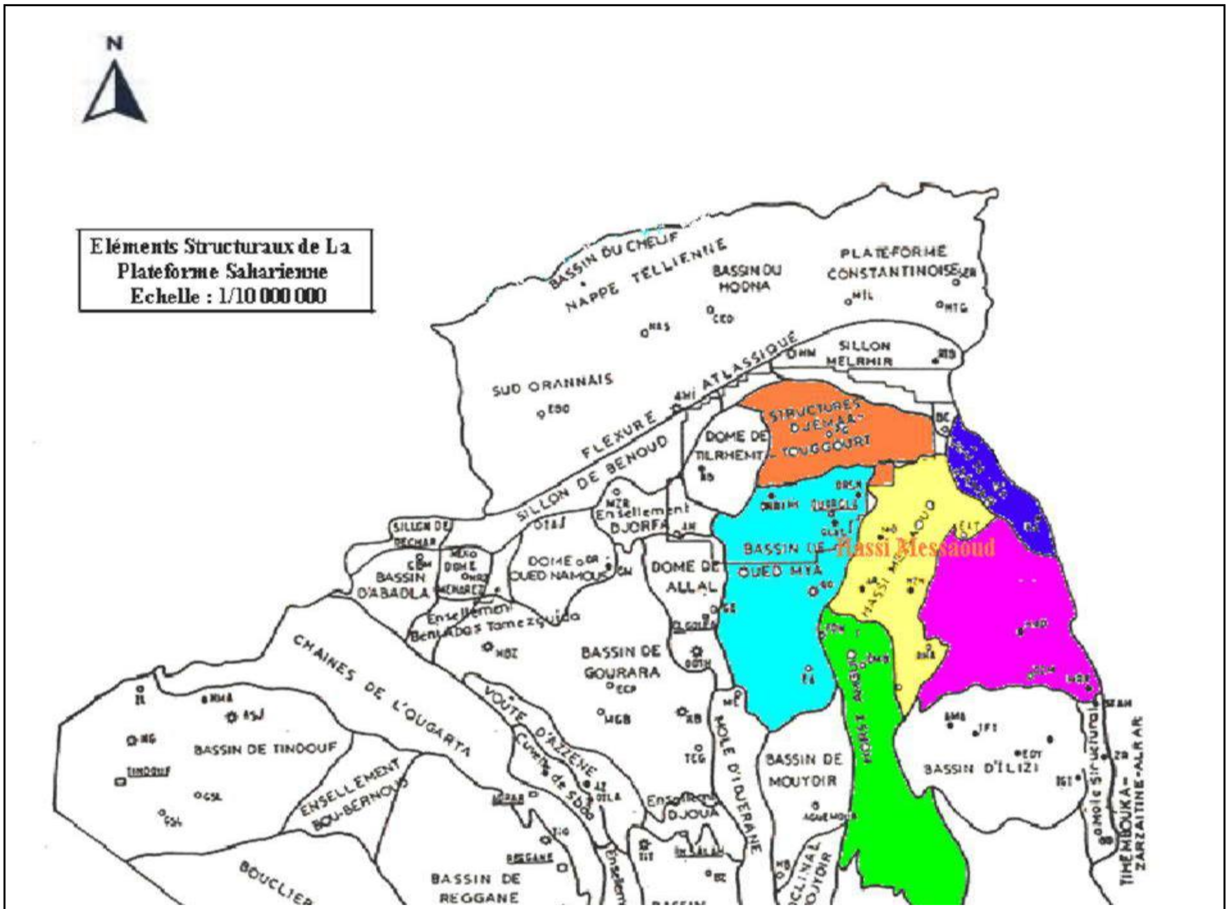


Figure I-1 : Geological situation of the Hassi Messaoud field. (WEC, 2007)

The Hassi Messaoud structure develops into a vast sub-circular anticline 45 km in diameter, direction: North – East / SOUTH-West. It is partially cracked and the cracks are due to plate tectonic movements which caused the structure to become anticlinal. The reservoirs have undergone natural hydraulic fracturing(TRABELSI, 2019).

Accidents affecting the reservoir are of two types:

- The faults in the submeridian direction and as well as the other faults, perpendicular in the northwest/southeast direction, highlight the tectonic character of the region,
- Breaks without releases which have a great effect on reservoir fracturing

From the reservoir characteristic point of view, the Hassi Messaoud deposit is defined in a perfect trilogy:

- Heterogeneous on the vertical and horizontal plane.
- Discontinuous from the point of view of fluid flow.
- Anisotropic: by the presence of silt and the existence of a matrix of small grains.

Hassi Messaoud field sub-divisionThe Hassi Messaoud field divided into Hassi Messaoud North and Hassi Messaoud South, Currently, the field is subdivided into 25 production zones (figure I-2). These zones are relatively independent, corresponding to a set of wells which communicate with each other lithologically and behave in the same way from a pressure point of view.(TRABELSI, 2019)

The Hassi Messaoud field is divided from east to west into two distinct parts: The field South and the North field, each has its ownpseudoname .

I.2.1 North Field

It is a geographical numbering supplemented by a chronological numbering,

example: Omo38, Onm14, Ompz12

Where:

- **O:** Capital letter, Ouargla permit.
- **m:** area of the oil zone: 1600 km2.
- **o:** Tiny, surface area of the oil zone of 100 km2.
- 3: x, and 8: y.

I.2.2 South Field

The numbering of the zones is chronological. Ex: MD1, MD2, MD3,…MDZ509*, MDZ557* (see Figure I-2).

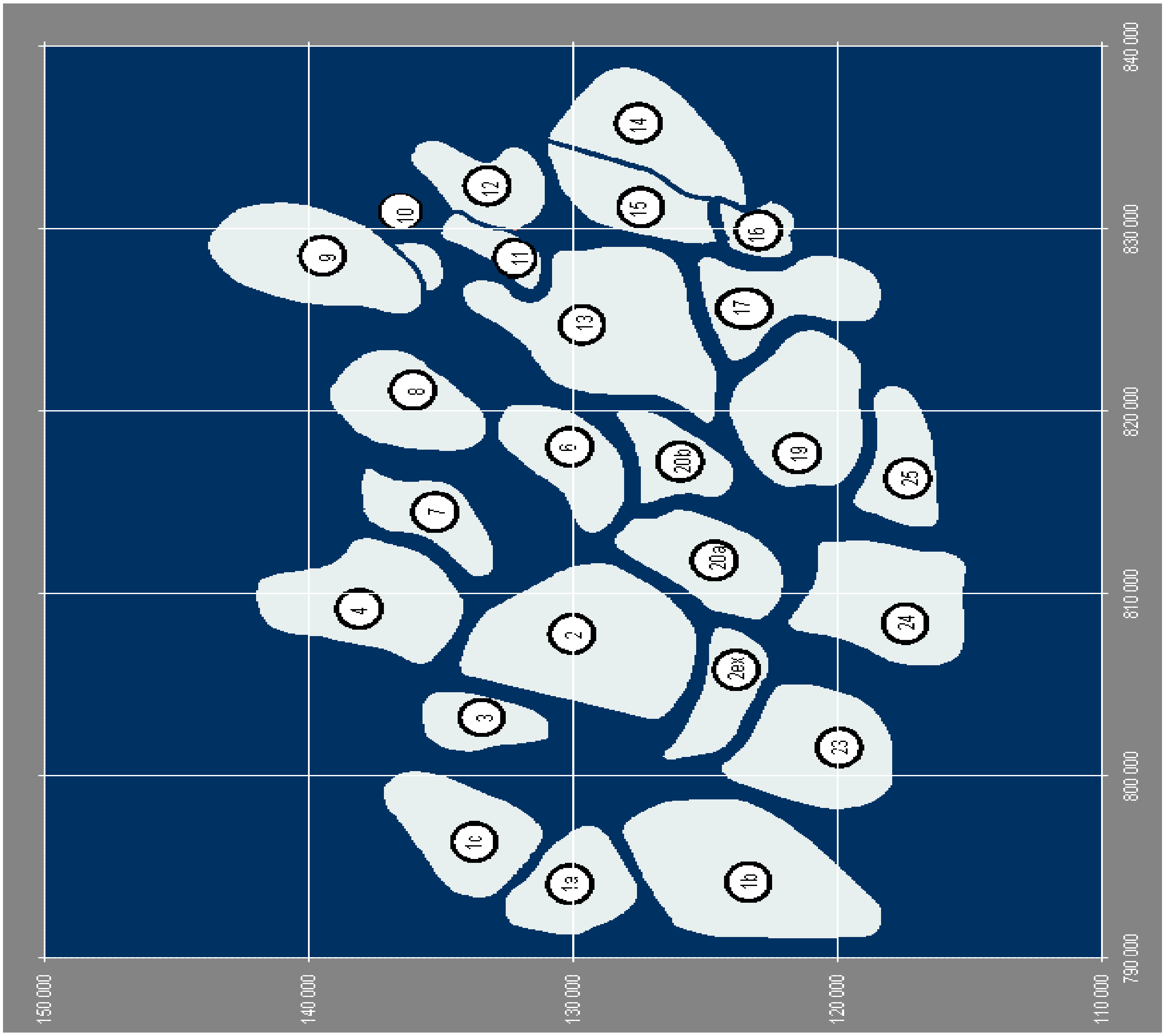


Figure I-2 : Hassi Messaoud reservoir zones (WEC, 2007)

I.3 Field historical background

The Hassi Messaoud deposit was discovered on January 16, 1956 by the first MD1 drilling, installed following a seismic refraction survey not far from the Hassi Messaoud camel well.

