PETROPHYSICAL EVALUATION USING MACHINE LEARNING MODELS FOR THE PREDICTION OF POROSITY

By

VAL-IZEVBIGIE SOPHIA ADESUWA

PSC1809092

DEPARTMENT OF PHYSICS

FACULTY OF PHYSICAL SCIENCES

UNIVERSITY OF BENIN

BENIN CITY.

September, 2023.

PETROPHYSICAL EVALUATION USING MACHINE LEARNING MODELS FOR THE PREDICTION OF POROSITY

By

VAL-IZEVBIGIE SOPHIA ADESUWA

PSC1809261

A THESIS SUBMITTED TO THE DEPARTMENT OF PHYSICS

FACULTY OF PHYSICAL SCIENCES, UNIVERSITY OF BENIN, BENIN CITY.

IN PARTIAL FULFLMENT OF THE REQUIREMENTS FOR THE AWARD OF A BACHELOR OF SCIENCE (B.Sc.) DEGREE IN APPLIED GEOPHYSICS.

September, 2023

CERTIFICATION

This is to certify that this research work **"PETROPHYSICAL EVALUATION USING MACHINE LEARNING MODELS FOR THE PREDICTION OF POROSITY"** was carried out and presented by VAL-IZEVIGIE SOPHIA ADESUWA. Of the Department of Physics, Faculty Of Physical Sciences, University Of Benin City, Edo State, Nigeria.

………………………… ………………………...

DR M.O. IKPONMWEN **DATE**

Project Supervisor

………………………… …………………………

PROF. O.D OSAHON **DATE**

(Head of Department)

………………………… 14-09-23 …………………………

External Examiner **Date**

(External Supervisor)

DEDICATION

I would like to dedicate this research to GOD, who has been my rock and absolute support system throughout my academic journey. I would also like to dedicate this research to my family whose encouragement and love knew no bounds throughout this journey; to my course mates and lecturers who made my stay in this school worthwhile; and finally to my peers who supported me in one way or the other.

CERTIFICATION ON DISSERTATION ON PLAGIARISM

We the undersigned attest and declare that the dissertation VAL-IZEVIGIE SOPHIA ADESUWA titled **PETROPHYSICAL EVALUATION USING MACHINE LEARNING MODELS FOR THE PREDICTION OF POROSITY** has passed the anti-plagiarism test and does not violate any copyright regulations.

……………………………   …………………………

DR M.O. IKPONMWEN.   **DATE**

(Project Supervisor)

……………………….   ………………………..

PROF. O.D OSAHON   **DATE**

(Head of Department)

ACKNOWLEDGEMENTS

I am deeply grateful to my supervisor, DR M.O. IKPONMWEN, for his guidance and support throughout the course of this project; His expertise and patience have been invaluable to me and I have benefited immensely from his advice and insights.

I would also like to thank THE UNIVERSITY OF BENIN for providing me with the resources and support needed to complete this project. I am grateful for the opportunity to be a part of this wonderful community and to learn from/with my fellow students and superiors.

ABSTRACT

A key branch of artificial intelligence is machine learning models. The incorporation of these models into petrophysical analysis has gained popularity since it gives a more cost-effective and efficient method of acquiring petrophysical parameters. Porosity prediction was performed for this study utilizing machine learning models and 10 well log data from Niger Delta X-Fields wells. The well data from well 02 was used to train four machine learning models. The Ridge Regression model, Bagging Regressor model, ExtraTrees Regressor model, and Xgboost model were employed. The model that predicted porosity the best was chosen and used to forecast missing permeability logs from nine (9) other well log data sets. The available log data include Caliper, Gamma, Res_Deep (Resistivity), Density, PHIE (Porosity), SW (Water Saturation), VSH (Volume of Shale), and PERM (Permeability) logs. The bagging model was selected as it was the most effective, with a mean absolute error of 0.003, a root mean squared error of 0.010, and a mean absolute percentage error of 2.3%. This in turn enabled the prediction of porosity logs for the aforementioned amount of wells with a very low percentage error. Predictions were carried out using mainly the Permeability and Density logs as they provide a very strong correlation to Porosity. It was discovered that the difference between AIC value and mean absolute error value cannot be used as the only method of model evaluation; hence, the entire error margin, as well as the visualization using subplots must be taken into consideration when evaluating model performance. It should also be noted that, the percentage error of the various models differ slightly; however, the model with the smallest error margin should be used.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

1.1 INTRODUCTION

The word "Petrophysics" was coined to convey the physics of rocks related to petrology in the same way that geophysics is related to geology (Badure S, 2021). Petrophysics is the study of the physical and chemical properties of rocks as well as their interactions with fluids in subterranean reservoirs. Petrophysical studies are used to determine rock parameters such as lithology, porosity, permeability, and fluid content using well log data, core samples, and other geological information. These investigations are then integrated with geological and geophysical studies, in addition to reservoir engineering, resulting in a complete picture of the reservoir. Petrophysics insights are critical for assessing reservoir quality, guiding drilling and production operations, and developing accurate geological models for resource exploration and recovery. Petrophysics is the science that studies the fundamental chemical and physical properties of porous media, specifically reservoir rocks and their contained fluids. These properties include storage and flow properties (porosity, permeability, and fractional flow), fluid identification, fluid phase distribution within gross void space (saturation), interactions of surface forces existing between the rock and the contained fluids (capillary pressure), pressure measurements, and scalability. (Mesini, 2020)

With the world delving more and more into the exploration of the subsurface, a supplementary method of analyzing these rather large and complex datasets was needed. A way to create a model that could recognize and analyze trends or patterns in the datasets without being explicitly programmed to or supervised was in increasing demand. Machine learning models are programs that use an input dataset to learn patterns or relationships in order to generate predictions or perform specified tasks on previously unknown datasets (Sen, 2021). Machine learning models can identify patterns in highly dimensional datasets which enables them to make predictions better than the human brain in record time and without direct supervision.

Machine learning is a subset of Artificial Intelligence (AI) that covers statistical analysis. It teaches computers to solve problems by analyzing thousands of examples, learning from them, and using that experience to solve the same problem in different situations. It is essentially building a model and teaching it to make predictions without any programming. Machine learning (ML) is a subset of AI that focuses on building algorithms that learn—or enhance performance—based on the data they ingest.

Machine learning has emerged as a significant tool in petrophysics, allowing for faster and more accurate processing and interpretation of subsurface data. Machine learning has applications in predicting petrophysical parameters, categorizing lithological facies, assessing reservoir quality, and analyzing seismic data. Machine learning aids in exploration and production decision-making by quantifying uncertainty. Its combination with petrophysics allows geoscientists to gain useful insights from a variety of data sources, resulting in more efficient and informed oil and gas operations.

1.2 STATEMENT OF PROBLEM

This project addresses the problem of determining the porosity of reservoir rocks from well log data using machine learning models.

1.3 AIM AND OBJECTIVES

This study aims to accurately predict reservoir porosity from well log data using machine learning models.

The objectives of this study include:

1. To correlate and identify the logs related to porosity.
2. To apply several machine learning models for porosity prediction.
3. To assess and compare the performance of the constructed models in terms of accuracy and precision.
4. Make recommendations for the most effective machine learning approach for predicting porosity.

1.4 SCOPE OF THE STUDY

The goal of this research is to use machine learning algorithms to predict porosity values in the study area. The project will gather and preprocess a diverse dataset of material samples with known porosity values. Random forests and other machine learning methods will be created and tested for porosity prediction. The accuracy and precision of each model's performance will be compared. To find the most informative features, feature selection and engineering techniques will be used. The study will provide insights into the important parameters influencing the accuracy of porosity prediction and will recommend the most effective machine learning approach for porosity prediction.

1.5 AREA OF INTEREST

The data was acquired in the offshore Niger Delta. The offshore Niger Delta covers a large area. It is located in the southern part of Nigeria on the Gulf of Guinea on the west coast of central Africa.

Figure 1.1 Map of the Niger Delta *(Ideozu, 2016)*

1.6 GEOLOGY OF THE STUDY AREA

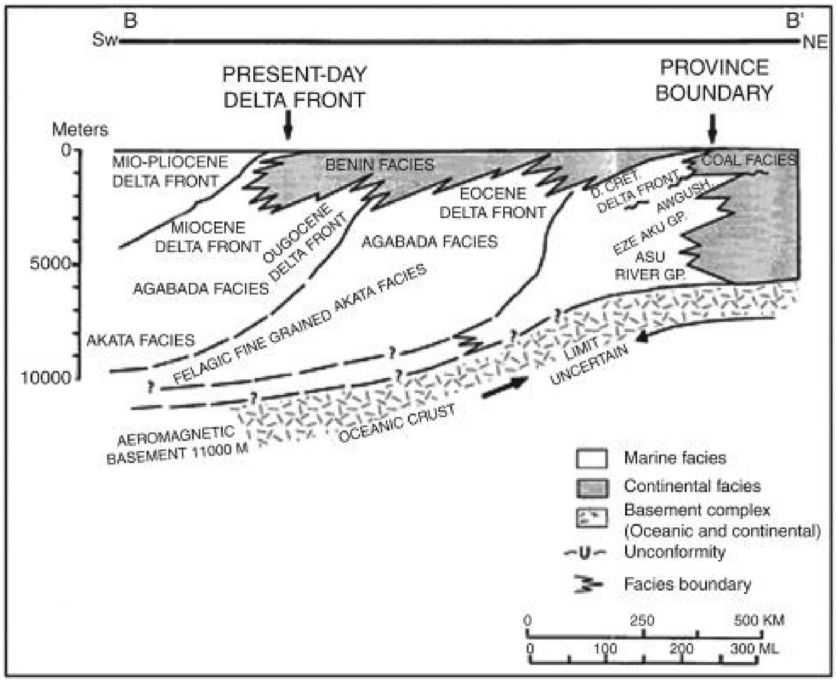

Figure 1.2 Geology of the Niger Delta Basin *(O.C. Ekwenye, 2014)*

The Niger Delta basin covers an area of about 75,000 km^2 and is composed of an overall regressive clastic sequence that reaches a maximum thickness of 9,000m to 12,000m (29,500ft to 39,400ft). The basin is composed of various sedimentary rocks deposited by the Niger River and other tributaries that were deposited during the Neogene and Quaternary periods of the Cenozoic Era. It is divided into three major formations, representing prograding depositional facies that are distinguished mostly based on their sand-shale ratios.

