Master's Thesis :

APPLICATION OF MACHINE LEARNING USING WELL LOGGING DATA TO PREDICT RESERVOIR PROPERTIES, SIRT BASIN

BY: REEMA OMAR KHALLEEFAH

منشورات البرنامج الوطني لتطوير المجلات العلمية

The Libyan Academy –Ajdabiya Branch Libya

School of Applied Sciences and Engineering

Department of Chemical and Petroleum Engineering

Application of Machine Learning Using Well Logging Data to Predict Reservoir Properties, Sirt Basin

By

Reema Omar Mohammed Khalleefah

Dr. Adel Alzwi

Associate Professor

Thesis Was Submitted in Partial Fulfillment of The Requirement for The Degree of Master of Science in Petroleum Engineering

At the Libyan Academy-Ajdabiya Branch Libya

On- 13\ January\ 2025

الإقرار

أقرّ أنا، ريما عمر محمد خليفة، بأن ما اشتملت عليه هذه الرسالة هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة، ككل أو جزء منها، لم تُقدَّم لنيل أي درجة علمية أو بحث علمي لدى أي مؤسسة تعليمية أو بحثية أخرى. كما أن للأكاديمية حق توظيف الرسالة أو الأطروحة والاستفادة منها كمصدر مرجعي للمعلومات، ولا أعارض الاطلاع عليها أو نشرها، بما لا يتعارض مع حقوق الملكية الفكرية المقررة وفق التشريعات النافذة.

التوقيع.....

التاريخ: 13\1\2025م.

Declaration

I, Reema Omar Mohammed Khalleefah, the undersigned, hereby confirm that the work contained in this thesis, unless otherwise referenced, is my own and has not been previously submitted to meet the requirements of an award at this Academy or any other higher education or research institution. Furthermore, I cede the copyright of this thesis in favor of the Libyan Academy.

Name: Reema Omar Mohammed Khalleefah (213041)

Signature:……

Date: 13-01-2025

ACKNOWLEDGEMENT

I shall begin by thanking ALLAH, the Almighty: without His will, I would have never found the right path. His mercy was with me throughout my life and ever more in this study. I thank Him for enlightening my soul.

I want to express my sincere gratitude to my research supervisor, Dr. Adel, for allowing me to conduct this research and providing invaluable guidance throughout the process. Dr. Adel's enthusiasm, vision, and dedication have been a constant source of inspiration. Their expertise in research methodology has been instrumental in shaping the direction and clarity of this work. Learning from and contributing to Dr. Adel's research environment has been a privilege.

I am obliged to staff members of (AGOCO)&(SOC), for the valuable information provided to them in their respective fields. I am grateful for their cooperation during the period of my assignment.

We would like to Mrs. Entisar Al-Baghdadi from the Information Technology Department for their invaluable support and guidance throughout this project. Additionally, we extend our heartfelt thanks to the faculty members of Petroleum Engineering for their constant encouragement and mentorship, which were instrumental in the successful completion of this work.

I am extremely grateful to my mom for her love, prayers, caring, and sacrifices in educating and preparing me for my future. In addition, I express my thanks to my sisters and brothers for their support, and Special thanks go to my friends for the keen interest shown in completing this thesis successfully.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Table of Contents

List of Figure

List of Tables

NOMENCLATURES

Symbol	Description	Unit
RCA	Routine Core Analysis	-
ML	Machine Learning	-
BOPD	Barrel oil per day	-
AI	artificial intelligence	-
mD	millidarcy’s	-
D	Darcey’s	-
HVPV	Hydrocarbon pore volume.	-
Swirr	irreducible (connate) water saturation	fraction
Sw	water saturation	fraction
So	Oil saturation	fraction
Sg	Gas saturation	fraction
CT	computerized tomography	-
NMR	Nuclear magnetic resonance	-
Cp	compaction factor	-
RHG	Raymer-Hunt-Gardner's	-
RMA	reduced major axis	-
OOIP	original oil in place	-
OGIP	original gas in place	-
CEC	cation exchange capacity	
SCAL	Special Core Analysis	-
ANN	artificial neural network	-
KNN	K-nearest neighbour	-
RF	random forest	-
DT	decision tree	-

GR	Gamma Ray	API
ILD	deep induction resistivity log	Ohm.m
ILM	Medium resistivity log	Ohm.m
RHOB	bulk density	g/cm3
DT	sonic transit time	μsec/ft
NPHI	neutron porosity	fraction
ANN	artificial neural networks	-
SVM	Support Vector Machine	-
ANFIS	Adaptive neuro-fuzzy inference system	-
R	correlation coefficient	-
AAPE	average absolute percentage error	-
GMDH	group method of data handling	-
GMDH-ED	group method of data handling- Differential evolution	-
SGR	spectral gamma-ray	API
RLA1	apparent resistivity focusing mode	Ohm.m
VSH	shale volume of rock	fraction
PHIE	effective porosity	fraction
PERFZ	photoelectric factor	barns per electron
RLA5	apparent resistivity focusing mode 5	
RHOZ	standard resolution formation density	g/cm3
CALI	caliper log	in
TNPH	thermal neutron porosity	fraction
MLR	multiple linear regression	-
RBF	Radial Basis Function Neural Network	-
AAE	average absolute error and	-
RMSE	root-mean-square error	-
E	experience	-

T	tasks	-
P	Performance	-
RBF	radial basis function	-
GAN	generative adversarial networks	-
LWD	logging while drilling	-
SP	self-potential	mV
CNL	compensated neutron log	fraction
EDA	Exploratory Data analysis	-
R^2	Square of Correlation Coefficient	-

ABSTRACT

Application of Machine Learning Using Well Logging Data to Predict Reservoir Properties, Sirt Basin

By

Reema Omar Mohammed Khalleefah

Supervisor:

Dr. Adel Alzwi

Prediction of reservoir properties such as porosity, permeability, and water saturation are critical for reservoir characterization, determining the original hydrocarbon in place, and optimizing hydrocarbon production. However, predicting these properties is challenging due to the uncertainty associated with the empirical equations used with petrophysical interpretation and the excessive cost and time-intensive associated with RCA.

The main objective of this study is to develop a time-efficient, economical, and reliable technique for predicting reservoir properties (phi, SW, and K). The focus is on selecting the optimum model and comparing machine learning results with the actual results from the traditional method. Two tree-based models, decision trees and random forests, have been developed for this purpose.

The Training dataset contains 734 points, and the blind dataset contains 202 points collected from eight wells in the Libyan oil field. These wells contain measured and calculated well-logging data (phi, SW, and K) that were correlated to core data.

The tree-based models were trained on the dataset used. The evaluation metric was conducted on the separate blind well (A-05).

The results of this study indicate that both models provide good results, but the Random Forest model demonstrates robust performance on blind tests.

Findings from this study using Machine Learning can provide a more accurate result, with the advantage of reducing time and cost compared with traditional methods. These findings align with the study's objective of identifying effective machine-learning models for reservoir properties prediction.

However, for future work, it’s important to use high data quality, use other features and more complex models to get more accurate results especially for permeability prediction.

Keywords: Reservoir Properties, Machine Learning, Well Logging, Mature Oil Field, Sirt Basin.

CHAPTER ONE

1. INTRODUCTION

1.1 Introduction

In the petroleum industry, porosity, permeability, and fluid saturation play a significant role in understanding the characteristics of reservoir formation. Precisely calculating these properties is the key parameter in determining the hydrocarbon volume. However, these properties cannot be available directly from petrophysical logs.

Routine Core Analysis (RCA) is a direct method that is used to determine the most accurate result of any reservoir properties. [1]Although RCA gives useful information, it is not always possible to core every well in the petroleum field. That is because RCA is limited by excessive cost and time intensive.[2] As a result, coring is usually reserved for rare situations, thereby leaving a significant number of wells without direct measurements.

To address this challenge, standard oil fields used the traditional method of well logging tools to analyse petrophysical parameters these methods are economical that is because these log data is collected during drilling and completion operations.[2]

The well-logging analysis is an indirect approach to determining reservoir parameters like porosity, permeability, and water saturation. Porosity is related to well logging including neutron log, sonic, and bulk density while fluid saturation is related to resistivity logs and permeability is related to sonic and nuclear magnetic resonance. However, the interpretation of these logs is difficult due to uncertainties surrounding empirical parameters and the adaptability of response equations.[3], [4] Therefore, it is significant to establish a more confident, cost-effective, and efficient time technique to predict reservoir parameters.

In recent years, Machine Learning (ML) has been used as a powerful data analysis approach which is used to develop rapid and reliable models to predict these properties. ML defined by Tom Mitchell, is enhanced by computer programs automatically from experience by studying computer algorithms.[5]

Using ML algorithms, it is possible to improve the accuracy and reliability of reservoir property estimations even in cases when direct measurements are not obtainable.

Figure 1. 1: ETP definition of machine learning (Mitchell, 1997)[5]

1.2 Statement of the Problem

The approach to this study problem starts with finding a method to obtain an accurate prediction of reservoir properties. The time and costs associated with traditional methods such as RCA and uncertainty related to well logging interpretation to calculate reservoir properties are the key challenges in reservoir engineering studies. Due to these challenges related to the traditional methods it's important to find more confidence and rapid methods to estimate accurate reservoir properties. With the advancement of machine learning techniques, can develop a reliable and fast model to predict reservoir properties based on well logging and core data. Accordingly, this study has two main goals. First, is to investigate and understand applications of machine learning for estimating the reservoir properties using well-logging data. Secondly, is to compare the results with traditional methods.

1.3 Objective of the Study

The main objective of this study is to develop and understand Machine Learning algorithms and use them to predict reservoir properties based on well logging and core data. This can be divided into:

1. Propose a more confident, economical and quick Machine Learning model using multiwell logging and core data.
2. This study focuses specifically on the usage of Machine Learning to estimate reservoir properties. Porosity, Permeability, and Fluid Saturation help the successful exploration and production of hydrocarbons.
3. Compare selected machine learning algorithms and determine the optimum prediction algorithms.

4. Comparison between Machine Learning models and traditional methods for estimating reservoir properties.

1.4 Study Area

A mature oil field is considered one of the largest fields in Libya in the Sirt basin. In the early 1960s, the field was discovered on a seismically defined structure. The first four years continued the development drilling then after completing the pipeline and loading terminal the production was started. The initial production rate of this field was about 8000BOPD, but in various were may reach a high rate of 20,000BOPD. However, to maintain the production the downhole pumps were used. Table 1. 1, describe the major features of the selected field.[6]

Table 1. 1: General information about the selected field.[6]

Criteria	Description
Basin	Sirt
Basin Type	Rifted Basin
Reservoir Rock Type	Sandstone
Environment Of Deposition	Deltaic
Reservoir Age	Cretaceous
Petroleum Type	Oil
Trap Type	Horst Block

Generally, on the western edge of the Calanscio Sand Sea in southern Cyrenaica lies this oil field. The upper Cretaceous-Tertiary Sirt basin contains all the major oil fields of Libya, including this one, which is located at the southeastern edge of it.[6]

This oil field has enough open-hole logging data which will facilitate achieving the study objectives. Appendix A shows a summary of the available data for each well in this study. Figure 1. 3, display the well location within the study area.

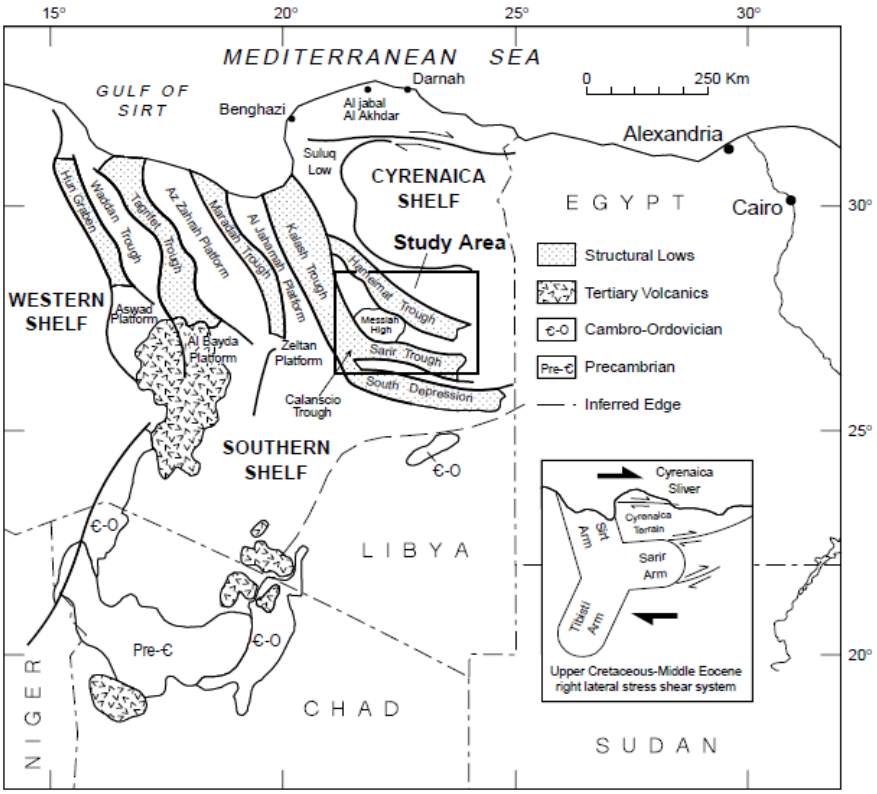

Figure 1. 2: Location of studied oil field - Sirt Basin, Libya.[7]

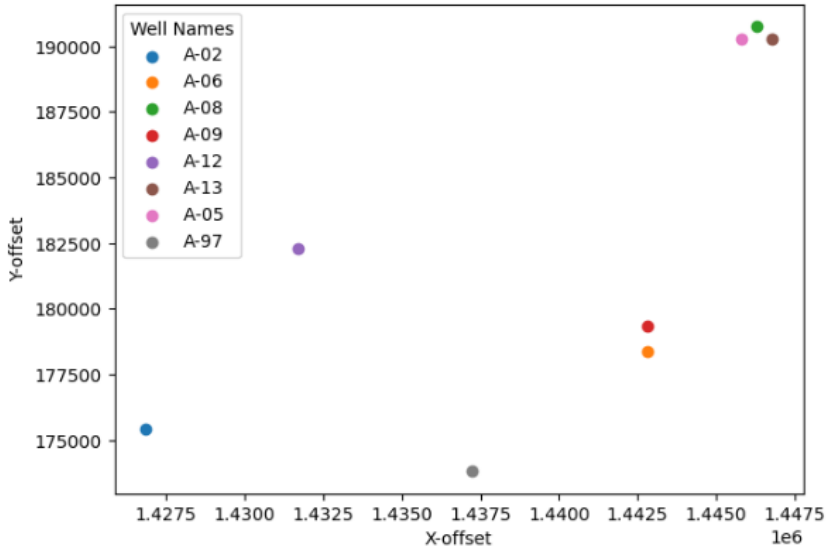

Figure 1. 3: Location of studied well within the study area.

1.4.1 Reservoir Formation

The main reservoir in this field is basal sandstone, it is a hitrogenite formation divided into five members, two of which are clean sandstone which separates the three shaly or tight sandstones. Thus, this distribution and differentiation in their properties affect the distribution of oil within the trap.[6] In the other meaning, this formation contained different interbedded such as sandstones, siltstones, shales and conglomerates.[8]

According to the previous study about the geology of the Sirt basin, this formation was dated from the Late Cretaceous. [7]

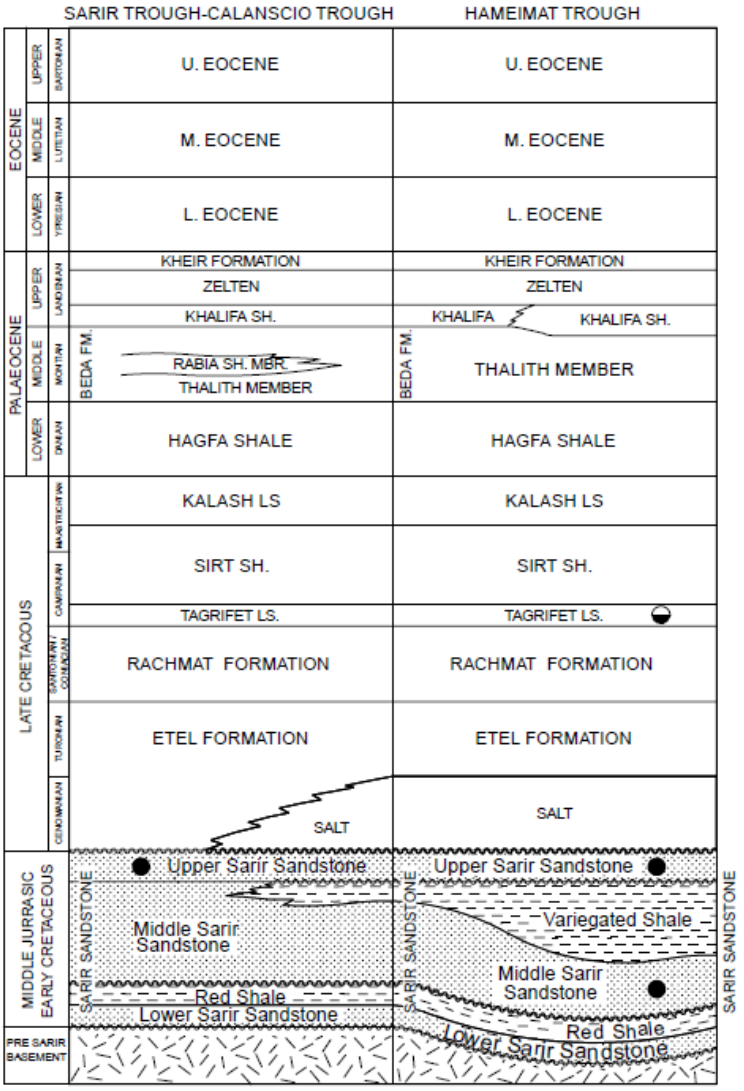

Figure 1. 4: stratigraphic column of the studied oil field, Sirt basin.*[7]*

1.5 Study Plan

This study will include the following chapters:

Chapter 1: introduction to the purpose and the objectives of this study.

Chapter 2: discusses the literature survey and previous work related to the point of research.

Chapter 3: discusses the well-logging approach and the machine-learning approach.

Chapter 4: exhibits the methodology of the study concerning the program.

Chapter 5: the obtained results will be discussed in detail.

Chapter 5: brief discussion of the conclusion of this study and recommendations for future study.

CHAPTER TWO

2. LITERATURE REVIEW

2.1 Introduction

Reservoir properties can be estimated by using different approaches. Some of these are direct approaches, such as Routine core analysis (RCA). However, there are other approaches considered indirect methods, like well-logging techniques. With the development of artificial intelligence (AI) in recent years, some algorithms of AI and ML have been used in petroleum engineering, especially in reservoir properties prediction. This chapter outlines the basis of reservoir properties and reviews the traditional method that is used to estimate these properties. Also will discuss the application of machine learning models in reservoir properties prediction.

2.2 Reservoir properties

2.2.1 Porosity

Porosity is a critical property in reservoir engineering. It is defined as the storage capacity of a reservoir.[9]Mathematically, porosity is defined as the ratio of the total pore volume to the bulk volume of the rock, representing the proportion of void spaces within the rock matrix. [10]Porosity can be calculated using Eq 2. 1.

$$\emptyset = \frac{Vp}{Vb} \qquad \textit{Eq 2. 1}$$

where:

Ø: porosity, fraction.

V_p: pore volume, cm^3.

V_b: bulk volume, cm^3.

However, it can calculate the reservoir porosity from the rock matrix and total bulk volume. Eq 2. 2 and Eq 2. 3 are used to calculate porosity when pore volume is unknown.[10]

$$\emptyset = \frac{Vp}{Vb} = \frac{Vb - Vm}{Vb} \qquad \textit{Eq 2. 2}$$

where:

V_m: rock matrix.cm3.

There are several classifications of porosity one of these classifications is mathematical classification. In this classification, Porosity is categorized into two main types: total porosity and effective porosity. Total porosity, also known as absolute porosity (Øt), total porosity refers to the ratio of the total void space within a rock to its bulk volume. At the same time, the effective porosity (Øe) is the portion of reservoir rock that is effective in storing the fluid without the effect of clay. However, there is another classification of porosity which is a geological classification. In this classification, porosity is classified as primary and secondary porosity. Primary porosity is developed during the deposition process. While secondary porosity is developed after deposition.[11]

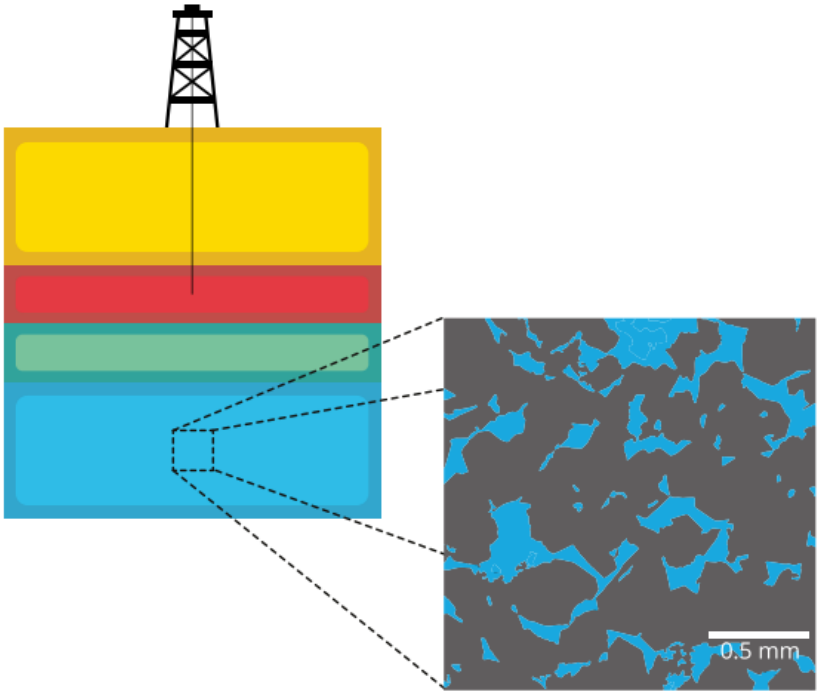

Figure 2. 1: Illustrate the pore space in a reservoir in a micro-scale. the blue color shows the water while the black color represents the matrix.*[10]*

2.2.2 Permeability

Permeability is an essential property in porous media. It is characterized as a rock's capacity to transmit fluid through its porous media.[12] Permeable formations such as sandstone typically have interconnected pores, which form permeable channels. However, other rocks, like shale, have tiny and less interconnected pores.

In 1856, Hanry Darcy introduced the first mathematical definition of permeability, which is called the Darcy law Eq 2. 3. Figure 2. 3 shows the Darcy experiment equipment.

$$q = \frac{kA\Delta P}{\mu L} \qquad \textit{Eq 2. 3}$$

q: steady-state flow, cm^3/sec.

k: proportionality constant, or permeability, Darcy's.

A: cross-section area, cm^2.

μ: viscosity of the flowing fluid, cp.

Δp: pressure drop, atm.

L: length of sample, cm.

By rearrangement of Eq 2. 3, the permeability can be computed by Eq 2. 4:

$$k = \frac{A\Delta P}{\mu q L} \qquad \textit{Eq 2. 4}$$

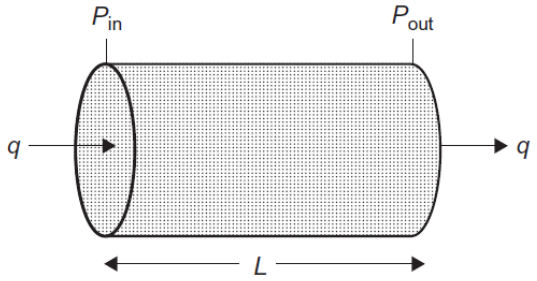

Figure 2. 2: core sample.[11]

The permeability is usually reported as millidarcy's (mD), but it's measured in Darcey's(D). The Eq 2. 4 referred to absolute permeability.

So, absolute permeability is when one fluid (single phase) flows in the rock. Effective permeability (Keff) it denotes the permeability when one fluid flows in the presence of another fluid in the reservoir. The relative permeability (Kr) is mathematically defined as the ratio between effective permeability and absolute permeability as shown in Eq 2. 5. It's also related to the flow of one or more fluids at the same time in the formation at a specific saturation. [12][13]

$$kr = \frac{keff}{k} \qquad \textbf{Eq 2. 5}$$

K_r: relative permeability.

K_{eff}: effective permeability, mD.

K: absolute permeability, mD.

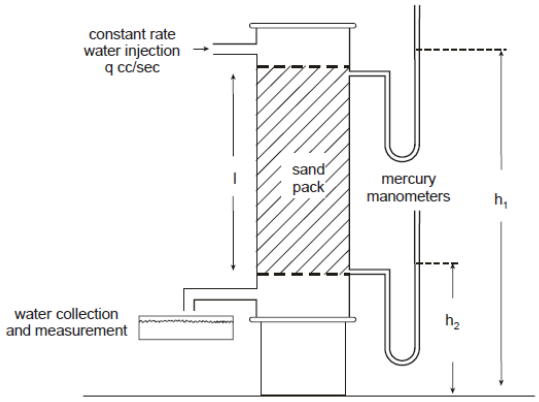

Figure 2. 3: Illustrate Darcy's experiment.[14]

2.2.3 Fluid Saturation

Fluid saturation especially water saturation is a crucial parameter used to estimate the original oil in place, which depends on the porosity and fluid saturation.

Fluid saturation is defined as the percentage of pore volume occupied by a specific fluid and is mathematically expressed in Eq 2. 6.[9]

$$fluid\ saturation = \frac{total\ volume\ of\ fluid\ phase}{pore\ volume} \qquad \textbf{Eq 2. 6}$$

The hydrocarbon pore volume refers to the total volume in the reservoir occupied by hydrocarbons.

$$HCPV = VtØ(1 - Swirr) \qquad \textit{Eq 2. 7}$$

Where:

HVPV: Hydrocarbon pore volume.

V_t: total volume.

S_{wirr} irreducible (connate) water saturation, fraction.

The saturation of rock refers to the interconnected pore space within a reservoir that is filled with specific fluids, such as oil, water, or gas.

$$Sw + So + Sg = 1 \qquad \textit{Eq 2. 8}$$

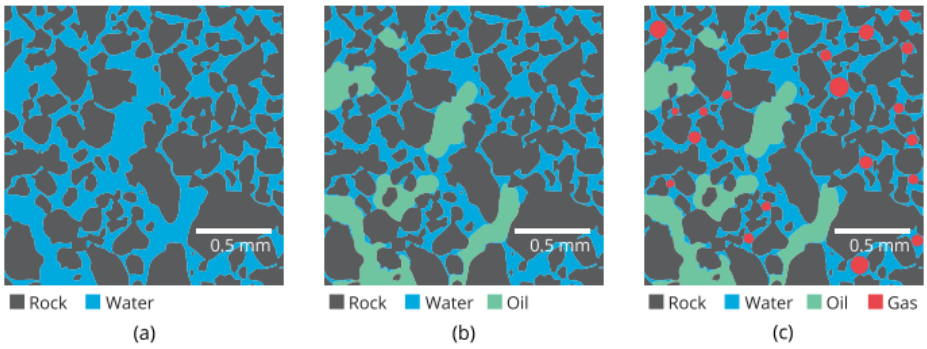

Figure 2. 4: Cross-section in micro-scale illustrates the pore space filled with fluid. (a) show the pore space filled with water (S_w=1),(b)represent the pore space filled with water and oil (S_w =1-S_o), (c) show the pore space that is filled with water oil and gas (Sw+So+Sg =1).[10]

2.3 Traditional methods

Reservoir properties can be determined by several methods. Porosity can be determined from neutron and density logging, sonic logging, Nuclear magnetic resonance (NMR) logging, computerized tomography (CT) scan, and the actual porosity obtained from core analysis in a laboratory.[15] while accurate permeability can be obtained from either core laboratory or well testing. However, permeability can be determined indirectly from well-logging data using empirical formulas.[16]

Water saturation is the main target of well-logging analysis, in which the resistivity logs are critical parameters for saturation determination for several empirical equations. Also, water saturation can be available directly from core data. [17]

This study will focus on using conventional logging techniques and correlated core analysis. Well-logging techniques use different logs and empirical correlations to estimate reservoir properties. On the other hand, the more accurate method to calculate these properties is by using laboratory measurements of a core sample.

2.3.1 Well-Logging Techniques

1 Porosity Determination

- **Neutron and Density combination**

The neutron and density logs are frequently used in well logging interpretation due to their wide range of applications. From neutron and density, cross-plot can estimate porosity, define lithology, and detect the gas zone from their crossover. [18]

Neutron and density combination is the most confident method for calculating porosity from well-logging techniques. However, complex lithology yields less accurate formation porosity. [19]

In gas bring formation both tools have opposite responses that cause the neutron and density porosity not to be equal. [20] Hydrocarbons like water contain hydrogen with different concentrations depending on the hydrocarbons' density in the reservoir. Gas has less hydrogen content per unit volume than liquids, leading to low neutron-derived porosity based on the hydrogen amount. On the other hand, the porosity derived from the density log is too high because the gas is less dense than liquid. [21]This principle has much effect on both neutron and density logging tool response. However, if both tools are displayed on compatible scales, they are widely separated in gas bring formation.[21] The neutron density porosity is affected by the presence of gas in a reservoir and therefore affects the fluid saturation and volume of hydrocarbons in place. Therefore, density porosity is calculated using the following equation:

$$\emptyset_D = \frac{\rho_{ma} - \rho_b}{\rho_{ma} - \rho_{fl}} \qquad \textit{Eq 2. 9}$$

$\emptyset_D$: Density porosity.

ρ_{ma}: matrix porosity, g/cm3.