On June 15 of that same year, this drilling discovered oil in the Cambrian sandstones at a depth of 3,338 meters. In May 1957, 7 km north-northwest of MD1, the OM1 borehole, drilled by the C.F.P.A, confirmed the existence of a very significant quantity of oil in the Cambrian sandstones.(KECHAR, 2020)

The deposit was therefore covered by two distinct concessions:

- In the North the C.F.P.A
- To the south the SN. REPAL.

The boundary cuts the field in the East - West direction into two approximately equal parts.

I.4 Reservoir Description:

The Hassi Messaoud field is part of the eastern province of the Saharan platform. This province contains the main hydrocarbon accumulations of the Sahara, the reservoirs are mainly made up of different sandstone levels from the Cambro-Ordovician and Triassic. Its depth varies between 3100 and 3380 m. Its thickness is up to 200 m. The oil is light with an API rating of 45.4. Its initial pressure evaluated in the well is 482 kg/cm2 for a bubble point between 140 kg/cm2 and 200 kg/cm2.(TRABELSI, 2019)

At Hassi Messaoud the hydrocarbons are found in the Cambro-Ordovician which is subdivided from top to bottom into:

- Sandstone from Hassi Messaoud.
- EL-GASSI sandstone (lower part of the clay-sandstone of Oued Maya). Due to the Hercynian unconformity, a large part of it has been eroded and it is the salt Triassic which constitutes the cover of the reservoir.

The black Silurian clays, 40 km northwest of Hassi Messaoud, rich in Kerogen (organic matter), are supposed to be the source rock.

I.5 Description of the stratigraphy series

From a stratigraphic point of view (Figure I-3).

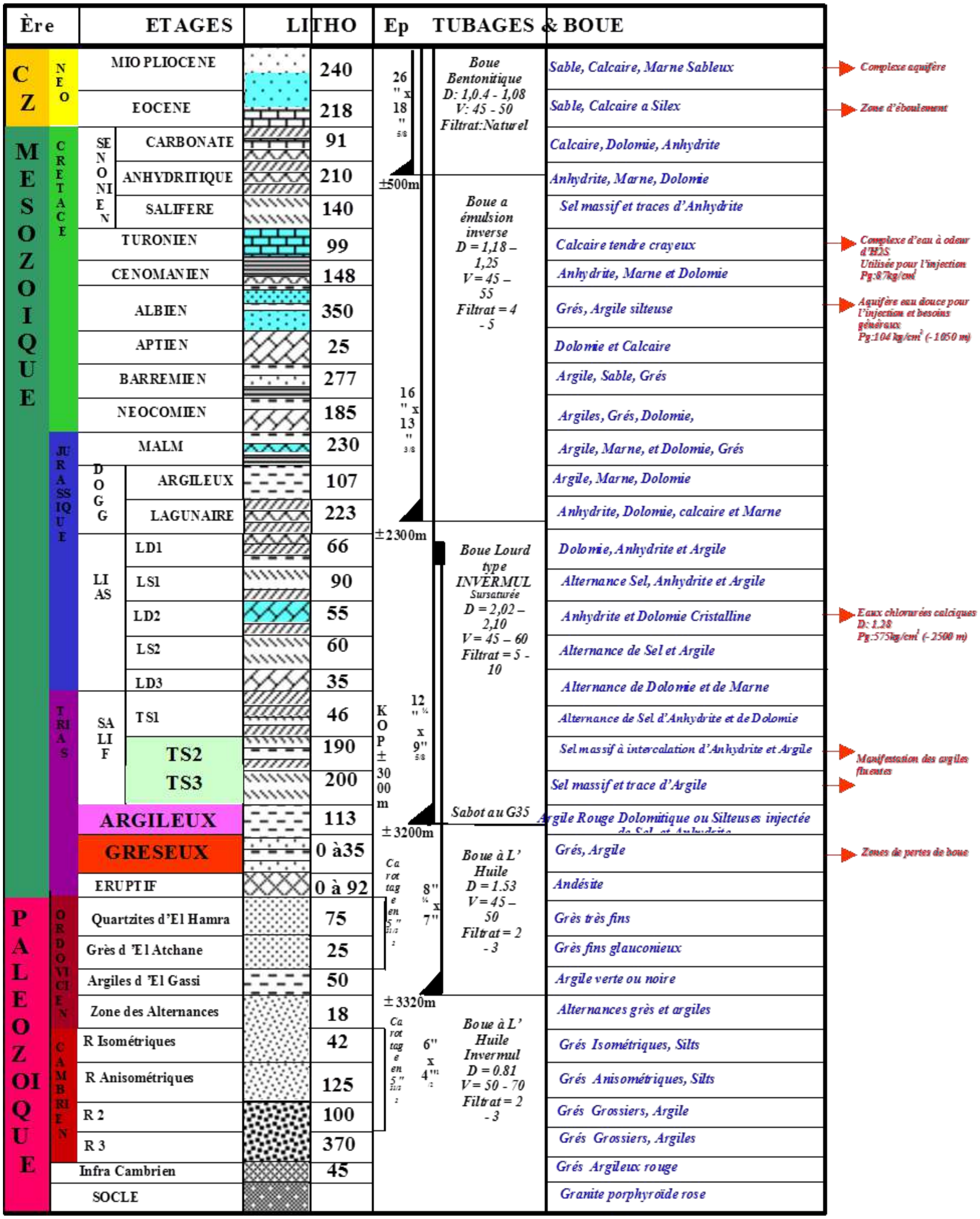


Figure I-3 : Stratigraphic in the Hassi-Messaoud reservoir (SONATRACH)

The Hassi Messaoud area comprises several geological formations from the Paleozoic to the Cenozoic eras. The Paleozoic formations, posterior to the Ordovician, are absent in the central part but present on the periphery. These formations include the Basement, Infracambrian, Cambrian, and Ordovician.

Three lithozones (R1, R2, R3) are distinguished:

Level R3:

- Average thickness of 270m.

- Composed of poorly sorted sand and gravel with quartz, feldspar, mica, siderite, and heavy minerals.

- Contains 30% clay (illite and kaolinite).

- Low permeability.

- Water saturation between 70-80%.

Level R2:

- Divided into R2C and R2AB.

- Average thickness of 80m.

- Poorly sorted grains with improved sphericity.

- Contains 17% clay.

- Improved permeability in R2AB.

- Water saturation considered.

Level Ra:

- Subdivided into three sedimentological zones.

- Average thickness of 120m.

- Deposition in a coastal alluvial plain with flat topography.

- Consists of coarse, well-sorted sand deposits with low clay content, and finer, less sorted sand deposits with clay intercalations.

- Improved permeability compared to other levels.

- Transition marked by increased Gamma Ray, Induction, and Neutron log responses due to changes in clay content and porosity.

Zone I (D1, ID, D2):

- Lower coarse zone.

- Drain D: Poorly sorted coarse sandstones and silts with clay intercalations.

- Inter-Drain (ID): Fine sandstones, silts, and clays.

- Drain D2: Good reservoir characteristics.

Zone II (D3):

- Intermediate fine zone.

- Smaller, better-sorted grains with lateral continuity.

- Clay and silt layers.

Zone III (D4):

- Upper coarse zone.

- Coarser sandstone layers at the base, similar to Zone I.

Level Ri (D5):

- Average thickness of 45m.

- Deposited after erosion, possibly removing D4 in the East and Southeast.

- Calm deposition environment with 30% clay.

- Small grain size, good sorting, and low permeability.

The Mesozoic:

- Triassic: Subdivided into Eruptive, Sandy, Clayey, and Saliferous Triassic.

- Jurassic: Average thickness of 844m with clayey-sandy sequence, limestone intercalations, and alternating lagoonal and marine facies.

- Cretaceous: Average thickness of 1620m with seven stages: Neocomian, Barremian, Aptian, Albian, Cenomanian, Turonian, and Senonian.

The Cenozoic:

- Eocene: Dolomitic limestone.

- Miocene-Pliocene: Sandy cover.

- Miocene-Pliocene.

Chapter II: Basics of Mudlogging

II.1 Definition of Mudlogging

The term "mudlogging" is composed of two words: "mud," meaning the drilling fluid or mud, and "logging," meaning the recording of data. Technically, it involves recording the data or information carried by the drilling mud (Aouimer, 2005). It is one of the important activities in drilling operations; it serves as a safety device as well as a means of receiving information gathered by the services. This unit mainly consists of three parts: geological control, mud and drilling parameters control (which is done using sensors), and gas detection instruments.

II.2 Objectives of mud logging

There are several broad objectives targeted by mud logging: identify potentially productive hydrocarbon-bearing formations, identify marker or correlatable geological formations, and provide data to the driller that enables safe and economically optimized operations. The actions performed to accomplish these objectives include the following: (Haywood, 1940)

- Collecting drill cuttings.
- Describing the cuttings (type of minerals present).
- Interpreting the described cuttings (lithology).
- Estimating properties such as porosity and permeability of the drilled formation.
- Maintaining and monitoring drilling-related and safety-related sensing equipment.
- Estimating the pore pressure of the drilled formation.
- Collecting, monitoring, and evaluating hydrocarbons released from the drilled formations.
- Assessing the productivity of hydrocarbon-bearing formations.
- Maintaining a record of drilling parameters.

Mud logging service first focused on monitoring the drilling mud returns qualitatively for oil and gas content. This included watching the mud returns for oil sheen, monitoring the gas evolving from the mud as it depressured at the surface, and examining the drill cuttings to determine the rock type that had been drilled, as well as looking for indication of oil on the cuttings. Detection of the onset of abnormal formation pressures using drilling parameters was proposed with the

كخcentury, saw its introduction in mud logging in the 1970s when electronics became sufficiently compact, rugged, and robust to be used at rig sites..