1. **AKATA FORMATION:**

   The Akata formation underlies the entire base of the delta. It is composed of marine pro-delta shale and turbidite sands that have been deposited since the Paleocene. It is only estimated that the formation is 7,000m thick. The Akata formation is a significant source rock for oil and gas in the Niger Delta and it is estimated to contain large reserves of hydrocarbons.

2. **AGBADA FORMATION**

   The Agbada Formation overlies the Akata Formation. It contains paralic siliciclastics and is considered to be the main deltaic sequence. The formation contains the most economically exploitable hydrocarbon in the region. In the delta field, the top of the Agbada formation is approximately 3,000ft below sea level. The base of the formation is 8,000ft below sea level.

3. **BENIN FORMATION**

   The Benin Formation overlies the Agbada Formation. The Benin Formation is a fluvial and upper coastal plain facies that have been deposited since the Oligocene. Like the Akata and Agbada formations, the Benin formation extends across the entire delta. It consists of sands that are up to 2,000m thick at some places in the delta. The Benin Formation is an important aquifer in the Niger Delta and is also a seal rock for oil and gas reservoirs.

1.7 SOURCE OF DATA

The data was sourced from Shell Petroleum Development Company of Nigeria Limited (SPDC).

1.8 DATA STATISTICS

The data consists of ten (10) wells plotted strategically across the field. All well data was provided in the ASCII format. The well logs, deviation data, well tops, and well headers were provided for each well. The various logs were provided:

1. **Gamma Ray Log:** A gamma ray log is a well logging tool used to detect the natural gamma radiation released by drilling rock formations. It contains useful data on lithology, stratigraphy, minerals, and hydrocarbon potential. The log aids in the identification of distinct rock types, the correlation of formations between wells, the detection of hydrocarbon reservoirs, the assessment of mineral composition, and the determination of wellbore location.
2. **Deep Resistivity Log:** A deep resistivity log is a tool used in well logging to determine the electrical resistivity of subterranean rocks. It generates a resistivity profile using electromagnetic induction principles. By detecting resistivity contrasts, the log aids in the identification of hydrocarbon reserves, the evaluation of formation fluids, the evaluation of lithology and reservoir quality, and the establishment of stratigraphic correlations.

3. **Density Log:** A density log is a tool used to measure the bulk density of subsurface formations during well logging. It gives useful information regarding the composition, porosity, and fluid content of the rock. The log measures the density of the formation surrounding the borehole using gamma-gamma or neutron-gamma techniques. The log density data are utilized to define lithology, identify fluid types, estimate porosity, and assess reservoir quality.
4. **Neutron Log:** A neutron log is a well logging tool used to assess the hydrogen concentration of subsurface formations. It provides information on porosity, lithology, and fluid saturation. The log operates by releasing high-energy neutrons into the formation, which interact with the rock's hydrogen atoms. The measurement of scattered or caught neutrons aids in the determination of porosity, the identification of fluid types, and the estimation of the presence of hydrocarbons.
5. **Water Saturation Log:** A water saturation log, also known as a saturation tool, is a well logging tool that is used to quantify the amount of water in subsurface formations. It measures the proportion of pore space occupied by water versus other fluids such as oil or gas. To determine water saturation, the log integrates observations from numerous logging instruments, including resistivity, neutron, and density logs. It is an important criterion in reservoir appraisal because it influences the potential productivity and overall quality of hydrocarbon reserves.
6. **Porosity Log:** A porosity log is a tool used in well logging to determine the porosity of subsurface rocks. It indicates the percentage of pore space within the rock that can potentially contain fluids like water, oil, or gas. The log measures porosity using a variety of methods, including sonic, density, and neutron porosity tools. The log porosity data are critical for evaluating reservoir quality, measuring fluid storage capacity, and estimating hydrocarbon potential.
7. **Permeability Log:** A permeability log is a tool used in well logging to determine the permeability of subsurface rocks. It provides information on a rock's ability to allow fluids to flow through it. To calculate permeability, the log uses measurements from multiple tools, including porosity, resistivity, and formation pressure logs. Permeability measurements are crucial in understanding the flow characteristics inside the formation and analyzing the reservoir's ability to produce fluids.

8. **Caliper Log:** A caliper log is a tool used in well logging to measure the diameter or thickness of a borehole. It provides information on the wellbore's size and shape. Caliper logs are critical for measuring wellbore stability, casing or tubing wear, and suitability for various downhole operations.
9. **Volume of Shale (VSH) Log:** A VSH (Volume of Shale) log is a tool used in well logging to assess the volume or proportion of shale in subsurface formations. It provides data on the composition and dispersion of shale, which is useful for reservoir characterization. The log uses gamma ray or resistivity data to differentiate between shale and non-shale strata. The VSH log aids in establishing the presence of possible hydrocarbon-bearing zones, identifying sealing barriers, and determining reservoir quality overall.

CHAPTER TWO

LITERATURE REVIEW

Machine learning (ML) is the subset of artificial intelligence (AI) that focuses on building systems that learn—or improve performance—based on the data they consume. Machine learning is a subset of Artificial Intelligence (AI) that covers statistical analysis. It teaches computers to solve problems by analyzing thousands of examples, learning from them and using that experience to solve the same problem in different situations. It is essentially building a model and teaching it to make predictions without any programming. The term machine learning refers to the automated identification and analysis of meaningful patterns in data (Ben-David, 2014).

Machine learning is a broad term that refers to computer approaches that use experience to improve performance or make accurate predictions. In this context, experience refers to previous knowledge available to the learner, which is often in the form of electronic data collected and made available for analysis. This data could include digitized human-labeled training sets or other sorts of information received through interaction with the environment. In all circumstances, its quality and magnitude are critical to the success of the learner's predictions (Mehryar Mohri, 2018).

Machine learning is a powerful alternative data-driven strategy for performing various petrophysical tasks using subsurface data. It can extract information from big, rich data sets and infer relationships, rules, and knowledge buried within the data. When the physics behind data becomes exceedingly complex, inexplicit, or even unclear/unknown, machine learning approaches outperform traditional physics-based interpretation models in terms of flexibility and application. Furthermore, machine learning can be used to help with numerous labor-intensive human interpretation tasks like faulty data identification, facies categorization, and geo-features segmentation from images data. (Chicheng Xu, Lei Fu, Tao Lin, Weichang Li, Shouxiang Ma , 2022).

2.1 BACKGROUND OF THE STUDY

For purposes of estimating petroleum volumes, economics, and decision-making, accurate estimation of reservoir properties such as porosity or water saturation is crucial. Such characteristics are determined via expensive, time-consuming core analyses or the analysis of petrophysical logs. In a field, not every well is cored, and there are only a certain number of fully cored wells.

One of the important rock parameters that must be considered for formation evaluation is porosity. Given that it specifies the interest zone storage potential for oil and natural gas, the proper estimation of porosity is crucial for reservoir rock. (Wakeel Hussain, 2023).

In order to quickly and economically predict reservoir characteristics from intricately coupled relations to indirect measures, in this case represented by well logs, machine learning techniques can be used. When the most modern technology is used, a better estimation of porosity can be made.

There have been many advancements in the petrophysical world since the introduction of machine learning in the early 1990s. Machine learning methods such as neural networks and decision trees were used in the early years to improve porosity and lithology forecasts based on well log observations. These methods showed encouraging results in terms of improving formation evaluation and reservoir characterization. Since then, tremendous progress has been made in the combination of machine learning and petrophysics. The increased availability of processing resources, improved algorithms, and the accumulation of big datasets has hastened the development of machine learning approaches in petrophysics.

In 2021, Roberto Ruiz, Anna Roubickova, Cyrille Reiser, and Neelofer Banglawala investigated the feasibility of mining a large petrophysics and rock physics well database in the Norwegian Sea using advanced machine learning algorithms to estimate reservoir elastic properties, and what this could mean for the optimization of petrophysical and rock physics workflows. They made use of two different ML algorithms in their study: Perceptron and Decision Tree based models. A perceptron is a mathematical abstraction of a single neuron in the human brain that may be chained and stacked to increase the model's complexity and predictive power; also referred to as Neural Networks (NN). Tree-based models, on the other hand, predict based on the recorded target properties of a collection of observations that are similar to the current sample, with the similarity rules defined during the training process. Their study illustrated the possibility of utilizing ML algorithms to reliably estimate porosity, hydrocarbon saturation, and Vsh from observed well logs in the Norwegian Sea, aided by a thorough petrophysical and rock physics atlas. This method, unlike previous empirical methodologies, does not require inputs such as mineralogy or fluid saturation to make a reliable forecast. (Roberto Ruiz, 2021).

In 2022, Pål Østebø Andersen, Carita Augustsson and Miranda Skjeldal explored Machine Learning Based Prediction of Porosity and Water Saturation. In this study, a time-efficient and economical method to estimate porosity, water saturation and hydrocarbon saturation is employed. Two Least Squares Support Vector Machine (LSSVM) machine learning models, optimized with Particle Swarm Optimization (PSO), were developed to predict these reservoir parameters, respectively. The porosity model was built around three features: density, deep resistivity, and gamma ray logs. It was discovered that logs traditionally linked with predicting porosity, such as neutron porosity and sonic, were deemed less meaningful throughout the feature selection process. (Pål Østebø Andersen, 2022).