ρ_b: formation bulk density, g/cm3.

ρ_{fl}: fluid density, g/cm3.

On the other hand, to calculate porosity in gas bring formation use the following equation:

$$\emptyset_{NDgas} = \sqrt{\frac{\emptyset_N^2 + \emptyset_D^2}{2}} \qquad \textit{Eq 2. 10}$$

$\emptyset_N$: neutron porosity.

- **Porosity From Sonic Log**

Porosity can be estimated from a sonic log using various methods, including Wyllie's time-average equation (1958) and Raymer-Hunt-Gardner's (RHG) equation (1980).

Eq 2. 11 illustrates the Wyllie time average used to calculate sonic porosity especially intergranular porosity in consolidated sandstones and carbonates. [22] On the other hand, porosity determined in carbonate formation from the Wyllie equation is too low due to the limitation of using the Wyllie equation to estimate secondary porosity from vuggy or fracture porosity. This limitation is because the sonic log measures matrix porosity rather than secondary porosity. [18]

However, the Wyllie equation has another limitation theoretical and practical. Theoretically, the model differs from reality and looks simplified. In practice, without introducing variable matrix transit times for single-mineral lithology, or in some cases, applying empirical correction factors for assumed under-compaction effects, there is often no satisfactory match between the porosities derived from the time average equation and those determined experimentally.[22]

$$\emptyset s = \frac{\Delta tlog - \Delta tmat}{\Delta tfl - \Delta tmat} \qquad \textit{Eq 2. 11}$$

where:

$Ø_{S:}$ sonic porosity, fraction.

Δt_{log}: interval transit time in the formation

Δt_{mat}: interval transit time in the matrix, Table 2. 1.

Δt_{fl}: interval transit time in the fluid in the formation (freshwater mud = 189 μsec/ft; saltwater mud = 185 μsec/ft)

In unconsolidated sands, it is essential to add an empirical compaction factor (Cp) to the Wyllie equation (1958) when calculating sonic porosity. As shown in Eq 2. 12 and Eq 2. 14

$$Øs = \left(\frac{\Delta tlog - \Delta tmat}{\Delta tfl - \Delta tmat}\right)\left(\frac{1}{cp}\right) \qquad \textbf{\textit{Eq 2. 12}}$$

where:

Cp = compaction factor.

The compaction factor is derived from the Eq 2. 13:

$$Cp = \frac{\Delta tsh * C}{100} \qquad \textbf{\textit{Eq 2. 13}}$$

Where:

Δ_{tsh} = represents the interval transit time in the shale adjacent to the formation of interest.

C = is a constant typically set to 1.0 (Hilchie,1978).

Table 2. 1: Sonic velocities and interval transit times for different matrix types are represented by constants used in the sonic porosity formulas mentioned above (after Schlumberger, 1972). *[18]*

Lithology/ Fluid	Matrix velocity ft/sec	Δtmatrix or Δtfluid (Wyllie) μsec/ft [μsec/m]	Δtmatrix (RHG) [μsec/ft [μsec/m]
Sandstone	18,000 to 19,500	55.5 to 51.0 [182 to 168]	56 [184]

Limestone	21,000 to 23,000	47.6 [156]	49 [161]
Dolomite	23,000 to 26,000	43.5 [143]	44 [144]
Anhydrite	20,000	50.0 [164]	-
Salt	15,000	66.7 [219]	-
Casing (iron)	17,500	57.0 [187]	-
Freshwater mud filtrate	5,280	189 [620]	-
Saltwater mud filtrate	5,980	185 [607]	-

In 1980 Raymer-Hunt-Gardner presented a time average equation (sonic porosity) transform, Eq 2. 14.[23] RHG equation provides more accurate results when matching with core results, especially in a carbonate reservoir. [24]

$$\emptyset = -a - \sqrt{[a2 + \left(\frac{\Delta tm}{\Delta t}\right) - 1]} \qquad \textit{Eq 2. 14}$$

Where:

$$a = \left(\frac{\Delta tm}{2\Delta tf}\right) - 1 \qquad \textit{Eq 2. 15}$$

2 Permeability Determination

To understand the fluid flow in petroleum reservoirs at different scales, it's essential to determine reservoir permeability accurately.[25] Reservoir permeability is generally high when the interconnected or effective porosity is high. Therefore, this leads to a higher production rate of petroleum reservoirs. Also, can help to detect production well location and optimum well path. [26]

Permeability from well-logging techniques is determined by using several empirical correlations. These correlations rely on porosity, permeability, and irreducible water saturation. [27]

- **Kozeny-Carmen**

Kozeny-Carmen 1927-1937 is the first empirical equation used to estimate permeability in early research. This equation assumes that permeability changes with porosity changes. Due to the porosity being a scalar property, it's adopted that the permeability changes are equal in all directions.[28], [29]

$$K = A_1 \frac{Ø^3}{So^2\ (1 - Ø)^2} \qquad \textit{Eq 2. 16}$$

Where:

A_1: The empirical constant, referred to as the Kozeny constant.

S_o: surface area per unit volume of solid material.

- **Tixier Model**

In 1949, Tixier introduced an empirical method for calculating permeability based on the resistivity gradient. Lateral or focused logs, are used to determine the resistivity gradient.

This model establishes the relationship between resistivity and water saturation, water saturation and capillary pressure, and capillary pressure and permeability. However, the application of this model is limited by poor density determination as it exists in a reservoir.

Also, there are no well logs that can detect exact oil-water contacts. On the other hand, the resistivity gradient for the average corresponding zone is used to calculate permeability.[27] [29]

$$K = C(a \frac{2.3}{\rho_w - \rho_o})^2 \qquad \textit{Eq 2. 17}$$

$$a = \frac{\Delta R}{\Delta D} \frac{1}{R_o} \qquad \textit{Eq 2. 18}$$

where:

C: It is a constant, typically around 20.

ΔR: It is the change in resistivity (ohm-m), and ΔD represents the change in depth (ft) corresponding to ΔR.

ρ_w: is formation water density (g/cm^3)

ρ_o: is hydrocarbon density (g/cm^3).

- **Timur**

The first introduction to the Timur equation was in 1968 based on the Kozeny equation. Timur used parameters A, B, and C these parameters were selected statistically. He utilized 155 sandstone samples from three distinct fields in North America. According to his laboratory measurements, a reduced major axis (RMA) method was used to analyse sandstone samples. He selected the relationship with the highest correlation coefficient and the lowest standard deviation from five identical permeability relationships, Eq 2. 19. The assumption for this equation is applicable When residual water saturation is present. Also, for the cementation factor (m), Timur proposed that is 1.5 for all cases.[27] [29]

$$K = 0.136\frac{Ø^{4.4}}{S_{wi}2} \qquad \textit{Eq 2. 19}$$

- **Coates & Dumanoir**

In 1974 Coates & Dumanoir introduced their equation that was proposed to be used in oil-bearing formation with specific oil density which was about 0.8 g/cm^3. Coates & Dumanoir adopted their exponents for saturation and cementation factors as w, which is the same value m = n = w by support core and log studies.[27] [29]

$$K^{1/2} = \frac{C}{w^4}\frac{Ø^{2w}}{R_w/R_{ti}} \qquad \textit{Eq 2. 19}$$

Where:

$$C = 23 + 465\rho_h - 188\rho_h 2 \qquad \textit{Eq 2. 20}$$

$$w^2 = (3.73 - Ø) + 0.5\left[\log_{10}\left(\frac{R_w}{R_t}\right) + 2.2\right]^2 \qquad \textit{Eq 2. 21}$$

This equation is applied when the formation is at irreducible water saturation. If the value of Rt is less than Rti, it indicates that the formation is not at irreducible water saturation, which results in an inaccurate estimation of the w value, Eq 2. 19,Eq 2. 20 and Eq 2. 21Coates & Dumanoir proposed a methodology to determine whether a formation is at irreducible water saturation based on their equation. Based on their methodology, they corrected the formations that were shaly or that were not at irreducible water saturation. According to this correction, this equation can be used for shaly formations or formations that are not at irreducible water saturation. [27][29]

- **Coates and Denoo**

Coates and Denoo in 1981, introduced their equation for permeability estimation. Similar to the other equations, this equation is used when permeability is zero at zero porosity and when Swirr = 100%. The formation must be at irreducible water saturation.[27] [29]

$$K^{1/2} = 100\frac{Ø^2(1 - S_{wirr})}{S_{wirr}} \qquad \textit{Eq 2. 20}$$

3 Fluid Saturation Determination

Water saturation (S_w) is the critical parameter in reservoir and petrophysical studies. It assists in obtaining more precise estimates for the original oil in place (OOIP) or original gas in place (OGIP). As well as other reservoir properties, Routine core analysis (RCA) or well-logging method can estimate water saturation. Using well logs are more conventional method for determining water saturation, well logs are simple and more available than core measurements.[30] Different empirical models can be used based on well-logging methods such as the Archie model, Simandoux model, Indonesia model, etc.., it depends on if the formation is clean or shaly formation.

- **Archie model**

In 1942, Archie introduced his equation, which is based on the relationship between porosity (Ø), rock resistivity (Rt), and water saturation (Sw). Also, he proposed other constant

parameters a, n, and m depending on reservoir rock, Eq 2. 21. However, he showed that rock resistivity when fully saturated with brine R_w, R_o, is related to R_w, Eq 2. 22. [31]

Archie assumes that his model is used for clean sandstone formation. Archie proposed that the rock matrix is non-conductive. Furthermore, the existence of shale in the reservoir zone influences the use of the Archie model and gives an overestimation of water due to shale conductivity. [32]

Therefore, the Archie model does not apply to shaly-sand formation. However, based on Archie's models several models were developed which consider the presence of shale in the reservoir.

$$Sw^n = \frac{a + R_w}{\emptyset^m * R_t} \qquad \textbf{Eq 2. 21}$$

Where:

a: tortuosity of the rock

m: Cementation exponent

n: saturation exponent

∅: Porosity

Sw: Formation water saturation

Rw: Resistivity of formation water

Rt: Formation resistivity

$$Ro = FRw \qquad \textbf{Eq 2. 22}$$

Where F is the formation resistivity factor.

- **Indonesia equation.**

In 1971 Poupon and Leveaux introduced their equation for shaly sand reservoir with freshwater salinity less than 20,000 NaCl ppm. Like other equations, this one relies on the true resistivity of the formation (Rt). however. Shale resistivity (R_{sh}) determines the shale zone from the resistivity log. [33]

$$\frac{1}{Rt} = Sw^{n/2}\left[\frac{Vsh^{(1-\left(\frac{Vsh}{2}\right))}}{\sqrt{Rsh}} + \left(\frac{\emptyset e^{\frac{m}{2}}}{\sqrt{a * Rw}}\right)\right] \qquad \textit{Eq 2. 23}$$

Where:

Rt: true (deep) resistivity in the un-invaded zone (Ohm-m)

Sw: un-invaded zone water saturation (fraction)

n: saturation exponent

V_{sh}: volume of shale (fraction)

R_{sh}: resistivity of shale (Ohm-m)

$Ø_e$: effective porosity (fraction)

m: cementation factor

a: constant multiplier

R_w: formation water resistivity (Ohm-m).

- **The Simandoux model**

In 1963 Simandoux developed his model which proposed the influence of shale volume on reduction in rock matrix conductivity and water saturation. He conducted experiments to examine the effect of shale volume on homogeneous mixtures of sorted sand and natural clay in different proportions. [33][34]

$$Sw = \frac{a * R_w * (1 - V_{sh})}{2Ø^m}\left[-\frac{V_{sh}}{R_{sh}}\right] + \left\{\sqrt{\left(\frac{VSh}{Rsh}\right)^2 + \left(\frac{4Ø^m}{a * R_w * R_t * (1 - V_{sh})}\right)}\right\} \qquad \textit{Eq 2. 24}$$

Where:

Sw: un-invaded zone water saturation (fraction)

Vsh: volume of shale (fraction)

Rsh: resistivity of shale (Ohmm)

Rw: Formation water resistivity (Ohm-m)

Rt: true deep resistivity in the un-invaded zone (Ohm-m)

Ø: effective porosity (fraction)

(a) and (m) are constants representing multipliers and cementation factors, respectively.

- **The Waxman–Smits model**

Waxman–Smits introduced their model, which depends on the shale content and cation exchange capacity (CEC). They utilized effective porosity and shale content derived from logs to correlate with CEC measurements obtained from core samples. These measurements are critical parameters for calculating accurate water saturation. [33][34]

$$\frac{1}{R_t} = \frac{Sw^2}{F * R_w} + \frac{BQv * R_w}{F} \qquad \textit{Eq 2. 25}$$

Where:

Rt: true resistivity in the un-invaded zone (Ohm-m)

Sw: the un-invaded zone water saturation (fraction)

Rw: the formation of water resistivity (Ohm-m)

F: formation factor. ($1/Ø^{m*}$) where $m^* \neq m$

Qv: cation exchange capacity volumetric (Unite per litre)

B: Equivalent cationic conductance of a sodium ion (L S/m) can be calculated by the next equation to relate B to formation temperature (T) Celsius degrees

Rw: water resistivity

T: formation temperature Celsius degrees.

$$B = -1.28 + 0.255 * T - 0.0004059 * T^2)(1 + (0.04 * T - 0.27) * R_w^{1.23} \qquad \textit{Eq 2. 26}$$

- **The dual-water model.**

The dual-water model proposes that the presence of free water within the pore spaces of the reservoir rock, along with bound water within the clay matrix, accounts for the influence of

clay minerals on the resistivity of the reservoir rock. This model is built upon the Waxman–Smits model, which suggested that the conduction geometry of free water and clay counterions is identical. Therefore, their model is designed to account for the conductivity occurring near and within the double layer, as well as the conductivity in the layer unaffected by the presence of clay. [34]

$$\frac{1}{R_t} = \frac{Swt^n}{Fo}\left[\frac{1}{R_w} + \frac{V_Q Q_V}{S_{wT}}\left(\frac{1}{R_{cw}} - \frac{1}{R_w}\right)\right] \quad \text{Eq 2. 29}$$

3.3.2 Core data analysis

Precise results for porosity, permeability, and saturation were directly obtained from core data. The primary challenge with core data is its limited availability across all wells in a field. Even when available, it typically covers only small intervals due to the high cost and time-intensive nature of core operations. Reservoir properties, including porosity, permeability, and saturation, can be determined in the core laboratory through Routine Core Analysis (RCA) or Special Core Analysis (SCAL).

- **Routine Core Analysis (RCA)**

Routine Core Analysis (RCA), also known as conventional core analysis, is used to measure the petrophysical properties of core samples using single-phase fluids. Its main advantages are that it is less expensive and quicker compared to Special Core Analysis (SCAL).[35]

Porosity, absolute permeability, Klinkenberg permeability, grain density, and water saturation are the main properties obtained from this analysis. Obtaining porosity and permeability from RCA from laboratory procedures has several limitations.

Gases are used on preserved, clean, dried core samples under room conditions, rather than reservoir conditions, to determine porosity and permeability.[36]

Routine core analysis is used to establish the relationship between porosity and permeability. In wells without core samples, routine core data and porosity are derived from well logs to estimate permeability.. [36]

- **Special Core Analysis (SCAL)**

Special Core Analysis (SCAL) is a more advanced testing method using flow experimental of two-phase fluids to measure relative permeability, capillary pressure, wettability, and remaining oil saturation. [36]

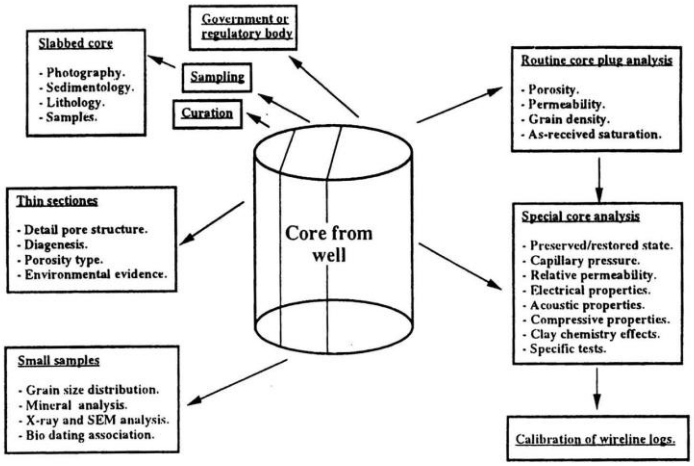

Figure 2. 5 : Data obtained from cored well.[37]

2.4 Machine Learning Technique

2.4.1 Porosity Estimation Using Machine Learning

Porosity is the main determinant for permeability estimating which is the void space in the formation that provides the storage for hydrocarbon accumulation. [38]

Porosity can be measured in a laboratory, which is the most costly method. Alternatively, it can be estimated from well logging techniques such as density, neutron, sonic, and NMR logs, which are less costly and time-consuming. However, the main challenge related to this method is the uncertainty in the porosity estimation due to the logging environment and tool effects and it depends on the statistical analysis. To determine the accurate value of reservoir porosity from several porosity logs validated with laboratory measurements using artificial intelligence (AI) is the best method.

Al-Fakih et al [39], used machine learning techniques including the artificial neural network (ANN), K-nearest neighbour (KNN), random forest (RF), and decision tree (DT) to predict reservoir properties (total porosity, effective porosity, and shale volume) derived from well-logging data from the North Sea sedimentary basin. Researchers used well logs from six wells as inputs for their models: Gamma Ray (GR), bulk density (RHOB), sonic transit time (DT), and neutron porosity (NPHI). The author found that Artificial Neural Networks (ANN) and Random Forest (RF) models provide the best result for reservoir property predictions. However, researchers found that machine learning models can provide accurate results for reservoir properties in challenging sedimentary basins with limited data available even when core data are scarce.

Ifrene et al. [40] showed that using machine learning in combination with well-logging analysis provides more reliable fracture porosity by integrating image logs and sonic scanner data in the Ahnet field in Algeria. The authors of this study used Pure Artificial Neural Networks (ANN) and hybrid models. During this study author found that using basic well logs for fracture porosity estimation is cost-effective and isn't needed to acquire expensive and time-consuming as advanced logs which also helps reduce and mitigate the risks associated with unstable wells during logging procedures.

Elkatatny et al. [41]used the effectiveness of artificial neural networks (ANN), Support Vector Machine (SVM), and Adaptive neuro-fuzzy inference system (ANFIS) to predict the porosity using 1700 field measurements with corresponding log data. The author shows that bulk density, neutron porosity and compressional time are the main inputs required to estimate accurate porosity using the AI technique. The results demonstrated that both ANN and ANFIS exhibited high correlation coefficient (R) and low average absolute percentage error (AAPE).

2.4.2 Permeability Estimation Using Machine Learning

Previous studies demonstrate the efficiency of using machine learning models to predict permeability by using different input data. Mkono et al [42], used the group method of data handling (GMDH), Differential evolution (GMDH-ED) and random forest (RF) algorithms in permeability prediction of the Wangkwar formation in the Gunya oilfield, Northwestern Uganda. The authors used a combination of measured and calculated logs including spectral

gamma-ray (SGR), apparent resistivity focusing mode (RLA1), shale volume of rock (VSH), effective porosity (PHIE), photoelectric factor (PERFZ), apparent resistivity focusing mode 5 (RLA5), standard resolution formation density (RHOZ), caliper log (CALI) and thermal neutron porosity (TNPH) which used 1953 samples for training and 1563 for testing. Their result showed that GMDH- DE provide more accurate results compared to other algorithms.

Topór et al. [43] used two machine learning models i.e. multiple linear regression (MLR) and random forests (RF) to predict permeability under confined stress conditions for samples from tight sandstone formations models trained on a dataset containing different data such as basic petrophysical data, and rock type from a visual inspection of the core material. The author found that both MLR and RF models can predict permeability accurately. Also, he discovered that porosity was the most important feature that affects model performance. On the other hand, depth also influenced permeability prediction.

Zolotukhin et al. [[44] explain using machine learning to determine the permeability of the porous media using laboratory experiments. The author used a neural network to predict the permeability and found a good match between predicted and calculated permeability. He also obtained an analytical expression describing a fluid flow in a porous medium which enabled the use of this information in a reservoir simulator.

Talebkeikhah et al. [45] Researchers have discussed using different machine learning i.e. Radial Basis Function Neural Network (RBF), Multi-Layer Perceptron Neural Network (MLP), decision tree (DT), Support Vector Regression (SVR), and random forest (RF)model to predict permeability and compare models' accuracy. For input data, researchers used random forests to investigate well-logging data. Random Forest was used to analyse real well-logging data, selecting five potent variables: CGR, SGR, NPHI, and RHOB, with permeability as output which is derived from core analysis. Researchers found that Support Vector Regression (SVR), and decision tree (DT) models provided more accuracy than others. Based on this comparison researchers found that artificial intelligence outperformed traditional methods to predict reservoir permeability.

El-Sebakhy et al [46]The author used functional networks as a novel approach and linear/nonlinear regression, neural networks, and fuzzy logic inference systems to predict

permeability in carbonate reservoirs using well-logging. The author used DT, MSFL, NPHI, PHIT, RHOB, and SWT as features for the selected model. Researchers found that functional networks show reliable permeability prediction compared with used models.

2.4.3 Saturation Estimation Using Machine Learning

Existing literature has identified the importance of using different machine-learning techniques to predict water saturation with high accuracy. A previous study by Hadavimoghaddam et al.[17] used five different machine learning models which were XGBoost, LightGBM, Adaboost, Catboost and Super Learner to predict water saturation without depending on resistivity log data. During this study, authors used two datasets which are logs data one of them containing (Gamma-ray, density, neutron, sonic, PEF) and another without (PEF) and then compared the results with laboratory data. The authors concluded that AdaBoost was considered the best result while XGBoost and Super Learner produced negligible errors.

The study by Baziar et al. [47] used Support Vector Machine, Multilayer Perceptron Neural Network, Decision Tree Forest, and Tree Boost to predict water saturation of Mesaverde tight gas sandstones located in the Uinta Basin. The authors used data from four wells including a gamma ray log (GR), a neutron porosity log (NPHI), a deep induction resistivity log (ILD), a bulk density log (RHOB) and a sonic travel-time log (DT) as input and compared between used methods by using three error indexes including correlation coefficient, average absolute error and root-mean-square error. The authors found that Support Vector Machine provides better results than other models.

The study by Hamada et al. [48] uses a neural network technique to predict formation porosity and water saturation. The authors used two neural networks, one of them to predict porosity and another to predict water saturation. For porosity model used 6 well logs data inputs (GR, LLD, RHOB, NPHI, PEF and Δt) while for water saturation model used 5 well logs inputs (GR, LLD, RHOB, NPHI and PEF). The authors found the results have shown a good match between neural network models and core data. Also, they found that neural networks are used to predict accurate reservoir properties and hydrocarbon potential in new wells.

CHAPTER THREE

3 MACHINE LEARNING & WELL LOGGING FUNDAMENTAL

3.1 Introduction

This chapter provides a foundation for two critical domains: machine learning (ML) and well-logging tools. Machine learning is a branch of artificial intelligence (AI), it's solved several problems in various industries with its ability to extract useful insights from data. In the context of reservoir engineering and characterization, ML offers opportunities to improve well-logging interpretation and analysis. Well-logging tools play an important role in gathering subsurface data to understand geological formation and reservoir properties. This chapter aims to comprehensively attempt to examine machine learning fundamental techniques and well-logging tools. This chapter will explore machine learning classification and its principles. Furthermore, I will discuss the features and applications of various well-logging tools.

3.2 Machine Learning Definition

Generally, machine learning can be defined as a field of computer science that focuses on studying algorithms and methods for automating solving problems which are difficult to handle with traditional programming.[49] There are different perspectives on defining machine learning (ML). However, Tom Mitchell [50]describes machine leasing as "A computer program is said to learn from experience E concerning some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". Mohri et al [51] Machine learning is described as a computational technique that leverages experiences to improve performance or make accurate predictions. Typically, it is used to predict undefined future scenarios for computers.

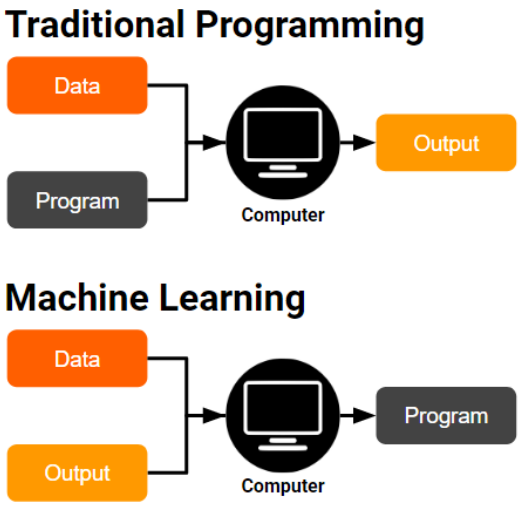

Figure 3. 1: Traditional Programming versus Machine Learning.[52]

3.3 Machine Learning Classification

Machine learning (ML) models can be categorized into three techniques: supervised learning, unsupervised learning, and semi-supervised learning.

Typically, ML algorithms are powerful because of their ability to learn from the available data.[49] This section will introduce the most widely used learning models.

3.3.1 Supervised learning

Supervised learning also referred to as predictive learning (e.g., regression and classification), deals with a dataset that is used to develop the model by training data which contain features and outcomes; Afterward, the trained model is assessed using testing data, and ultimately, the model is employed to predict outcomes based on the features of unseen data. [53] Figure 3. 2, shows the basic principle of supervised learning.

In this type of learning, a labelled dataset which is a large set of data points with answers (outputs) is provided to the learning algorithms. The algorithms need to discover the key features of each data point and figure out the answers. In other words, algorithms learn the data pattern or function that allows them to map input variables to their corresponding outputs for each specific instance. [5] Therefore, when the algorithms are obtained with unseen data points (new data points) they can predict the outcome based on the key features. [49]

Generally, supervised learning is divided into two types: Classification and Regression. During classification problems, it's able to classify data into classes or categories. Therefore, this type is used in the petroleum industry to solve problems such as facies, fractures, faults, etc. However, regression problems, refer to the capability to forecast a value of continuous numbers. In the oil and gas industry, this type of learning is used to solve various problems, such as predicting porosity, permeability, shear wave velocity, and more.[5]

Typically, the accuracy of the output prediction heavily depends on the quality of the data. If the input data used to train the algorithms is of high quality, a well-trained supervised learning algorithm can predict the output accurately for unknown or unobserved data instances. Nevertheless, low-quality training data could result in overfitting problems, which affects the algorithms' inability to generalize the underlying predictor-target relationships, leading to regression failure.[54]

Furthermore, this study will use regression models to predict reservoir properties using different logs, all of which values are in continuous form.

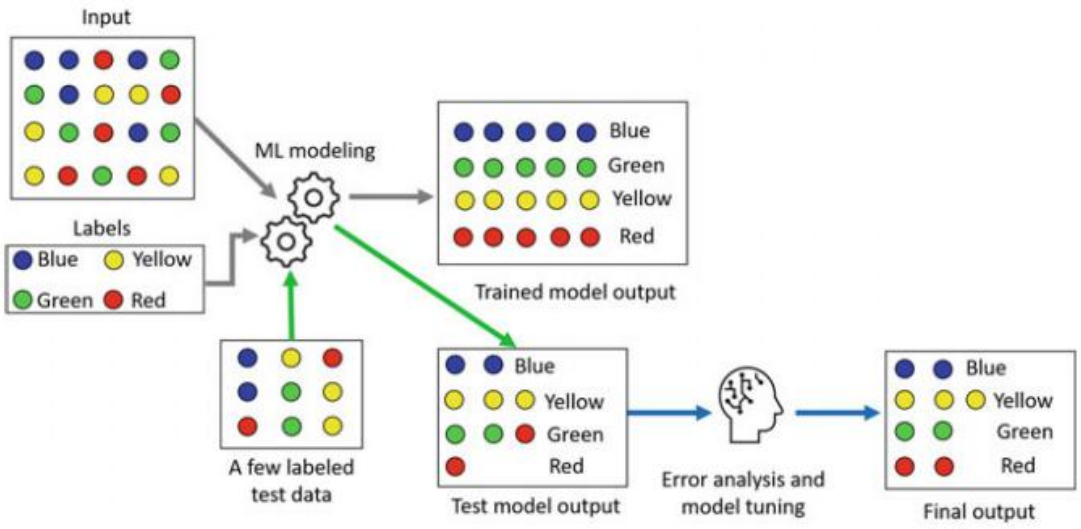

Figure 3. 2 basic concepts of supervised ML.*[5]*

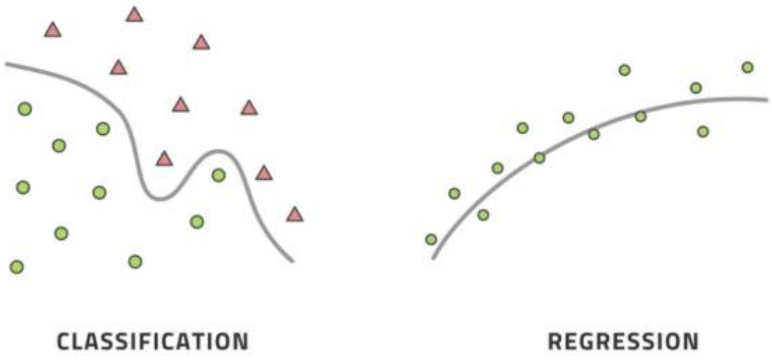

Figure 3. 3: the difference between classification and regression methods.*[55]*

- **Artificial Neural Networks**

Artificial neural networks it recognized as supervised machine learning that perform two main tasks: regression and classification. [56] It is popularly known as brain-like computation. It is built on the concept of decision-making capabilities in the human brain.[57] According to Haykin [58] ANN is a "massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use".