II.3 Results of Mudlogging

The reports generated from mudlogging during drilling operations include:

- Ensuring the safety of personnel and the well by predicting blowouts.
- Reducing drilling costs by avoiding additional operations (fishing, side tracking, plugging with cement) through continuous monitoring of drilling parameters.
- Enhancing the understanding of reservoir levels and real-time characterization of these levels.
- The capability to transfer acquired data from all sensors (contractors) in real time via WITS (Wellsite Information Transfer Specification).
- The final well report, which provides information on all engineering operations conducted during well construction, the formations encountered, the intervals cored and tested, and the problems and events encountered during drilling (stuck pipes, mud losses, kicks, etc.).

II.4 The Mudlogging Cabin

II.4.1 Definition

A mud logging cabin is a specialized mobile unit used on drilling sites, particularly in the oil and gas industry. It is designed to house equipment and personnel involved in the process of mud logging, which involves the collection, analysis, and monitoring of drilling fluids (mud) and geological samples. Here are the main features and functions of a mud logging cabin:

Data Collection and Analysis: The cabin is equipped with various instruments to collect data on drilling parameters, such as rate of penetration (ROP), mud weight, gas levels, and lithology (rock types). This data helps in making real-time decisions about drilling operations.

Sample Examination: Mud loggers examine cuttings brought to the surface by the drilling mud to determine the type of rock being drilled and to identify any hydrocarbons present. This involves both visual inspection and the use of microscopes and other analytical tools.

Safety and Communication: The cabin serves as a communication hub, allowing mud loggers to relay critical information to the drilling crew and other stakeholders. It also provides a controlled environment for safe handling of potentially hazardous materials. (Houcine A, 2018)

II.4.2 The Role of the Mudlogging Cabin

The main role of Mudlogging cabin i :

- Monitor all drilling parameters in real time.
- Report all anomalies during drilling operations to the following personnel:

    - Drilling supervisor
    - Shift leader
    - Site manager
    - Mud engineer
    - Other relevant personnel

- Create a stratigraphic log during drilling and provide a lithological description of each formation.
- Save all parameters in a database.
- Prepare a final drilling report for the client.

II.5 Geological Monitoring

II.5.1 Sampling

Cuttings are collected at the shale shaker, taking into account the "lag time". The sample of cuttings collected by the Mudlogger must be representative of the entire interval between two samplings.

The collection of cuttings is done using a sieve and a cup. The size of the cuttings depends on the bit's teeth. Therefore, they need to be found on the shale shaker. The shale shaker should be equipped with a board fixed at the slope of the screen to collect all the cuttings brought up during the sampling interval (10m, 5m, or other). The sampling interval is defined in the sampling program with the technical details of the samples. Generally, for exploration wells, the client requires several types of samples. (AOUIMER, 2005)

II.5.2 Mudlogging Sensors

II.5.2.1 Weight On Hook Sensor

The measurement of hook load is performed using tension measurements from the dead line through a hydraulic pressure cell. Typically, the sensor used is directly connected to the driller's measurement circuit (Figure II-1). The tension applied to the cable is converted into pressure in a hydraulic circuit. The sensor, consisting of a hydraulic strain gauge installed on this circuit, generates an electrical signal that can be calibrated to indicate weight. (Rahmouni.H, 2012)

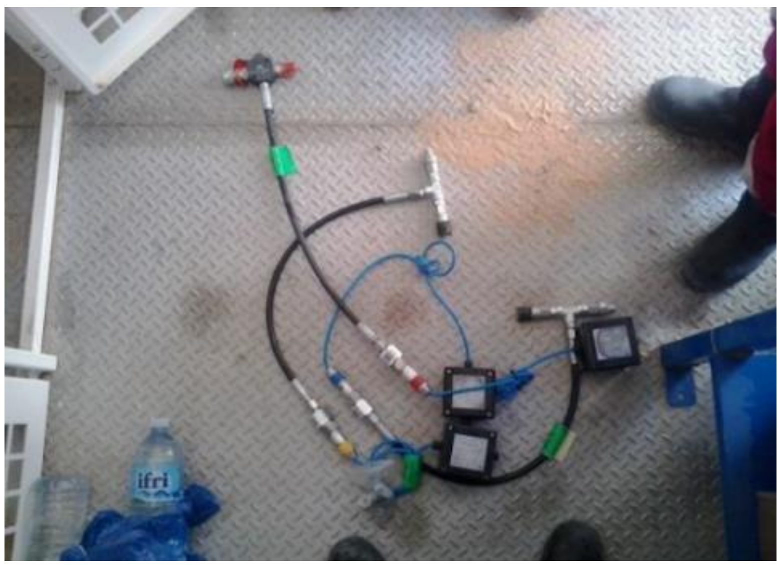


Figure II-2:WOH Sensor

II.5.2.2 Stand Pipe Pressure Sensor

The mud pressure is measured using sensors on the floor manifold to obtain the input value

Figure II-3:SPP Sensor

II.5.2.3 Casing Pressure Sensor

On choke manifold to get the output value

Figure II-4:Casing Pressure Sensor

II.5.2.4 Pump Sensor (SPM)

The easiest method is to count the number of pump strokes. Knowing the volume injected at each stroke and the efficiency of the pump, the flow rate can be calculated. It is easy to measure the number of pump strokes by proximity switches (Figure II-5) or contactors electrical. (Rahmouni.H, 2012)

Figure II-6:Pump Sensor (SPM)

II.5.2.5 Flow Out sensor (flow paddle)

The flow rate coming out of the sludge is measured using a sensor placed at the level of the chute (Figure II-7), the sensor is composed of two main elements, a potentiometer and a paddle (paddle). During the circulation, the mud pushes the pellet up which turns the potentiometer, then a signal will be transmitted to the acquisition system (Rahmouni.H, 2012).

Figure II-8:Flow Out output flow sensor

II.5.2.6 Depth sensor

The proximity sensor (depth sensor) is placed on the winch (Figure II-9), indicating the depth.

Figure II-10:Depth sensor

II.5.2.7 IN and OUT Density Sensors

The most common devices use the differential hydrostatic pressure between two membranes of different heights in a mud column. (Rahmouni.H, 2012)

Figure II-11:IN and OUT Density Sensors

II.5.2.8 Tank level sensors

The measurement of the tank level is carried out using ultrasonic sensors (Figure II-12), which send sound waves downwards to detect the level of the fluid, which will then be converted by volume by the acquisition system. (Rahmouni.H, 2012)

Figure II-13:Tank level sensors

II.5.2.9 INlet and OUTlet temperature sensors

The temperature of the sludge at the inlet and outlet is recorded using rods Thermometers with platinum filament protected by a stainless-steel sheath (Figure II-14).

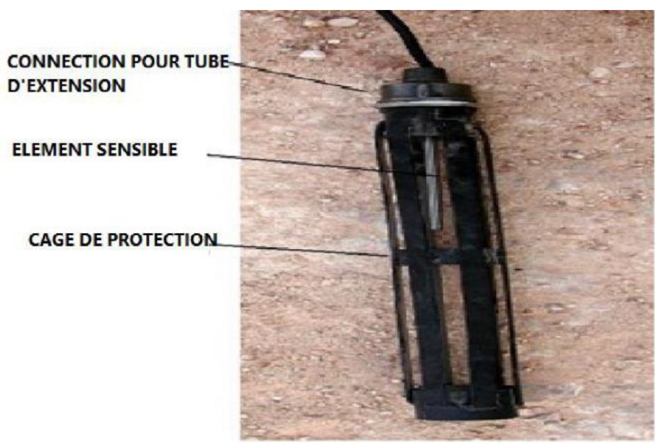


Figure II-15:IN and OUT temperature sensors

II.5.2.10 Rotation sensor

At the level of the rotation table, as the name suggests, the rotation sensor works when a metal object passes near its nose (Figure II-17), causing a circuit closure internal, which will subsequently give an event.

Figure II-18: Rotation Sensor

II.5.2.11 Torque sensors

At the level of the power cable of the generator (Figure II-19) which drives the table of rotation. The torque parameter is of considerable importance during drilling, it gives us an idea on the condition of the tool, the drill pipe, and the nature of the drilled formation.

Figure II-20:Torque sensors

Chapter III: Data-Driven Technologies

Introduction

Since the early days of the oil industry, accurately determining formation types in real-time has been crucial for safe and efficient drilling. This information is essential and it maintaining wellbore stability. Drilling engineers use rate of penetration (ROP) , formation cuttings, and mud logging to identify formation types. However, these methods have drawbacks such as high costs, lower accuracy, and significant labor, as well as temporal or depth-related delays, limiting their effectiveness and feasibility.

III.1 Data mining

Data mining can be defined as the exploration and analysis of large quantities of data to discover meaningful patterns and rules (Berry & Linoff, 2008). It is a powerful tool for uncovering knowledge that might otherwise remain hidden. Traditionally, oil and gas companies relied on educated guesses to analyze information. For example, an engineer might correlate well depth with pressure, assuming that deeper wells would have higher pressure. Data mining transforms this approach by systematically analyzing all available data collected during oil exploration and production without preconceived notions.

In the context of lithology prediction, data mining can analyse diverse data sources to accurately predict rock types and their properties, thereby enhancing drilling efficiency and safety. This predictive capability is crucial for optimizing drilling parameters, reducing non-productive time, and improving overall well performance. The integration of data mining into lithology prediction represents a significant advancement in the oil and gas industry, enabling smarter, data-driven decisions that improve operational outcomes.