In 2023, Ghoulem Ellah haithem Ifrene, Doina Irofti, Ruichong Ni, Sven Egenhoff and Prasad Pothana studied New Insights in Fracture Porosity Estimation using Machine Learning and Advanced Logging Tools. Their study made use of Pure Artificial Neural Networks Models and hybrid model (SVM-ANN). The pure Artificial Neural Network (ANN) model focused on regression analysis, whereas the hybrid model (SVM-ANN) focused on the combination of regression and classification analysis, also known as Support Vector Machine. The results were then validated against logging data using a combination of the Machine Learning technique and advanced logging technologies. Following that, the results are sent into two machine-learning algorithms. Pure Artificial Neural Networks and hybrid models were utilized to generate complete findings, which were then verified to ensure the models' accuracy. The results of the two approaches show that the hybridized model had lower Root Mean Square Error (RMSE) than pure ANN. Their findings clearly suggested that adding hybridized machine learning methods in fracture porosity calculations can help to construct more reliable static reservoir models in simulation programs. (Ghoulem Ellah haithem Ifrene, 2023).

In 2023, Tam Tran and Dinh Hoang Truong Thanh studied the Estimation of Petrophysical Properties by Using Machine Learning Methods. In this study, the authors estimate porosity and permeability using petrophysical data using both traditional petroleum engineering approaches and modern machine learning methods. They compared an artificial neural network (ANN) model to a Least-squares support-vector machines (LSSVM) model and an empirical model for predicting porosity and permeability. The results reveal that the ANN model has the highest R2 (coefficient of determination) of 0.9997 and the lowest MSE (mean squared error) of 6.7769 in predicting porosity and permeability. (Thanh, 2023).

2.2 HISTORY OF MACHINE LEARNING

The field's origins may be traced back to the 1950s, when researchers began investigating the use of computers to process and analyze massive volumes of data. We cannot claim that machine learning, particularly advanced ML algorithms, was invented by a single person. Many brilliant people contributed to its development.

When considering when ML was invented, though, a particular individual stands out. The phrase "Machine Learning" was coined in 1952 by Arthur Samuel, an IBM computer scientist and pioneer in AI and computer gaming.

By the early 1960s, Raytheon Company had built an experimental "learning machine" with punched tape memory dubbed Cybertron by the early 1960s to analyze sonar waves, electrocardiograms, and speech patterns using rudimentary reinforcement learning. It was repeatedly "trained" to recognize patterns by a human operator/teacher and was outfitted with a "goof" button that caused it to re-evaluate bad decisions. Nilsson's book on Learning Machines, which dealt mostly with machine learning for pattern categorization, was a prominent book on machine learning research throughout the 1960s.Pattern recognition remained a popular topic in the 1970s, as detailed by Duda and Hart in 1973. A report on employing teaching methodologies to teach a neural network to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal was presented in 1981 (Machine Learning Theory, 2023).

Franken Rosenblatt's (1958) discovery of the Perceptron, a sort of artificial neural network, in the late 1950s was one of the field's early and most impactful accomplishments. The Perceptron was created to process and classify incoming data, ushering in the modern era of machine learning.

In the 1960s and 1970s, researchers began investigating the use of decision trees, a form of algorithm that allows computers to make decisions based on a set of rules, for machine learning applications. In the 1980s, the discipline of machine learning saw the introduction of innovative algorithms such as vector machines and k-means clustering.

Due to the availability of enormous amounts of data and the development of new techniques such as boosting and random forests in the 1990s and early 2000s, machine learning witnessed a resurgence in popularity. With the introduction of deep learning algorithms and the broad acceptance of machine learning in a range of industries in the 2010s, the discipline continued to grow and evolve in popularity.

2.3 MACHINE LEARNING MODELS

1. **SUPERVISED LEARNING**

   The most frequent type of machine learning algorithm is supervised learning. A data scientist serves as a guide in this model, instructing the algorithm on what conclusions it should reach. This is the most useful type of machine learning. It is intended to learn from examples or from labeled data presented. It can be used to predict new outputs based on the input. It is used in petrophysics for facie prediction, permeability prediction, and other purposes. The purpose of the supervised learning paradigm is to infer a function from a set of data such that

   $$yn' = (xn, \Theta^*)$$

   Where $y'n$ denotes output at nth instance and $xn,^*$ denotes input and set of parameters.

2. **UNSUPERVISED LEARNING**

   Unsupervised learning is employed when there are no labels in the data. The data is fed into the machine, which determines the relationships between the data. Unsupervised learning techniques that are commonly used include clustering algorithms like K-Nearest Neighbor and K-Means clustering, anomaly detection methods like isolation forest, and association rule learning algorithms like the apriori algorithm (Tufail S, 2023). It can be used for face prediction and outlier detection. Unsupervised learning can be a goal in and of itself (finding hidden patterns in data) or a means to an end.

3. **REINFORCEMENT LEARNING**
   This is a sort of learning in which a computer program interacts with a dynamic environment to attain a certain goal. As it navigates its issue space, the software receives input in the form of incentives, which it attempts to maximize. Its applicability can be demonstrated in adjusting the video bit rate from high to low bit rate based on machine learning system estimates.

Another popular sort of machine learning that has not been addressed above is goal-oriented **Regression Learning**. In order to acquire total objectivity, it learns to make decisions based on interactions with its surroundings without the need for supervision. It attempts to optimize the benefit of each iteration by leveraging experience from previous iterations.

2.4 APPLICATIONS OF MACHINE LEARNING IN PETROPHYSICS

Several decades of hydrocarbon exploration have resulted in the acquisition and storage of vast volumes of well-related measurements, which have been utilized to describe the underlying geology and its hydrocarbon potential. The potential of these massive amounts of data has been increasingly realized over the last few decades as processing power and the use of new machine learning methods has increased. Machine learning has been applied in the petrophysical space to accelerate workflows, describe geology into distinct electrofacies, create forecasts, and much more. The following are a few examples of how machine learning has been utilized to assist with various areas of the petrophysical workflow:

1. **Automated Outlier Detection**: In well log measurements, outliers are data points that fall beyond the range of the dataset's normal or expected statistical distribution. Outliers can happen for a number of different reasons. Measurement and sensor mistakes, borehole washout, drilling vibrations that affect logging, unforeseen events and geology are only a few of the causes. It's crucial to recognize outliers and deal with them properly.

Manual techniques, including the use of Z-Scores, boxplots, and standard crossplots (scatterplots), can be used to do this. Domain knowledge should be applied to definitively determine whether a point or set of points are outliers. This will lessen the possibility of misclassifying points that could actually be meaningful data. Examples of applications of machine learning in automated outlier detection are: Unsupervised Outlier Detection Techniques for Well Logs and Geophysical Data (Siddharth Misra, 2020), Accelerating and Enhancing Petrophysical Analysis With Machine Learning: A Case Study for an Automated Well Log Outlier Detection And Reconstruction (Akkurt, et al., 2018), Novel Methodology For Automation of Bad Well Log Data Identification And Repair (Banas, McDonald, & Perkins, 2021).

2. **Well Log Repair**: Once outliers and flawed data have been found, they may typically be eliminated or fixed before a petrophysical interpretation or the use of a machine learning model is carried out. Numerous writers have looked at approaches to automate or semi-automate this process because it can take a lot of time to find and check data quality concerns in a project. Models are trained on "good" portions of the well, and they are then applied to forecast over the problematic interval. Examples of applications of machine learning in well log repair are: Novel Methodology for Automation of Bad Well Data Identification and Repair (Ryan Banas, 2021) , Machine Learning Assisted Petrophysical Logs Quality Control, Editing and Reconstruction (Singh, et al., 2020).

3. **Well Log Normalization**: The normalization of well logs is a typical step in the petrophysical workflow. It is used to eliminate data inaccuracy brought on by a variety of problems, such as disparities in tool and sensor technology, variations in the environment surrounding the borehole, and problems with tool calibrations. The process of "normalizing" involves scaling or calibrating the well logs to make them consistent with other logs from other wells in the same field or region. By manually doing a single-point normalization (linear shift) or a two-point normalization ('stretch and squeeze') on the necessary curve, this can be accomplished. Gamma-ray logs are normally normalized, but they can also be normalized for neutron porosity, bulk density, acoustic, and spontaneous potential logs. In general, resistivity logs are not normalized unless there is a compelling

reason to. Examples of applications of machine learning in well log normalization are: Machine Learning For Well Log Normalization (Akkurt, et al., 2019) , Petrophysics: Gamma Ray Normalization in Python (McDonald, Petrophysics: Gamma Ray Normalization in Python, 2020).

4. **Prediction Of Missing Well Logs:** When performing subsurface characterization, it is vital that datasets are as thorough as feasible; nevertheless, there are instances in which well logging measurements are lacking. This can include varied well log vintages, logging speeds that are too fast for tool sampling, environmental concerns in boreholes, incorrect data management, high acquisition costs, and casing impacts. In these cases, data can be filled in using empirical connections between the last few logging values or by utilizing machine learning models from the present well and/or offset wells. Examples of applications of machine learning in prediction of missing well log data are: Visualizing Well Data Coverage Using Matplotib (McDonald, Visualizing Well Data Coverage Using Matplotib, 2020) , Synthetic Well Log Generation using Machine Learning Techniques (Mohammad Khan, 2021), Machine Learning Techniques for the Prediction of Shear Wave Velocity Using Petrophysical Logs (Shi & Zhang, 2021).

5. **Depth Alignment of Logging Measurements:** All well logs are connected to depth, which is a crucial measurement for the effective development and completion of a well. It enables the choice of hole intervals and the proper depth setting for packers. As a result, it is crucial that a constant depth reference be used for all measurements. Different depth references between wireline and logging while drilling passes, cable stretching, different sampling rates between tool types and logging passes, and even variations in weather conditions and sea surface swell on semi-submersible drilling platforms can cause well log measurements to differ in depth from one another. In the oil and gas business, ensuring that well logs from several passes and runs are consistent has long been a problem. Data alignment has historically been done manually, comparing two or more logging curves from different passes and adding pins where necessary. Then, these features are positioned so as to produce data that is in-depth with one another. This method can frequently be biased and time-consuming to complete. Cross-correlation and

covariance measures between two logging curves have been used to construct semi-automatic and automated methods over the years. Recently, though, several writers have made an effort to automate this procedure and eliminate any bias using machine learning models. Examples of applications of machine learning in depth alignment of logging measurements are: Reinforced Learning Technique for Multi-Well Logs Depth Matching Yield Better Reservoir Delineation (Bittar, Wang, Chen, & Wu, 2020) , A Machine Learning Framework For Automated Well Log Depth Matching (Le, Liang, Zimmermann, Zeroug, & Heliot, 2019).