Typically, the architecture of an artificial neural network (ANN) consists of three types of layers: the input layer, hidden layer, and output layer. The input layer receives the data that the algorithms learn from, while the hidden layer (which may include multiple hidden layers depending on the task) processes the input data through mathematical computations. The output layer contains output neurons that convert the results of the neural network into a form understandable by the user. [38][59][60]

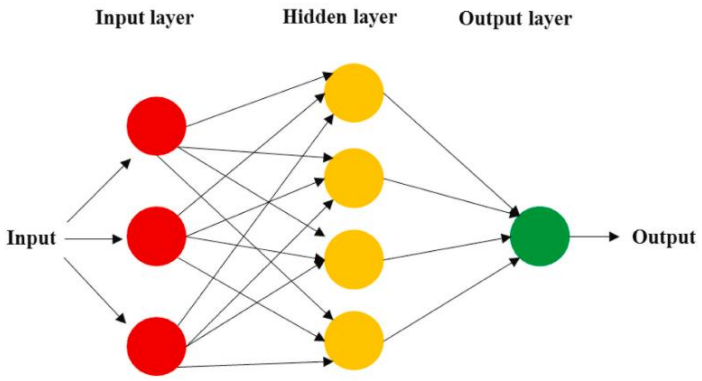

Figure 3. 4: basic configuration of neural network.*[61]*

Figure 3. 4, represents a basic ANN architecture where each layer contains a specific number of nodes connected by links (weights), with each link having an associated weight and activation level.[1] The number of input nodes can be determined based on the available training data. The number of nodes' hidden layer(s) may also change. However, the increased number of nodes can increase computation complexity, while the limited number of nodes may lead to impaired ANN learning capability. [38]

Consequently, ANN operates on the concept of a weighted sum of inputs, where each input is multiplied by its associated weight before the results are summed. An activation function is then applied to the weighted sum to determine the output of each node. This process is repeated for each node in the network, with the output of one node serving as the input for the next, until the final output is achieved. Sigmoid, the most widely used activation function, ReLU (rectified linear unit), and tanh (hyperbolic tangent) are a few examples of common activation functions. [2], [62]

- **K Nearest Neighbour**

One of the most widely used machine learning algorithms for resolving regression and classification issues is the KNN algorithm. KNN is an instance-based, lazy, non-parametric learning algorithm.

It’s known as lazy learning algorithms which means that to generate a model the algorithms don’t need any training data point. During this algorithm, the testing phase takes more time and memory than training data which is faster, that is because the testing phase uses all training data. However, it required more memory to store training data and more time to scan all data points. Hence, more effort was focused on the testing phase, with less emphasis on the training phase. While non-parametric means the model structure is selected based on the dataset. In other words, the underlying data distribution isn’t presumptively. However, instance-based indicates that the model doesn’t explicitly learn. Alternatively, it is determined to memorize training instances that are utilized after that as knowledge during the prediction stage. [63], [64]

The concept of determining K's nearest neighbour for a given data point is very simple. The supposition is that similar items are to be close to one another. In this context, close refers to Euclidean distance which is a measure of distance between two points. If the test data

belongs to a class that means that is identified as the largest class of items that are like test data. [49]

In other words, KNN algorithms find K nearest neighbouring points in the sample data using the similarity measure. Then, forecast the output category according to the class that it appears the most frequently.[57]

Nevertheless, KNN has several limitations, it's not suitable for datasets with imbalanced classes. That's because the class composition, the majority decision among the K neighbors is mostly in favor of the class that has the most instances in the training set. Another shortcoming is that the KNN algorithm considers all features that are collected even if some of them are completely insignificant. [49]

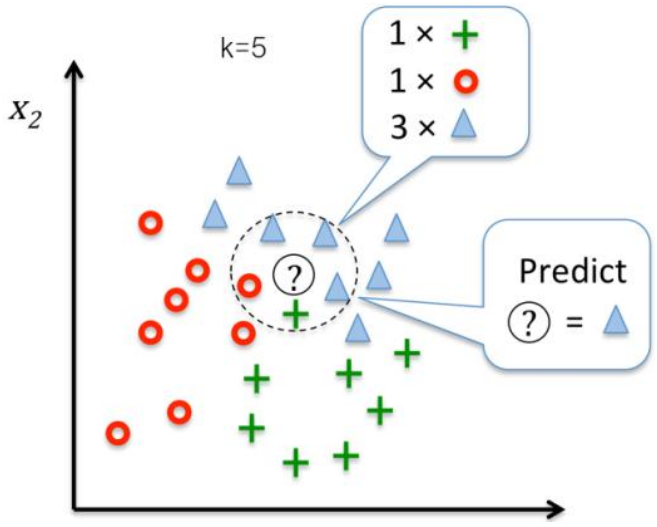

Figure 3. 5: k nearest neighbour example.*[65]*

- **Support Vector Machine**

The support vector machine (SVM) is a commonly used supervised learning technique. Vapnik et al [66] initially suggested SVM as one of the practical models that is used for identifying the patterns. It is applied for solving regression and classification of real-world problems.[49] Also, it's a statistical learning theory that reduces risk by using a strong mathematical foundation in both regression (SVR) and classification (SVC) models.[67]

Typically, SVM is an algorithm that seeks to divide data points into classes by utilizing a hyperplane with the maximum margin as shown in Figure 3. 6. [68] Hence, it has three key elements in the SVM support vector, hyperplane, and margin. [69]

A hyperplane is a decision boundary that divides the data points into multiple categories in a feature space, Figure 3. 6. Support vectors are the closest data points to the hyperplane, and they play a significant role in determining the hyperplane and margin. The gap between the hyperplane and the support vectors is called margin. The basic goal of support vectors is to maximize the margin. A larger margin indicates higher categorization performance. [70] Kernel is the mathematical function, which is utilize for high- dimensional mapping. These functions assist in exhibiting the outcomes when the initial data cannot see the data in a high-dimensional feature space this is helpful when the data involved is not linearly separable in the original input space, Figure 3. 7.[5] [70]

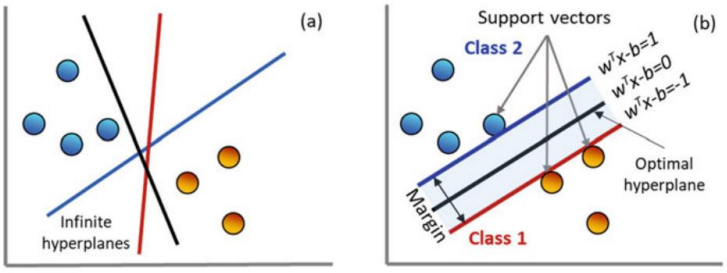

Figure 3. 6: Schematic of support vector machine for classification.*[5]*

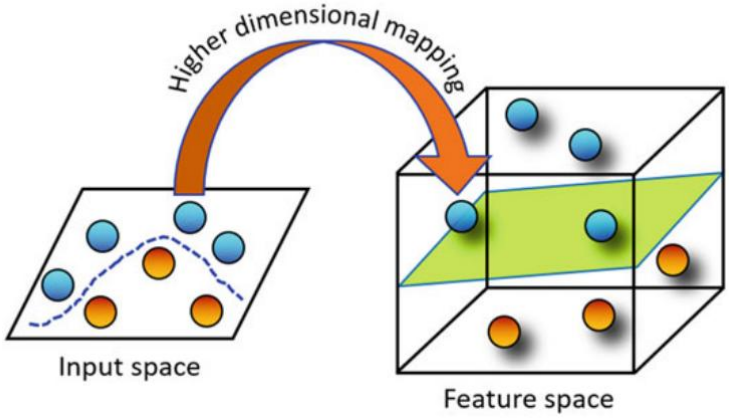

Figure 3. 7: Support vector machines' use of the higher-dimensional mapping notion.*[5]*

A variety of kernel functions are available, including linear, polynomial, radial basis function (RBF, often known as Gaussian kernel), sigmoid, and mixed. Table 3. 1, illustrate their mathematical expressions.

Table 3. 1: illustrating various type of kernels.[5]

Kernels	Expressions	Parameters
Linear	$K(x_i, xj) = x_i . xj$	
Polynomial	$K(x_i, xj) = (x_i . xj+1)^d$	d: degree of polynomial
Radial basis function	$K(xi, xj) = (\exp(-\Upsilon \| x_i - xj\|^2$	$\Upsilon > 0$
Sigmoid	$K(xi, xj) = \tanh(m(x_i, x_j) + b)$	m (slop) and b(intercept)

Kernels can be Chosen according to the characteristics of the dataset. Customarily, the optimal Kernel is not clear. Thereby, one needs to try simple linear Kernels if the performance of the classification is not efficient then select a nonlinear kernel. Nevertheless, a significant portion enhances model performance when feature or data dimensionality reduction is applied. Typically, for a dataset select which kernel performs best using cross-validation. [49]

Support vector machine (SVM) provides an optimal method of classification with small input data and high features. [49]

- **Decision Tree**

A Decision Tree (DT) is a machine learning model constructed by a sequence of decisions based on variable values to choose one path or another. [49] The non-parametric technique is utilized for classification and regression.[71] A classification tree provides a categorial output (class, text, color etc..) whereas a regression is a decision tree that provides a numerical value. [1]

The algorithms create a tree structure which is like a flowchart. Figure 3. 8, Show the tree components, which are separated into three categories: decision nodes, branches, and leaf nodes. The DT begins with the root node, located at the top of the tree, and finishes with multi-leaf nodes. It uses the specific input feature condition to partition the data into recursively segments.

Essentially, all nodes in the tree perform a test on an attribute of the instance, and the node represents a possible attribute value in each branch descending. The root nodes and internal nodes are joined by branches which show possible outcomes. DT split the data into two groups at each node according to a selected feature and certain thresholds. At the new node, the process is repeated for the subtree rooted. [5] [50]

Typically, to build the optimum decision tree in the beginning should ask, which attribute (Feature) the best is to split the data based on the feature. To solve this question, it's important to evaluate each instance attribute statistically which helps to assess the accuracy in classifying training data. The best attribute is chosen to be at the root node in the tree. Then, for the interior node, the process is repeated to select the best attribute.[50]

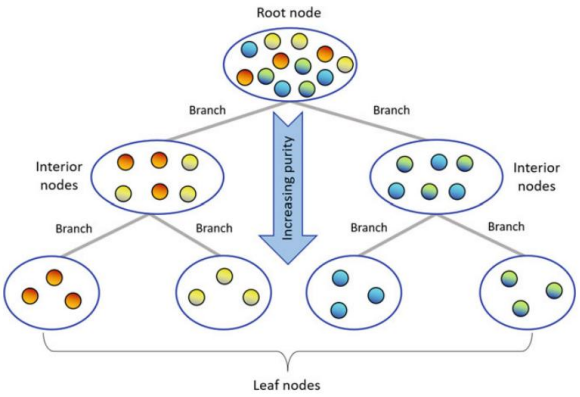

Figure 3. 8: the decision tree diagram.*[5]*

Conversely, the features that are more sensitive in the tree affect the purity of the tree for that reason this feature is selected to be at each node. Generally, there are more than metrics to measure the impurity of the splitting dataset such as Entropy, information gain, and Gini impurity.

Entropy is the amount of information disorder in the dataset. On the other hand, it is a measure of the amount of uncertainty present in the value of random variables. Hence, it's also used to investigate the quality of the model and its potential to make reliable predictions. [72] Eq 3. 1, represent the entropy.

$$Entropy(p) = -\sum_{i=1}^{N} Pi \ \log_2 pi \qquad \textit{Eq 3. 1}$$

pi is the probability of occurrence of ith possible value.

n is the total number of values.

The lower the entropy the less uniform the distribution of the pure node. However, the entropy in the node depends on the number of random data in that node and is calculated for each node. In decision trees, the goal is to look for trees that have the smallest entropy in their node. The entropy is used to calculate the homogeneity of the sample in that node. Entropy is zero when the samples are completely homogeneous. If the samples are equally divided it has an entropy is one.

The other criterion in the decision tree is the information gain which refers to the information that enhances the certainty after partition.[73] The optimum split of the decision in the decision tree will get with the highest information gain. [49] However, the information gain is the opposite of the entropy, which means when the entropy increases the information gain decreases and vice versa. Hence, the best DT provide a high information gain for the splitting data.

$$Information\ Gain = (Entropy\ Before\ Split) - (Weighted\ Entropy\ after\ split) \quad \textbf{Eq 3. 2}$$

Gini is the statistical approach used to evaluate the distribution in a population used to infer the inequalities.[5] Also, it's a measure of the amount of random data points that are incorrectly classed.[69] The Gini index value ranges between 0 and 0.5 the best data split provides a 0 Gini index, otherwise, the value 0.5 means the worst split.[69][73] In contrast to information gain the lower Gini value performs the best decision-making. Eq 3. 3 used to calculate the Gini index.

$$G = 1 - \sum_{i=1}^{k} pi^2 \quad \textbf{Eq 3. 3}$$

pi is the probability of the data point belonging to the ith class label.

k accounts for different class labels.

In summary, several metrics are used in DT to measure the accuracy of the model. However, DT is prone to overfitting this occurs when the tree becomes very deep.

- **Random Forest**

Random Forest (RF) is a supervised machine learning algorithm that is built on an ensemble decision tree. This approach can be applied to classification and regression problems.[5], [57] Generally, RF algorithms create multiple decision trees using randomly selected subsets from the dataset, and the final prediction for the classification task is computed by majority voting, whereas for the regression task, the result is from the average result from each decision tree. [74] Figure 3. 9, illustrate the basic concept of the Random Forest algorithm.

RF model focuses on the overfitting problem associated with the decision tree.[5] RF algorithms differ from DT algorithms which start with creating multiple parallel decision trees to train and predict the dataset with bagging or bootstrap aggregation. The term bagging refers to the short form of bootstrap aggregation. In machine learning algorithms with poor performance or overfitting issues, it is useful to use bagging that trains the dataset in parallel.

During this process, the prediction for regression is provided through averaging while voting through a classification. This procedure minimizes the chances of overfitting and yields a more stable model. However, bootstrap aimed to reduce underfitting by training the model previously trained model with low performance this approach provides the best model performance. [57]

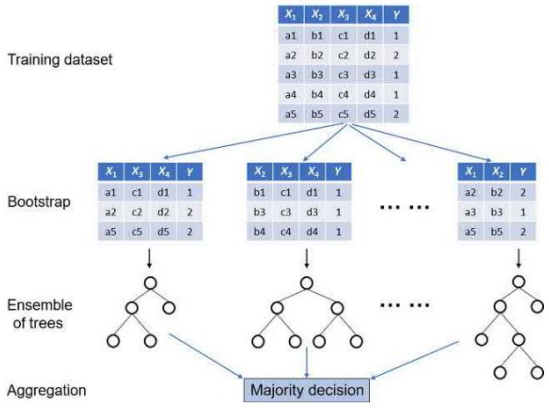

Figure 3. 9: basic concept of the Random Forest model.*[53]*

- **Linear Regression**

Regression analysis is a statistical approach for assessing and modelling the relationship between variables.[75] Linear Regression is the simplest type of supervised machine learning which predicts output is continuous and has linear slop. The regression task depends on the number of input variables which are classified into simple regression and multiple regression learning.[76] Simple linear regression expressed by Eq 3. 4 it is containing one independent and dependent variable.

$$y = \beta 0 + \beta 1x \qquad \textit{Eq 3. 4}$$

Where:

β_0 is intercept and β_1 represent the slop.

y is the dependent variable.

x is the independent variable.

Eq 3. 5, represent multiple regression which contains more than one independent variable and one dependent variable.[77]

$$y = \beta_0 + \beta_1 X + \beta_2 X + \cdots \ldots \ldots \beta_n X \qquad \textit{Eq 3. 5}$$

Where:

β0 is intercept and β1 represent the slop.

y is the dependent variable.

X1, X2, X is the independent variable.

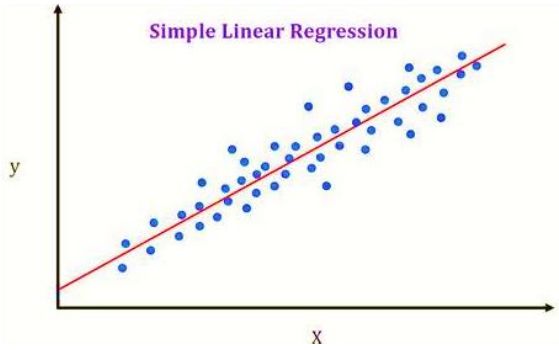

Figure 3. 10: Scatter plot for x and y with Straight-line relationship between x and y.*[78]*

Figure 3. 10 , illustrate the distribution of data points in the x-axis and y-axis. The distribution of data points isn't along full perfectly on a straight line, the goal of regression is to get the best fit as shown in Figure 3. 10 .

3.3.2 Unsupervised learning

In unsupervised learning or descriptive learning (e.g., clustering and transformation), in this type of machine learning, the datasets lack labeled data for training. The model learns to generate outcomes based on features without any information about the corresponding outcomes.[46]

Figure 3. 11, Demonstrate the concept of unsupervised learning, where the model divides the data into distinct groups or clusters based on the natural patterns in the data and the spatial differences between the groups.[4]

As previously stated, the algorithms try to identify clusters with the same properties in the data sample. After training the clustering algorithms, the prediction of any new observation is assigned to any cluster based on the identification cluster from the original data.[57]

Generally, during the beginning of the project or when don't have much data about a system, unsupervised learning is used. It doesn't depend on prior geological knowledge and minimizes human bias, which can have advantages and disadvantages depending upon the situation. This type of learning is particularly suited for facies classification tasks in extensive regions with 3D seismic data and limited borehole and outcrop information. Although, when the core and cutting data are not available apply this technique for classification-based well-logging.[5]

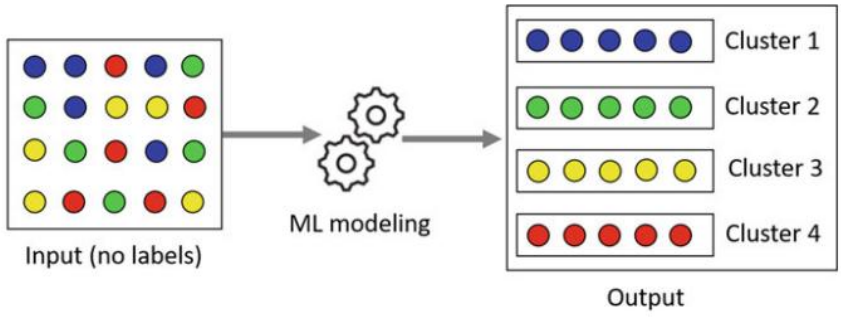

Figure 3. 11: the basic approach of unsupervised learning.*[5]*

3.3.3 Semi-Supervised Learning

In semi-supervised learning (e.g., generative adversarial networks (GAN) and reinforcement learning), incorporating both supervised and unsupervised learning, the model in this approach is trained on a dataset that includes a small portion of labeled data and a large portion of unlabeled data.[5][79]

In this context, the algorithm supplied a large dataset containing limited labelled data. A clustering algorithm is utilized to detect a group within a dataset and identify a few labelled data within each cluster to assign labels to the remaining data points within that cluster. The main benefit of this technique is that it reduces the time and effort needed to label each data point, which can be a time-consuming process.[49]

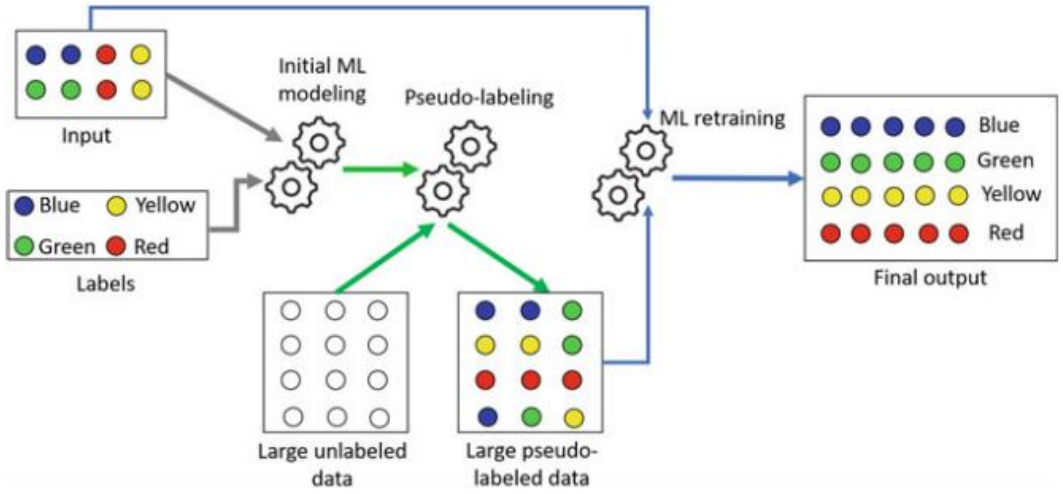

Figure 3. 12: A principle of semi-supervised learning.[5]

Table 3. 2: The difference between ML algorithms.[80]

Algorithms	**Type**		
	Input data	**Complexity**	**Accuracy**
Supervised	Labelled	Simpler	Higher
Unsupervised	Unlabeled	Complex	Lesser
Semi-supervised	Partially labelled	Depends on the use of supervised or unsupervised	Lesser

Table 3. 2, illustrate the comparison between different types of ML algorithms. The major difference between ML algorithms is the kind of data required and how it can used. As shown in Figure 3. 13, in supervised learning, it is important to use fully labelled datasets this increases the accuracy of the models which may be laborious and costly. But, in unsupervised learning works without labelling datasets, the algorithm discovers the pattern and structure of the data, making it more complicated to evaluate the output. Although semi-supervised learning used datasets containing labelled and unlabeled data that combined the strength of supervised and unsupervised learning.

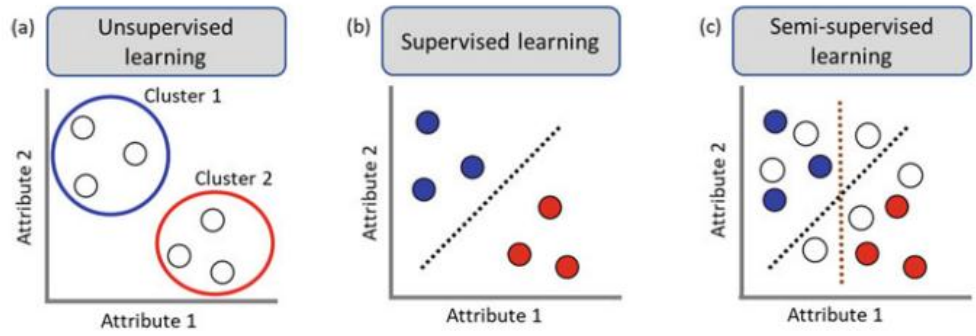

Figure 3. 13: The different types of ML.

3.4 Well Logging Tools

Well-logging tools run in holes after the drilling operation stops which run one or more than devices. In advanced operations, well-logging tools run during drilling operations this is called logging while drilling (LWD). This study predicts rock properties based on conventional well logging tools (e.g. Gamma Ray, Spontaneous potential, deep resistivity, density, neutron, sonic etc.) this provides useful information about geological and engineering investigation. Well logging is a continuous record as a function of depth which is an indirect measurement of formation parameters.

3.5 Well Logging Objectives

Well-logging is the among most critical datasets, as they offer a continuous logging of wellbore measurements that may help to understand the fluid distribution in the reservoir, reservoir characteristics and reservoir deposited environment; in other words, do these rocks

store hydrocarbons and will they flow? [81] The primary usage of well logging is summarized below:

- Identification of Lithology.
- Determining reservoir properties (e. g porosity, saturation, permeability).
- Identify the difference between reservoir and non-reservoir rocks.
- Determining the type of fluid present in the pore space of reservoir rock (gas, oil, water)
- Detection of productive zones.
- Estimation of the depth and thickness of productive zones.
- Identifying reservoir fluid contacts.
- Well-to-well correlation for determining the lateral extent of subsurface geological cross-sections.

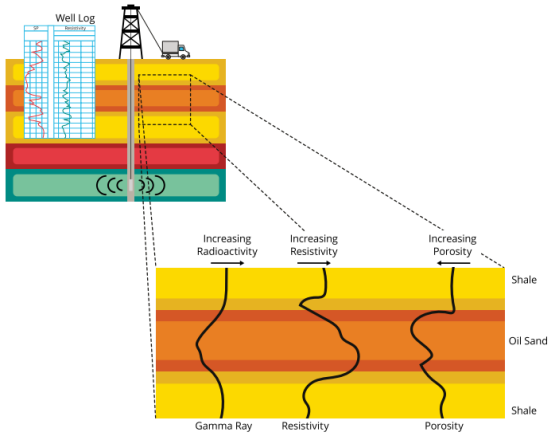

Figure 3. 14: Schematic diagram of a typical set-up for running wireline logs and typical response of three main properties.*[10]*

3.7 Types of Well Logging

Various types of logs are available to identify well tops, sequence the zones, and assess the properties of the rocks, these logs are Caliper logs, Gamma-ray logs, Spontaneous potential, Density logs, resistivity logs, Neutron logs, sonic logs they will be explained in detail in this study.

3.7.1 Gamma Ray Log (GR)

A gamma-ray log is one of the key well-logging tools. It measures the natural radiation in the penetrated formations. A formation that is rich in uranium, potassium and thorium emits natural radiation. Shale riches with these radioactive materials which increases gamma ray response. [18][81]

Clean sandstone contains a low concentration of radioactive materials, resulting in low gamma ray readings. However, if the sandstone contains heavy minerals like potassium feldspars, micas, glauconite, or is exposed to uranium-rich waters, it may produce a high gamma-ray log response. [18][81]

If heavy minerals are found in the clean sandstone, it's preferred to use a spectral gamma ray log.

A gamma-ray log is used to identify permeable and impermeable zones, calculate shale volume, and establish correlations between formations. [18] Typical gamma-ray responses for common lithologies are listed in Table 3. 3.

Table 3. 3: Typical gamma-ray responses for common lithologies.[81]

Lithologies	API units
Limestone	5-10
Dolomite	10-20
Sandstone	10-30
Shale	80-140
Organic-rich shale	150-200+

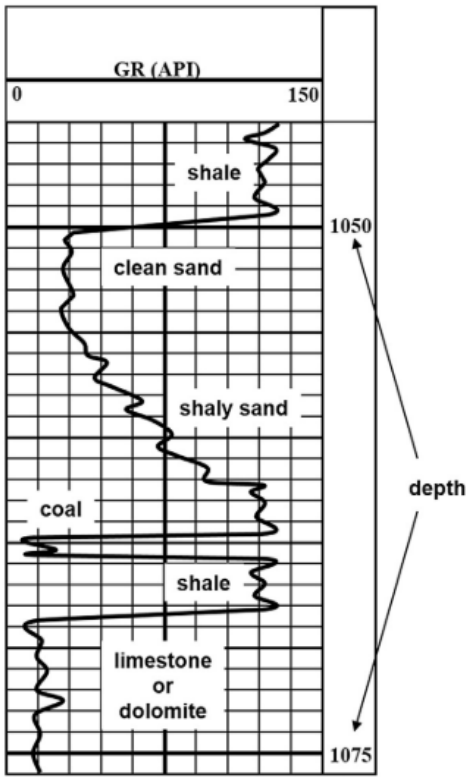

Figure 3. 15: Typical gamma ray response.*[82]*

3.7.2 Spontaneous Potential Log (SP)

Spontaneous potential log, also referred to as the self-potential log (SP) is one of the oldest logging measurements. It's a continuing record that measures the natural potential between the wellbore and the surface. [24]

The natural potential occurred due to electrochemical differences between drilling fluids and fluid within reservoir formation caused by salinity variations, the electrokinetic potential generated by the flow of mud filtrate through the mud cake.[81] In other words, The electrical potential arises from the connection between different formations (membrane potential), the interaction between fluids (fluid junction potential), and the movement of fluids through the formation (electrokinetic potential).[82]

- **Membrane Potential**

Shale zones consist of clay minerals that are impermeable to (cl-) ions but permeable to (Na+) ions. The positive ions move to low-saline fluid (e.g. less saline mud) from high-saline fluid (e.g. saline formation water). [82]

Shale plays a critical role in determining the SP response because the tool needs to conductive wellbore fluids. [18]

- **Fluid Junction Potential**

The electrical potential is generated when the mud filtrate comes into contact with the formation water. The negative ions (Cl-) are more mobile than the positive ions (Na+), leading to the production of a negative current charge. [82]

- **Electrokinetic Potential**

When the electrolyte flows through a permeable formation it creates electrokinetic potential and current. It may occur in both formation and mud cake. [82]

The SP log is used to identify permeable formations and calculate the formation water resistivity (Rw), identify the depositional environment, and determine of shale content. In reading the SP log measured in millivolts (mV), it is recommended to first establish a shale baseline, Figure 3. 17. This is the typical SP level for shales and can be found by comparing the SP log with the GR log response. [18]

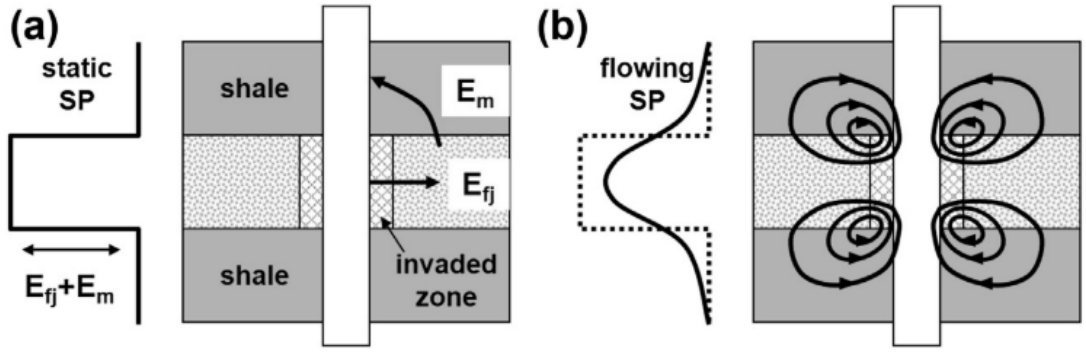

Figure 3. 16: Static potential (a) and measured flowing potential (b) for a fresh mud interacting with a saline formation. The arrows illustrate the magnitude and direction of the potential gradient. *[82]*

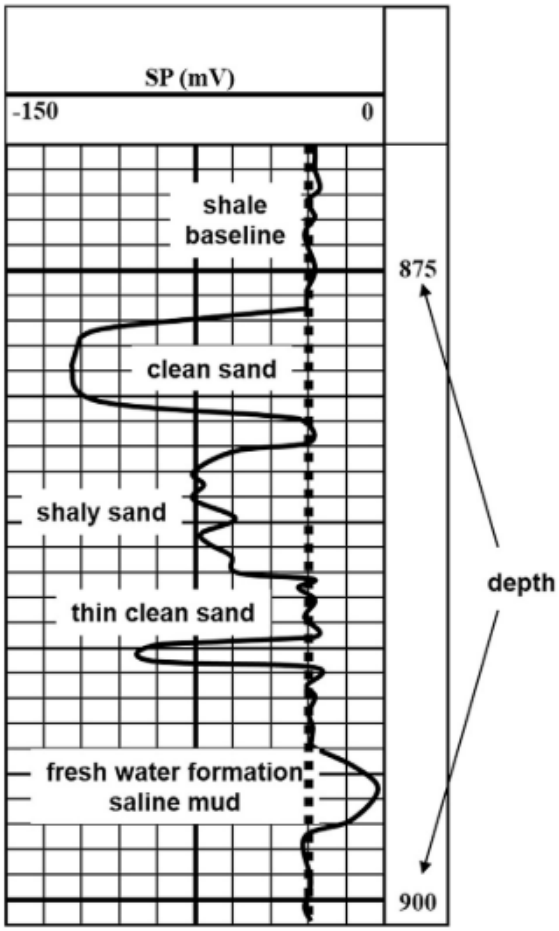

Figure 3. 17 Typical response of SP log. *[82]*

3.7.3 Neutron Log (NPHI)

The neutron log detects the hydrogen content within the formations. The neutron log tool contains a neutron source that emits neutrons and measures the amounts of scattered neutrons or gamma rays emitted (depending on the tool used) after the formation is exposed to neutron sources which collide with the nuclei and the neutrons lose their energy. When occurred sufficient collision the neutron is absorbed by the nucleus which resulting the greatest energy loss when colliding with hydrogen atoms, Figure 3. 18, shown the neutron log tool. [13] [42]

The neutron depends on the hydrogen atoms that fill the pore. Thus, more concentration of hydrogen atoms in the porous formations means more energy loss which is related to

formation porosity. However, the porosity that is calculated in shale formation can be overestimated value due to boron, chlorine, and other rare earth elements that are present in shale formation that affect on neutron capture rate.[18] [82]

The neutron log records the apparent porosity units of dolomite, sandstone, or limestone porosity. The apparent porosity matches the true porosity when the formation consists of limestone. However, the apparent porosity needs to adjust if the formation is sandstone or dolomite, Figure 3. 20, shows chart for correct neutron porosity to lithology. [18]

Neutron log recorded low porosity if the gas present in the formation this can be utilized to identify gas zones. However, when neutron logs combined with density logs can be used to detect fluid contact.