III.1.1 Data Mining Process

1. Data Preparation: The first step involves collecting and cleaning the data. This ensures the information is accurate, consistent, and ready for analysis.

2. Data Exploration: Data mining algorithms then delve into the prepared data, identifying interesting trends, outliers, and potential relationships between different data points.

3. Pattern Recognition: Sophisticated algorithms analyze the data to uncover hidden patterns and relationships that might not be obvious through traditional methods.

4. Model Building: Based on the discovered patterns, data mining can build models to predict future outcomes or identify areas for improvement.

5. Evaluation and Deployment: The effectiveness of the models is evaluated, and if successful, they are deployed to guide decision-making within the oil and gas company.

III.2 Artificial intelligence

Artificial intelligence is a suite of innovative analytical methodologies striving to mimic life. These AI techniques demonstrate the capacity to learn and adapt to novel circumstances (Zurada et al., 1994). Key technologies falling under the umbrella of artificial intelligence include artificial neural networks, evolutionary programming, and fuzzy logic, all of which exhibit various reasoning attributes like generalization, discovery, association, and abstraction (Eberhart et al., 1996).

Over the past decade, artificial intelligence has evolved into a refined arsenal of analytical tools enabling the resolution of previously daunting or insoluble problems. Presently, these AI tools are effectively employed, often integrated with traditional statistical analysis methods, to construct intricate systems capable of tackling complex challenges.

The widespread utilization of these tools' spans across diverse domains, permeating into commercial applications. Artificial intelligence finds application in various sectors, including medical diagnosis, credit card fraud detection, bank loan approval, smart home devices, public transportation systems, automotive technologies like automatic transmissions and driverless cars, financial management, robot navigation, among others. In industries such as oil and gas, artificial intelligence aids in addressing issues pertaining to pressure transient analysis, well log interpretation, reservoir characterization, and the identification of suitable well candidates for stimulation, among myriad other tasks.(Shahab D. Mohaghegh, 2017)

III.2.1 Artificial Neural Network

Artificial Neural Networks (ANNs) represent a prominent technique within the realm of Artificial Intelligence (AI) research. These systems draw inspiration from the structural organization of the human brain (McCulloch & Pitts, 1943). The simplest definition of ANN is provided by the inventor of one of the first neurocomputers, Dr. Robert Hecht-Nielsen. He defines a neural network as:

a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.

Inspired by the human brain's structure, artificial neural networks (ANNs) are computational models mimicking biological neurons within AI research. The groundwork for ANNs was laid in 1943 through the groundbreaking work of McCulloch, a neuroscientist, and Pitts, a logician. Their conceptual model introduced the artificial neuron, a unit that, mirroring a biological neuron, receives input signals, processes them, and generates an output (McCulloch & , 1943).

Artificial neural networks (ANNs) mimic the brain's structure with interconnected processing units called "neurons" or "nodes." These nodes utilize "activation functions" and work in parallel across layers to solve problems. Learning in ANNs involves adjusting the connections between neurons, represented by numerical weights. These weights determine the output signal based on new input data. Information flows from the input layer receiving patterns, through hidden layers where complex processing occurs, and finally to the output layer where the final result is generated, as illustrated in Figure III-1.

Here, x_j are input nodes, w_{ij} are weights from the input layer to the hidden layer, and vi and y denote the weights from the hidden layer to the output layer and the output node, respectively. The ANN method has been established as a powerful technique to solve a variety of real-world problems because of its excellent learning capacity.

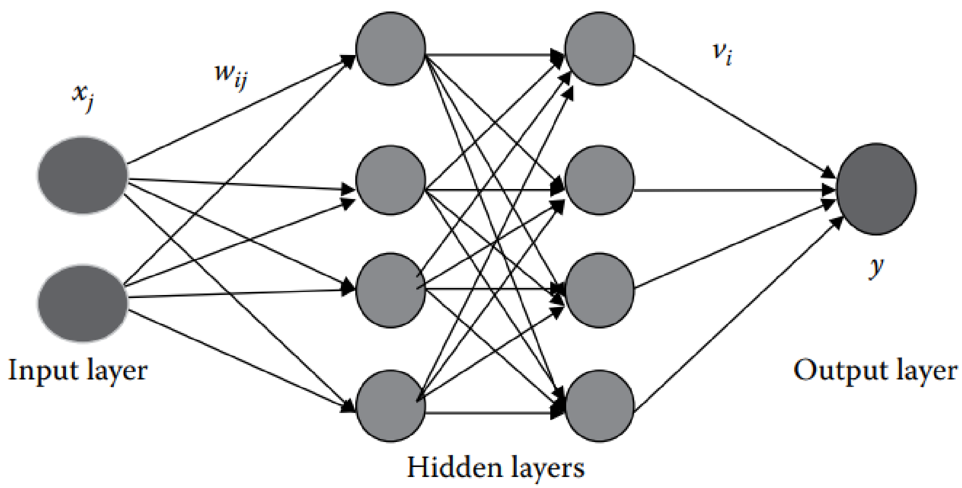


Figure III-1 : Structure of artificial neural network.

III.2.2 Architecture of ANN

Neural computing is a mathematical model inspired by the biological model. This computing system is made up of a large number of artificial neurons and a still larger number of interconnections among them. According to the structure of these interconnections, different classes of neural network architecture can be identified, as discussed next.

III.2.2.1 Feed-Forward Neural Network

Feed-forward neural networks arrange neurons in layers. Each layer receives input from the one before it and sends its output to the next layer, with no feedback loops. Information flows strictly forward from input to output neurons. As shown in Figure III-2, the network processes an input vector (X), with each element representing a feature (x_1, x_2, ..., x_n). Activation functions (f) transform these values within each neuron. Weights (W), organized in a connection matrix, determine the influence of each input on a neuron. The net input value is calculated by multiplying the weight matrix (W) with the input vector (X).(Zurada, 2004)

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}$$

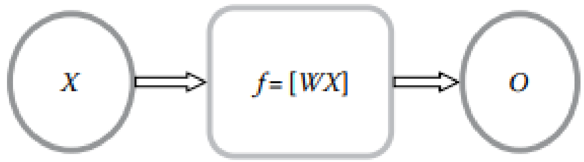


Figure III-2 : Block diagram of feed-forward ANN.

III.2.2.2 Feedback Neural Network

Unlike feed-forward networks, feedback neural networks allow connections to loop back, enabling signals to travel in both directions. This flexibility makes them powerful but potentially complex (Figure III-3). All connections between neurons are permitted. These networks excel at optimization problems, constantly adjusting their internal state until they reach an equilibrium point, making them ideal for scenarios where the best arrangement of interconnected factors needs to be identified.(Zurada, 2004)

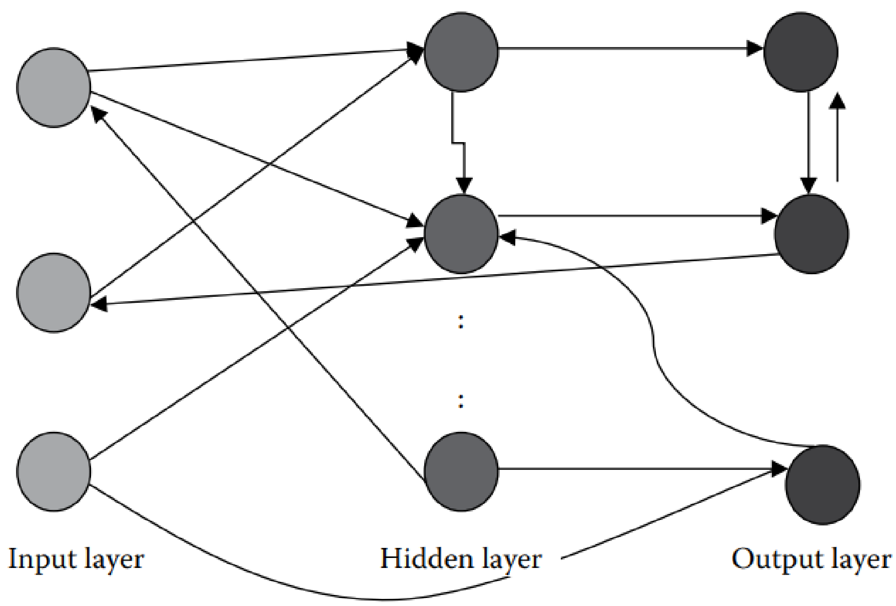


Figure III-3 : Diagram of feedback neural network.

III.2.3 Paradigms of Learning

ANNs possess a powerful capability: learning and generalizing from training data. This learning comes in two primary forms – supervised and unsupervised.

III.2.3.1 Supervised Learning

Supervised learning is a powerful training method for ANNs. Here, the network receives both the input data and the desired output. It then compares its generated output to the correct one, calculating an error. This error guides adjustments to the network's internal parameters, ultimately improving its performance. In simpler terms, the network learns by continuously correcting its responses based on provided examples.(Zurada, 2004)

III.2.3.2 Unsupervised Learning

Unsupervised training throws ANNs into a world of unlabeled data. Unlike supervised learning, the network receives only inputs, and must identify patterns and relationships on its own. This "self-organization" allows the network to discover hidden structures within the data, making it a powerful tool for tasks like data exploration and clustering.(Zurada, 2004)

III.2.4 Learning Rules or Learning Processes

These networks store knowledge within the connections' weights, which are adjusted through "learning rules" based on the input patterns. The delta learning rule, commonly used in backpropagation neural networks (BPNNs), is a popular example. As the name suggests, backpropagation involves propagating the error signal backward through the network to fine-tune these weights and improve performance. There are, however, various learning rules available for ANNs(Bishop, 1995) :

1. Error Back-propagation (Delta Rule): This widely used rule for multi-layer networks propagates the error signal backward, adjusting weights to minimize the difference between desired and actual output.