6. **Prediction of Continuous Curves from Discrete Core Measurements:** Permeability is one of the main outputs from a petrophysicist. It serves as a guide to the ease with which fluids, particularly hydrocarbons, can pass through a rock or reservoir. Downhole formation permeability cannot currently be measured directly by logging tools; instead, it must be inferred from tool responses or empirical connections. Many of the empirical correlations may not fully apply to other areas because they were obtained from core measurements taken in particular geological and geographical regions. As a result, it is customary to apply the link discovered between core permeability and porosity to log-derived porosity. Core data may only be present in a small number of wells within a given area and is not always accessible. This is a result of how costly and time-consuming coring operations are. In order to forecast permeability or porosity in all other wells, various authors have used machine learning models that are trained on important wells. Examples of applications of machine learning in prediction of continuous curves from discrete core measurements are: Porosity-Permeability Relationships Using Linear Regression in Python (McDonald, Porosity-Permeability relationships using linear regression in python, 2020) , Synthetic Well Log Generation using Machine Learning Techniques (Akinnikawe, Lyne, & Roberts, 2018).

7. **Prediction of Facies:** A key task in geoscience and petrophysics is comprehending the subsurface lithology. One of the responsibilities frequently given to the petrophysicist is determining a lithology flag or mineral quantities. Well log readings are frequently categorized into several lithological groups, called facies, using machine learning

methods. Both supervised learning and unsupervised learning algorithms can be used to accomplish this procedure. Exploratory data analysis (EDA), which divides the data into groups based on shared qualities or properties, is frequently used to cluster the data unsupervised. The same cluster contains data points that are comparable to one another, and the other cluster contains points that are different from one another. Examples of applications of machine learning in predictions of facies are: How to Use Unsupervised Clustering on Well Log Data with Python (McDonald, How To Use Unsupervised Clustering On Well Log Data With Python, 2021) , Machine Learning in Rock Facies Classifications: An Application of XGBOOST (Licheng Zhang, 2017).

2.5ADVANTAGES OF MACHINE LEARNING IN PETROPHYSICS

In contrast to traditional physics-based interpretation models, machine learning approaches have the advantage of being more flexible and having a larger range of applications when the physics underlying the data becomes exceedingly complex, oblique, or even unclear/unknown. Additionally, machine learning can be used to help with a variety of labor-intensive human interpretation tasks, including the segmentation of geo-features from imagery data, the classification of faces, and the identification of faulty data.

In addition, the following list includes some specific areas where machine learning helps with the examination of petrophysical data:

1. **Machine Learning Solutions**: As its use in petrophysics rapidly expands, machine learning has shown to be an effective tool for a variety of issues. The goal of employing machine learning is to identify a solution to aid in corporate decision making, regardless of the model or approach employed. Sometimes, a solution doesn't have to be perfect to be effective, especially when taking into account the uncertainties associated with the data's collection and representativeness. From a physics standpoint, problems might not be addressed if a physical model is lacking or too complex. In this situation, machine learning can typically offer an answer with quantifiable uncertainty because it is a data-driven technology. All models are incorrect, as stated, yet some are really helpful. It is helpful if the machine learning solution can support corporate decision making.

2. **Machine Learning Automation for Productivity and Consistency**: Another significant benefit of adopting machine learning is the increased productivity achieved when tasks like data labeling, grouping, and well log correlation are carried out more effectively by machines as opposed to people. Human interpretation is dependent on knowledge, experience, and abilities while machine learning relies on mathematically based models. As a result, machine learning results are more reliable than human-produced goods, which may be subjective and prejudiced. As a result, machines can complete some monotonous tasks far more quickly and consistently than humans. It was also asserted that workflow automation considerably increased project time efficiencies by offering a structured method of evaluating big, varied, and complex datasets, which may potentially free the petrophysicist from spending the majority of the project's time doing manual repetitive work.

3. **Deciphering High-Dimensional Data:** High-dimensional data such as arrays, images, waveforms, and 3D volumes are challenging to the capability of human recognition. Even with the assistance of modern 3D visualization, it is still hard for human to accurately label and interpret high-dimensional data. For example, it is nearly impossible to pin-point different minerals on a high-resolution thin section image or trace every pore on a 3D volume CT scan of a rock. In these cases, we must resort to machine to find a solution.

2.6 LIMITATIONS OF MACHINE LEARNING

1. **Data Correctness and Representativeness:** Even if the data are correct, they may not be representative (Amabeoku, 2014) . Several factors, including as tool calibration, tool resolution, samples acquisition, and physical sample modification prior to testing, might lead to data that is not representative. If the data used is not indicative of the desired problem, both ML and human analysis will provide results that are unreliable. Data outliers can be identified by applying ML approaches (Akkurt, et al., 2018) ; then use feedback from subject matter experts to fix the data that has been detected. However, it is not anticipated that a machine would be able to notice this data quality issue, which will have an impact on the predicted outcomes, if uncorrected non-representative data is

utilized in ML modeling. Before incorporating the data in the machine learning model, it is advised that incorrect outliers be eliminated and all necessary modifications be made to the data.

2. **Data Size, Quality, and Relevance:** To train the model, most ML algorithms require the so-called "big data". The adage "the bigger the better" might be accurate when data is accurate and representative. However, in the field of petrophysics, the data collected are frequently sparse; for instance, measurements on a few hundred core plugs are normal for specific rock parameters. Data quality drastically declines as data sets grow larger due to the possibility that the measurements made for each set may not follow the same process. Since geological conditions may have altered between wells and fields, the relevance of data collected far from the area under study decreases. When developing a straightforward model, it is often possible to get better results by employing a small but highly relevant training dataset.
3. **Labeling of Data with Accuracy:** The accuracy of the labeled data has a significant impact on the ML model built using that data. The ML model will be misleading if the labeled data is incorrect. For instance, core data are frequently used as the basis for calibrating well log interpretation or training machine learning models that are based on well logs. The well log-based model will also be wrong if the core measurements are handled incorrectly and produce inaccurate findings (for example, permeability values without applying the proper reservoir stresses). The labeled data from the core cannot be verified by a machine to determine its accuracy. Facies are classified by geologists based on their personal experiences, and this classification can be prejudiced. As a result, the bias will also be present in the ML model that was trained using the human-labeled data.

CHAPTER THREE

METHODOLOGY

3.1 MATERIALS

For this project, the following materials were used:

1. 10 wells with composite logs including; gamma ray, deep resistivity, density, water saturation, neutron log, porosity log and permeability log.
2. **Jupyter Notebook:** The Jupyter Notebook is an open-source web application used to make and share documents with live code, equations, visuals, and narrative text for a variety of purposes, including data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and more.
3. **Python Programming Language:** Python is frequently used for creating websites and applications, automating repetitive tasks, and analyzing and displaying data. Learning is comparatively simple. Many professionals who are not programmers, such accountants and scientists, have started using Python for a variety of routine activities like managing finances.
4. **Sklearn Library:** Scikit-Learn, sometimes referred to as Sklearn, is a Python toolkit used to implement statistical modeling and machine learning algorithms. It can be used to create multiple machine learning models for clustering, classification, and regression with scikit-learn, and also, use statistical tools to analyze these models. Additionally, it offers dimensionality reduction, feature extraction, feature selection, ensemble approaches, and built-in dataset functionality.
5. **NumPy Library:** The NumPy library is an essential building block for learning Machine Learning. NumPy may be used to perform a variety of array-based mathematical tasks. It expands Python with complex data structures for efficient array and matrix calculations, as well as a large library of high-level mathematical functions that operate on these arrays and matrices.
6. **Panda Library:** Pandas is a Python library for manipulating data sets, focusing on tabular data. It offers data analysis, cleansing, exploration, and manipulation, including sorting, subsets, summary statistics, and data transformations. It seamlessly integrates

with popular PyData ecosystem programs like NumPy, Matplotlib, Seaborn, Plotly, and other visualization packages.

7. **Matplotlib Library:** Matplotlib is a Python package that enables the generation of static, animated, and interactive visualizations. Matplotlib is the Python programming language's basic visualizing or plotting package. Matplotlib is a useful tool for doing a wide range of activities. It can generate many graphical reports such as line plots, scatter plots, histograms, bar charts, pie charts, box plots, and many more. This package also allows for three-dimensional plotting.
8. **Seaborn Library:** Seaborn is a matplotlib-based Python data visualization package. It offers a high-level interface for creating visually appealing and informative statistical visuals. Seaborn assists in exploring and comprehending data. Its charting functions operate on data frames and arrays containing entire datasets, performing the necessary semantic mapping and statistical grouping internally to generate useful graphs.
9. **Lasio Library:** Lasio's primary purpose is to read and write data and metadata to and from LAS files. It is intended to read as many LAS files as possible, even those with common faults and invalid formatting.
10. **Missingno Library:** The missingno library is a Python data visualization package that is used to visualize missing data in a dataset (null or NaN values). It creates several graphical representations of missing data patterns to provide a simple and informative approach to determine the completeness of your data.