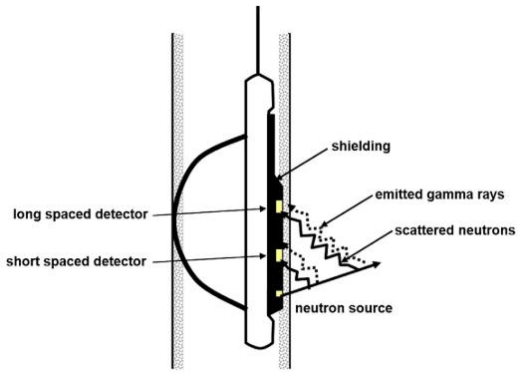

Figure 3. 18: neutron log tool.*[82]*

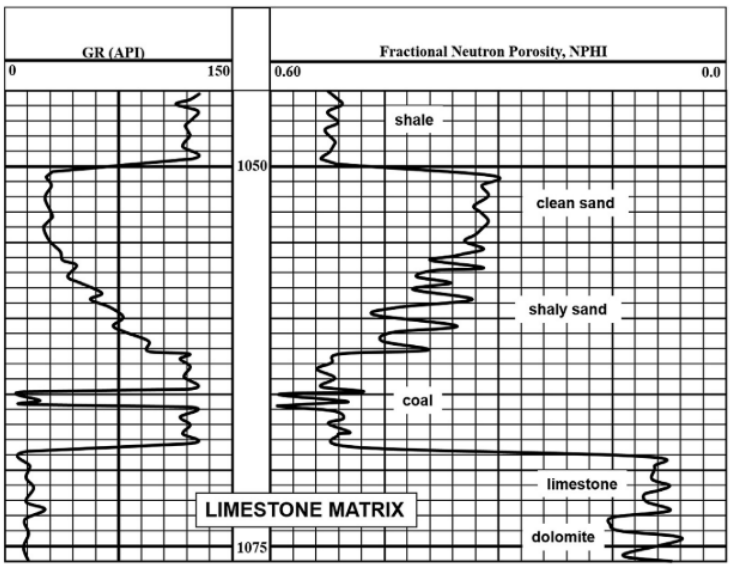

Figure 3. 19:neutron porosity log response.[82]

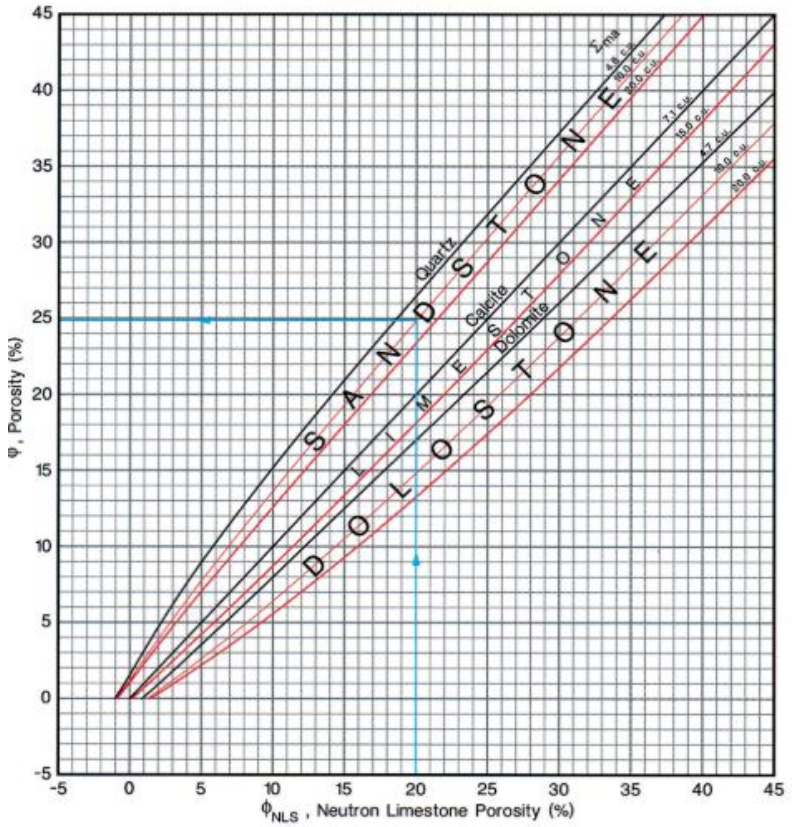

Figure 3. 20: chart to correct limestone porosity.[83]

3.7.4 Density Log (RHOB)

Density log quantifies the formation bulk density which is used to derive the formation porosity. However, it's also used to detect the gas-bring formation. The density log tool contains a radioactive source which is used to bombard the formation and sensor to measure how radiation returns. Density log used gamma-ray as a radioactive element. Gamma-ray reach the formation, it interacts with the electrons atoms that make up the formation and undergo Compton scattering, Figure 3. 21.[84]

When the bulk density of the formation is high, the sensor detects a low gamma-ray count rate, indicating that the formation contains a large number of density electrons, which significantly reduces the gamma-ray. However, a high gamma-ray count rate is recorded in the sensor when the formation bulk density is low and has a low number of density electrons, resulting in minimal gamma-ray attenuation. [84]

The density log tool was used to calculate the bulk density of the matrix with the effects of mud cack and wellbore irregularities.

However, the bulk density of a formation is the difference between its matrix density and that of the mud fluid. on the other hand, the presence of shale in the formation influenced the porosity. The shale density ranges between 2.20 and 2.85 g/cc depending on the clay mineral in the formation.[11] Table 3. 4, shows the typical matrix density of different lithologies.

Table 3. 4: Typical matrix density of different lithologies.[82]

Lithologies	ρ_{matrix} (g/cm^3)
Carbonate(limestone)	2.71
Calcareous sand	2.69
Consolidated sand	2.65
Unconsolidated sand	2.60
Shaly sand	2.6
Sand	2.2-2.85

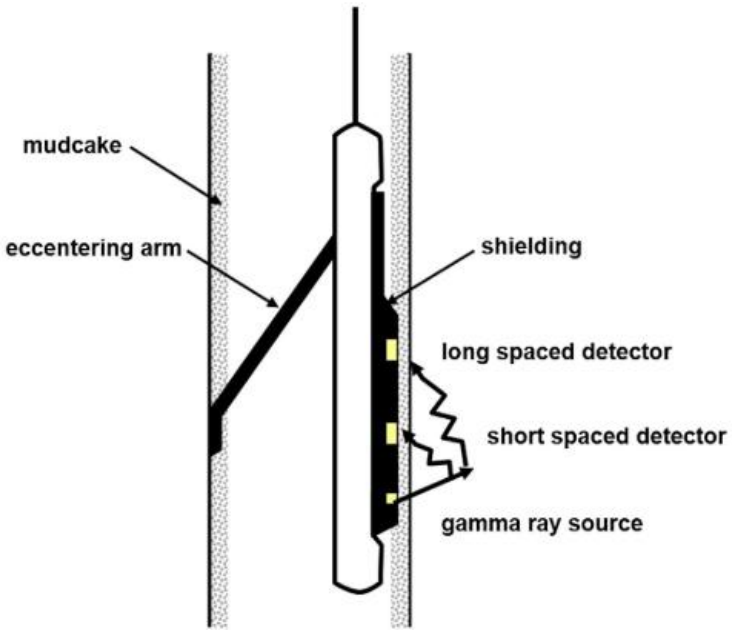

Figure 3. 21: density log tool. *[82]*

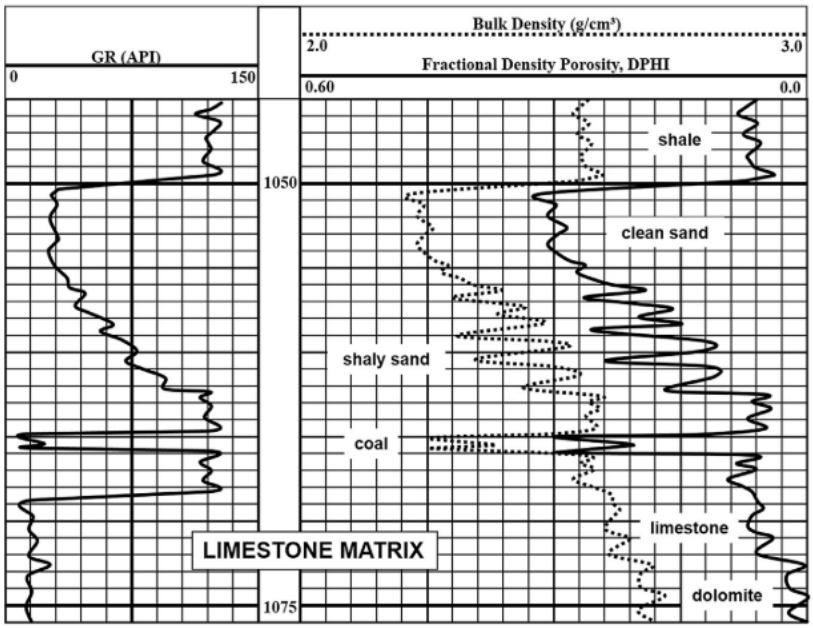

Figure 3. 22: bulk density with density log in different lithologies.*[82]*

3.7.5 Sonic Log (Dt)

Sonic logs, also called acoustic logs, are utilized to measure the time it takes for sound waves to travel through a formation. the sonic log tool consists of two transmitters and two or four receivers which are used to overcome the borehole size variations. The sonic log used to derive formation porosity, Eq 2. 11 and Eq 2. 12. However, it is also used to calibrate seismic data with petrophysical data. Lithology and porosity are the main factors affecting interval travel (transit) time.[11] [18] Table 3. 5, shows the formation lithologies and transit time.

Table 3. 5: common transit time and lithologies.[18]

Lithology	**Δt_{ma} µ/ft.**
Oil	232
Water	189-200
Sandstones	55.6
Dolomite	42
Anhydrite	50
Carbonates	43.5-47.6
Shales	62.5-167
Salt	15

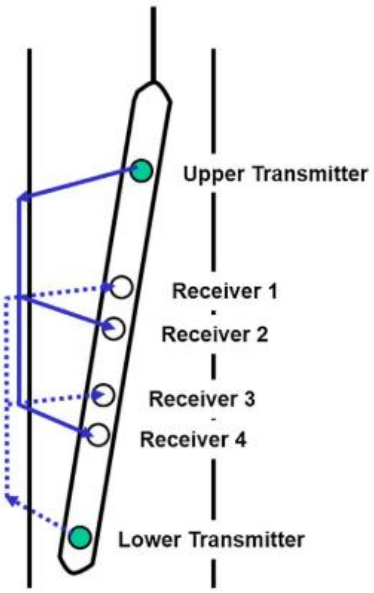

Figure 3. 23: Schematic of the sonic logging tool.*[82]*

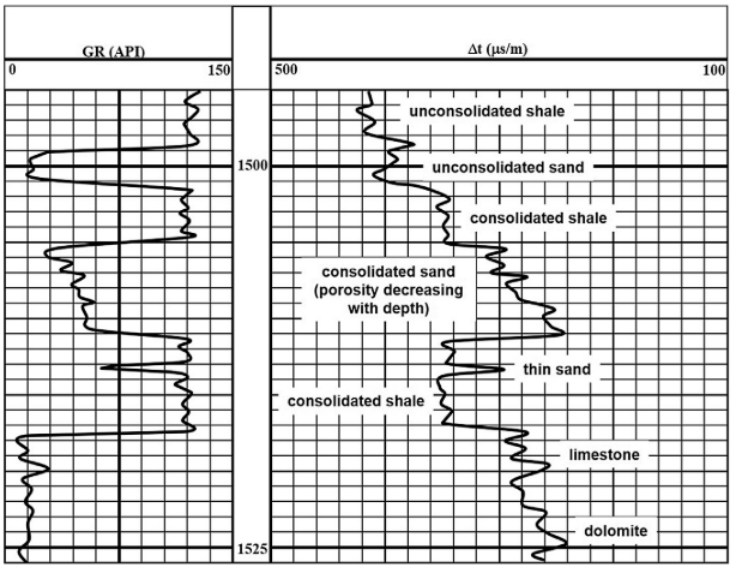

Figure 3. 24: sonic log records in different lithologies.*[82]*

3.7.6 Caliper Log (CALI)

A caliper log measures the form and diameter of a borehole. In other words, a caliper log is used in open-hole wells to measure the variation in wellbore diameter.[85] The tool has two, four or six arms, as shown in Figure 3. 25. The use of more arms will provide more accurate measurements of the well shape.

Bit size may be presented with a caliper log track, which is used to identify mud cack, wash-out, and hole roughness. Also, used in cementing calculations to determine the volume of the wellbore. [82]

When the caliper log tool is moved from the borehole, this movement allows the arms to move in and out and is then translated into an electrical signal by a potentiometer. [84]

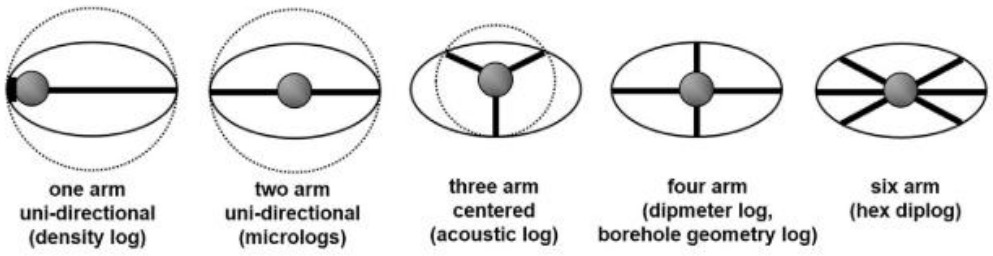

Figure 3. 25: Types of calipers configurations and typical logging applications.[82]

The combination of caliper log and bit size displays various seniors that can be summarized as follows:

- **Caliper log reading smaller than the bit size**

The presence of mudcake in the formation will result in a caliper log reading that is less than the bit size. This occurs when the lithology is porous and permeable, such as permeable sandstone, or in swelling shale formations. [84]

- **Caliper log reading larger than the bit size**

When the caliper log reading exceeds the bit size, it indicates an unconsolidated or soft formation. Drilling in this formation may result in caving or sloughing, such as unconsolidated sands, fragile shales, or salt formations drilled with freshwater drilling mud. [84]

- **Caliper log reading equal to bit size reading**

In some formations, the caliper log reading is identical to the bit size reading; these formations are well-consolidated and non-permeable, such as huge sandstones, calcareous shales, igneous rocks, and metamorphic rocks. [44]

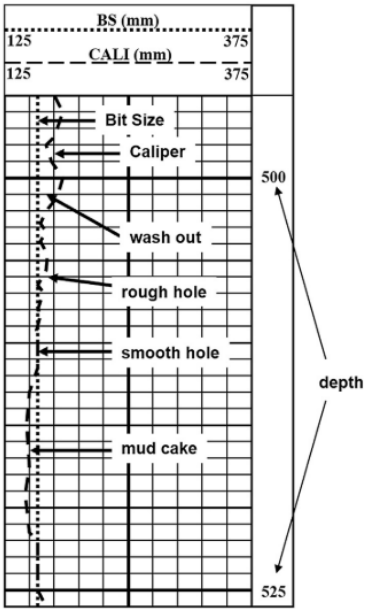

Figure 3. 26: caliper log recording. *[82]*

3.7.7 Resistivity Log (R_D & R_M)

To identify hydrocarbon-bearing or water-bearing zone resistivity logs are mostly used. Any presence of hydrocarbons, rock matrix and grains in pores, is non-conductive, the ability to carry the electricity in the pores is almost reliant on the presence of water. The resistivity of the formation increases if the hydrocarbon saturation increases as well as water saturation decreases.[18]

Three main types of resistivity logs are used to measure the formation resistivity. The following section will discuss them.

- **Conventional logging tool**

The electrode logging tool measures the voltage drops between electrodes when a current flows through the formation, as shown in Figure 3. 27. The measuring electrode is connected to the surface at a galvanometer in a normal tool arrangement. Generally, 16 in is the spacing between the recording electrode and the current electrode this is (short-normal) and also can be 64 in (long-normal) or 18 ft 8 in for the lateral log which measures the voltage drops between two electrodes.

However, it's important to note that a greater depth of investigation is achieved with greater spacing between electrodes. [82], [86]

- **Laterolog**

The current flows horizontally into the formation using single electrodes. Above and below the current electrodes placing two guard electrodes helps to create horizontal flow. A sheet current flows through the formation by balancing the guard electrode current with the central electrode, Figure 3. 27 shown later log scheme. The voltage of the central and guard electrode is measured as the sone is raised. In general, later logs are used in low-resistivity and salt mud. [86]

- **Induction logs**

This type of resistivity tool is used for freshwater or oil-based mud with low resistivity. A high-frequency alternating current is transmitted using transmitter and receiver coils placed at both ends of the sonde. A magnetic field is created by these currents which flows through the formation. Thus, according to formation resistivity the measured current is received, as shown in Figure 3. 27. [86]

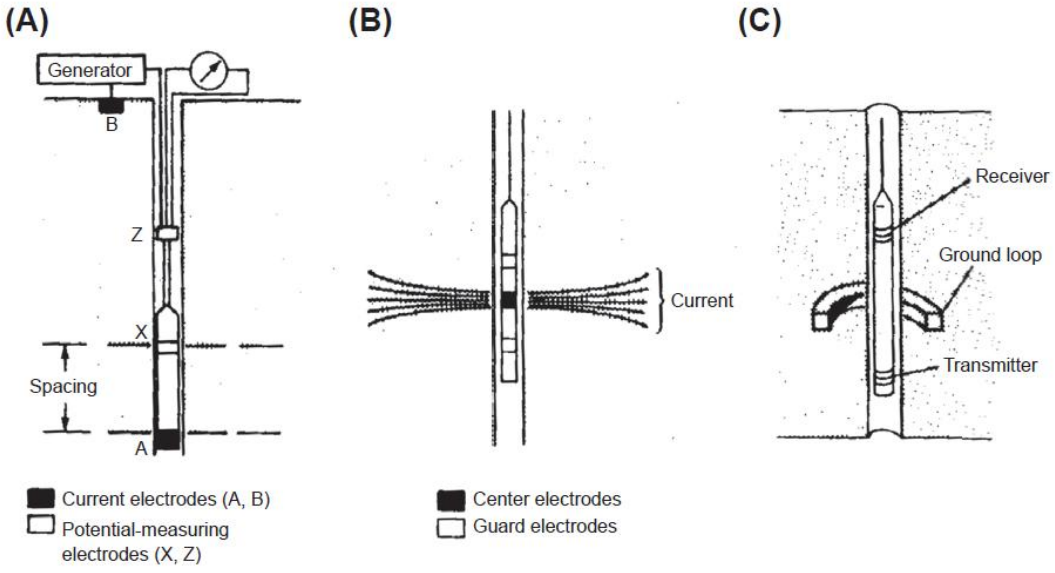

Figure 3. 27: Illustrations of three different types of resistivity logging equipment. (A) A typical resistivity logging device. The distance from the borehole where the resistivity is measured is determined by the variation in the spacing between A and X. (B) A diagram of the later log approach. (C) Diagram illustrating the induction principle.*[86]*

CHAPTER FOUR

4 STUDY METHODOLOGIES

4.1 INTRODUCTION

This chapter provides a detailed methodology approach and data that is employed throughout this study. The primary goal of this chapter is to outline the source of the dataset and describe the analytical techniques applied to achieve the study objective outlined in earlier chapters. Thereby, this chapter provides a foundation for understanding how the data was collected and preprocessed which involves converting raw data to a clean dataset which can the effectively visualized and analyzed, and ensuring reliability and validity of this study's findings.

The data utilized in this study comprise petrophysical well-logging data (conventional well-logging data) and correlated core data (porosity, permeability, and water saturation) which is the continuous dataset. Hence, this data was chosen from the Libyan field due to availability.

The methodology framework applied in this study involves data cleaning, visualization, model generation, model evaluation, and model selection and the final step is comparison with traditional methods.

The following section of this chapter is composed as follows: section 4.2 discusses the methodology workflow which provides a detailed description of data collection and visualization. Followed by data preprocessing which involves data cleaning techniques and normalization. Thereafter model generation starts with the split dataset to training and test set which used the selected feature to predict the labels (porosity, permeability and water saturation) then the model was evaluated using evaluation metric correlation coefficient (R^2) and Root mean square error (RMSE). Finally, the ML model result will be compared based on evaluation metrics to select the best ML model then each model compared with the actual result of the label selected.

4.2 Methodology

- **Python Programming**

Recently, python programming became a popular language for building machine learning models. The Python programming language was discovered to have the following advantages: simple syntax, which makes the code easy to create, read, and maintain; a big standard library, which supports object-oriented programming and other paradigms; cross-platform compatibility, which supports practically all modern systems. The software interface is user-friendly, so it's easy to design software solutions for the oil and gas business.[87]

- **Study Workflow**

Throughout this study, the main workflow is to apply machine learning models that were used to estimate reservoir properties using a variety of Python libraries, including Scikit-Learn (machine learning library), NumPy (numerical library), Pandas (DataFrame library), Matplotlib (graphic and visual library) etc. The general workflow of this study is illustrated in Figure 4. 1.

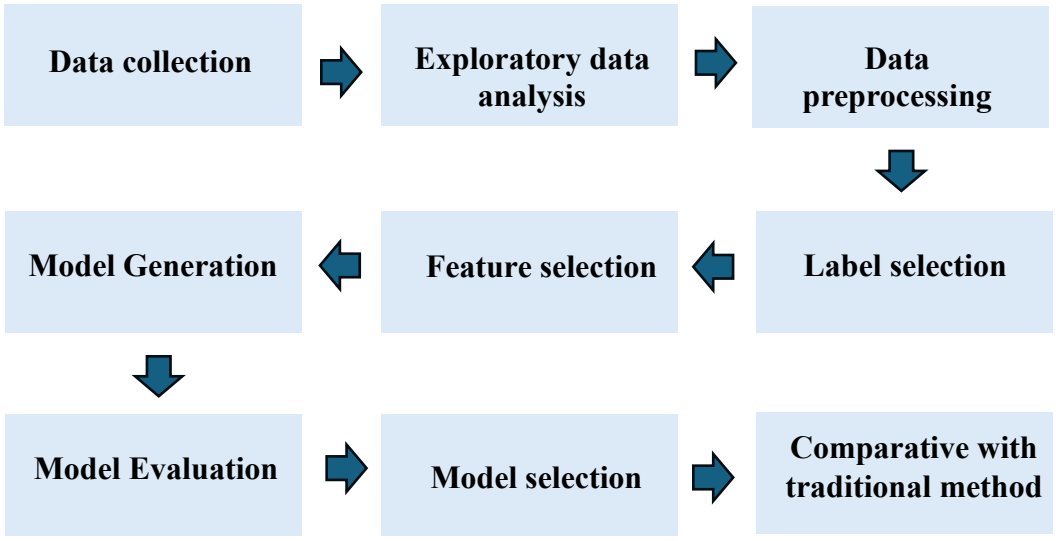

Figure 4. 1: Illustrate the workflow of the study.

To apply the workflow will use Anaconda Jupyter Notebook to write and execute Python code. Anaconda is a platform for data sciences which includes a wide range of libraries that are already installed in the notebook that is necessary for data sciences and machine learning

Figure 4. 2, shows the Jupyter Notebook window. However, if there is a library that is uninstalled in the notebook it is easy to install a used command prompt or Anaconda command prompt and use (pip install library name).

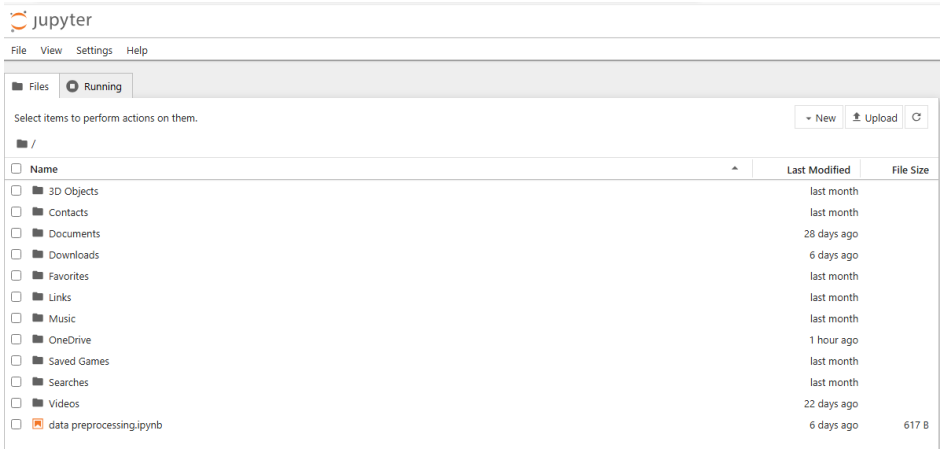

Figure 4. 2: window of Jupyter notebook.

4.2.1 Data collection

This section addresses, principal information about used well-logging and correlated core data. The well-logging data is provided in an Excel sheet (CSV format). Before starting the analysis, the data quality must be thoroughly checked to ensure the reliability and accuracy of the predictions. High-quality data plays a crucial role in enhancing the performance of machine learning models, as it directly impacts the accuracy of the reservoir property predictions. Data quality involves managing missing data; however, when available data is complete and includes all high and low values variables, it can enhance the model's ability to learn all patterns in the reservoir zone. High-quality, comprehensive data allows machine learning models to make more accurate predictions by capturing the full range of variability within the dataset. Additionally, it can decrease the time required to test various machine learning algorithms and perform hyperparameter tuning in the model development process.

For wells that have specific well-logging data such as (CAL, ILD, RHOB, GR, ILM, MSFL, CNL, BADHOLE, VSHALE, DT, SP) and correlated core data (phi, K, Sw) these wells used in training and testing dataset as well as in blind well. However, due to limitations in core data for the selected areas, the label selection used was correlated to core data these data were used as labels for machine learning models.

Hence, seven wells (A-13, A-12, A-02, A-06, A-08, A-09, A-97) were used in this study as a training dataset while well (A-05) was used as a blind well to test the model accuracy, as shown in Figure 1. 3.

The lithology of these wells was mainly sandstone with some interbeds of shale siltstone and mudstone.

The well-logging data measures function in depth which is in feet unit. The selected depth used to train and test the ML model was at the reservoir zone which is the zone of interest in any reservoir study. The depth for well A-13 was from 8686 ft to 8808 ft and the lithology for this interval was shaly sand with some interbeds of siltstone and mudstone.