2. Hebbian Learning: Inspired by biology, this unsupervised rule strengthens connections based on the activity of connected neurons.

3. Perceptron Learning Rule: A foundational supervised rule used in single-layer perceptrons, adjusting weights based on the desired and actual output of the network.

4. Widrow-Hoff Learning Rule: An adaptation of the Perceptron rule for linear networks, adjusting weights based on a learning rate and the error signal.

5. Winner-Take-All Learning Rule: Used in competitive networks, this unsupervised rule strengthens connections of the neuron with the highest activation, while weakening others.

III.2.5 Multilayer ANN Model

The chosen architecture for this multi dimension is a three-layer artificial neural network (ANN) as depicted in Figure III-4This network consists of:

- Input Layer: This layer has a single input node (x) representing the raw data fed into the network.

- Hidden Layer: This layer acts as the core processing unit, containing a variable number of neurons. Unlike the single input node, the number of hidden layer neurons isn't predetermined. A trial-and-error approach will be used to identify the optimal number for the task.

- Output Layer: The final layer houses a single output node, responsible for generating the network's prediction or classification based on the processed data.

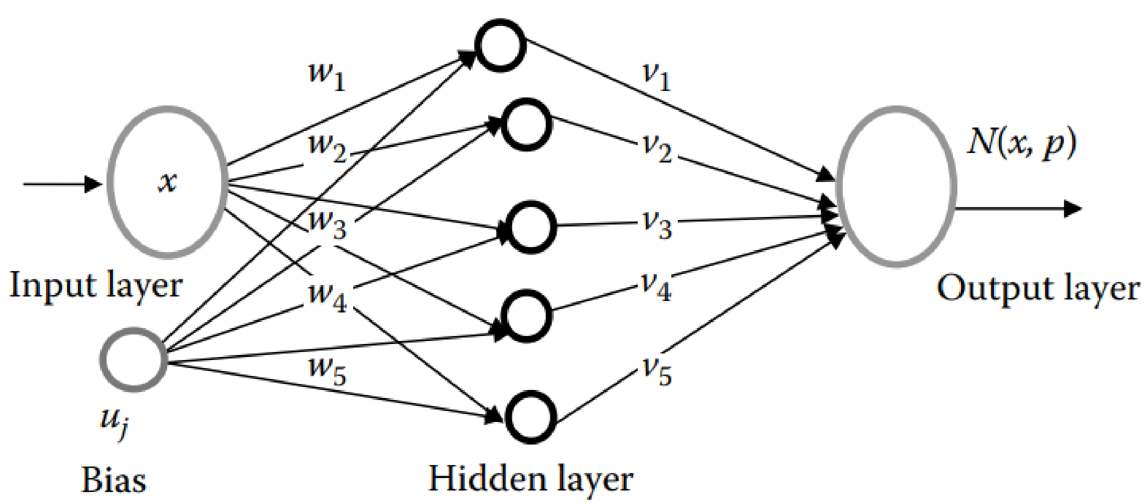


Figure III-4 : Structure of multilayer ANN.

III.2.5.1 Enhancing Learning: Biases and Weight Initialization

To improve the network's learning ability, biases (uj) are incorporated into each hidden layer neuron. These biases act like constant inputs, allowing the network to shift its activation function and create a wider range of possible outputs.

Furthermore, the initial weights (wj) connecting the input layer to the hidden layer and (vj) connecting the hidden layer to the output layer are assigned random values. This random initialization helps the network avoid getting stuck in local minima during training, allowing it to explore the solution space more effectively.(Nielsen, 2015)

III.2.5.2 Optimizing the Hidden Layer

The number of neurons within the hidden layer plays a crucial role in the network's performance. A limited number of neurons might restrict the network's ability to learn complex patterns, while too many could lead to overfitting. To address this, a trial-and-error method will be employed. Starting with a small number of hidden layer neurons, the network's performance will be evaluated on a validation set as the number is gradually increased.

The validation set is a separate data pool used to monitor the network's ability to learn and generalize without influencing the training process. By monitoring the validation performance, the optimal number of hidden layer neurons that achieves the best balance between learning and generalization can be identified.(Nielsen, 2015)

This approach allows for fine-tuning the network architecture for the specific problem at hand, maximizing its potential to achieve accurate results.

Chapter IV: Practical Part

IV.1 The methodology

The methodology initiates with the acquisition of field data, succeeded by meticulous data processing to secure a resilient dataset. The process of feature selection entails scrutinizing correlation matrices to pinpoint pivotal variables. During the model training stage, appropriate machine learning algorithms are selected, the dataset is partitioned into training, validation, and testing subsets, and model performance is enhanced via hyperparameter adjustments. Subsequent to the evaluation of the model's precision, should the outcomes be suboptimal, the procedure cycles back to data processing for additional enhancements. This systematic strategy guarantees dependable and actionable forecasts for discerning the lithology of formations, aiding in informed decision-making as the Figure IV-1 shows.

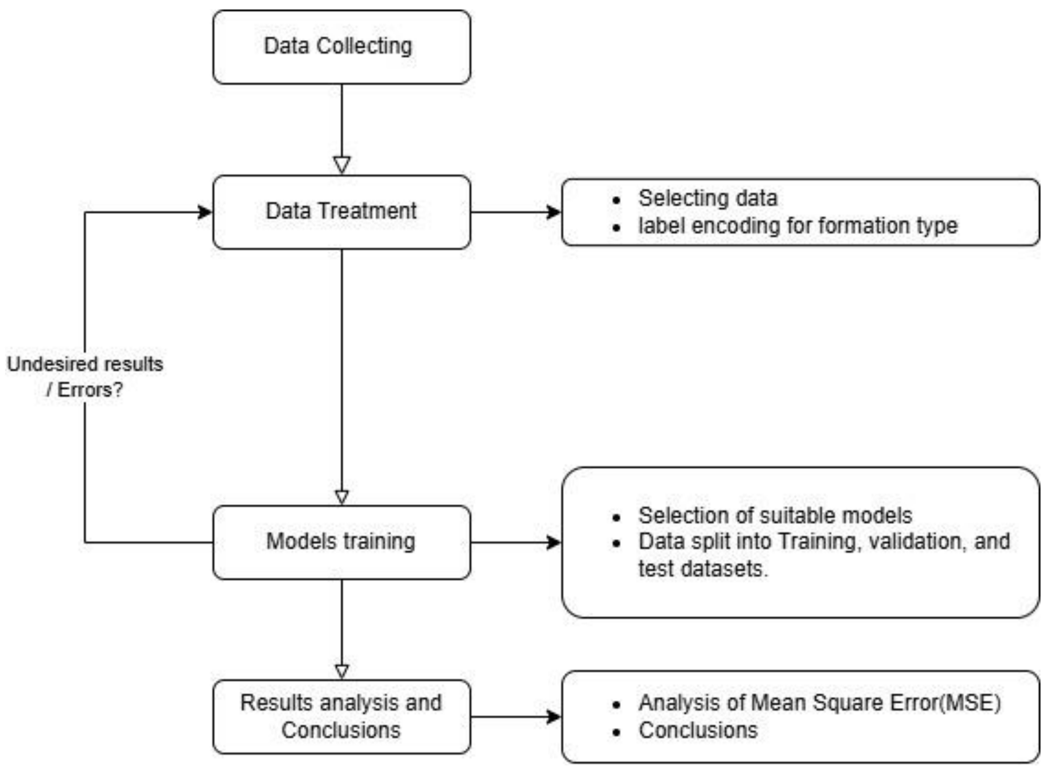


Figure IV-1 : Workflow for Predicting Formation Type Using Machine Learning

IV.2 Data preparation and treatment

The pre-processing stage is a critical component of machine learning projects, where raw data is transformed into a structured format amenable to ML algorithms. During this phase, we aim to reveal underlying patterns and rectify any inconsistencies or errors within the data, preparing it for subsequent machine learning models. The data, imported as an Excel file, pertains to various wells and is organized in Table IV.1

Table IV-1:Features description

Feature	Description	Unit
Name	Formation type	/
Depth	Depth at which drilling is taking place	m
RPM	Rotation per minute at drill bit	rpm
WOB	Weight applied to the drill bit	tons
TORQUE	Average surface torque	Ft*Lb
SPP	The mud pressure	psi
FLWpmps	mud flow rate	l/mn
ROP	Average rate of penetration	m/h

IV.2.1 Data treatment

To treate our dataset and address the unusual values, we first need to identify and remove the outliers. The describe() function in pandas reveals that while the percentiles seem reasonable as showed

	TOT_Depth	ROP	RPM	WOB	SPP	TORQUE	FLW_pmps
count	1195.000000	1195.000000	1195.000000	1195.000000	1195.000000	1195.000000	1195.000000
mean	660.148117	34.063598	122.116318	8.363180	1193.230126	6945.056067	2830.788285
std	376.737670	23.425417	38.777604	5.782952	254.598481	4711.915175	291.103091
min	13.000000	0.000000	36.000000	0.000000	0.000000	179.000000	605.000000
25%	311.500000	11.100000	85.000000	3.000000	1096.000000	1118.000000	2620.500000
50%	688.000000	32.600000	146.000000	8.000000	1268.000000	8260.000000	2911.000000
75%	986.500000	49.250000	151.000000	12.000000	1356.000000	10806.500000	2951.000000
max	1285.000000	116.400000	165.000000	25.000000	1524.000000	15044.000000	3387.000000

Table IV-2:Features statistics

Next, we have written a python function that counts the number of negative values of our dataset, Since negative values for WOB are not physically meaningful, we will remove any rows with WOB less than 0.