3.2 METHOD

3.2.1 DATA COLLATION AND IMPORTATION

The well data is firstly converted into .csv format. This will enable the IDE to analyze and process it easily. Jupyter's Notebook, an Integrated Development Environment (IDE), is used for analytical and predictive procedures. These libraries contain graphic and mathematical software.

```
In [1]:  import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns
         import numpy as np
         %matplotlib inline
         import lasio
         import missingno as msno
         import warnings
         warnings.filterwarnings('ignore')
         %config InlineBackend.figure_format= 'svg'
```

Figure 3.0.1 Data Importation using libraries

3.2.2 DATA CLEANING AND VISUALIZATION

Data Cleaning is an important stage in the machine learning process. It entails getting the data ready for use in a predictive model. This procedure entails detecting and correcting data mistakes, as well as dealing with missing values and outliers. Missing data is frequently encountered in data sets, and it was solved by removing the rows with missing values entirely. Data cleaning has a substantial impact on the predictive model's performance. A model that has been trained on clean data is more likely to be accurate and precise than one that has been trained on dirty data.

```
In [6]:  well= well_02.df()
         well.head()
Out[6]:
```

	CALIPER	GAMMA	RES_DEEP	DENSITY	PHIE	SW	VSH	PERM
DEPTH								
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 3.0.2 Identification and removal of missing values using missingno library

```
In [7]:  missing_well = well.isnull()
         for column in missing_well.columns.values.tolist():
             print(column)
             print(missing_well[column].value_counts())
             print("")
```

```
CALIPER
False    16857
True      8045
Name: CALIPER, dtype: int64

GAMMA
False    16855
True      8047
Name: GAMMA, dtype: int64

RES_DEEP
False    16855
True      8047
Name: RES_DEEP, dtype: int64

DENSITY
False    16844
True      8058
Name: DENSITY, dtype: int64

PHIE
False    16674
True      8228
Name: PHIE, dtype: int64

SW
False    16674
True      8228
Name: SW, dtype: int64

VSH
False    16839
True      8063
Name: VSH, dtype: int64

PERM
False    16674
True      8228
Name: PERM, dtype: int64
```

Figure 3.0.3 Data Visualization

3.2.3 DATA CORRELATION

Correlation is a measure of the inter-dependence between variables. In machine learning models, correlation measures the statistical link between variables, which is defined by the correlation coefficient "r" ranging from -1 to 1. It is critical for feature selection, model evaluation, and data relationship comprehension. Correlation analysis aids in the identification of essential traits, but it must be used with caution to avoid misinterpretation because correlation does not indicate causation, and other evaluation metrics should be evaluated alongside it.

- r = 1 : Total positive linear correlation

- r = 0 : No linear correlation
- r = -1 : Total negative linear correlation

3.2.4 DATA SPLITTING

It is customary in machine learning to divide a data set into two parts: training and testing. The training set is used to train the model, whereas the test set is used to assess the model's performance. The train-test split process entails randomly dividing the data into a training set and a test set in a 70:30 or 80:20 ratio. This approach enables the model to be trained and evaluated on many datasets, which can help to avoid over-fitting or the problem of the model performing well on training data but badly on fresh, unknown data. The train-test approach is an essential aspect of the machine learning process since it allows the model's performance to be assessed in a controlled environment. It also contributes to the model's ability to generalize well to new, previously unknown data, which is vital for real-world applications of the model.

```
In [14]:  x= well.drop(["PHIE"], 1)
          y= well[["PHIE"]]

In [15]:  from sklearn.model_selection import train_test_split as tts

          x_train, x_test,y_train, y_test = tts(x,y, test_size= 0.15, random_state=1)
          print("number of test samples :", x_test.shape)
          print("number of train samples :", x_train.shape)

          number of test samples : (2502, 7)
          number of train samples : (14172, 7)

In [16]:  density= x_test[['DENSITY']]
```

Figure 3.0.4 Data splitting into test and training data

3.2.5 FEATURE SELECTION

In machine learning, feature selection entails selecting a subset of the most relevant features from a dataset in order to improve model performance, reduce overfitting, and improve interpretability. It is divided into filter, wrapper, and embedding approaches, with factors such as domain expertise and data exploration playing a significant part in the selection process. Feature selection is an iterative and domain-specific procedure that aids in the optimization of models, the reduction of computational costs, and the prevention of over-complexity in high-dimensional datasets.

3.2.6 MODEL TRAINING AND SELECTION

Model testing and training are critical processes in the creation of machine learning models. Data is prepared during training, and the model learns patterns and relationships from a training set. A validation set is used to tune hyper parameters for best performance. A different test set is used to assess the model's ability to generalize to previously unseen data. This iterative method ensures the development of precise and robust models suitable for use in real-world applications. Some of the tests on which the model was tested include:

- Mean absolute error
- Mean squared error
- Root mean squared error
- Mean absolute percentage error

Machine learning models are computational algorithms or mathematical constructs that are designed to learn and predict or make judgments from data automatically. A total of four models were trained and tested on the data. They include:

- **Ridge Regression Model:** Ridge regression, also known as Tikhonov regularization, is a linear regression technique used to reduce model complexity and handle the issue of multicollinearity (strong correlation between predictor variables). It is an extension of ordinary least squares (OLS) regression in which a penalty term is introduced into the loss function.
- **Bagging Regression Model:** Bagging, short for bootstrap aggregation, is an ensemble learning technique that integrates numerous models, often known as base learners, to improve the ensemble's overall predictive performance and stability. It is often used for classification as well as regression applications.
- **ExtraTree Regression Model:** Extra Trees Regressor, also known as Extremely Randomized Trees Regressor, is a regression method version based on the Random Forest algorithm. It is an ensemble learning method that makes predictions by combining numerous decision trees. While Random Forest chooses a subset of characteristics at each split at random, Extra Trees goes a step further by randomly selecting split sites. The

splitting thresholds in Extra Trees are chosen at random rather than being established by the optimization method used in conventional decision trees and Random Forest.

- **XGBoost Regreesion Model**: XGBoost, which stands for Extreme Gradient Boosting, is a popular gradient boosting framework in machine learning competitions and real-world applications. It is intended to provide exceptional performance and accuracy for both regression and classification workloads. Gradient boosting is a method of ensemble learning that combines numerous weak predictive models (usually decision trees) to build a strong predictive model. XGBoost improves efficiency, scalability, and predictive performance by extending gradient boosting.

3.2.7 EVALUATION OF MODEL PERFORMANCE USING SUBPLOTS

The dividing of a single figure or canvas into numerous smaller plots or charts is referred to as a subplot. They enable you to exhibit and compare various data sets or visualizations within the same figure. Subplots are often used in data visualization and data analysis to exhibit numerous pieces of related information at the same time. Using subplots to evaluate model performance is a popular practice in data visualization, especially when comparing many models or different aspects of a single model's performance. Subplots enable the display of many plots or visualizations within a single figure, making it easier to examine and compare results.

3.2.7 APPLY THE SELECTED MODEL TO NEW FRESH DATA

The model with the least percentage errors, the least difference between the Akaike Information Criterion (AIC) and mean absolute error, as well as the best visualized model should be selected.

```
In [53]: Baggingpred = Bag.predict(x)

In [54]: Deploy['Predicted_PHIE'] = Baggingpred
         Deploy
```

Figure 3.0.5 Porosity Prediction using the Bagging Regression Model

CHAPTER FOUR

RESULTS AND DISCUSSION

Table 4.1 Table showing the training scores, test scores and percentage evaluation of the different models

S/N	MODELS	TRAIN SCORE	TEST SCORE	AIC	MEAN ABSOLUTE ERROR	ROOT MEAN SQUARED ERROR	MEAN ABSOLUTE PERCENTAGE ERROR
1	Ridge Regression	0.885876	0.888025	-10004.896174	0.023385	0.032747	10.926906
2	Extra Trees Regreesion	0.999478	0.988004	-15605.971006	0.003210	0.010719	2.472291
3	XGBoost	0.996210	0.990684	-16229.266037	0.004020	0.009446	2.546310
4	Bagging	0.998138	0.988455	-15705.720365	0.003080	0.010515	2.331187

When choosing the best machine learning model, the following indices are taken note of as seen in table 4.1

4.1 TRAINING ACCURACY

The accuracy of a model on training data is referred to as training accuracy in machine learning. It assesses the model's ability to predict the proper outcome from a given input in the training set of data. The training accuracy of a machine learning model is an important parameter to consider since it reflects how well the model can learn from training data. A machine learning model's training accuracy evaluates how well the model performs on the data it was trained on. It is commonly quantified using metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) on the training data for classification and the percentage of correctly categorized training instances for regression. Although high training accuracy implies a strong match to the training data, it does not guarantee successful generalization to new data. To ensure that the model generalizes successfully, it should be assessed on a separate validation or test dataset. Building durable machine learning models requires balancing training and validation accuracy.

The Mean Squared Error (MSE) is a popular statistic for evaluating regression model performance. In a dataset, it computes the average of the squared discrepancies between expected and actual values. Lower MSE values imply a better model-to-data fit, which is important for model comparison and selection. MSE, on the other hand, is susceptible to outliers and measures mistakes in squared units, therefore it is not always the best interpretable metric on its own.

The Root Mean Squared Error (RMSE) is a metric used to evaluate the performance of regression models. It calculates the square root of the average of the squared differences between predicted and actual values. RMSE is expressed in the same units as the target variable, making it more interpretable. Lower RMSE values indicate better model performance, and it is commonly used for model comparison and selection. However, RMSE is sensitive to outliers, so it's essential to consider its limitations in the presence of extreme data points.

The Mean Absolute Error (MAE) is a common statistic for assessing regression models. It computes the average absolute difference between predicted and actual values, resulting in a simple and understandable measure of model performance. Lower MAE values indicate better model accuracy, and they are expressed in the same units as the target variable, making them simple to grasp. MAE is resistant to outliers, making it appropriate for datasets with extreme values. It's often used for model comparison and selection, and it's simple to understand for non-technical stakeholders.

The Mean Absolute Percentage Error (MAPE) is a statistic used to evaluate forecasting or regression models, particularly those that deal with percentages or proportional data. It expresses the average percentage difference between expected and actual values as a percentage. Lower MAPE values suggest that the model is more accurate. MAPE is a scale-invariant and interpretable model that may be used for model comparison and communication with non-technical stakeholders. It is, however, sensitive to zero values in the data and must be handled with care in such instances.