While well A-12 the interested zone extends from 8672 ft to 8740 ft which the lithology was sandstone with the presence of shale.

As well as, well A-02 which extends from 8767 ft to 8790 ft the lithology was sandstone with the presence of shale. Whereby, in well A-06 the depth was from 8760 ft to 8830 ft and the lithology was sandstone with some percentage of shale. Whereas, well A-08 extend from 8680 ft to 8776 ft and the lithology was sandstone with some interbeds of shale. Whilst, well A-09 from 8750 ft until 8826 ft as well as the other wells the lithology was sandstone with some percent of shale and siltstone. Through, well A-97 start from 8674 ft until 8772 ft which the lithology was sandstone with some percentage of shale. As well as, the blind well A-05 extends from 8640 ft to 8740.5 ft as well as other wells the main lithology was sandstone with some percentage of shale.

However, in these reservoir depths, there were some missing values so it was removed before starting the analysis. Table 4. 1, shows the depth with the formation description.

The well logging data show different readings of gamma-ray as well as neutron density which indicate the presence of sandstone with some percentage of shale. Figure 4. 3 and Figure 4. 10 illustrate correlated core and logging data in the reservoir zones in studied wells. However, all the following figures were created using Python and their library.

Table 4. 1: the depth of the reservoir zone with formation description.

Well name	depth (ft)	Lithology

A-13	8686 - 8808	Sandstone with thin Shale siltstone and mudstone intercalations
A-12	8672 - 8740	
A-02	8767 - 8790	
A-06	8760 - 8830	
A-08	8680 - 8776	
A-09	8750 - 8826	
A-97	8674 - 8772	
A-05 (Blind)	8640 – 8740.5	

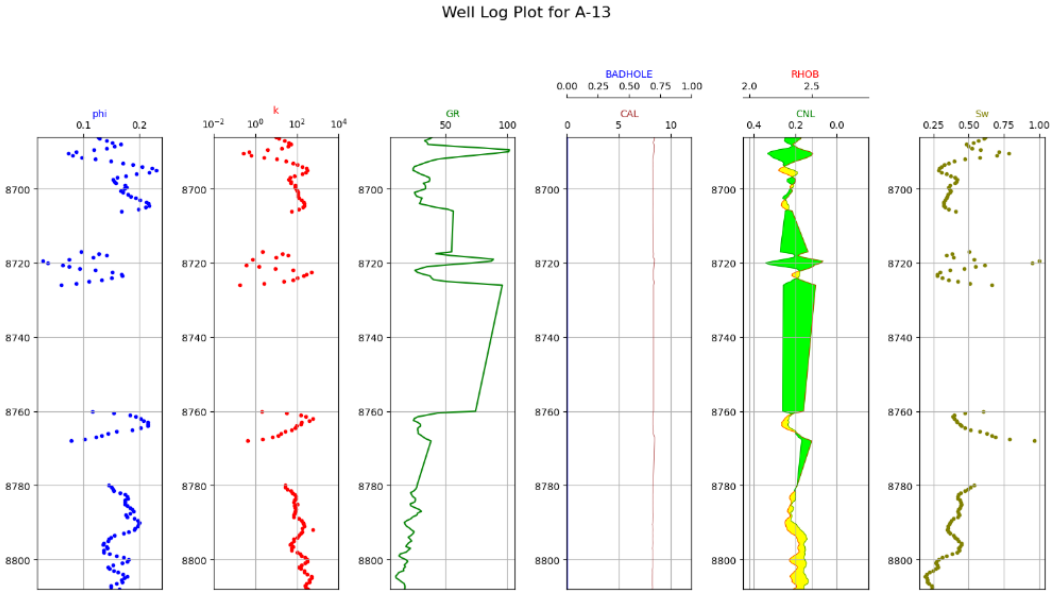

Figure 4. 3: the display of well logging data with core data, well A-13.

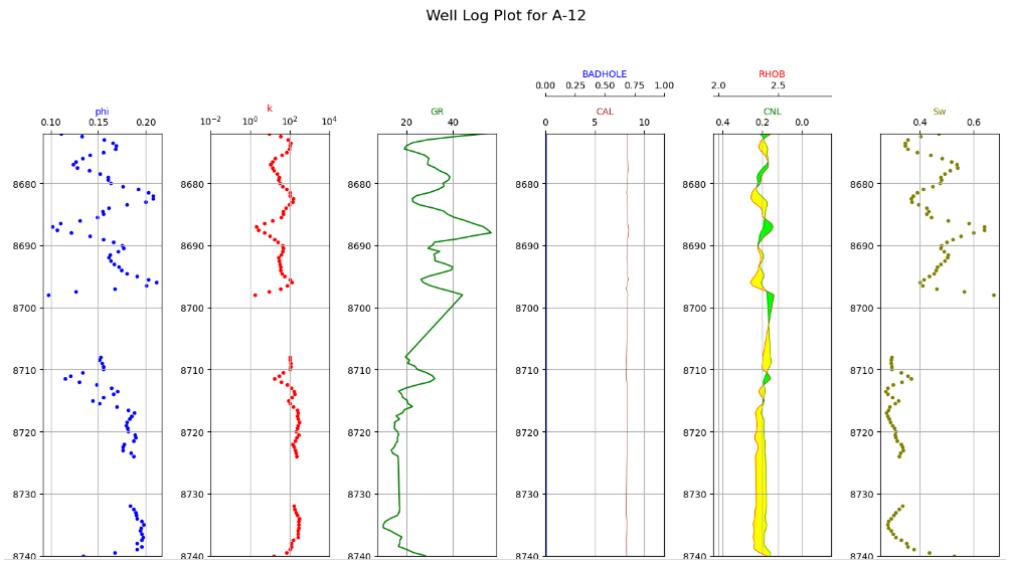

Figure 4. 4: the display of well logging data with core data, well A-12.

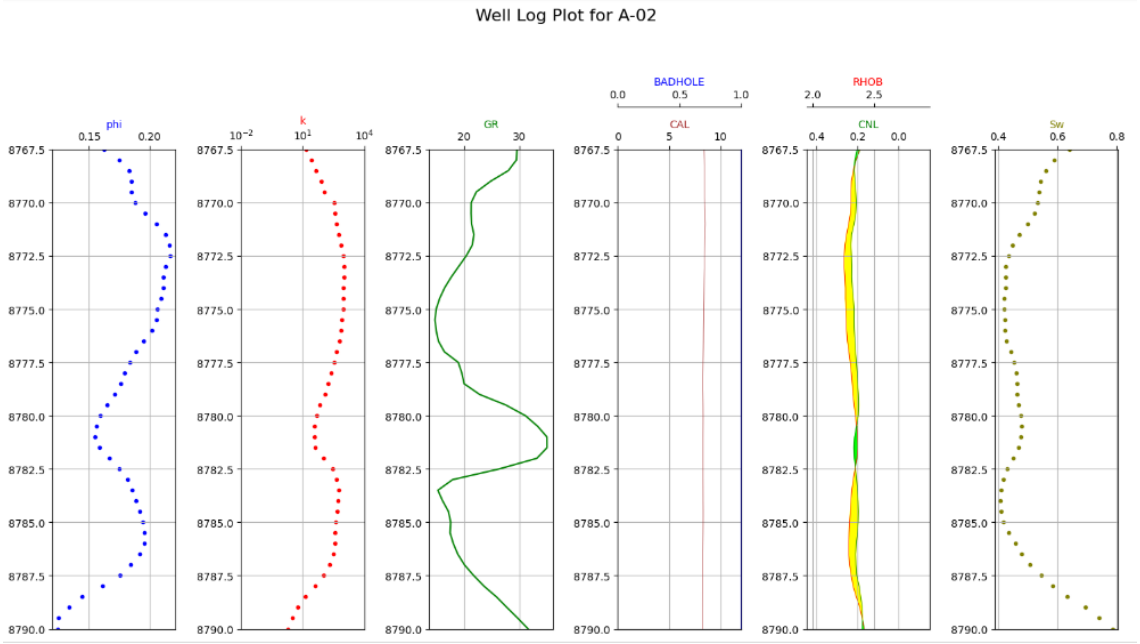

Figure 4. 5: the display of well logging data with core data, well A-02.

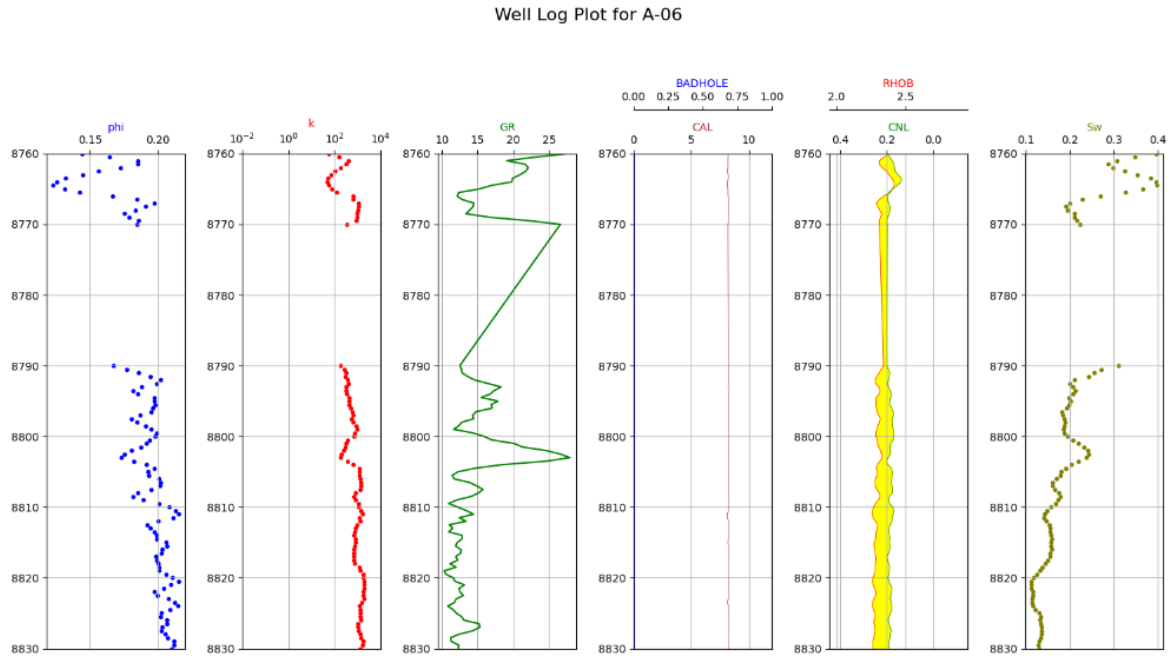

Figure 4. 6: the display of well logging data with core data, well A-06.

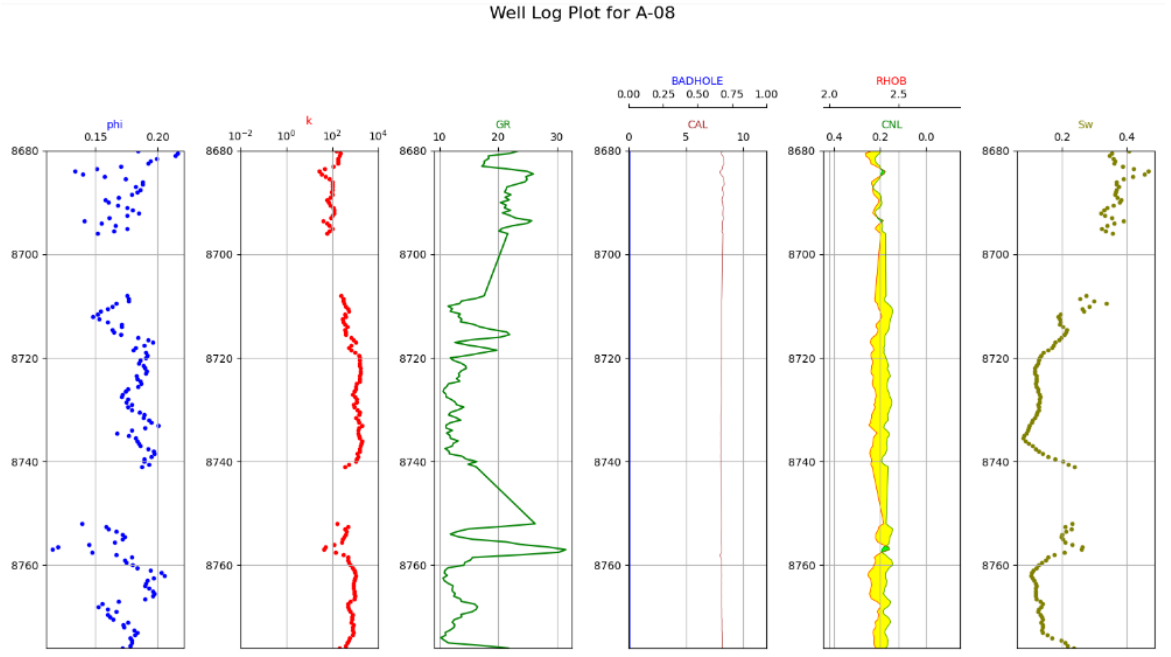

Figure 4. 7: the display of well logging data with core data, well A-08.

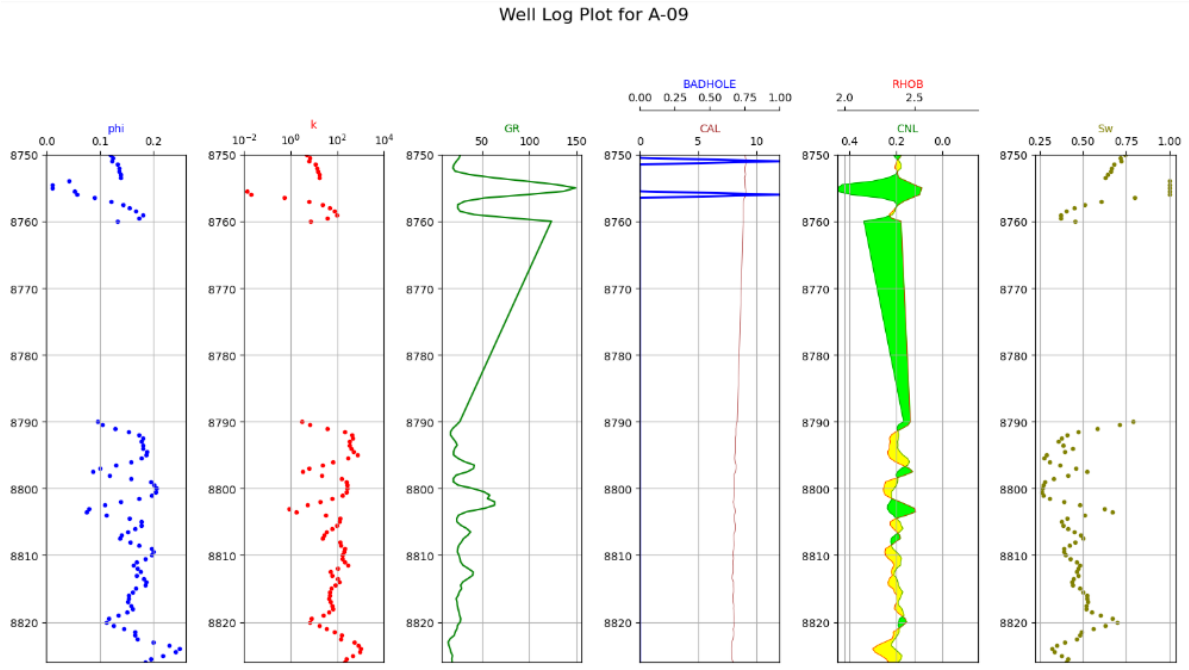

Figure 4. 8: display of well logging data with core data, well A-09.

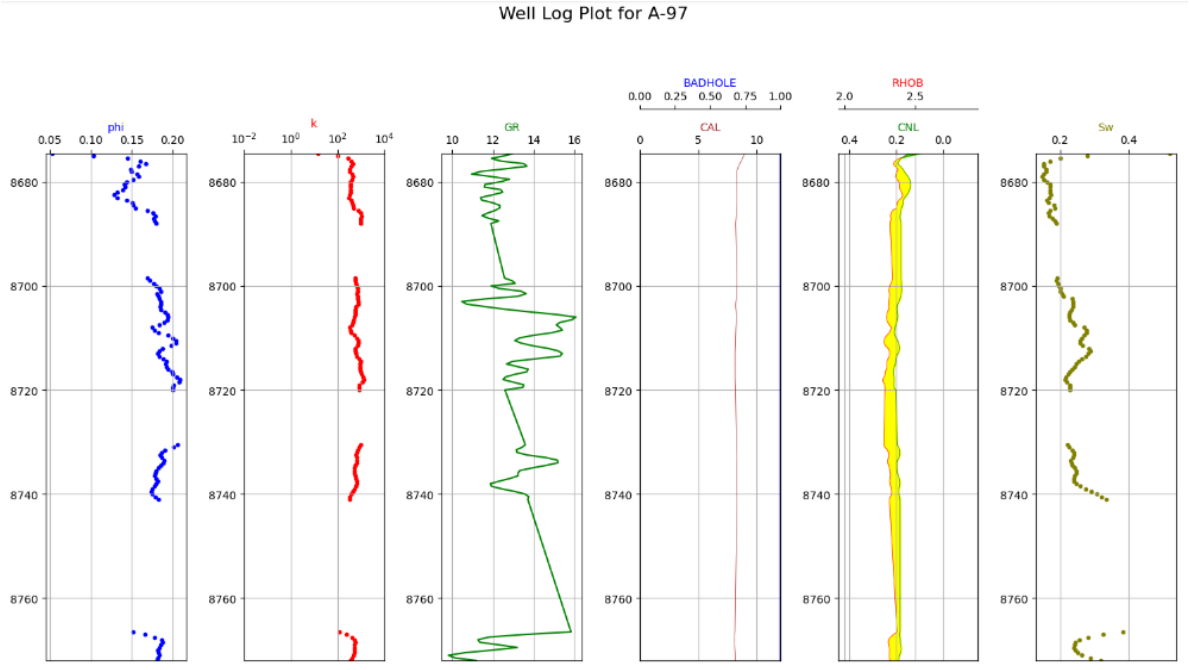

Figure 4. 9:display of well logging data with core data, well A-97.

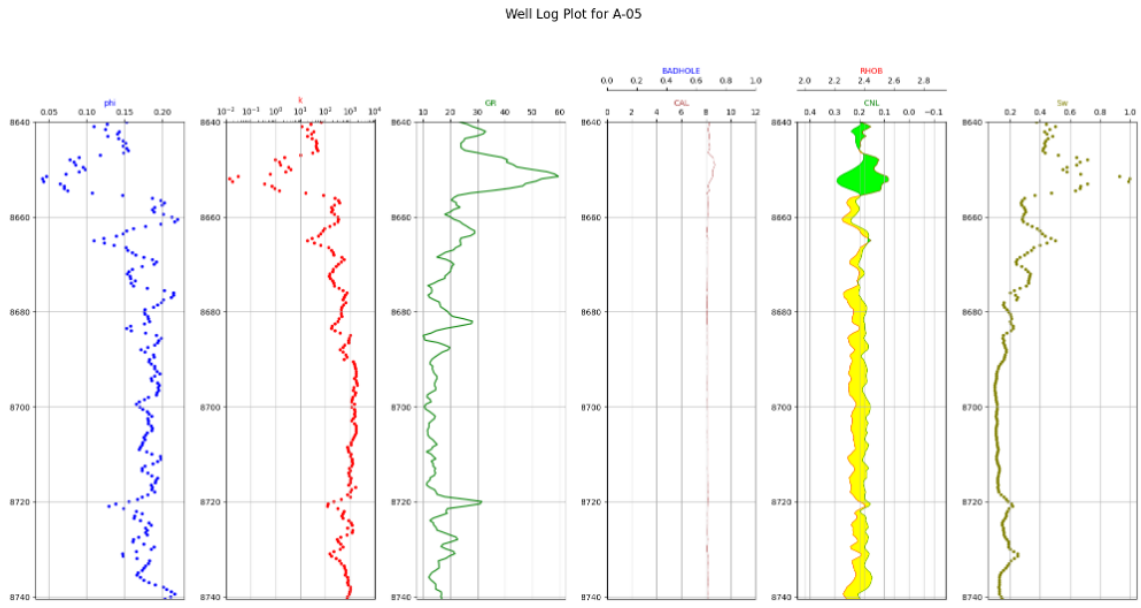

Figure 4. 10: the display of well logging data with core data, well A-05(blind well).

4.2.2 Exploratory Data Analysis

The data available in the real world are not available in perfect format and are not prepared for use in machine learning (ML) models as well as petrophysical data. Thus, the critical step before starting training ML models is Exploratory Data analysis (EDA) which helps in understanding and preparing data for ML.

EDA includes understanding the relationship between the variables, identifying inconsistencies, and checking assumptions.[57]

Therefore, during this step, Pandas and NumPy libraries were used to summarize statistics and for visualization techniques python libraries such as matplotlib and Seaborn were used and detect the nature of the dataset and answer the questions about the available data. Table 4. 2, illustrate the statistics summary of the dataset. The total number of data points for training data is about 734 points, all of these points were at the reservoir zone. The number of data points in blind well A-05 is about 202 points as shown in Appendix B. Based on Table 4. 2, the training dataset doesn't have a null value that is because all null values deleted from the CSV file before starting in analysis. However, almost all data points have well-logging reading and correlated core data as shown in Figure 4. 11.

Data visualization is a critical method in EDA that is used to understand the data distribution and identify the outliers. An outlier is a parameter that has a range out of the reasonable.

The boxplot was used to visualize the dataset which lead to detecting the outliers using Eq 4. 2, Eq 4. 3 and Eq 4. 3. Figure 4. 12, shows the outlier for the training dataset.). However, all outlier in the dataset is related to the geology. Hence, the removal of any of them affects the model reality.

$$\boldsymbol{IQR = Q3 - Q1} \qquad \textbf{Eq 4. 1}$$

$$Rmin = Q1 - 1.5 * IQR \qquad \textit{Eq 4. 2}$$

$$\boldsymbol{Rmax = Q3 + 1.5 * IQR} \qquad \textbf{\textit{Eq 4. 3}}$$

Where:

IQR = the inter quartile range.

Q3 = 75th percentile.

Q1 = 25th percentile.

R_{min} = minimum outliers.

R_{max} = maximum outliers.

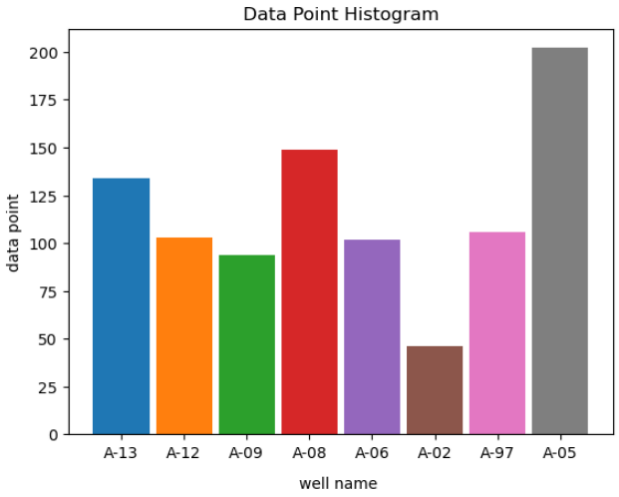

Figure 4. 11: The number of data points for each well.

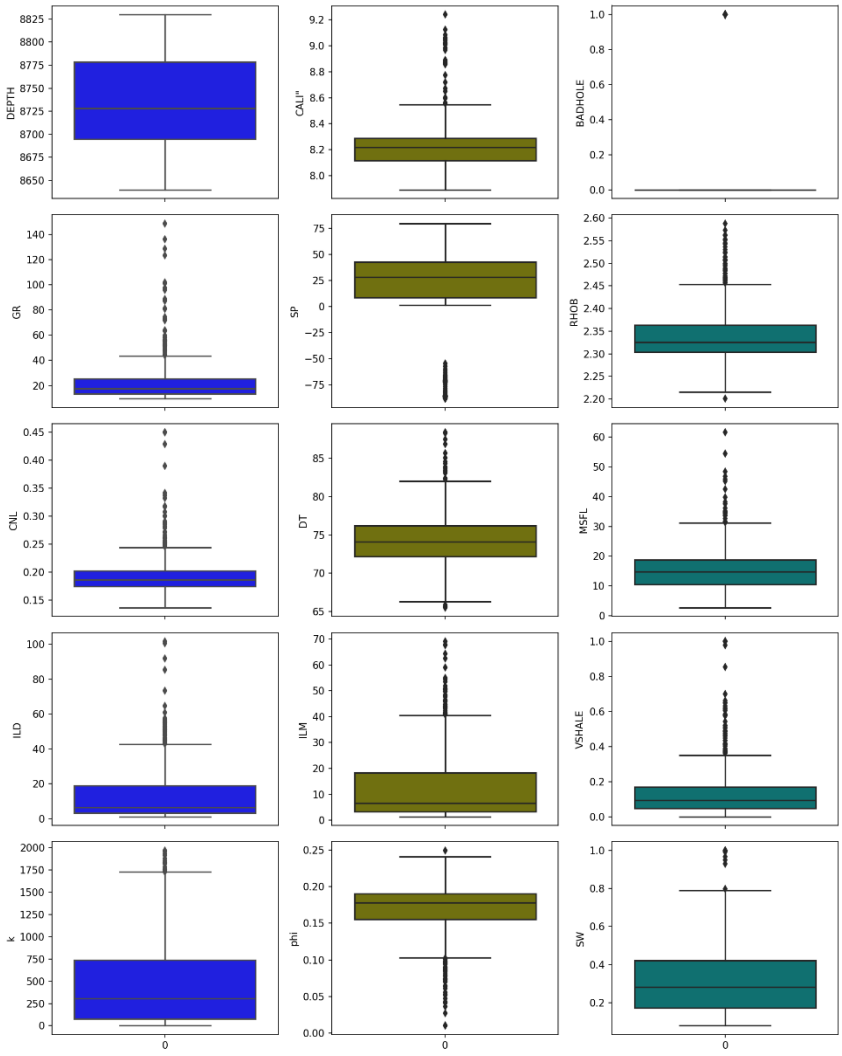

Figure 4. 12: dataset boxplot.

Table 4. 2: statistics summary for the training dataset.

Index	count	mean	std	min	25%	50%	75%	max
DEPTH	734	8748.846	45.700723	8672	8709.5	8753.25	8791	8830
CAL	734	8.2443557	0.1722234	7.891	8.14945	8.234	8.2998425	9.242
ILD	734	11.104149	13.263816	1.10609	2.68175	5.592	14.129	101.608
RHOB	734	2.3371622	0.0533698	2.201	2.302	2.325	2.364	2.588
GR	734	22.915112	15.895796	9.594	13.232555	18.04628	26.594	148.625
ILM	734	10.642556	11.243764	1.391	3.06975	5.961	14.10525	69.044
MSFL	734	15.808876	7.2606582	2.709	10.0065	14.8675	19.348508	61.531
CNL	734	0.1927011	0.0330029	0.136	0.176	0.189	0.204	0.45
BADHOLE	734	0.2098093	0.40745	0	0	0	0	1
K	734	423.59333	451.16039	0.0008	71.130425	258.39099	662.99569	1968.7157
phi	734	0.170704	0.0317291	0.01	0.1554	0.178065	0.191175	0.2495
VSHALE	734	0.1435133	0.1477706	0.001	0.0538	0.102395	0.18475	1
SW	734	0.3299444	0.1657051	0.0792	0.198825	0.30915	0.4275275	1
DT	734	74.604067	3.569266	65.5	72.25	74.69331	76.6	88.4
SP	734	3.1850791	44.383595	-88.07013	4.5	13.75	34.461	71.25

4.2.3 Data Preprocessing

Data preprocessing is a vital step for building a machine learning model as it involves converting raw data to a clean dataset that can be effectively visualized and analyzed. Corrections for this purpose include repairing null/NaN values, allocating missing values, scaling features, normalizing samples, removing anomalies, encoding qualitative/nominal category features, and reformatting data. [5], [57]

Data preparation is crucial for assuring the quality and dependability of the input data for machine learning models. That can confirm better results from machine learning models. Thus, the scikit learn library was used during this study to perform data preprocessing and machine learning models.

The data preprocessing includes label selection, feature selection, data normalization/scaling and standardization, data splitting and cross-validation.

- **Label selection**

The label selection is a part of data preprocessing which is the variables that the model is trying to predict, porosity, water saturation and permeability are selected as the labels for this study.

- **Feature selection**

Feature selection is a method to select the needed feature used to build the model and contribute to predicting the output. The main benefit of feature selection is to reduce overfitting. However, feature selection has another advantage increases accuracy where modeling accuracy increases with fewer false data. As well as, Shortens Training Time (STT) which algorithms can learn more quickly with less data. [88]

Typically, increasing model performance reduces the error rate and avoids overfitting by selecting good features for each prediction (porosity, permeability and water saturation).

There are several methods to select the best features for model generation such as the filter method which selects the features based on statistical methods.

Heat map is used to understand the relationship between variables which used the seaborn library to plot such map.

A heat map is a two-dimensional representation of data based on colour. Moreover, with plotting a heat map, it's essential to use Pearson correlation coefficient to detect the collinear variables.[89] Eq 4. 4, used to calculate Pearson correlation.

$$P_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \qquad \textit{Eq 4. 4}$$

Where:

cov (X, Y) is the covariance between parameters X and Y.

σX is the standard deviation of parameter X.

σY is the standard deviation of parameter Y.