IV.2.2 Label encoding formation names

Since our data is categorized and distinguished by the formation name feature, one-hot encoding can be particularly beneficial for this feature for several reasons. It allows machine learning algorithms to effectively handle categorical variables, enables the model to capture potential interactions between different formation names, and facilitates the assessment of each formation's importance in influencing the target variable.

In this process, we used the LabelEncoder library in Python to streamline the work. Ultimately, we encoded each dataset with a digit instead of a formation name, with the first formation being encoded as the number 1, and so forth. (Figure IV-2)

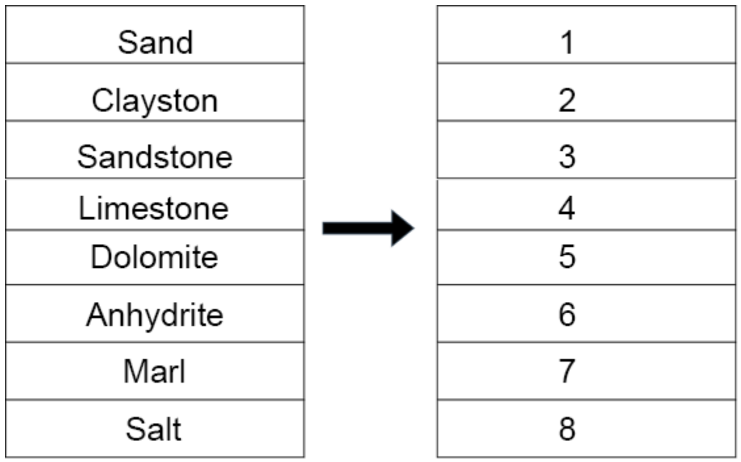


Figure IV-2 : Label encoding for formation name

IV.2.3 Selection of Test Algorithms

To select the most suitable model for our training, we wrote a code that evaluates the performance of several regression algorithms on our data. The model with the lowest mean squared error (MSE) was identified as the best fit for predicting our target variable.

The regression algorithms that are used are:

- LinearRegression
- xgboost.XGBRegressor (from XGBoost library)

IV.2.4 Data Splitting

In addition, to ensures the model is evaluated on unseen data, we have to divide the data into training and testing sets using a technique like “train_test_split” from scikit-learn, we used the recommended data split percentage which is:

- 60% for training data
- 20% for validation

- 20% for testing data

where the testing data is used for studying how much is our model is accurate when trained on training data.

IV.3 Model training and results

To calculate the Mean Squared Error (MSE) for each formation, we need to loop through each formation and compute the mean MSE. The training results, presented in the table, indicate that XGBoost is the best model for our study due to its lowest MSE. In contrast, Linear Regression have higher MSE values, leading to less accurate models.

To accurately determine the Mean Squared Error (MSE) for each formation, we need to systematically loop through each formation and compute the mean MSE. This process involves calculating the individual MSE for each data point within the formation and then averaging these values to obtain a representative mean MSE for the formation.

The training results, presented in the accompanying table, reveal significant insights into the performance of various models. Among the models evaluated, XGBoost emerges as the most effective model for our study. This conclusion is based on its remarkably low MSE, indicating superior predictive accuracy and reliability. The lower the MSE, the closer the predicted values are to the actual values, signifying a more accurate model.

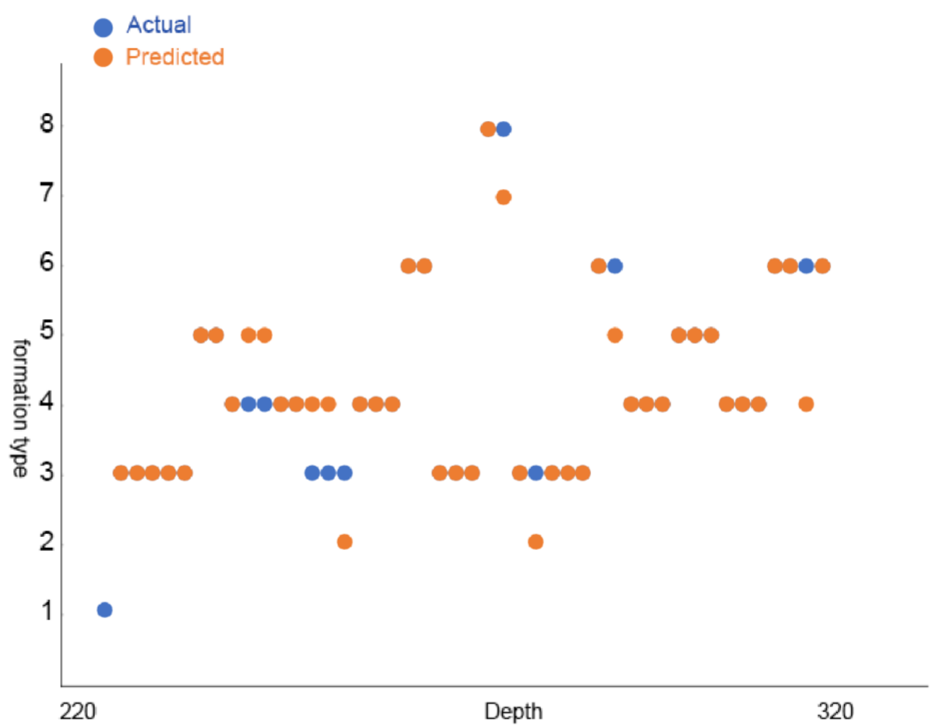
In contrast, the other models, specifically Linear Regression, exhibit considerably higher MSE values. These elevated MSE values suggest that these models have greater prediction errors, making them less reliable and less accurate compared to XGBoost.

Therefore, based on the comprehensive analysis of the MSE values, XGBoost is identified as the best model for our study. It consistently demonstrates the lowest MSE across various formations, reinforcing its status as the most accurate and dependable model for our predictive tasks.

Table IV-3 : MSE comparison of different models

Model	MSE
XGBoost	0.419
Linear Regression	7.514

Figure IV-3: Actual vs. Predicted formation



After developing our model using data as the foundation, we thoroughly evaluated its performance on the same dataset. The assessment involved comparing the predicted and actual formation type results, revealing the model's impressive accuracy in capturing lithological patterns (Figure IV 3). The model achieved a success rate of 70 %. These results demonstrate the model's ability to learn meaningful relationships from the provided drilling parameters and effectively predict the majority of formation.

Conclusions

Conclusion

In thiswork , we investigated the use of Artificial Intelligence (AI), specifically through Artificial Neural Networks (ANNs), for real-time identification of formation types using drilling data. Our objective was to develop a model that accurately predicts formation types, enhancing drilling operations and decision-making. Accurate predictions are essential for informed drilling strategies, mitigating risks, and reducing costs.

Our ANN model achieved a promising 70% prediction accuracy, demonstrating its ability to identify patterns in drilling data. While this is commendable, there is potential for improvement through advancements in neural network architectures and training methodologies. Enhancing data preprocessing and incorporating more varied datasets can further refine the model's accuracy, making it a vital tool for the drilling industry.

This research underscores AI's transformative role in revolutionizing drilling operations by providing precise, real-time predictions. Machine learning, particularly ANNs, can significantly improve drilling efficiency, cost-effectiveness, and sustainability.

Finally, the "Ama" (AI Mud Logger Assistant) software can offer even more effective assistance by incorporating additional data, such as lithological and geophysical information. This will make predictions more specific, potentially enabling precise reservoir location and further enhancing drilling efficiency. The success of our AI model in determining reservoir tops in real time marks a significant advancement in drilling technology and precision.

References

Abou-Kassem, J. H. (Jamal H., Farouq Ali, S. M. (Syed M., & Islam, R. (2006). *Petroleum reservoir simulation : a basic approach*. Gulf Pub. Co.

Ahmed, T. (2010). *Reservoir Engineering Handbook*. Elsevier Science.

Ahmed, T., & McKinney, P. (2011). *Advanced Reservoir Engineering*. Elsevier Science.

Arnold, K., & Stewart, M. (1998). *Surface Production Operations, Volume 1:: Design of Oil-Handling Systems and Facilities*. Elsevier Science.

Berry, M. J. A., & Linoff, G. S. (2008). *MASTERING DATA MINING: THE ART AND SCIENCE OF CUSTOMER RELATIONSHIP MANAGEMENT*. Wiley India Pvt. Limited.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press.

Cossé, R. (1993). *Basics of Reservoir Engi...* Editions OPHRYS.

Eberhart, R. C., Simpson, P. K., Dobbins, R. C., & Dobbins, R. W. (1996). *Computational Intelligence PC Tools*. AP Professional.

Hans, V., Yuing, X., & Ian, P. (2002). Assessment Of Several Sand Prediction Models With Particular Reference To HPHT Wells. *Proceedings of the SPE/ISRM Rock Mechanics in Petroleum Engineering Conference*. https://doi.org/10.2118/78235-MS

KECHAR. (2020). *ETUDE DES CHITINOZOAIRES DE L'ORDOVICIEN. EXEMPLE DU PUITS DE BORDJ NILI-2 (NL-2) DANS LE BASSIN D' OUED MYA*.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Moitra, A. K., Bhattacharya, J., Kayal, J. R., Mukerji, B., & Das, A. K. (2021). *Innovative Exploration Methods for Minerals, Oil, Gas, and Groundwater for Sustainable Development*. Elsevier Science. https://books.google.dz/books?id=TEY6EAAAQBAJ

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Determination Press.