In terms of how well the model fit the data, the hierarchy goes thus: Ridge Regression model > ExtraTrees Regression Model > Bagging Regression Model > XGBoost Regression Model. In terms of accuracy, the XGBoost regression model and the Bagging regression model performed very well.

4.2 CORRELATION RESULTS

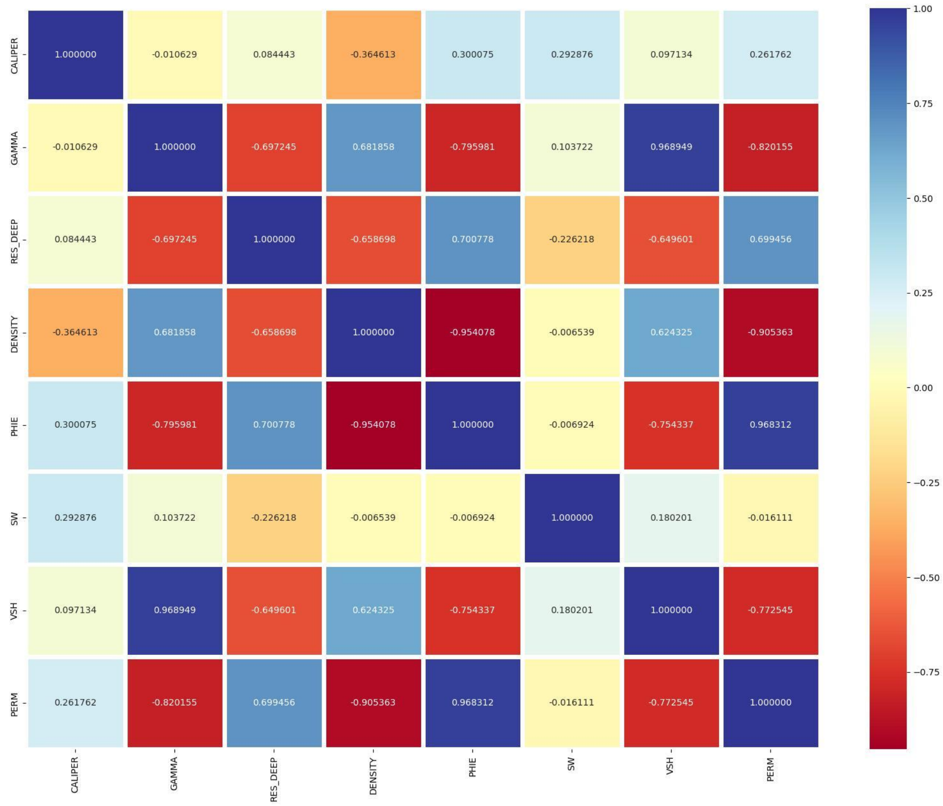

Figure 4. 1 Spearman Correlation

Table 4.2 Table showing the correlation values of Porosity with the other logs

S/N	WELL LOG	CORRELATION VALUE
1	Caliper	0.300075
2	Gamma	-0.775981
3	Res_Deep	0.700778
4	Density	-0.954078
5	Porosity	1.000000
6	Water Saturation	-0.006924
7	Permeability	0.968312

A positive correlation indicates that when one variable increases, the other variable tends to increase too. A negative correlation value indicates that when one variable increases, the other variable tends to decrease.

4.3 FEATURE SELECTION RESULTS

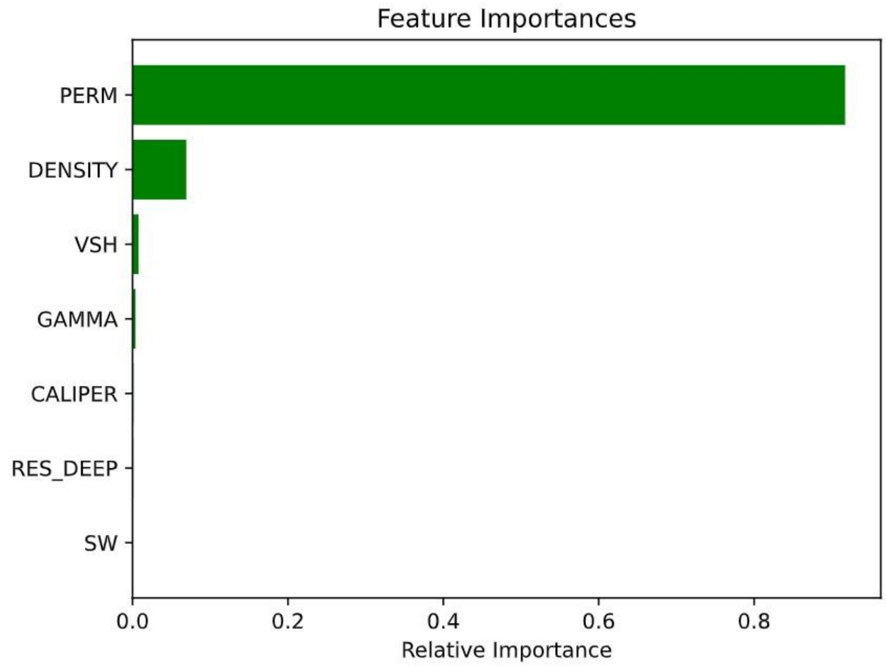

Figure 4.2 Feature Selection

This chart shows that Porosity can be computed directly from Permeability (with a feature selection score of about 90%), Density (with a feature selection score of about 15%), VSH (with a feature selection score of less than 10%) and gamma (with a feature selection score of far less than 10%). In terms of relative importance, the Permeability log and Density logs will used to predict porosity.