The Pearson correlation coefficient varies between -1 to 1 which measures how a strong or weak relationship between variables. Thus, the positive values indicate as one variable increases the other value increases while the negative value is related to an increase in one value and a decrease in another value and vice versa. [89] The heat map for the training dataset is illustrated in Figure 4. 13. A heat map with a Pearson correlation coefficient helps detect relevance between features for each prediction process and reflects the possibility of using a regression model. However, during this study features were selected according to specific requirements for each property even if there is a low linear relationship.

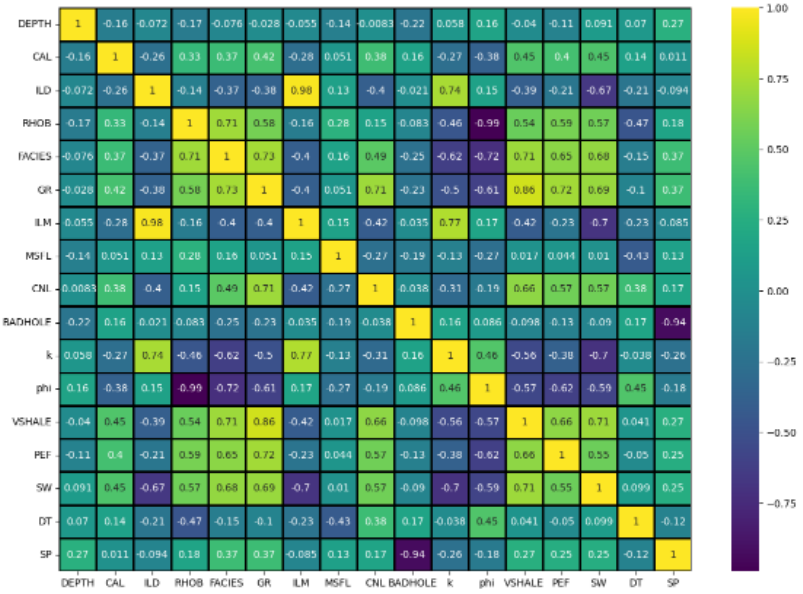

Figure 4. 13: Heat map plot for the full dataset.

- **Data normalization, scaling, and standardization**

Various machine learning algorithms try to find trends in the data points by comparing data point features. However, when the features are completely on different scales considered there is a problem. Thus, the aim of normalization is to make all data points uniformly at the same scale so all feature has equal influence.[90]

Typically, making all features on the same scale can speed up the optimization algorithms. On the other hand, not all ML algorithms are affected by normalization such as Tree-based models (e.g. DT and RF) these models are not distance-based so can handle variations of data such as petrophysical data.

There are several approaches to normalization and standardization. The popular approaches are shown in the following section:

- **Min\Max normalization**

It is one of the most popular approaches to normalize the data. Eq 4. 5 show the minimum\maximum normalization. By using this equation all variables range between 0 and 1. [5] However, this method is used when the maximum and minimum bound of data are known, and the data have little or no outliers. Thus, need to remove outliers before applying this method. [57] In Python can normalize the dataset by using Scikit-Learn using the MinMaxScaler class.

$$\boldsymbol{X_n = \frac{Xvalue - min\ (X)}{max(X) - min\ (X)}} \qquad \textit{Eq 4. 5}$$

Where:

X_n: is the normalized feature of variable x.

Min(x): is the minimum value.

Max(x): is the maximum value.

- **Stander scaler**

Standardization also known as z-Score normalization this approach rescaled the data by converting the Normal distribution feature to Normal distribution with zero mean and one standard deviation. [5], [89]

Thus, this approach is used to minimize the effects of low standard deviation and outliers. Eq 4. 6, can perform to standardization of the features. In Python can Standardized the dataset by using Scikit-Learn using the StandardScaler class.

$$X' = \frac{X - \mu}{\sigma} \qquad \textit{Eq 4. 6}$$

Where:

X'_i is the standardized data point.

μ: is the mean of the data set.

σ: is the standard deviation of the data set.

- **Robust scalar**

Both previous methods have a common limitation which is sensitive to the outliers. The presence of the outliers affects the computation of mean and standard deviation. [91] So, the Robust approach uses the quantile ranges from 25% to 75% to scale the dataset and remove the median. Eq 4. 7, can used to apply a Robust scalar. In Python can Standardized the dataset by using Scikit-Learn using the RobustScaler class.

$$\hat{z}_{i=}\frac{value-m}{QR} \qquad \textit{Eq 4. 7}$$

Where:

$\hat{z}_i$: is the Robust scalar.

m: is the median

QR: is the quantile range (e.g. IQR (Q3-Q1))

The selection of any normalization method used is mainly based on the dataset and ML algorithms used. However, this study didn't use any normalization methods due to the use of Tree-based models (DT and RF).

4.2.4 Model Generation

Since the predicted output (porosity, water saturation and permeability) is a continuance data, the supervised regression technique was used in this study, as discussed previously. In this study, the model was generated by using Python script which used Scikit-Learn library which contained several ML models.

The following section discussed the workflow for the generation of the machine learning models:

- **Data splitting**

To generate a machine learning model the dataset must be split into three sections after preprocessing it which are training, validation, and testing. Splitting datasets in machine learning can be done in a variety of ways. One of the most used approaches is to split the dataset randomly in a specific ratio. Every one of these segments has a different percentage of the dataset. Typically, use 20–40 percent of the data for testing and validation, and 60–80 percent for training. The entire dataset can be assembled into a single workbook (e.g., TableauTM, AccessTM, ExcelTM, etc.) and then split the data into segments for training, testing, and validation at random.[5]

Typically, the quality of the training and testing dataset can affect the model generation which increases or decreases the model performance. In this study, the selected dataset (input and output) from seven wells was saved in an Excel worksheet (CSV file) which was used after that to train and test ML models using Scikit-Learn library by import train_test_split function. Hence, the blind well (A-05) used for the validation of ML models was saved in a separate Excel worksheet (CSV file).

Typically, the training dataset is split randomly which uses 25 – 75 percent of the data for testing and training.

- **Model evaluation**

Model evaluation is an important step in the machine-learning workflow. Which uses quantitative metrics to evaluate the model performance to comprehend how well the model generalizes on an unseen dataset. [53], [57]

There are several evaluation metrics used for model evaluation which depend upon the problem. However, for regression algorithms coefficient of determination (R^2), and root-mean-squared error (RMSE) are widely used. The following section discusses the equation for R^2 and RMSE:

- **Square of Correlation Coefficient**

This is the most widely evaluated metric method that is used in regression models. This method is also known as the coefficient of determination (R^2). This can be stated as a ratio that represents the strength of the linear relationship between the actual and predicted responses.[5], [57] In other words, the coefficient of determination (R^2) is a measure of the variability in dependent variables (target value) that can be anticipated by the independent variable (model prediction).

Eq 4. 8, used to calculate the coefficient of determination (R^2). Typically, python uses the Scikit-Learn library to import metrics and then import the r2_score function.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad \textit{Eq 4. 8}$$

Where SS_{res} and SS_{tot} are calculated using Eq 4. 9 and Eq 4. 10.

$$SS_{res} = \sum (y_i - y_{reg})^2 \qquad \textit{Eq 4. 9}$$

$$SS_{tot} = \sum (y_i - \bar{y})^2 \qquad \textit{Eq 4. 10}$$

Where:

SS_{res}: is the sum of squares residual.

SS_{tot}: is the total sum of squares.

y_i: is the value of each data point.

y_{reg}: is the predicted value by the regression model.

$\bar{y}$: is the mean value.

The better R^2 value is achieved if the R^2 for the training and testing dataset is 1 or close to 1 as well as R^2 for a blind dataset which means that the model well captures the data point.

- **Root Mean Square Error**

The root mean square error (RMSE) is defined as the square root of the average of the squared difference between actual and anticipated values. RMSE resemble to standard deviation.[53] In contrast with R^2, the less error model and better performance were achieved with low RMSE.[5] Eq 4. 11, used to calculate RMSE. Python used the Scikit-Learn library to import metrics and then import root_mean_squared_error regression loss.

$$RSME = \sqrt{\frac{\sum_1^n (\hat{y}_i - y_i)^2}{n}} \qquad \textit{Eq 4. 11}$$

Where:

$\hat{y}_i$: is the predicted value by the regression model.

n: is a number of data points.

In this study, R^2 and RMSE were used to evaluate the model accuracy.

- **Overfitting and Underfitting**

Typically, the main target for training machine learning models is to generate predictions for unseen datasets. Thus, unseen dataset prediction should provide the same level of accuracy provided by the training dataset. However, in some circumstances, the model's accuracy for an unseen dataset may be limited due to overfitting or underfitting. [57]

Overfitting: indicates that the model provides high accuracy for the training dataset due to their ability to learn all patterns for the training dataset but for unseen data, models provide low accuracy due to the inability to capture well unseen data. [57]

Underfitting: occurs when the model provides low accuracy for both training dataset and unseen data. However, that's means that model unable to learn the data pattern for the training dataset.[57]

In this study, this problem is detected using R^2 for both the training dataset and the blind dataset.

- **Model selection**

For model selection until now there's nobody can select specific machine learning algorithms for specific problems. It can use different algorithms to compare the results and select the best model based on model accuracy. [69] On the other hand, Tree-based models (DT and RF) don't need to normalization and handle data with little or no preparation also the nonlinear relationship between variables doesn't affect the model performance. [92] And according to Grinsztajn et al [93] found that Tree-based models surpass on deep learning method in medium data size.

Hence, due to limited data points, this study focuses on developing the Decision Tree (DT) Regression model and Random Forest (RF) Regression model for the prediction of reservoir properties and achieving the study objective.

4.2.5 Comparative Method

According to the sensitivity of the prediction methods in any reservoir study, it's a critical step to compare the predicted value from ML models with actual data. This study used correlated core porosity, water saturation and permeability to compare with ML prediction.

CHAPTER FIVE

5 RESULTS AND DISCUSSION

This chapter delves into the study result derived from the application different machine learning models to estimate reservoir properties using well-logging data. According to the foundation laid in previous chapters, methodology workflow this chapter aimed to outline the ML models results and discuss their importance in reservoir studies.

In this section, will introduce a comprehensive analysis of the obtained result for the application of ML models in this study focusing on the accuracy and model performance based on the R^2 value and comparing with the actual result provided with the actual data in blind well (A-15).

Thereafter, will present a detailed discussion about the importance of finding. During the discussion will explore machine learning models in predicting reservoir properties, such as porosity, permeability, and water saturation. As well as, will also critically analyze the advantages and limitations of the models, considering factors such as model complexity and data quality.

Therefore, Random Forest (RF) and Decision Tree (DT) models were used to predict reservoir properties (porosity, permeability, and water saturation) based on well-logging data and correlated core data.

This study, aimed to provide significant insights into how machine learning models can help and improve reservoir properties prediction, as well as recommend future studies and prospected applications, especially in reservoir study.

5.1 porosity prediction

5.1.1 Feature Selection

As discussed in previous chapters, the feature selection for training and blind datasets is highly influenced by model performance, it's critical to find which feature has a high impact on the prediction of reservoir properties. Hence, wrong feature selection will reduce model accuracy.

Using of Pearson correlation coefficient Eq 4. 4 provides insight into the correlation between features and selected labels. However, the Pearson correlation coefficient is used to understand the ability to use regression models. Although the Pearson correlation coefficient is important for feature selection in many cases, the non-linear feature impacts the model performance which may depend on domain-specific factors.

Although, the dataset contains both measured and calculated logs for porosity prediction select measured data as input for the machine learning model. Hence, bulk density (RHOB), sonic (DT), compensated neutron log (CNL), and gamma-ray (GR) were used. The selected features are commonly used to predict reservoir porosity due to their direct relationship with it.

RHOB presented a strong P-value which was -0.99 while GR was -0.63 whereby DT was 0.45 while CNL showed a lower value in the selected feature which was -0.29, as shown in Figure 5. 1.

Typically, the indirect relationship between RHOB and phi indicates that denser rocks have low porosity while GR inversely relationship with phi indicates that high GR is related to the presence of shale and then low porosity. On the other hand, the direct relationship between DT and phi which is important to detect porosity indicates the porosity increases then the travel time increases. However, CNL shows an inverse relationship with phi. The features selected and removed are shown in Table 5. 1

Table 5. 1: Feature selection for porosity prediction.

Feature	Pearson correlation coefficient
Feature selected	RHOB, GR, DT, CNL
Feature removed	DEPTH, SP, ILM, ILD, MSFL, CAL, BEDHOLE, K, SW, VSHALE

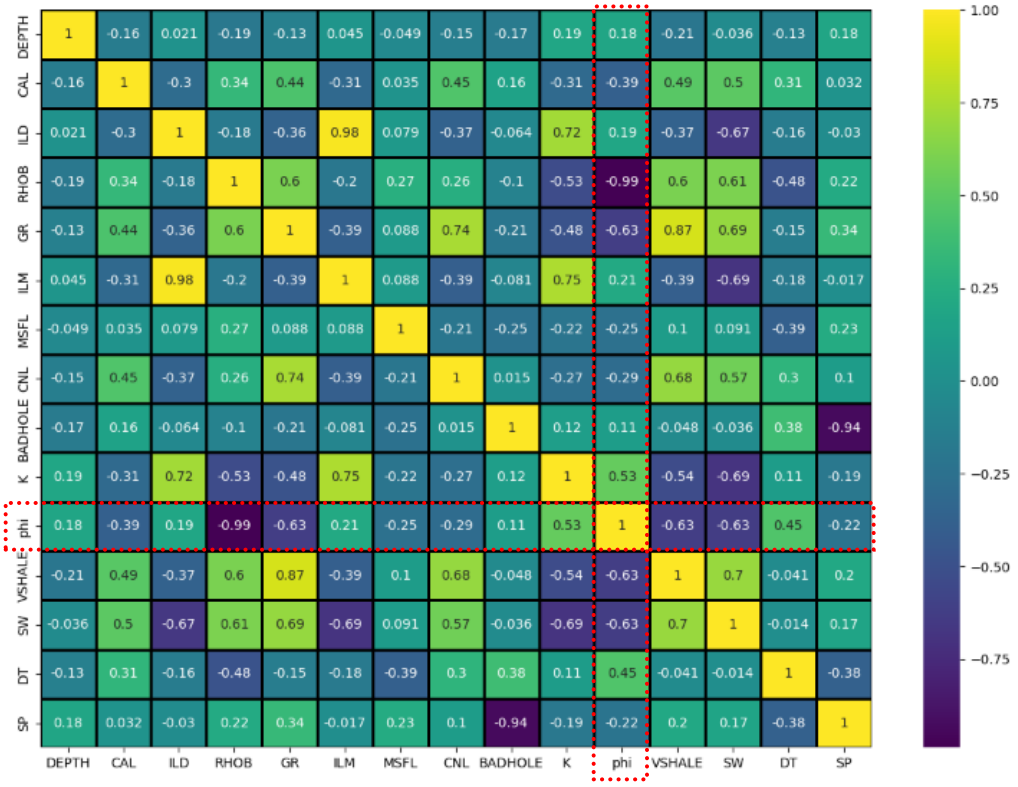

Figure 5. 1: Pearson correlation coefficient with heat map for porosity prediction.

5.1.2 Models Prediction

After preparing the dataset and selecting the feature, the next step is to select the optimal ML model to predict the reservoir porosity. This study selected the DT and RF as discussed in previous chapters which are built using well-logging data as input features and correlated core data as label output. The following section presents the obtained results for the ML models used.

The model starts by training it on the training dataset. Then, the model was evaluated based on the R^2 and RMSE values and compared with the actual value in the blind well (A-05).

Figure 5. 2, show the actual porosity plot compared to the prediction for the training dataset which the R^2 was about 0.935 using the DT model.

However, the R^2 and RMSE for the blind well (A-05) were about 0.993 and 0.002 respectively. Figure 5. 3, show the predicted porosity in the blind well (A-05) function in depth and compare it with the actual porosity (correlated core porosity).

In contrast, the RF model is built with the same steps as the DT model. However, there was a slight difference between the RF and DT results as shown in Figure 5. 3. The R^2 for the training dataset was about 0.976 while the R^2 and RMSE for a blind well (A-05) were about 0.999 and 0.001. Figure 5. 4, show the comparison between the actual porosity and predicted training dataset using the RF model. While Figure 5. 5, present the predicted porosity by RF model with the actual porosity for blind well.

Figure 5. 6, show the scatter plot between actual porosity and predicted using the DT model. The y-axis represents the predicted value while the x-axis represents the actual value. This plotting is used to evaluate model performance visually. The fitting line (red line) is plotted to analyze the prediction which presents the relationship between actual and predicted value. The high R^2 and low RMSE provided using the DT and RF models show that the DT and RF are effective in porosity prediction.

In comparison, the data point that shows the actual porosity and predicted from the Random Forest model more closely to the fitting line than the DT model which provided R^2 higher than DT as shown in Figure 5. 7. This indicates that the Random Forest model captures the majority of variability for porosity prediction more effectively than the Decision Tree, especially for low porosity values.

As well as, shown in Figure 5. 8 for porosity prediction the Random Forest (RF) model provides more accurate results compared with Decision Tree results for both training and blind datasets. On the other hand, Figure 5. 9 shows the comparison between RMSE for a blind well (A-05) using both models.

Table 5. 2: Evaluation metric for porosity prediction.

Model	Evaluation metric
R^2 (%)	
Decision Tree	99.3
Random Forest	99.9
RMSE (%)	
Decision Tree	0.28
Random Forest	0.12

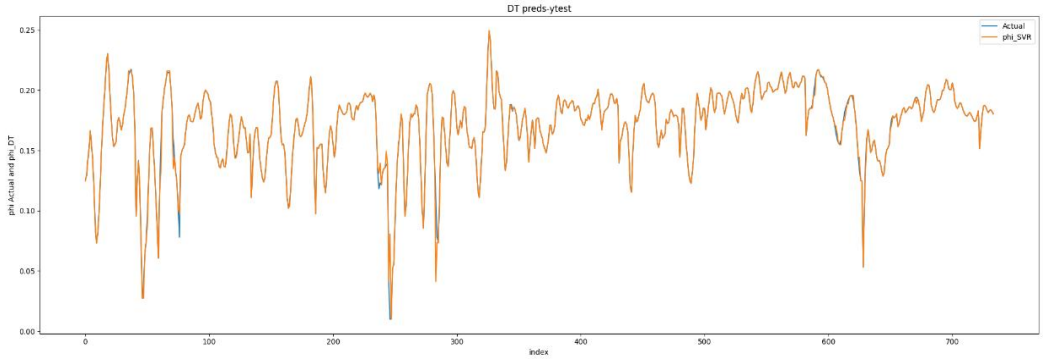

Figure 5. 2: predicted porosity versus actual porosity for training dataset using DT model.

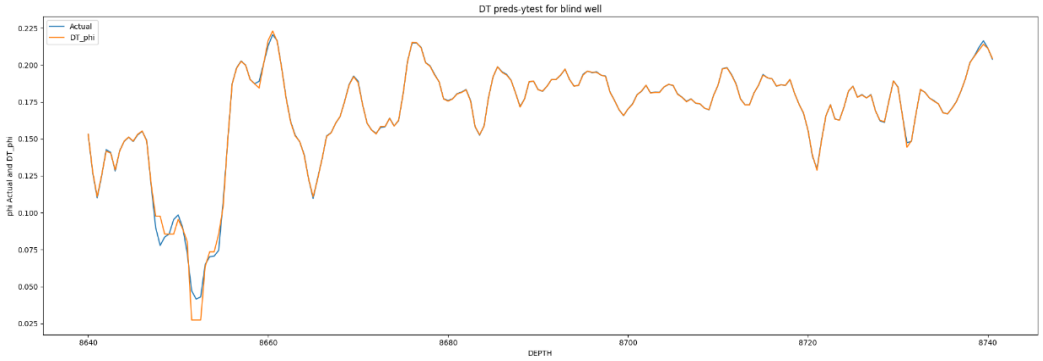

Figure 5. 3: predicted porosity versus actual porosity for blind dataset (well A-05) using the DT model.

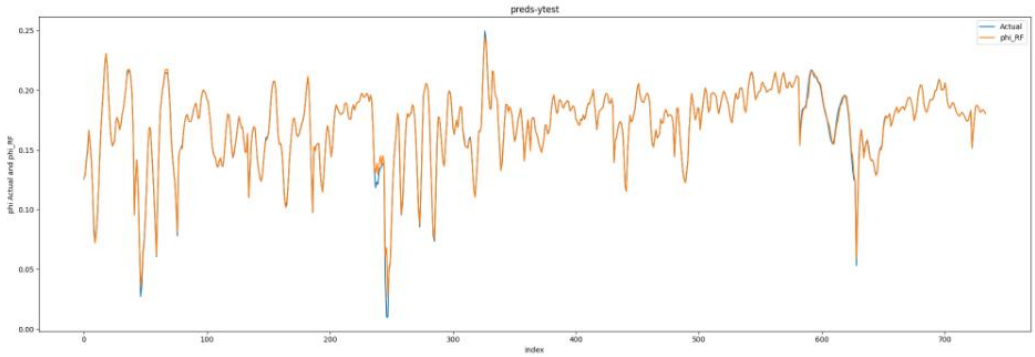

Figure 5. 4: predicted porosity versus actual porosity for training dataset using RF model.

Figure 5. 5: predicted porosity versus actual porosity for blind dataset (well A-05) using RF model.

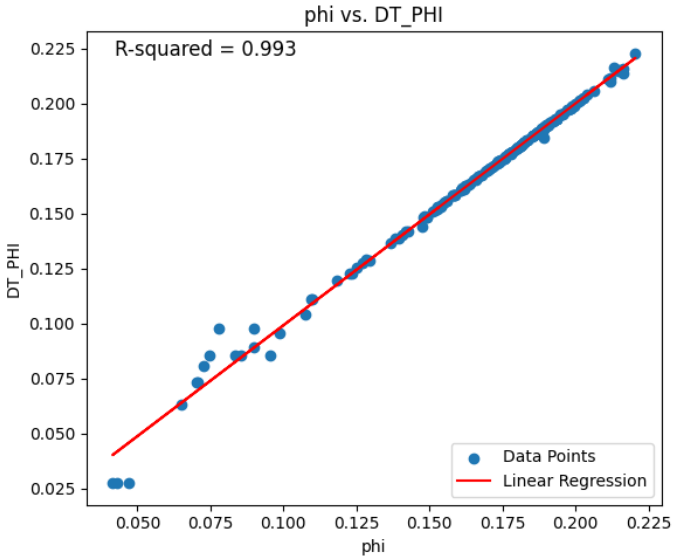

Figure 5. 6: actual porosity versus porosity prediction using DT.

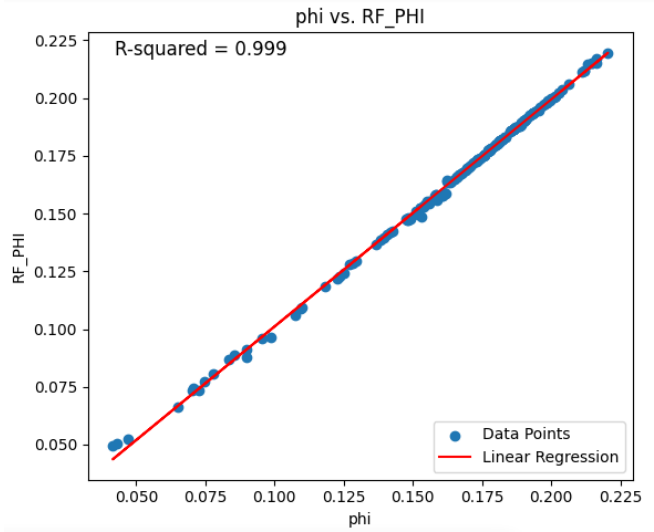

Figure 5. 7: actual porosity versus porosity prediction using RF.

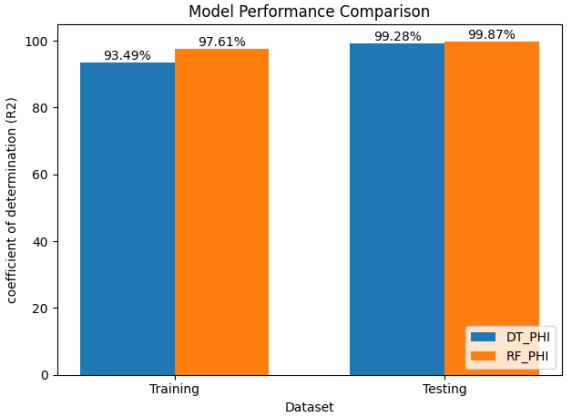

Figure 5. 8: R2 value for porosity prediction for the training dataset and blind dataset for both models.

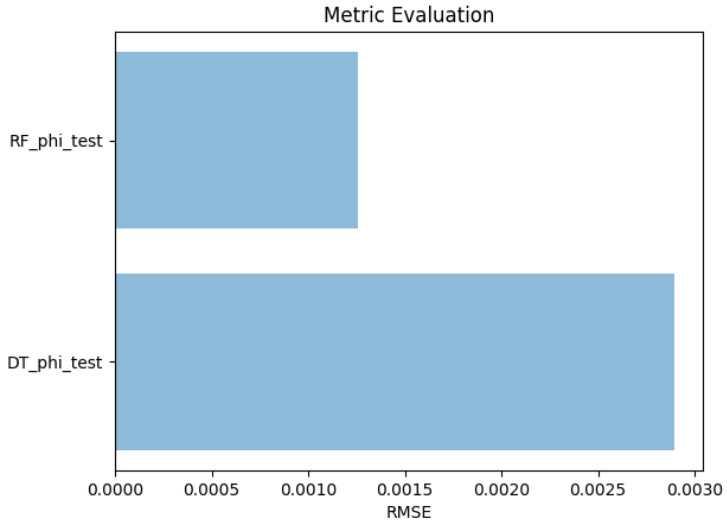

Figure 5. 9: RMSE value for porosity prediction for the blind dataset for both models.

5.2 water saturation prediction

5.2.1 Feature Selection

As discussed in previous chapters the selection of the most relevant features improved the model generalization. Hence, for water saturation prediction select deep resistivity (ILD), medium resistivity (ILM), bulk density (RHOB), compensated neutron log (CNL), and gamma-ray (GR).

The heat map illustrates the strong positive relationship with RHOB was about 0.61, as well as in the GR was 0.69 whereby the CNL was 0.57.

However, ILD, and ILM, show a strong negative relationship with water saturation. Consequently, The P-value was about -0.67 and -0.69 for ILD and ILM.

Hence, these results indicate that these input features have a strong linear relationship with water saturation. However, the selection was based on their impact on the model. Table 5. 3, illustrates the selected and removed features for training the model.

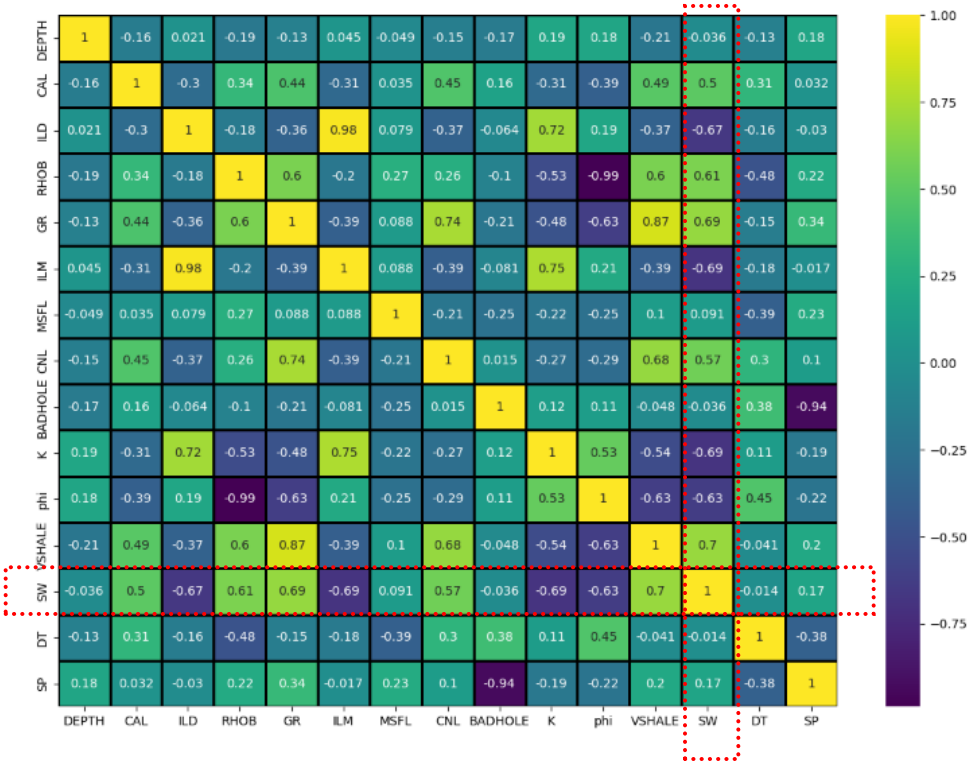

Figure 5. 10: Pearson correlation coefficient with heat map for water saturation prediction.

Table 5. 3: Feature selection water saturation prediction.

Feature	Pearson correlation coefficient
Feature selected	ILM, ILD, RHOB, GR, CNL
Feature removed	DEPTH, SP, MSFL, CAL, BEDHOLE, K, SW, VSHALE, DT

5.2.2 Model Prediction

As mentioned before, this study used DT and RF models as ML models. The models trained on the training dataset to predict the water saturation based on well logging data and correlated core data for the labelled output, then the models were evaluated using R2 and RMSE and compared with the blind well (A-05).

Figure 5. 11 and Figure 5. 13, show the actual porosity plot compared with the predicted porosity from the training dataset from DT and RF respectively. Where the R2 for the training dataset was about 0.93 using the DT model while using RF was about 0.94. However, for the blind dataset well (A-05) R2 and RMSE were about 0.973 and 0.028 by using the DT model and 0.969 and 0.03 by using the RF model. Figure 5. 12 and Figure 5. 14, show the comparison between actual water saturation and predicted using DT and RF models.