Nouri, A., Vaziri, H., Belhaj, H., & Islam, R. (2004). Sand Production Prediction: A New Set of Criteria for Modeling Based on Large-Scale Transient Experiments and Numerical Investigation. *SPE Journal*, *11*. https://doi.org/10.2118/90273-MS

Odeh, A. S. (1969). Reservoir Simulation ...What Is It. *Journal of Petroleum Technology*, *21*(11), 1383–1388. https://doi.org/10.2118/2790-PA

Scott, J. (2008). *Method and System for Automated Choke Control on a Hydrocarbon Producing Well*.

Shahab D. Mohaghegh. (2017). *Data-Driven Reservoir Modeling*. Society of Petroleum Engineers.

Terry, R. E., Rogers, J. B., & Craft, B. C. (2015). *Applied Petroleum Reservoir Engineering*. Prentice Hall.

TRABELSI, K. G. (2019). *Caractérisation Pétro-physique d'un Réservoir cambroordovicien de la zone 13 du champ HMD Par l'utilisation de Diagraphies et des Mesures sur Carottes*.

Zurada, J. M. (2004). *Introduction to Artificial Neural Systems*. Jaico Publishing House.

Zurada, J. M., Marks, R. J., & Robinson, C. J. (1994). *Computational Intelligence: Imitating Life*. IEEE.

Appendices

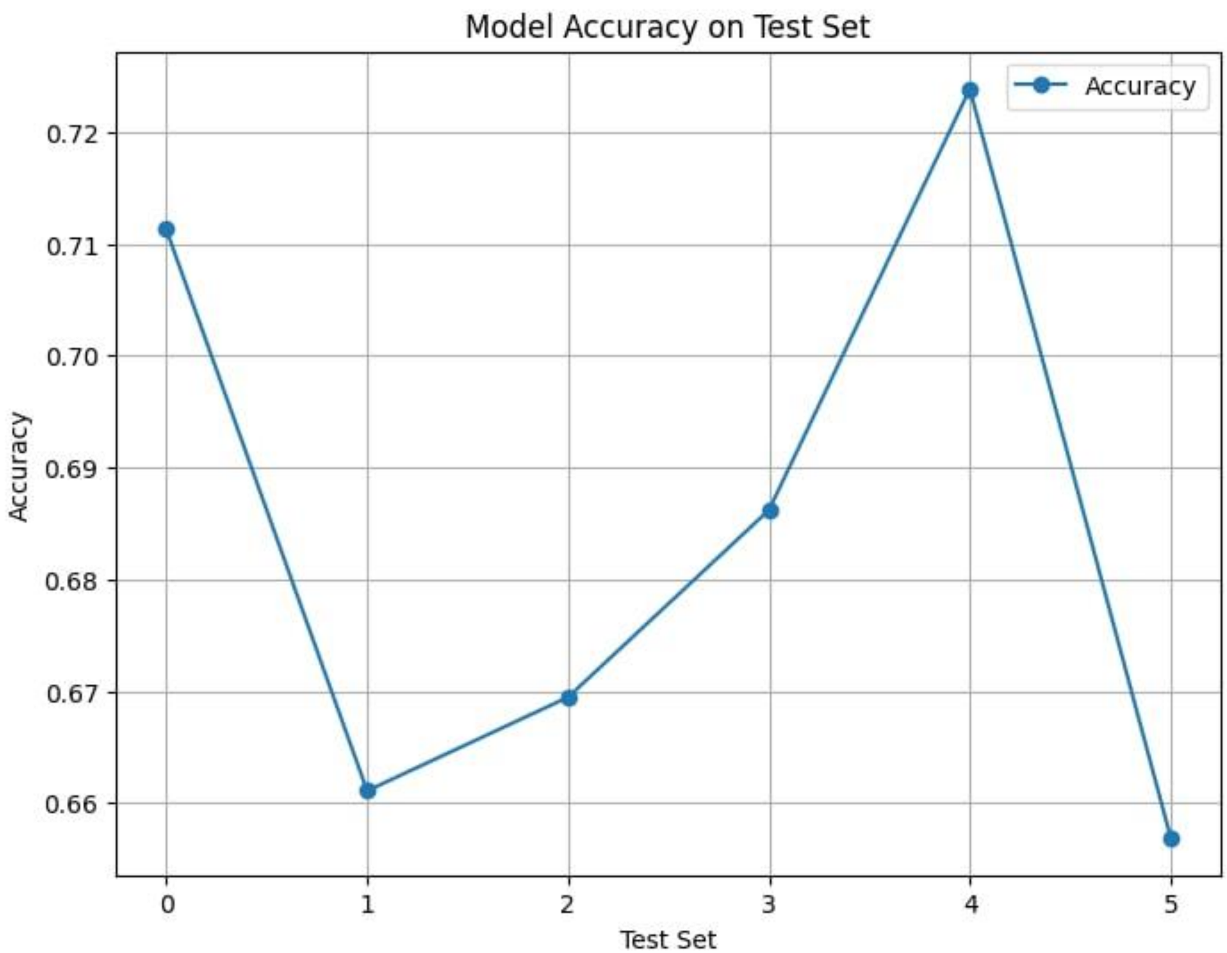
```python
import numpy as np
from IPython.display import Markdown, display
import pandas as pd
from sklearn.model_selection import train_test_split, StratifiedKFold
from xgboost import XGBClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, confusion_matrix
import matplotlib.pyplot as plt
data = pd.read_excel('https://onedrive.live.com/download?resid=DDFFB4F4FF07E504%214749&authkey=!ADE2v_B3JIY0Rdg&em=2"')
```

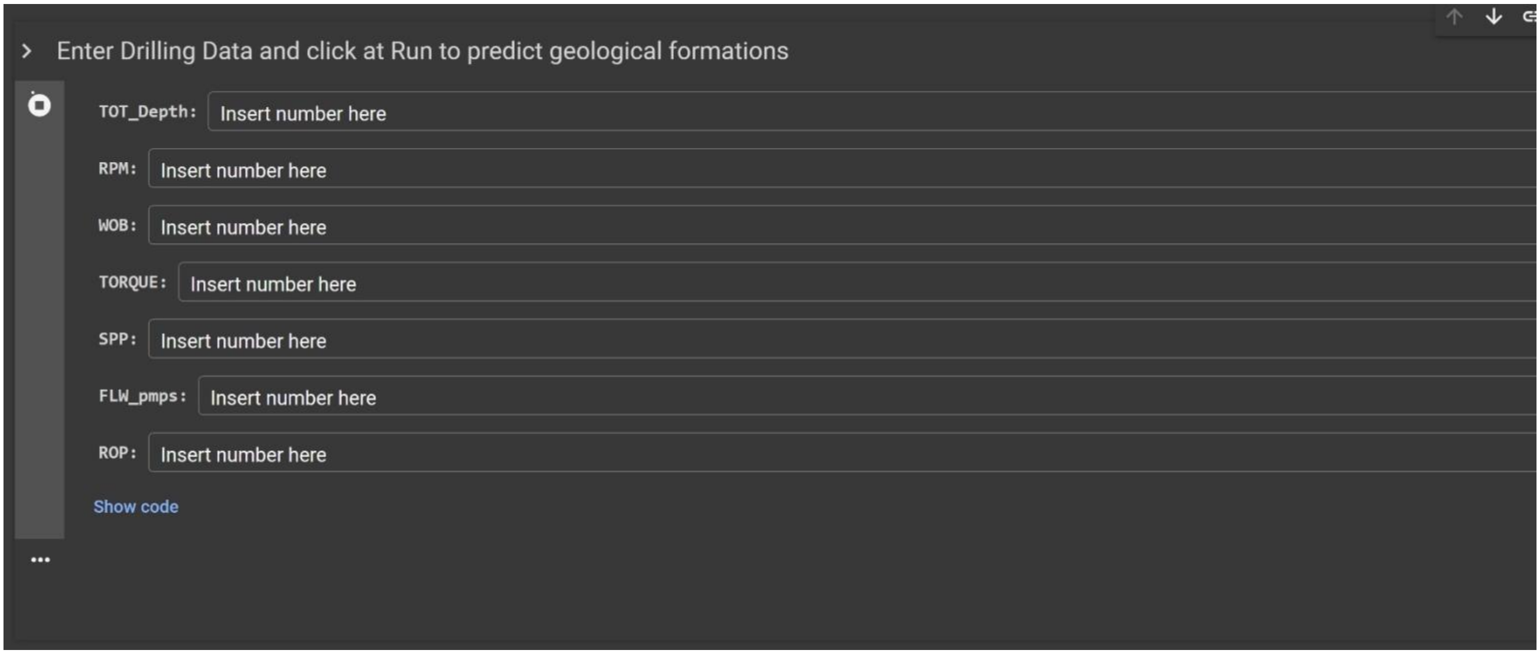
```python
features = ["TOT_Depth", "RPM", "WOB", "TORQUE", "SPP", "FLW_pmps", "ROP"]
target = "Description_geologique"

encoder = LabelEncoder()
data[target] = encoder.fit_transform(data[target])
X = data[features]
y = data[target]
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

accuracy_scores = []
for train_index, val_index in skf.split(X, y):
    X_train, X_val = X.iloc[train_index], X.iloc[val_index]
    y_train, y_val = y.iloc[train_index], y.iloc[val_index]

    model = XGBClassifier(objective="multi:class", num_class=len(encoder.classes_))
    model.fit(X_train, y_train)

    y_pred_val = model.predict(X_val)
    accuracy = accuracy_score(y_val, y_pred_val)
    accuracy_scores.append(accuracy)
```