Figure 5. 15, shows the scatter plot between actual water saturation and predicted using the DT model. The fitting line (red line) is plotted to analyze the prediction which presents the relationship between actual and expected value. The plot shows that the Decision Tree model predicts some constraints as straight lines and these were due to the outlier's effect. Thus, the DT model didn't capture the majority of the variability for high water saturation.

In comparison, the Random Forest model is more efficient in water saturation prediction compared to the Decision Tree model as shown in Figure 5. 16 even when have a slight difference in R^2 and RMSE. This indicates that the Random Forest model captures the majority of variability for water saturation prediction more effectively than the Decision Tree.

Figure 5. 17, the histogram shows the R2 value for both the ML model for the training dataset and the blind well (A-05). Table 5. 4, shows the evaluation metric for both models used to predict water saturation.

Based on this result for water saturation, RF provides better results in comparison with the DT model for both training and blind datasets.

Table 5. 4: Evaluation metric for water saturation prediction.

Model	Evaluation metric
R^2 (%)	
Decision Tree	97.3
Random Forest	96.9
RMSE (%)	
Decision Tree	2.8
Random Forest	3

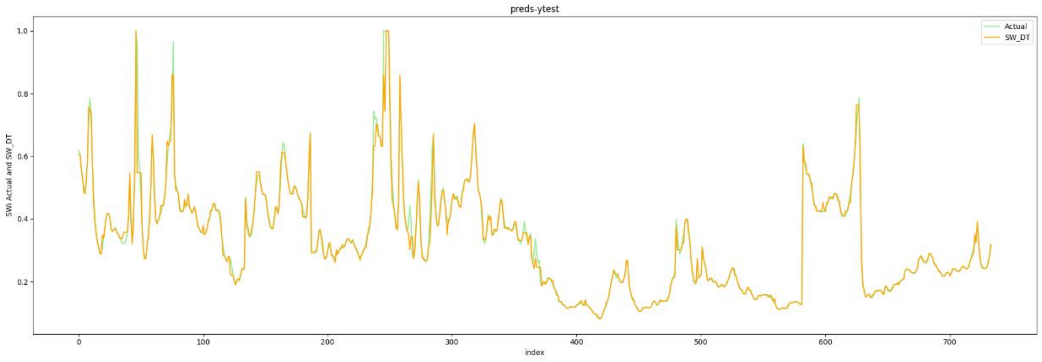

Figure 5. 11: predicted water saturation versus actual water saturation for the training dataset using the DT model.

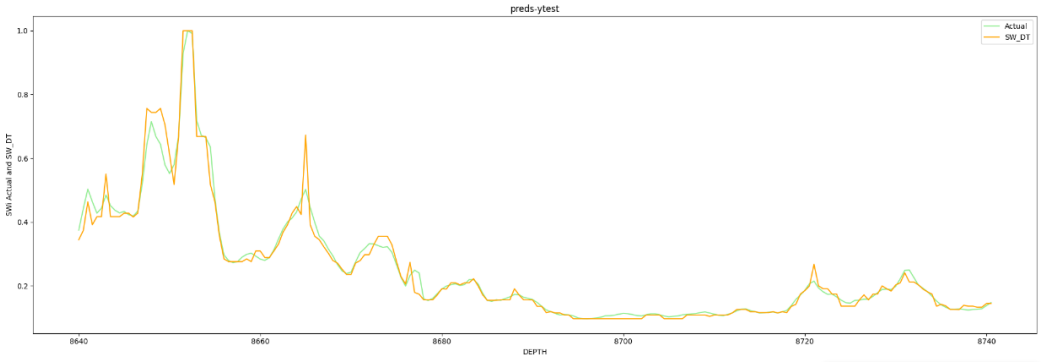

Figure 5. 12: predicted water saturation versus actual water saturation for a testing dataset using the DT model.

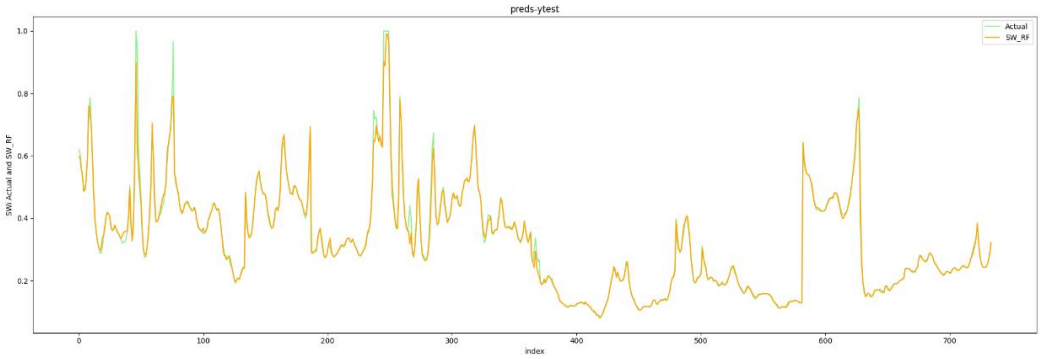

Figure 5. 13: predicted porosity versus actual water saturation for training dataset using RF model.

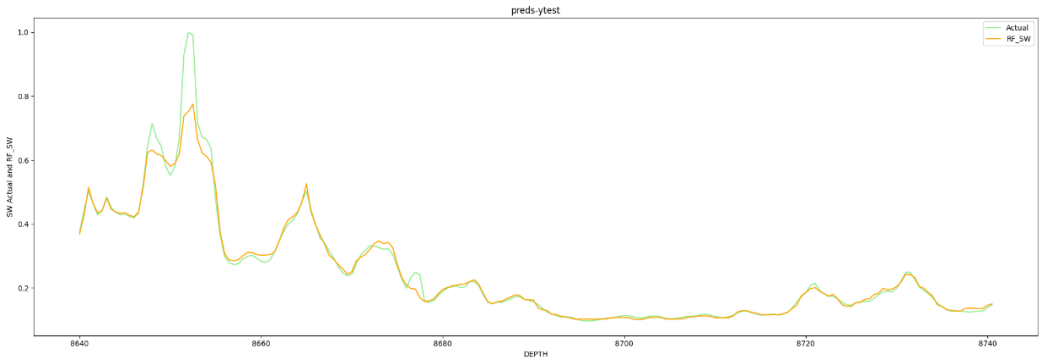

Figure 5. 14: predicted water saturation versus actual water saturation for blind dataset (well A-05) using RF model.

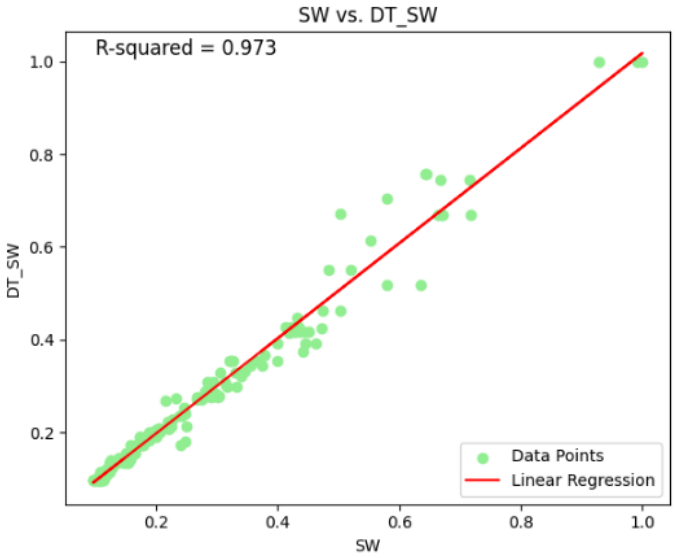

Figure 5. 15: actual water saturation versus water saturation prediction using DT.

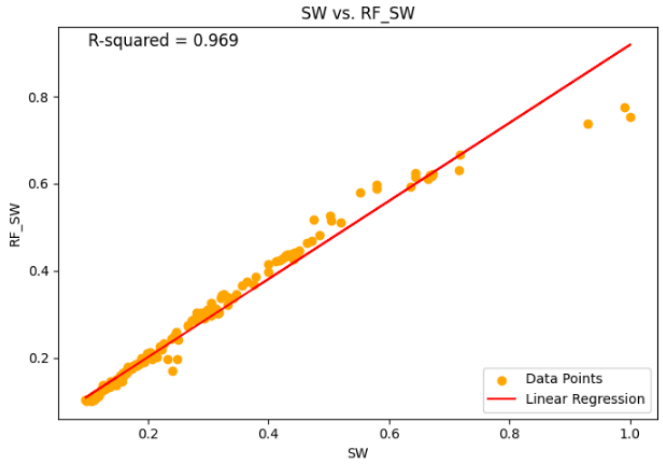

Figure 5. 16: actual water saturation versus water saturation prediction using RF.

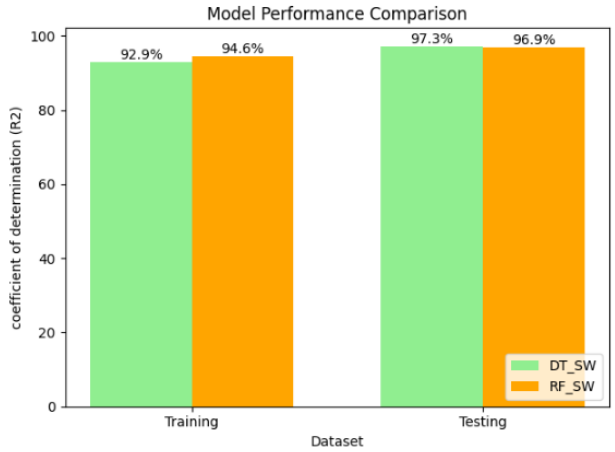

Figure 5. 17: R2 value for water saturation prediction for the training dataset and blind dataset for both models.

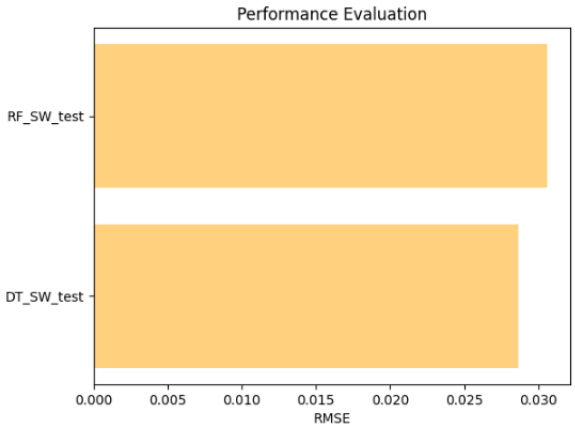

Figure 5. 18: RMSE value for water saturation prediction for the training dataset and blind dataset for both models.

5.3 permeability prediction

5.3.1 Feature Selection

Generally, permeability is a more complex property, especially for prediction in petrophysical studies. In contrast to porosity and saturation, no log measures permeability directly. Thus, for permeability prediction feature selection is more critical.

The input features used for permeability prediction were ILD, RHOB, PHI, VSHALE, SW, DT, and SP. As stated in the heat map shown in Figure 5. 19, ILD, PHI and DT have a direct relationship with permeability (K) while RHOB, VSHALE, SW, and SP have an inverse relationship with permeability (K). Thus, the P-values for ILD, PHI and DT were 0.72, 0.75, 0.53 and 0.11 sequentially. Whereby, RHOB, VSHALE, SW, and SP were -0.53, -0.62, -0.54, -0.69, and -0.19.Table 5. 5, shows the feature selection for permeability prediction.

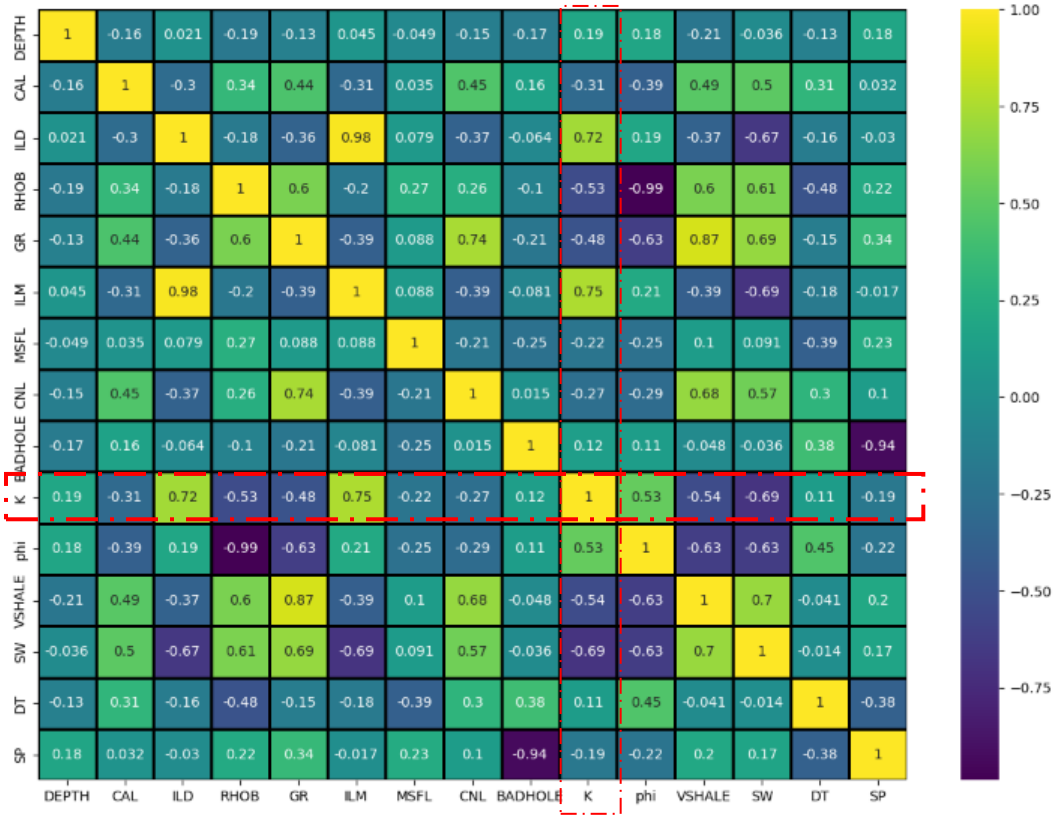

Figure 5. 19: Pearson correlation coefficient with heat map for permeability prediction.

Table 5. 5: Feature selection for permeability prediction.

Feature	Pearson correlation coefficient
Feature selected	ILD, RHOB, PHI, VSHALE, SW, DT, and SP.
Feature removed	MSFL, CAL, BEDHOLE, GR, CNL, DEPTH, ILM.

5.3.2 Model Prediction

The permeability was predicted using DT and RF models using previous features selected as shown in Figure 5. 19. The model generated by the trained dataset was then evaluated by using a blind dataset (A-05) which contained the same features used in the training dataset as well as the actual output that was used for comparison purposes.

Based on the DT model, the training dataset provided an R^2 of about 0.956, while for the blind well (A-05), the R^2 was about 0.939 and the RMSE was about 0.06. Figure 5. 20, show the predicted permeability for the training dataset compared with actual permeability which presents a good match between the actual and predicted permeability. However, Figure 5. 21, illustrates the result for the blind well (A-05) which displays the predicted permeability using the DT model with the actual permeability.

In contrast, the RF model provides an R^2 of about 0.969 for the training dataset while the R^2 and RMSE for the blind well (A-05) are about 0.973 and 0.045. The high R^2 for the training dataset and high R^2 for the blind dataset indicate that the RF model can capture all data points in the blind well as shown in Figure 5. 22 and Figure 5. 23.

Figure 5. 24, show the scatter plot between actual permeability and predicted using the DT model. The fitting line (red line) is plotted to analyse the prediction which presents the relationship between actual and predicted value. According to Figure 5. 21 and Figure 5. 24 the decision tree prediction can't capture the permeability pattern well due to the wide range of permeability, which makes the decision tree need more depth to get a good prediction as well as the effect of the outliers.

In comparison, the Random Forest model strongly in permeability prediction compared to the Decision Tree model which provided R^2 higher than DT.

As well as, shown in Figure 5. 26 for permeability prediction from the Random Forest (RF) model provides more accurate results compared with the Decision Tree (DT) result for a blind dataset. Figure 5. 27, provide the RMSE for permeability prediction which shows that RMSE for the DT is higher than RMSE for the RF model.

Table 5. 6: Evaluation metric for permeability prediction.

Model	Evaluation metric
R^2 (%)	
Decision Tree	93.9
Random Forest	97.1
RMSE (%)	
Decision Tree	6.8
Random Forest	4.5

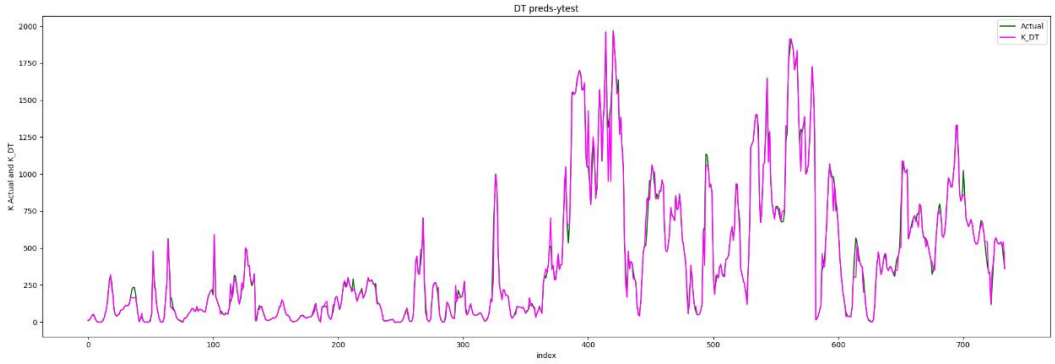

Figure 5. 20: predicted permeability versus actual permeability for training dataset using the DT model.

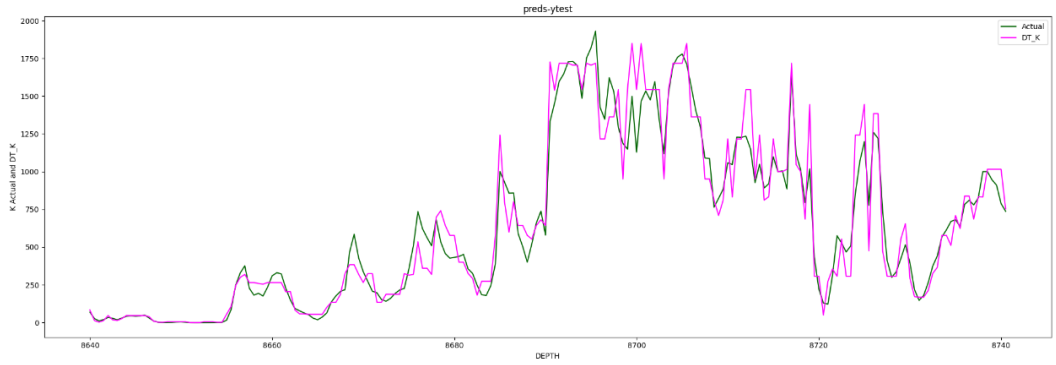

Figure 5. 21: predicted permeability versus actual permeability for blind dataset (well A-05) using the DT model.

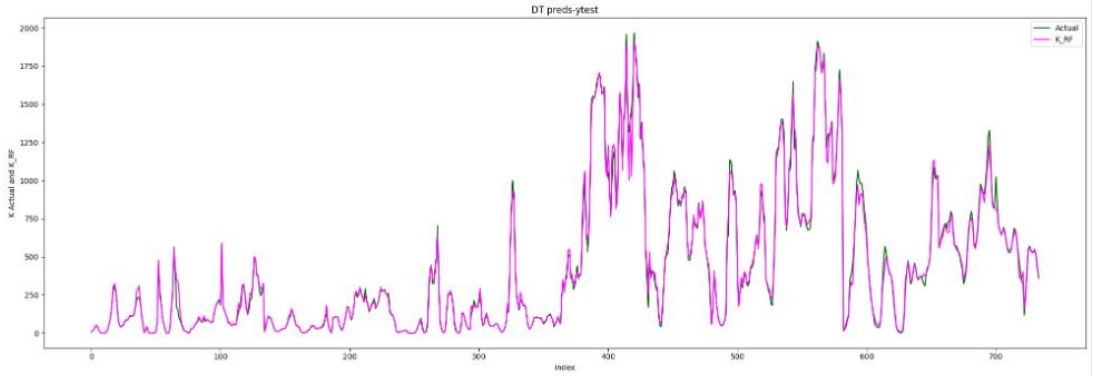

Figure 5. 22: predicted permeability verses actual permeability for training dataset using RF model.

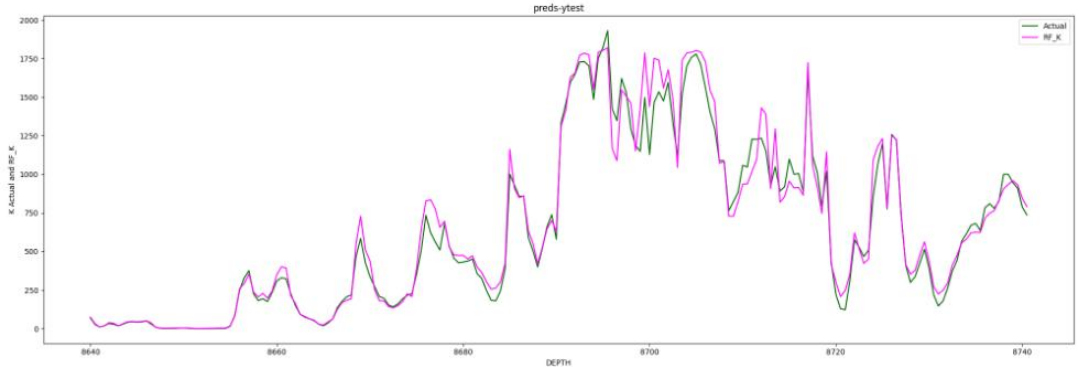

Figure 5. 23: predicted permeability versus actual permeability for blind dataset (well A-05) using the RF model.

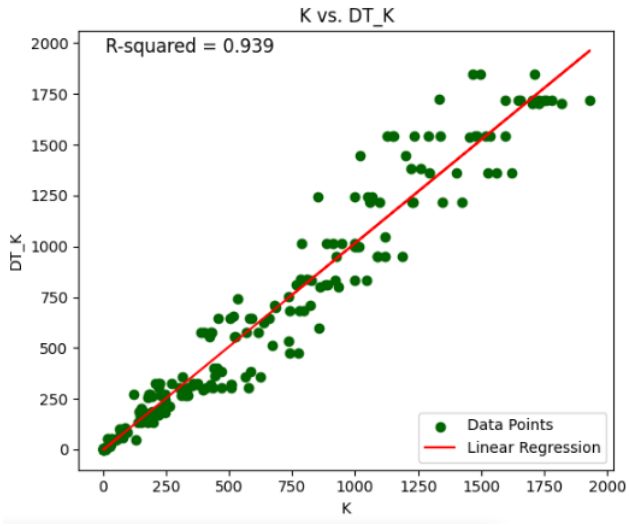

Figure 5. 24: actual permeability versus permeability prediction using DT.

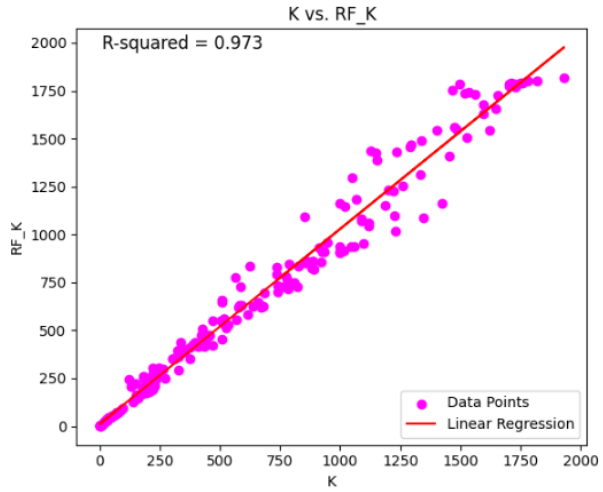

Figure 5. 25 : actual permeability versus permeability prediction using RF.

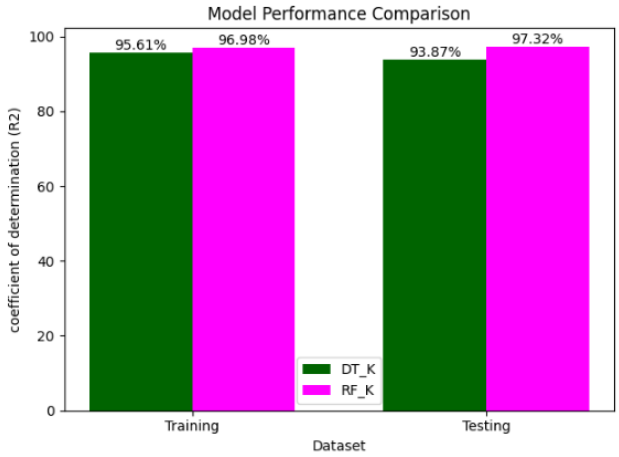

Figure 5. 26: R2 value for permeability prediction for the training dataset and blind dataset for both models.

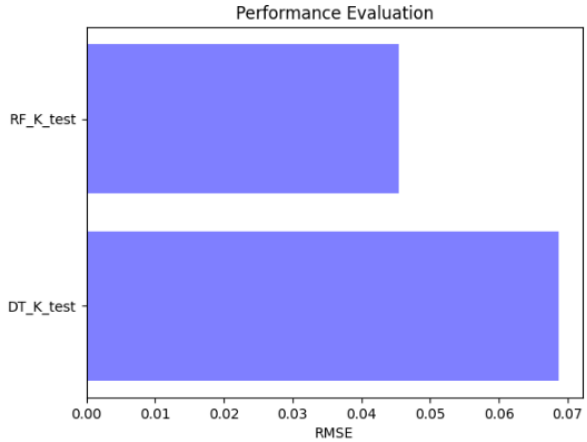

Figure 5. 27: RMSE value for permeability prediction for the blind dataset for both models.

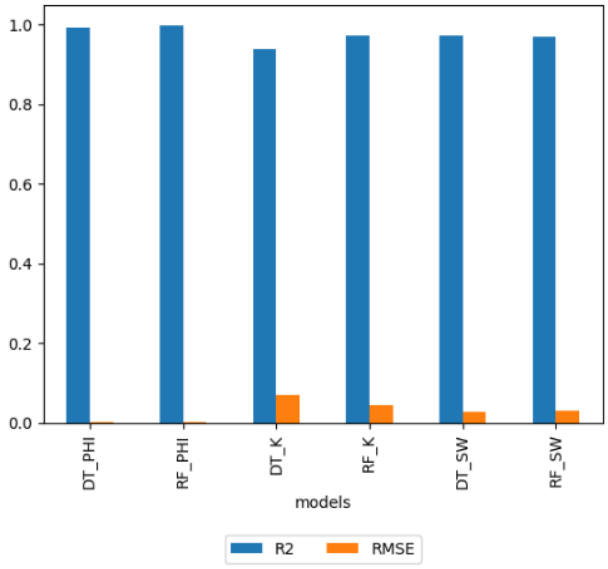

Figure 5. 28: Comparison between evaluation metrics for the blind dataset for all models.

As mentioned in previous chapters, the best model result provides high R^2 and low RMSE. According to Figure 5. 28, the best result for both models used in this study was for porosity prediction, in which both models (DT and RF) can generalize the data well in unseen datasets (blind well-A-05). Followed by water saturation prediction, RF models provide good results compared to the DT model. On the other hand, for permeability prediction, the RF model provides more accurate results than the DT model. Thus, the RF model provides a good match for all properties predicted.

5.4 Discussion

This study applied the Decision Tree (DT) and Random Forest (RF) regression model to predict reservoir properties including porosity, permeability and water saturation in a sandstone formation in a Libyan oil field in the blind well (A-05). The main goal was to apply effective models in time and cost as well as more confident models to predict reservoir properties select the best model for each property and compare the used model with the actual result.

Although feature engineering plays a critical role in machine learning studies, in this study the feature selection was guided by the specific consideration for each property and the nature of the data used. However, the use of the Pearson correlation coefficient provides

insight into the ability to use regression models and understand the relation between variables.

The dataset contains measured and calculated data. For porosity and water saturation, measured data was used to train the models while for permeability both measured and calculated were used. Caliper (CAL)and bad hole (BADHOLE) logs were neglected from features their usage was only to understand the borehole environment.

For porosity prediction used DT and RF models both models show better porosity prediction compared with other properties. Both models used the same features to predict the porosity. R^2 and RMSE for the DT model were about 99.3% and 0.28% respectively while the RF model R^2 and RMSE were about 99.9% and 0.12% sequentially. However, the slight difference in the R^2 and RMSE metrics is due to differences in model structure and how each model learns from data. However, the DT model was affected by the presence of outliers in porosity (low porosity) due to this the DT model can't be generalized well at this point. On the other hand, the RF model enhances the prediction for this point and provides a good match with actual porosity.

As discussed, two main ML models (DT and RF) to predict water saturation. Both models were trained and tested with the same features. Hence, the R^2 and RMSE for DT was about 97.3% and 2.8%while the R^2 and RMSE using RF was about 96.9% and 3%. However, both models provide a low prediction for water saturation at high values but the DT model showed some constrains (straight lines) due to the outlier effect. On the other hand, according to this result, the RF model showed a good match when compared to the actual water saturation.

With the same models (DT and RF) the permeability was predicted with a similar input feature. R2 and RMSE for DT and RF were about 94.5% and 6.8% while for the RF were about 97.1 and 4.5%. The main reason for the difference between the DT result and the RF result is that the DT model was affected by outliers as well as the different range for permeability. However, the RF model captured the pattern for data points for a blind well (A-05) providing a more accurate result.