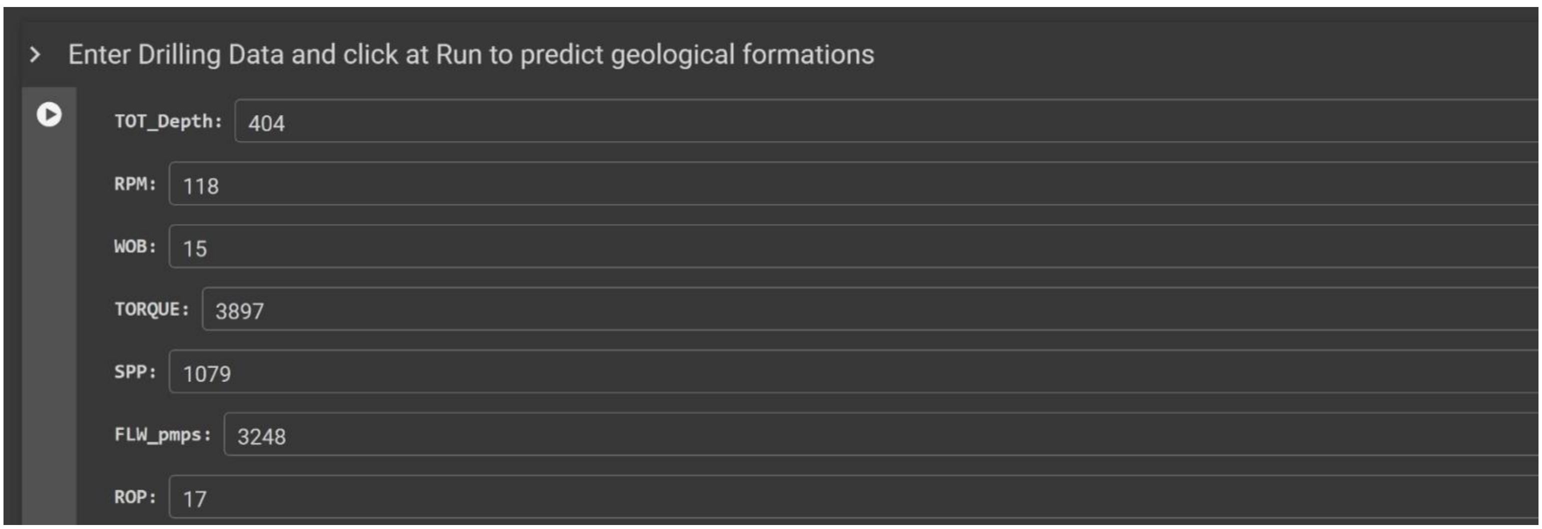
```python
average_accuracy = np.mean(accuracy_scores)
std_dev = np.std(accuracy_scores)
# print(f"Average Accuracy: {average_accuracy:.4f}")

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = XGBClassifier(objective="multi:class", num_class=len(encoder.classes_))
model.fit(X_train, y_train)

y_pred_test = model.predict(X_test)
test_accuracy = accuracy_score(y_test, y_pred_test)
print(f"Test Set Accuracy: {test_accuracy:.4f}")

def has_zeros(data):
  for value in data.values():
    if value == 0:
      return True
  return False

# @title Enter Drilling Data and click at Run to predict geological formations
TOT_Depth = 222  # @param {type:"number"}
RPM = 88  # @param {type:"number"}
WOB = 1  # @param {type:"number"}
TORQUE = 470  # @param {type:"number"}
SPP = 1182  # @param {type:"number"}
FLW_pmps = 3090  # @param {type:"number"}
ROP = 5  # @param {type:"number"}

test_sample = {
    "TOT_Depth": TOT_Depth,
    "RPM": RPM,
    "WOB": WOB,
    "TORQUE": TORQUE,
    "SPP": SPP,
    "FLW_pmps": FLW_pmps,
    "ROP": ROP,
}
```

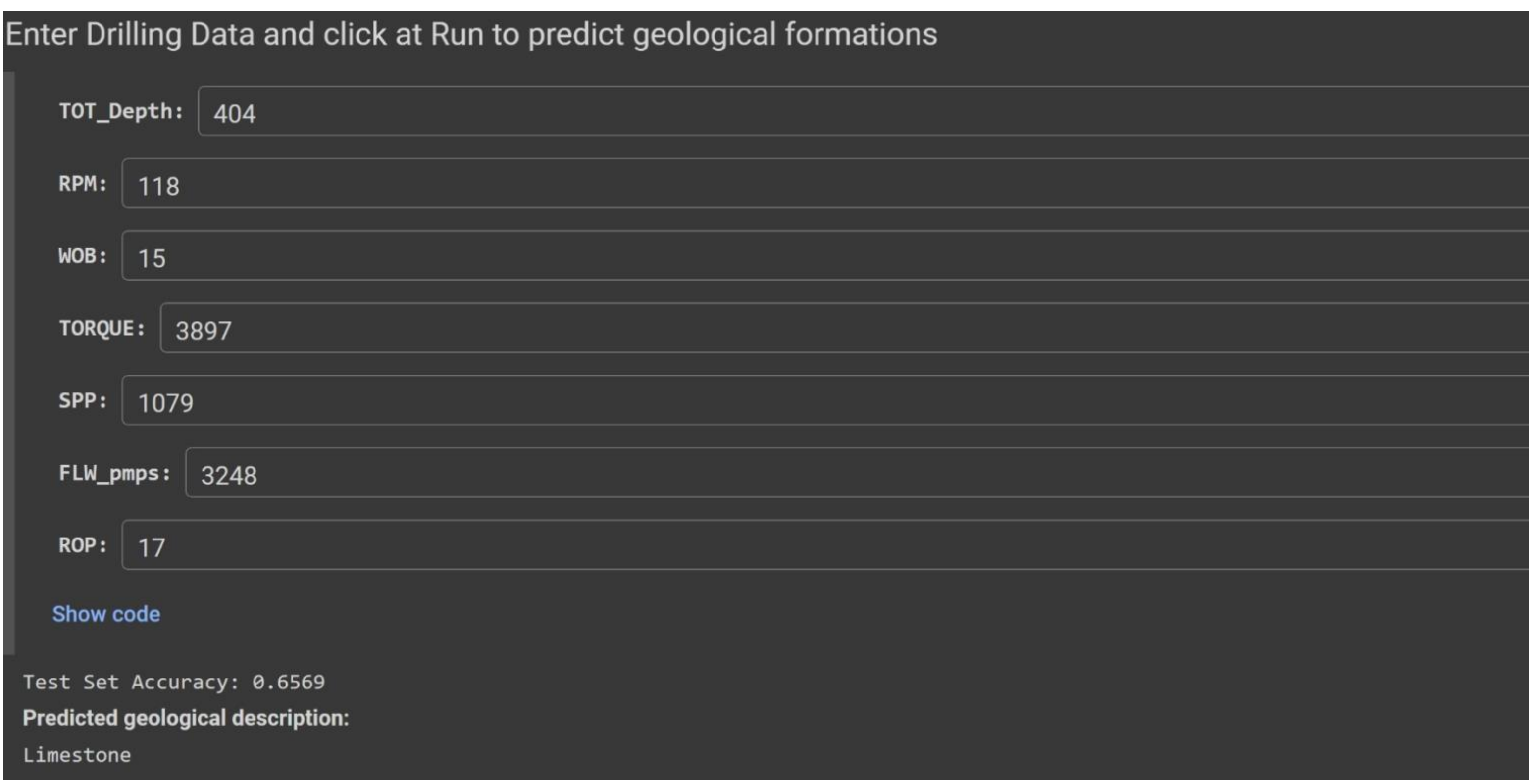
```python
else:
  test_sample_array = np.array([list(test_sample.values())])
  prediction = model.predict(test_sample_array)
  predicted_description = desc.iloc[round(encoder.inverse_transform(prediction)[0])-1]["descreption"]
  def printmd(string):
     display(Markdown(string))
  printmd('**Predicted geological description:**')
  print(predicted_description)
```

```python
plt.figure(figsize=(8, 6))
accuracy_scores.append(test_accuracy)
plt.plot(accuracy_scores, marker='o', label='Accuracy')
plt.xlabel('Test Set')
plt.ylabel('Accuracy')
plt.title('Model Accuracy on Test Set')
plt.grid(True)
plt.legend()
plt.show()
```




Enter Drilling Data and click at Run to predict geological formations
TOT_Depth: Insert number here
RPM: Insert number here
WOB: Insert number here
TORQUE: Insert number here
SPP: Insert number here
FLW_pmps: Insert number here
ROP: Insert number here
Show code


Enter Drilling Data and click at Run to predict geological formations
TOT_Depth: 404
RPM: 118
WOB: 15
TORQUE: 3897
SPP: 1079
FLW_pmps: 3248
ROP: 17


Enter Drilling Data and click at Run to predict geological formations
TOT_Depth: 404
RPM: 118
WOB: 15
TORQUE: 3897
SPP: 1079
FLW_pmps: 3248
ROP: 17
Show code
Test Set Accuracy: 0.6569
Predicted geological description:
Limestone