However, the model performance for reservoir properties predictions may increase by increasing the number of data points (data size) and data quality used as well as increasing

the number of wells. Moreover, if the number of data points increases models can improve the learning process.

Another factor that can influence the model's prediction is feature selection. Thus, could use other methods to select the optimum features.

As well as the data size, the model performance may increase with a change in hypermeter tuning to identify the data pattern and enhance the prediction.

Conclusively, according to the result obtained from this study, the usage of machine learning in the Libyan oil field can provide accurate results and help in reservoir properties predictions.

Finally, the advanced use of a Machine learning model

- Compared to the traditional method machine learning models are financial and time efficient to predict reservoir properties.
- Cost efficient for petrophysical services which can detect the critical petrophysical logs used for the studied models and remove less important logs.
- Can help reservoir modelling and evaluation study.

CHAPTER SIX

6 CONCLUSION AND RECOMMENDATION

6.1 Conclusion

This study presents the application of machine learning to predict the reservoir properties in the Sandstone reservoir in the Libyan field with well-logging measurement and correlated core data. The findings of this study are summarized in the following section:

This study was done by using Python programming with Jupyter Notebook. For data analysis, several libraries including Pandas, NumPy, Seaborne, Matplotlib, and SciPy for preprocessing and machine learning used the Scikit-Learn library.

The dataset for training ML models was selected from the Libyan field which contained about 734 points from different well wells (A-13, A-12, A-02, A-06, A-08, A-09, A-97) while the blind dataset well (A-05) contained 202points. This limitation in data points may affect model performance especially if the data quality is low.

The Exploratory Data analysis (EDA) was applied which helps to understand the relationship between variables and detect the outliers. However, the outliers detected in the data used in this study were related to the geology and reservoir formation so that wasn't removed.

For feature selection, selected based on specific considerations for each property. Using the Pearson correlation coefficient is the main way to understand the relation between different variables.

Both models Decision Tree (DT) model and the Random Forest (RF) model predict more accurate porosity in comparison with permeability and water saturation. The cause behind that the data range for porosity is almost the same in training and blind dataset well (A-05) with a slight difference.

For water saturation prediction the Random Forest (RF) model shows more accurate results than the Decision Tree (DT).

The permeability prediction provided good performance, especially for the Random Forest (RF) model. However, permeability data have a wide range. This affects capture generalization for the DT model.

The findings from this study conclude that the Random Forest model provides a good match with sandstone reservoir properties for the Libyan field.

The findings from this result show that Machine Learning models with relevant features and high data quality can enhance the reservoir properties prediction in efficient time and reduce the cost for RCA.

6.2 Recommendation

The investigation revealed that the application of machine learning is essential for the prediction of reservoir properties using well-logging data. On this basis, the future study should:

- Increasing the size and numbers of data points as well as data quality could enhance the Machine Learning model and provide high performance.
- Using Feature engineering methods such as feature importance may provide more accurate results.
- Developed machine learning and artificial intelligence algorithms with core data and well logging data from all wells available in the study area.
- Using more complex Machine Learning (ML) models such as Neural Networks.
- Using TensorFlow library which provides more complex deep learning and artificial intelligent models.
- For future studies explore the application of machine learning in different Libyan fields and reservoir properties.

References

[1] A. Erofeev, D. Orlov, A. Ryzhov, and D. Koroteev, “Prediction of Porosity and Permeability Alteration based on Machine Learning Algorithms,” Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.06525

[2] S. Alatefi, R. Abdel Azim, A. Alkouh, and G. Hamada, “Integration of Multiple Bayesian Optimized Machine Learning Techniques and Conventional Well Logs for Accurate Prediction of Porosity in Carbonate Reservoirs,” *Processes*, vol. 11, no. 5, May 2023, doi: 10.3390/pr11051339.

[3] P. Ø. Andersen, M. Skjeldal, and C. Augustsson, “Machine Learning Based Prediction of Porosity and Water Saturation from Varg Field Reservoir Well Logs,” in *Society of Petroleum Engineers - SPE EuropEC - Europe Energy Conference featured at the 83rd EAGE Annual Conference and Exhibition, EURO 2022*, Society of Petroleum Engineers, 2022. doi: 10.2118/209659-MS.

[4] Z. Zhong and T. R. Carr, “Application of a new hybrid particle swarm optimization-mixed kernels function-based support vector machine model for reservoir porosity prediction: A case study in Jacksonburg-Stringtown oil field, West Virginia, USA,” *Interpretation*, vol. 7, no. 1, pp. T97–T112, Feb. 2019, doi: 10.1190/INT-2018-0093.1.

[5] S. Bhattacharya, *A Primer on Machine Learning in Subsurface Geosciences*, vol. 1. Springer, 2021. doi: https://doi.org/10.1007/978-3-030-71768-1.

[6] C. J. Lewis, R. M. Sanford, and J. Gong, “SARIR FIELD SIRTE BASIN, LIBYA Desert Surprise Then-and Now Some Keys to Revisit of Libya Search and Discovery Article, no. 10005.,” *AAPG*, 2000.

[7] G. Ambrose, “THE GEOLOGY AND HYDROCARBON HABITAT OF THE SARIR SANDSTONE, SE SIRT BASIN, LIBYA,” 2000.

[8] F. T. Barr and A. A. Weegar, “STRATIGRAPHIC NOMENCLATURE OF THE SIRTE BASIN, LIBYA,” 1972.

[9] A. Y. Dandekar, *Petroleum reservoir rock and fluid properties*, Second. CRC press, 2013.

[10] N. Alyafei, *Fundamentals of reservoir rock properties*, Second. Hamad Bin Khalifa University Press, 2021. doi: https://doi.org/10.5339/Fundamentals_of_Reservoir_Rock_Properties_2ndEdition.

[11] Djebbar Tiab and Erle C. Donaldson, *Petrophysics, Theory and Practice of Measuring Reservoir Rock and Fluid Transport Properties*, Third edition. Elsevier, 2012.

[12] Tarek Ahmed, *Reservoir Engineering Handbook*, Second Edition. Gulf Publishing, 2001.

[13] Steve Cannon, *Reservoir Modelling: A Practical Guide*. John Wiley & Sons, 2018. doi: 10.1002/9781119313458.

[14] L.P Dake, *Fundamentals of Reservoir Engineering*. Elsevier, 1983.

[15] H. S. Kharraa, M. A. Al-Amri, S. Aramco, M. A. Mahmoud, and T. M. Okasha, "Assessment of Uncertainty in Porosity Measurements Using NMR and Conventional Logging Tools in Carbonate Reservoir," 2013.

[16] B. Rafik and B. Kamel, "Prediction of permeability and porosity from well log data using the nonparametric regression with multivariate analysis and neural network, Hassi R'Mel Field, Algeria," *Egyptian Journal of Petroleum*, vol. 26, no. 3, pp. 763–778, 2017, doi: 10.1016/j.ejpe.2016.10.013.

[17] F. Hadavimoghaddam, M. Ostadhassan, M. A. Sadri, T. Bondarenko, I. Chebyshev, and A. Semnani, "Prediction of water saturation from well log data by machine learning algorithms: Boosting and super learner," *J Mar Sci Eng*, vol. 9, no. 6, Jun. 2021, doi: 10.3390/jmse9060666.

[18] G. B. Asquith, C. R. Gibson, Steven. Henderson, and N. F. Hurley, *Basic well log analysis*. American Association of Petroleum Geologists, 2004.

[19] O. Ijasan, C. Torres-Verdín, and W. E. Preeg, "Interpretation of porosity and fluid constituents from well logs using an interactive neutron-density matrix scale,"

Interpretation, vol. 1, no. 2, pp. T143–T155, Nov. 2013, doi: 10.1190/INT-2013-0072.1.

[20] L. F. Qfintero, S. Oilfield, S. ; Es, and Z. Bassiouni, "Porosity Determination in Gas-Bearing Formations," *In SPE Permian Basin Oil and Gas Recovery Conference,SPE*, p. SPE-39774, Mar. 1998, doi: https://doi.org/10.2118/39774-MS.

[21] L. M. Anovitz and D. R. Cole, "Characterization and analysis of porosity and pore structures," in *Pore Scale Geochemical Processes*, De Gruyter, 2015, pp. 61–164. doi: 10.2138/rmg.2015.80.04.

[22] M. H. Kamel, W. M. Mabrouk, and A. I. Bayoumi, "Porosity estimation using a combination of Wyllie-Clemenceau equations in clean sand formation from acoustic logs." [Online]. Available: www.elsevier.com/locate/jpetscieng

[23] G. Tao', "POROSITY AND PORE STRUCTURE FROM ACOUSTIC WELL LOGGING DATA1," 1993.

[24] John Cubitt, *Practical Petrophysics*, vol. 62. Elsevier, 2015. doi: 10.1016/b978-0-444-63270-8.00016-5.

[25] F. Tapias and A. Hutin, "Porous media in reservoir engineering: Permeability," in *Porous media in reservoir engineering*, 1st ed., The little corner of porous media, 2022, ch. Permeability. doi: 10.5281/zenodo.6391195.

[26] A. Sheykhinasab *et al.*, "Prediction of permeability of highly heterogeneous hydrocarbon reservoir from conventional petrophysical logs using optimized data-driven algorithms," *J Pet Explor Prod Technol*, vol. 13, no. 2, pp. 661–689, Feb. 2023, doi: 10.1007/s13202-022-01593-z.

[27] S. Mohaghegh, B. Balan, and S. Ameri, "Permeability Determination From Well Log Data," *SPE Formation Evaluation*, vol. 12, no. 03, pp. 170–174, Sep. 1997, doi: 10.2118/30978-pa.

[28] R. C. K. Wong, "A model for strain-induced permeability anisotropy in deformable granular media," *Canadian Geotechnical Journal*, vol. 40, no. 1, pp. 95–106, Feb. 2003, doi: 10.1139/t02-088.

[29] B. Balan and S. Ameri, “State-Of-The-Art in Permeability Determination From Well Log Data: Part 1-A Comparative Study, Model Development,” *SPE*, p. SPE-30978, Sep. 1995, doi: https://doi.org/10.2118/30978-MS.

[30] K. M. Rahuma and B. M. Ben Ghawar, “Calculation and Comparison of Water Saturation in Carbonate Formation by Different Methods,” *THE INTERNATIONAL JOURNAL OF ENGINEERING AND INFORMATION TECHNOLOGY (IJEIT)*, vol. 5, no. 2, 2019, [Online]. Available: www.ijeit.misuratau.edu.ly

[31] G. M. Hamada and M. N. AL-Awad, “Evaluating Uncertainty in Archie’s Water Saturation Equation Parameters Determination Methods,” *SPE*, 2001, Accessed: Apr. 18, 2024. [Online]. Available: https://doi.org/10.2118/68083-MS

[32] G. E. Archie, “The Electrical Resistivity Log as an Aid in Determining Some Reservoir Characteristics.” [Online]. Available: http://onepetro.org/TRANS/article-pdf/146/01/54/2179020/spe-942054-g.pdf/1

[33] M. El-Bagoury, “Integrated petrophysical study to validate water saturation from well logs in Bahariya Shaley Sand Reservoirs, case study from Abu Gharadig Basin, Egypt,” *J Pet Explor Prod Technol*, vol. 10, no. 8, pp. 3139–3155, Dec. 2020, doi: 10.1007/s13202-020-00969-3.

[34] J. Sam-Marcus, E. Enaworu, O. J. Rotimi, and I. Seteyeobot, “A proposed solution to the determination of water saturation: using a modelled equation,” Dec. 01, 2018, *Springer Verlag*. doi: 10.1007/s13202-018-0453-4.

[35] C. McPhee, J. Reed, and I. Zubizarreta, *Core analysis : a best practice guide*. Elsevier, 2015. doi: http://dx.doi.org/10.1016/B978-0-444-63533-4.09989-3.

[36] J. H. Stiles, E. Spe, U. K. E&p, and J. M. Hutfilz, “The Use of Routine and Special Core Analysis in Characterizing Brent Group Reservoirs, U.K. North Sea,” *Journal of Petroleum Technology*, vol. 44(06), pp. 704–713, 1992, doi: https://doi.org/10.2118/18386-PA.

[37] M. Al-Saddique, G. Hamada, and M. N. J AI-Awad, “State of the Art: Review of Coring and Core Analysis Technology,” *Journal of King Saud University-*

Engineering Sciences, vol. 12, no. 1, pp. 117–138, Feb. 1999, doi: https://doi.org/10.1016/S1018-3639(18)30709-8.

[38] Y. Z. Ma, "Quantitative Geosciences: Data Analytics, Geostatistics, Reservoir Characterization and Modeling," *Springer International Publishing*, 2019, doi: https://doi.org/10.1007/978-3-030-17860-4.

[39] A. Al-Fakih, S. I. Kaka, and A. I. Koeshidayatullah, "Reservoir Property Prediction in the North Sea Using Machine Learning," *IEEE Access*, vol. 11, pp. 140148–140160, Nov. 2023, doi: 10.1109/ACCESS.2023.3336623.

[40] G. Ifrene, D. Irofti, R. Ni, S. Egenhoff, and P. Pothana, "New Insights into Fracture Porosity Estimations Using Machine Learning and Advanced Logging Tools," *Fuels*, vol. 4, no. 3, pp. 333–353, Aug. 2023, doi: 10.3390/fuels4030021.

[41] S. Elkatatny, Z. Tariq, M. Mahmoud, and A. Abdulraheem, "New insights into porosity determination using artificial intelligence techniques for carbonate reservoirs," *Petroleum*, vol. 4, no. 4, pp. 408–418, Dec. 2018, doi: 10.1016/j.petlm.2018.04.002.

[42] C. N. Mkono, C. Shen, A. K. Mulashani, and P. Nyangi, "An improved permeability estimation model using integrated approach of hybrid machine learning technique and shapley additive explanation," *Journal of Rock Mechanics and Geotechnical Engineering*, Sep. 2024, doi: 10.1016/j.jrmge.2024.09.013.

[43] T. Topór, "Application of machine learning algorithms to predict permeability in tight sandstone formations," *Nafta - Gaz*, vol. 2021, no. 5, pp. 283–292, May 2021, doi: 10.18668/NG.2021.05.01.

[44] A. B. Zolotukhin and A. T. Gayubov, "Machine learning in reservoir permeability prediction and modelling of fluid flow in porous media," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Nov. 2019. doi: 10.1088/1757-899X/700/1/012023.

[45] M. Talebkeikhah, Z. Sadeghtabaghi, and M. Shabani, "A Comparison of Machine Learning Approaches for Prediction of Permeability using Well Log Data

in the Hydrocarbon Reservoirs," *Journal of Human, Earth, and Future*, vol. 2, no. 2, 2021, [Online]. Available: www.HEFJournal.org

[46] E. A. El-Sebakhy *et al.*, "Functional networks as a new data mining predictive paradigm to predict permeability in a carbonate reservoir," *Expert Syst Appl*, vol. 39, no. 12, pp. 10359–10375, Sep. 2012, doi: 10.1016/j.eswa.2012.01.157.

[47] S. Baziar, H. B. Shahripour, M. Tadayoni, and M. Nabi-Bidhendi, "Prediction of water saturation in a tight gas sandstone reservoir by using four intelligent methods: a comparative study," *Neural Comput Appl*, vol. 30, no. 4, pp. 1171–1185, Aug. 2018, doi: 10.1007/s00521-016-2729-2.

[48] G. M. Hamada and M. A. Elshafei, "Application of Artificial Neural Network (ANN) in petrophysical evaluation of shaly sand reservoirs," *NAFTA*, pp. 58 (11) 551-555, 2007, [Online]. Available: https://www.researchgate.net/publication/357016445

[49] G. Rebala, A. Ravi, and S. Churiwala, *An Introduction to Machine Learning*. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-15729-6.

[50] T. M. (Tom M. Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997.

[51] Mohri Mehryar, Rostamizadeh Afshin, and Talwalkar Ameet, *Foundations of Machine Learning second edition*, Second edition. Massachusetts Institute of Technology, 2018.

[52] Sravya Reddysetty, "Traditional Programming vs Machine Learning," Medium . Accessed: May 13, 2024. [Online]. Available: https://sravya-tech-usage.medium.com/traditional-programming-vs-machine-learning-e9bbed5e491c

[53] S. Misra, H. Li, and Jiabo. He, "Machine Learning for Subsurface Characterization," 2020.

[54] J. Masapanta, "MACHINE AND DEEP LEARNING FOR LITHOFACIES CLASSIFICATION FROM WELL LOGS IN THE NORTH SEA," 2021. Accessed: Oct. 09, 2024. [Online]. Available: https://hdl.handle.net/11250/2786292

[55] Bangert Patrick, *Machine Learning and Data Science in the Oil and Gas Industry Best Practices, Tools, and Case Studies*. Gulf Professional Publishing of Elsevier, 2021. doi: 10.1016/b978-0-12-820714-7.00020-0.

[56] A. A. Shehata, O. A. Osman, and B. S. Nabawy, "Neural network application to petrophysical and lithofacies analysis based on multi-scale data: An integrated study using conventional well log, core and borehole image data," *J Nat Gas Sci Eng*, vol. 93, Sep. 2021, doi: 10.1016/j.jngse.2021.104015.

[57] Y. N. Pandey, A. Rastogi, S. Kainkaryam, S. Bhattacharya, and L. Saputelli, *Machine Learning in the Oil and Gas Industry: Including Geosciences, Reservoir Engineering, and Production Engineering with Python*. Springer Science+Business Media, 2020. doi: 10.1007/978-1-4842-6094-4.

[58] Simon Haykin, *Neural Networks-A Comprehensive Foundation*, Second. Prentice Hall, 1999.

[59] S. Mohaghegh, R. Arefi, S. Ameri, and D. Rose, "Design and development of an artificial neural network for estimation of formation permeability," in *Proceedings - Petroleum Computer Conference*, Publ by Society of Petroleum Engineers (SPE), 1994, pp. 147–154. doi: 10.2118/28237-pa.

[60] GeeksforGeeks, "Artificial Neural Networks and its Applications," GeeksforGeeks. Accessed: Jul. 21, 2024. [Online]. Available: https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/?ref=header_search

[61] D. A. Otchere, T. O. Arbi Ganat, R. Gholami, and S. Ridha, "Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models," May 01, 2021, *Elsevier B.V.* doi: 10.1016/j.petrol.2020.108182.

[62] S. S. Haykin, *Neural networks and learning machines*, Third. Prentice Hall/Pearson, 2009.

[63] Mustafa Murat ARAT, “A Complete Guide to K-Nearest-Neighbors with Applications in Python.” Accessed: May 28, 2024. [Online]. Available: https://mmuratarat.github.io/2019-07-12/k-nn-from-scratch

[64] D N Sachin, “K-Nearest Neighbors Algorithm,” Medium . Accessed: May 28, 2024. [Online]. Available: https://medium.com/analytics-vidhya/k-nearest-neighbors-algorithm-7952234c69a4

[65] Ajitesh Kumar, “K-Nearest Neighbors (KNN) Python Examples,” Analytics Yogi Reimagining Data-driven Society with Data Science & AI. Accessed: May 29, 2024. [Online]. Available: https://vitalflux.com/k-nearest-neighbors-explained-with-python-examples/

[66] C. Cortes, V. Vapnik, and L. Saitta, “Support-Vector Networks,” *Machine Leaming*, vol. 20, pp. 273–297, 1995, doi: https://doi.org/10.1007/BF00994018.

[67] A. Al-Anazi and I. D. Gates, “Support-Vector Regression for Permeability Prediction in a Heterogeneous Reservoir: A Comparative Study,” *SPE Reservoir Evaluation & Engineering*, vol. 485, 2010.

[68] A. T. Combs and M. Roman, *Python machine learning blueprints*, Second. Packt Publishing, 2019.

[69] H. Azizi and H. Reza, “Applied Machine Learning Methods for Detecting Fractured Zones by Using Petrophysical Logs,” *Intelligent Control and Automation*, vol. 12, no. 02, pp. 44–64, 2021, doi: 10.4236/ica.2021.122003.

[70] geeks for geeks, “Support Vector Machine (SVM) Algorithm,” geeks for geeks. Accessed: Jun. 08, 2024. [Online]. Available: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[71] scikit-learn user guide, “Decision Trees,” https://scikit-learn.org/stable/modules/tree.html. Accessed: Jul. 17, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html

[72] Neha Seth, “Entropy in Machine Learning: Definition, Examples and Uses,” Analytics Vidhya . Accessed: Jul. 17, 2024. [Online]. Available:

https://www.analyticsvidhya.com/blog/2020/11/entropy-a-key-concept-for-all-data-science-beginners/

[73] Sharath Manjunath, “Decision Tree,” Medium . Accessed: Jul. 17, 2024. [Online]. Available: https://sharathmanjunath.medium.com/decision-tree-8c33a0c17744

[74] Awad Mariette and Khanna Rahul, *Efficient Learning Machines Theories, Concepts, and Application for Engineers and System Designers*. Springer nature, 2015.

[75] DOUGLAS C. MONTGOMERY, ELIZABETH A. PECK, and G. GEOFFREY VINING, *Introduction To Linear Regression Analysis*, Fifth. WILEY, 2012. Accessed: Jul. 20, 2024. [Online]. Available: https://ia801407.us.archive.org/33/items/econometrics_books/Intro.%20to%20Linear%20Regression%20Analysis%20-%20D.%20C.%20Montgomery%2C%20E.%20A.%20Peck.pdf

[76] Abhay Jidge, “The Complete Guide to Linear Regression Analysis,” Medium . Accessed: Jul. 20, 2024. [Online]. Available: https://towardsdatascience.com/the-complete-guide-to-linear-regression-analysis-38a421a89dc2

[77] Geeks for Geeks, “Linear Regression in Machine learning,” Geeks for Geeks . Accessed: Jul. 20, 2024. [Online]. Available: https://www.geeksforgeeks.org/ml-linear-regression/

[78] Benjamin Obi Tayo Ph.D., “Linear Regression Basics for Absolute Beginners,” Medium. Accessed: Aug. 27, 2024. [Online]. Available: https://pub.towardsai.net/linear-regression-basics-for-absolute-beginners-68ed9ff980ae

[79] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, “Machine Learning for Fluid Mechanics,” *Annu. Rev. Fluid Mech. 2020*, vol. 52, pp. 477–508, 2019, doi: 10.1146/annurev-fluid-010719.

[80] Ravish Kumar, “Supervised, Unsupervised, And Semi-Supervised Learning,” Medium . Accessed: Jul. 20, 2024. [Online]. Available: https://medium.com/enjoy-algorithm/supervised-unsupervised-and-semi-supervised-learning-64ee79b17d10

[81] Cannon Steve, *Petrophysics: A Practical Guide*. John Wiley & Sons, Ltd, 2015.

[82] R. O. Baker, H. W. Yarranton, and J. Jensen, *Practical Reservoir Engineering and Characterization*. Gulf Professional Publishing, 2015. doi: https://doi.org/10.1016/C2011-0-05566-7.

[83] Schlumberger, “Log Interpretation Charts.”

[84] P. W. J. Glover, *Petrophysics*. University of Aberdeen UK, 2000.

[85] B. C. P Parsons and M. Aime, “Caliper Logging,” *Transactions of the AIME*, vol. 151(01), pp. 35-47., 1942, doi: https://doi.org/10.2118/943035-G.

[86] SELLEY C. RICHARD and SONNENBERG A. STEPHEN, *Elements of Petroleum Geology*, Third. Elsevier, 2015. doi: 10.1016/b978-0-12-386031-6.09001-9.

[87] E. I. Okon, D. Appah, and J. A. Ajienka, “Review of Python Applications in Solving Oil and Gas Problems,” *Journal of Engineering Research and Reports*, pp. 13–22, Oct. 2021, doi: 10.9734/jerr/2021/v21i317448.

[88] J. Brownlee, *Machine Learning Mastery With Python*. Machine Learning Mastery, 2016.

[89] Hoss Belyadi and Alireza Haghighat, *Machine Learning Guide for Oil and Gas Using Python*. Elsevier, 2021. doi: 10.1016/b978-0-12-821929-4.01001-5.

[90] Codecademy Team, “Normalization,” Codeacademy . Accessed: Aug. 14, 2024. [Online]. Available: https://www.codecademy.com/article/normalization

[91] G. Bonaccorso, *Mastering Machine Learning Algorithms*, Second Edition. Packt Publishing, 2020.

[92] IBM, “What is a decision tree?,” IBM. Accessed: Sep. 26, 2024. [Online]. Available: https://www.ibm.com/topics/decision-trees

[93] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?,” Conference on Neural Information Processing Systems, 2022.

Appendix A

Summary of available data for studied well in this study.

Well name	Well logging data											Label output		
	CAL	GR	SP	ILD	ILM	MSFL	RHOB	CNL	DT	VSHALE	BEDHOLE	phi	K	SW
A-12	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A_13	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A-06	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A-08	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A-09	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A-97	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A-02	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A-05	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Appendix B

statistics summary for the training dataset.

	count	mean	std	min	25%	50%	75%	max
DEPTH	202	8690.25	29.22827	8640	8665.125	8690.25	8715.375	8740.5
CAL	202	8.153015	0.106243	8.039	8.109	8.121	8.141	8.719
ILD	202	21.91548	17.56939	3.098	5.077	18.9375	37.96925	57.656
RHOB	202	2.343	0.059783	2.251	2.30625	2.327	2.36075	2.563
GR	202	19.91504	9.530534	10.094	13.721	16.5785	23.0355	59.594
ILM	202	20.1805	15.84678	3.047	4.93675	16.07	35.531	54.938
MSFL	202	14.9159	4.087093	7.43	12.05275	14.863	17.426	29.781
CNL	202	0.185619	0.024928	0.148	0.17	0.178	0.197	0.291
BADHOLE	202	0	0	0	0	0	0	0
K	202	632.1316	536.9654	0.0139	179.4141	505.7004	1016.919	1930.497
phi	202	0.167868	0.034306	0.0415	0.157825	0.1771	0.1889	0.2205
VSHALE	202	0.096411	0.090123	0.01	0.03705	0.06625	0.12145	0.4349
SW	202	0.248931	0.173697	0.0965	0.1232	0.17995	0.311075	1
DT	202	73.34109	2.421821	67	72	73.1	74.325	81.8
SP	202	49.93944	7.15742	40.687	45.562	48.937	51.578	79.75

الأكاديمية الليبية – فرع اجدابيا

مدرسة العلوم الهندسية والتطبيقية

قسم الهندسة الكيميائية والنفط

تطبيق نماذج التعليم الالي باستخدام بيانات تسجيلات الأبار للتنبؤ بالخصائص المكمنية، حوض سرت

إعداد

ريما عمر محمد خليفة

إشراف الدكتور

عادل الزوي

(أستاذ مشارك)

قدمت هذه الرسالة استكمالاً لمتطلبات الإجازة العالية الماجستير

في العلوم الهندسية النفطية

بتاريخ 13 رجب 1446 هـ الموافق 13 \ 1 \ 2025 م

الملخص

توقع خصائص المكمن مثل المسامية والنفاذية وتشبع الماء أمر بالغ الأهمية لوصف المكمن، وتحديد كمية الهيدروكربونات الأصلية في المكان، وتحسين إنتاج الهيدروكربونات. ومع ذلك، فإن التنبؤ بهذه الخصائص يمثل تحديًا بسبب عدم اليقين المرتبط بالمعادلات التجريبية المستخدمة في تفسير البيانات البتروفيزيائية، فضلاً عن التكلفة المرتفعة والوقت الطويل المطلوب لإجراء التحليل الروتيني للعينات الأساسية.(RCA)

الهدف الرئيسي من هذه الدراسة هو تطوير تقنية موثوقة وفعالة من حيث الوقت والتكلفة للتنبؤ بخصائص المكمن المساميةφ ، تشبع الماءSW ، والنفاذية K وتركز الدراسة على اختيار النموذج الأمثل ومقارنة نتائج التعلم الآلي بالنتائج الفعلية المستخرجة باستخدام الطريقة التقليدية. ولتحقيق ذلك، تم تطوير نموذجين يعتمدان على الأشجار، وهما شجرة القرار (Decision Trees) والغابات العشوائية.(Random Forests)

يحتوي بيانات التدريب على 734 نقطة، بينما يحتوي مجموعة البيانات العمياء على 202 نقطة تم جمعها من ثماني آبار في أحد الحقول النفطية الليبية. تتضمن هذه الآبار بيانات تسجيلات الآبار المقاسة والمحسوبةφ) ، SW، (Kوالتي تمت مقارنتها مع بيانات العينات الأساسية.

تم تدريب النماذج المعتمدة على الأشجار باستخدام مجموعة البيانات المتاحة، وتم تقييمها باستخدام بئر عمياء منفصلة (A-05).

تشير نتائج هذه الدراسة إلى أن كلا النموذجين قدما أداءً جيدًا، إلا أن نموذج الغابات العشوائية أظهر أداءً أقوى عند الاختبار على البيانات العمياء.

توضح نتائج هذه الدراسة أن استخدام التعلم الآلي يمكن أن يوفر نتائج أكثر دقة مع تقليل الوقت والتكلفة مقارنةً بالطرق التقليدية. تتماشى هذه النتائج مع هدف الدراسة في تحديد نماذج تعلم آلي فعالة للتنبؤ بخصائص المكمن.

ومع ذلك، فمن المهم في الدراسات المستقبلية استخدام بيانات ذات جودة عالية، وإضافة ميزات أخرى، واستخدام نماذج أكثر تعقيدًا للحصول على نتائج أكثر دقة، خاصةً فيما يتعلق بتنبؤ النفاذية.

الكلمات المفتاحية: خصائص المكمن، التعلم الآلي، تسجيلات الآبار، الحقول النفطية الناضجة، حوض سرت.