4.4 MODEL TRAINING AND SELECTION RESULTS

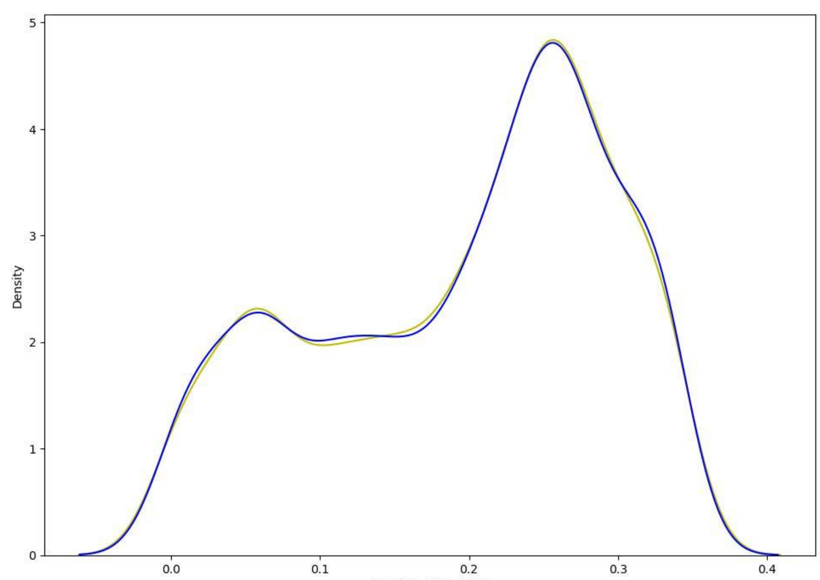

Figure 4.3 Subplot of Bagging Regression Prediction

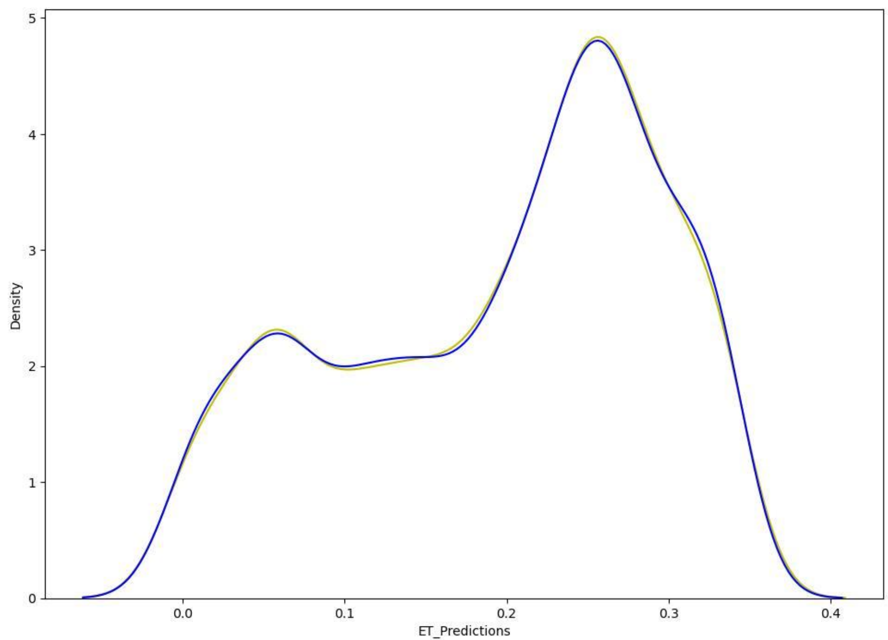

Figure 4.4 Subplot of ExtraTrees Regression Model Prediction

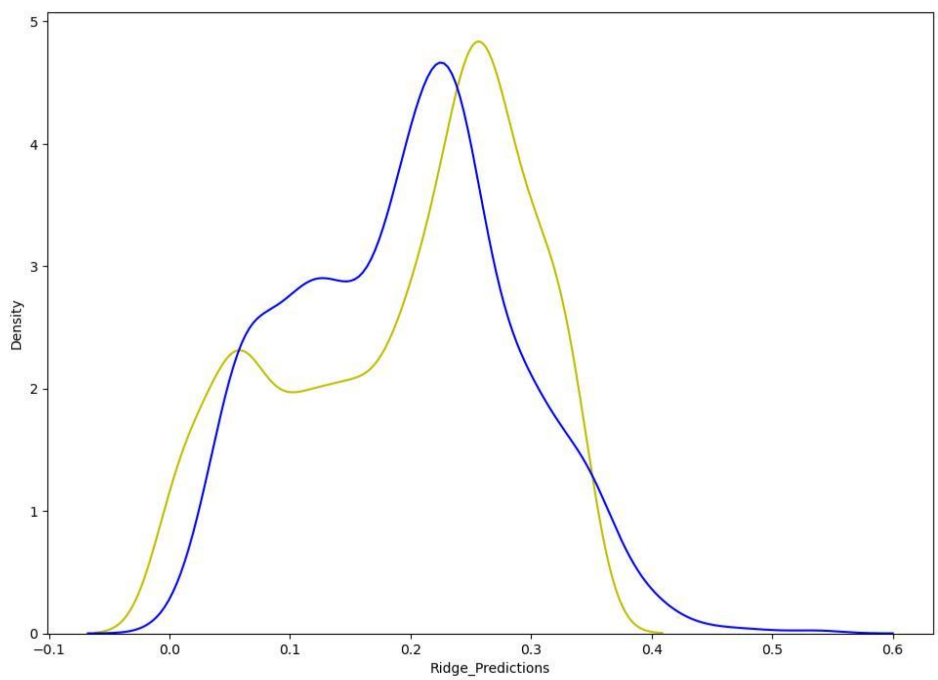

Figure 4. 5 Subplot of Ridge Regression Model Prediction

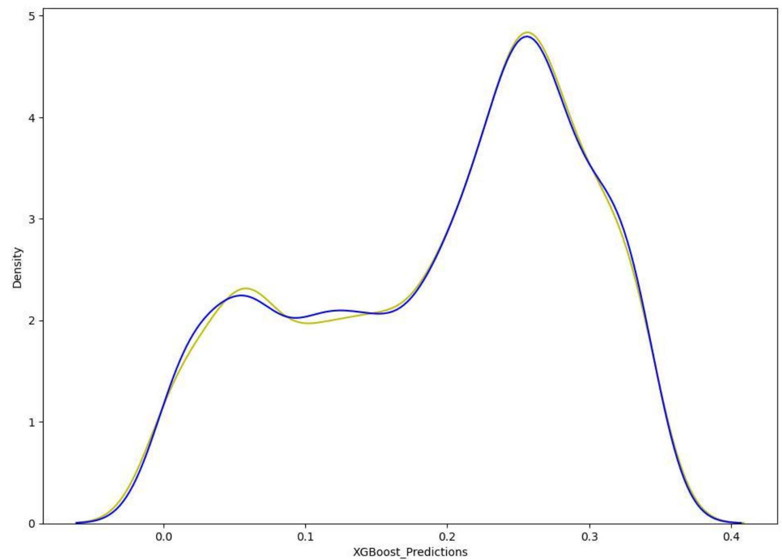

Figure 4.6 Subplot of XGBoost Regression Model Predictions

- The bagging model had the best fit and deviated the least from the actual porosity as shown in fig 4.3
- The ExtraTrees Regression model performed better than the ridge model; however, it under predicted for the most part as shown in fig 4.4
- The Ridge regression model performed very poorly; it over predicted and deviated very far from the actual porosity values as shown in fig 4.5
- The XGBoost model did not perform very well as it over predicted and under predicted at various points as shown in fig 4.6

4.5 APPLYING THE SELECTED MODEL TO NEW DATA

The models were tested and trained on Well 02. The Bagging Regression Model was selected as it satisfied all the parameters of selection and had the best results. The model was applied to predict the Porosity of nine (09) other wells. Their results are as follows:

1. **WELL 02**

Out[39]:

DEPTH	Actual	Ridge_Predictions	ET_Predictions	Bagging_Predictions	XGBoost_Predictions
9036.5	0.2426	0.215473	0.242605	0.242620	0.243464
9004.5	0.2386	0.207810	0.230064	0.227220	0.235441
4922.0	0.2517	0.220906	0.251710	0.251700	0.250004
11983.5	0.0029	0.037859	0.011530	0.007427	0.007406
4724.0	0.3049	0.288963	0.304909	0.304860	0.302115
...	...	...	...	...	...
11812.5	0.0747	0.087684	0.074706	0.074650	0.075430
9788.5	0.1004	0.160460	0.098379	0.100450	0.105724
8308.5	0.1720	0.150490	0.172068	0.172130	0.169226
7789.5	0.2288	0.207582	0.228942	0.229450	0.232854
7048.5	0.2675	0.241142	0.267544	0.264980	0.265616

2502 rows × 5 columns

Figure4.7 Predicted PHIE values for Well 02

2. **WELL 01**

In [54]:

```
Deploy['Predicted_PHIE'] = Baggingpred
Deploy
```

Out[54]:

DEPTH	CALIPER	Gamma	RES_DEEP	DENSITY	vsh	SW	PERM	Predicted_PHIE
0.0	-999.0	-999.000	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
0.5	-999.0	-999.000	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
1.0	-999.0	-999.000	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
1.5	-999.0	-999.000	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
2.0	-999.0	-999.000	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
...	...	...	...	...	...	...	...	...
12973.5	-999.0	134.760	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
12974.0	-999.0	135.675	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
12974.5	-999.0	128.070	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
12975.0	-999.0	105.885	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
12975.5	-999.0	-999.000	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589

25952 rows × 8 columns

Figure 4. 8 Predicted PHIE values for Well 01

3. **WELL 04**

```
In [61]: Deploy['Predicted_PHIE'] = Baggingpred
         Deploy
```

Out[61]:

DEPTH	CALIPER	Gamma	RES_DEEP	DENSITY	SW	vsh	PERM	Predicted_PHIE
2000.0	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0	-999.0	0.0589
2000.5	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0	-999.0	0.0589
2001.0	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0	-999.0	0.0589
2001.5	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0	-999.0	0.0589
2002.0	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0	-999.0	0.0589
...	...	...	...	...	...	...	...	...
13098.0	-999.0	32.9894	10.5660	-999.0	-999.0	-999.0	-999.0	0.0589
13098.5	-999.0	33.1646	10.5660	-999.0	-999.0	-999.0	-999.0	0.0589
13099.0	-999.0	32.4712	10.5723	-999.0	-999.0	-999.0	-999.0	0.0589
13099.5	-999.0	31.7778	10.5805	-999.0	-999.0	-999.0	-999.0	0.0589
13100.0	-999.0	31.0844	10.5888	-999.0	-999.0	-999.0	-999.0	0.0589

22201 rows × 8 columns

Figure 4. 9 Predicted PHIE Values for Well 04

4. **WELL 05**

```
In [68]: Deploy['Predicted_PHIE'] = Baggingpred
         Deploy
```

Out[68]:

DEPTH	CALIPER	Gamma	RES_DEEP	DENSITY	SW	vsh	PERM	Predicted_PHIE
1000.0	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0000	-999.0	0.0589
1000.5	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0000	-999.0	0.0589
1001.0	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0000	-999.0	0.0589
1001.5	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0000	-999.0	0.0589
1002.0	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0000	-999.0	0.0589
...	...	...	...	...	...	...	...	...
12208.5	-999.0	98.5327	3.0744	-999.0	-999.0	0.5184	-999.0	0.0589
12209.0	-999.0	96.3137	3.2902	-999.0	-999.0	0.4844	-999.0	0.0589
12209.5	-999.0	95.3689	3.6354	-999.0	-999.0	0.4706	-999.0	0.0589
12210.0	-999.0	95.3689	4.2573	-999.0	-999.0	0.4706	-999.0	0.0589
12210.5	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0000	-999.0	0.0589

22422 rows × 8 columns

Figure 4. 10 Predicted PHIE values for Well 05

5. WELL 06

In [75]:
```
Deploy['Predicted_PHIE'] = Baggingpred
Deploy
```

Out[75]:

DEPTH	CALIPER	Gamma	RES_DEEP	DENSITY	Neutron	SW	vsh	PERM	Predicted_PHIE
1000.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
1000.5	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
1001.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
1001.5	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
1002.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
...	...	...	...	...	...	...	...	...	...
12498.5	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
12499.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
12499.5	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
12500.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
12500.5	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589

23002 rows × 9 columns

Figure 4.11 Predicted PHIE values for Well 06

6. WELL 07

In [54]:
```
Deploy['Predicted_PHIE'] = Baggingpred
Deploy
```

Out[54]:

DEPTH	Gamma	RES_DEEP	DENSITY	SW	vsh	PERM	Predicted_PHIE
1000.0	42.9207	101.084	-9999.0	-999.0	-999.0	-999.0	0.0589
1000.5	39.9890	111.056	-9999.0	-999.0	-999.0	-999.0	0.0589
1001.0	35.1668	121.504	-9999.0	-999.0	-999.0	-999.0	0.0589
1001.5	31.7753	130.093	-9999.0	-999.0	-999.0	-999.0	0.0589
1002.0	31.2861	136.287	-9999.0	-999.0	-999.0	-999.0	0.0589
...	...	...	...	...	...	...	...
12498.5	-999.0000	-999.000	-9999.0	-999.0	-999.0	-999.0	0.0589
12499.0	-999.0000	-999.000	-9999.0	-999.0	-999.0	-999.0	0.0589
12499.5	-999.0000	-999.000	-9999.0	-999.0	-999.0	-999.0	0.0589
12500.0	-999.0000	-999.000	-9999.0	-999.0	-999.0	-999.0	0.0589
12500.5	-999.0000	-999.000	-9999.0	-999.0	-999.0	-999.0	0.0589

23002 rows × 7 columns

Figure 4.12 Predicted PHIE values for Well 07

7. WELL 08

```
In [61]: Deploy['Predicted_PHIE'] = Baggingpred
         Deploy
```

Out[61]:

DEPTH	CALIPER	Gamma	RES_DEEP	DENSITY	SW	vsh	PERM	Predicted_PHIE
3100.0	-999.0	46.1853	29.3189	-999.0	-999.0	0.1177	-999.0	0.0589
3100.5	-999.0	48.6909	28.4970	-999.0	-999.0	0.1310	-999.0	0.0589
3101.0	-999.0	56.5329	28.5206	-999.0	-999.0	0.1750	-999.0	0.0589
3101.5	-999.0	58.3044	30.0595	-999.0	-999.0	0.1855	-999.0	0.0589
3102.0	-999.0	53.4299	31.4757	-999.0	-999.0	0.1571	-999.0	0.0589
...	...	...	...	...	...	...	...	...
10153.5	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0000	-999.0	0.0589
10154.0	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0000	-999.0	0.0589
10154.5	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0000	-999.0	0.0589
10155.0	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0000	-999.0	0.0589
10155.5	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0000	-999.0	0.0589

14112 rows × 8 columns

Figure 4.13 Predicted PHIE values for Well 08

8. WELL 09

```
In [68]: Deploy['Predicted_PHIE'] = Baggingpred
         Deploy
```

Out[68]:

DEPTH	Gamma	RES_DEEP	DENSITY	SW	vsh	PERM	Predicted_PHIE
0.0	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589
0.5	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589
1.0	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589
1.5	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589
2.0	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589
...	...	...	...	...	...	...	...
10298.5	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589
10299.0	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589
10299.5	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589
10300.0	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589
10300.5	-9999.0	-999.0	-9999.0	-999.0	-999.0	-999.0	0.0589

20602 rows × 7 columns

Figure 4. 14 Predicted PHIE values for Well 09

9. **WELL 10**

```
In [75]: Deploy['Predicted_PHIE'] = Baggingpred
         Deploy
```

Out[75]:

DEPTH	Caliper	GAMMA	RES_DEEP	DENSITY	SW	vsh	PERM	Predicted_PHIE
0.0	-999.0	-999.0000	-999.0000	-999.0	-999.0	-999.0	-999.0	0.0589
0.5	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0	-999.0	0.0589
1.0	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0	-999.0	0.0589
1.5	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0	-999.0	0.0589
2.0	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0	-999.0	0.0589
...	...	...	...	...	...	...	...	...
10678.5	-999.0	70.0555	5.5215	-9999.0	-999.0	-999.0	-999.0	0.0589
10679.0	-999.0	66.3238	5.4413	-9999.0	-999.0	-999.0	-999.0	0.0589
10679.5	-999.0	60.5921	5.3948	-9999.0	-999.0	-999.0	-999.0	0.0589
10680.0	-999.0	53.1499	5.3627	-9999.0	-999.0	-999.0	-999.0	0.0589
10680.5	-999.0	-999.0000	-999.0000	-9999.0	-999.0	-999.0	-999.0	0.0589

21362 rows × 8 columns

Figure 4.15 Predicted PHIE values for Well10

10. **WELL 11**

```
In [82]: Deploy['Predicted_PHIE'] = Baggingpred
         Deploy
```

Out[82]:

DEPTH	CALIPER	GAMMA	RES_DEEP	DENSITY	SW	vsh	PERM	Predicted_PHIE
2460.0	-999.0	16.7456	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
2460.5	-999.0	16.6249	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
2461.0	-999.0	17.0050	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
2461.5	-999.0	18.9987	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
2462.0	-999.0	20.9076	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
...	...	...	...	...	...	...	...	...
8498.5	-999.0	43.5106	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
8499.0	-999.0	43.5834	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
8499.5	-999.0	43.6563	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
8500.0	-999.0	43.6008	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589
8500.5	-999.0	-9999.0000	-999.0	-999.0	-999.0	-999.0	-999.0	0.0589

12082 rows × 8 columns

Figure 4. 16 Predicted PHIE values for Well 11

CHAPTER FIVE

FINDINGS, CONCLUSION AND SUGGESTIONS FOR FURTHER STUDIES

5.1 FINDINGS

In this study, the following findings were made:

1. Porosity is directly related to density and permeability.
2. The ridge regression model had the worst performance, with a mean absolute error of 0.023. The ExtraTree Regression model performed well, with a mean absolute error of 0.003, a root mean squared error of 0.010, and a percentage error of 2.5%. The XGBoost model performed well, with a mean absolute error of 0.004. The bagging model was the most effective, with a mean absolute error of 0.003, a root mean squared error of 0.010, and a mean absolute percentage error of 2.3%.
3. Porosity logs were successfully predicted using the bagging model which can be used to infer the location of reservoir beds in the well.
4. A good fit between the training data and the model does not equal accurate predictions; hence the need to visualize the model's performance using subplots as shown in figs 4, 5, 6 and 7.

5.2 CONCLUSION

In conclusion, Machine learning models has been proven to be an effective means of predicting porosity given that there is sufficient density data provided. However, the performance of the models vary with the quality of the training data. It was discovered that the difference between AIC value and mean absolute error value cannot be used as the only method of model evaluation; hence, the entire error margin, as well as the visualization using subplots must be taken into consideration when evaluating model performance. It should also be noted that, the percentage error of the various models differ slightly; however, the model with the smallest error margin should be used.

5.3 SUGGESTIONS FOR FURTHER STUDIES

The following are the suggestions for further studies;

1. The model should be tested on a larger and more diverse data set to see if the results can be generalized to different datasets.
2. Machine learning should be applied to other aspects of petrophysics such as the prediction of other petrophysical parameters.

REFERENCES

- Akinnikawe, O., Lyne, S., & Roberts, J. (2018). Synthetic Well Log Generation using Machine Learning Techniques. *OnePetro*.
- Akkurt, R., Conroy, T. T., Psaila, D., Paxton, A., Low, J., & Spaans, P. (2018). Accelerating and Enhancing Petrophysical Analysis With Machine Learning: A Case Study of an Automated System for Well Log Outlier Detection and Reconstruction . *OnePetro*.
- Akkurt, R., Miller, M., Hodenfield, B., Pirie, I., Farnan, D., & Koley, M. (2019). Machine Learning For Well Log Normalization . *OnePetro*.
- Amabeoku, S. M. (2014). Core analysis with emphasis on carbonate rocks — quality assurance and control for accuracy and representativeness. *SEG Library*.
- Badure S. (2021). Petrophysics and Petrophysicist. *Journal of Petroleum & Environmental Biotechnology, Vol. 12 Iss. 8 No: 434*, 1.
- Banas, R., McDonald, A., & Perkins, T. J. (2021). Novel Methodology for Automation of Bad Well Log Data Identification and Repair . *OnePetro*.
- Ben-David, S. S.-S. (2014). *Understanding Machine Learning: From Theory To Algorithms.* New York City: Cambridge University Press.
- Bittar, M., Wang, S., Chen, J., & Wu, X. (2020). Reinforced learning technique for multi-well logs depth matching yield better reservoir delineation. *OnePetro*.
- Chicheng Xu, Lei Fu, Tao Lin, Weichang Li, Shouxiang Ma . (2022). Machine Learning in Petrophysics: Advantages and Limitations. *Artificial Intelligence in Geosciences 3*, 1.
- Ghoulem Ellah haithem Ifrene, D. I. (2023). New Insights in Fracture Porosity Estimation using Machine Learning and Advanced Logging Tools. *ResearchGate*.
- Ideozu, D. R. (2016). Map of the Niger Delta. *Sedimentological control on permeability anisotrophy and heterogeneity in shore face reservoir, Niger Delta*.
- Le, T., Liang, L., Zimmermann, T., Zeroug, S., & Heliot, D. (2019). A machine learning framework for automated well log depth matching. *OnePetro*.
- Licheng Zhang, C. Z. (2017). Machine Learning in Rock Facies Classifications: An Application of XGBOOST. *ResearchGate*.
- *Machine Learning Theory*. (2023, July 22). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Machine_learning#Theory
- McDonald, A. (2020). Petrophysics: Gamma Ray Normalization in Python. *TowardsDataScience*.

- McDonald, A. (2020). Porosity-Permeability relationships using linear regression in python. *TowardsDataScience*.
- McDonald, A. (2020). Visualizing Well Data Coverage Using Matplotib. *TowardsDataScience*.
- McDonald, A. (2021). How To Use Unsupervised Clustering On Well Log Data With Python. *TowardsDataScience*.
- Mehryar Mohri, A. R. (2018). *Foundations of Machine Learning: Adaptive Computation and Machine Learning series.* Massachusetts: MIT Press, 2018.
- Mesini, P. M. (2020). PETROPHYSICS AND RESERVOIR CHARACTERISTICS. *PETROLEUM ENGINEERING – UPSTREAM*.
- Mohammad Khan, Z. T. (2021). Synthetic Well Log Generation using Machine Learning Approaches. *ResearchGate*.
- O.C. Ekwenye, G. N. (2014). A paleogeographic model for the sandstone members of the Imo Shale, south-eastern Nigeria. *ScienceDirect*.
- Pål Østebø Andersen, C. A. (2022). Machine Learning Based Prediction of Porosity and Water Saturation from Varg Field Reservoir Well Logs. *SPE EuropEC - Europe Energy Conference featured at the 83rd EAGE Annual Conference & Exhibition.* Madrid, Spain: Society of Petroleum Engineers.
- Roberto Ruiz, A. R. (2021). Data mining and machine learning for porosity, saturation, and shear velocity prediction: recent experience and results. *First Break*.
- Ryan Banas, A. M. (2021). Novel Methodology for Automation of Bad Well Log Data Identification and Repair. *ResearchGate*.
- Sen, J. (2021). *Machine Learning - Algorithms, Models and Applications.* IntechOpen.
- Shi, L., & Zhang, J. (2021). Prediction of shear wave velocity using machine learning technique, multiple regression and well logs. *OnePetro*.
- Siddharth Misra, O. J. (2020). Unsupervised Outlier Detection Techniques for Well Logs and Geophysical Data. *Research Gate*.
- Singh, M., Makarychev, G., Mustapha, H., Voleti, D., Akkurt, R., Daghar, K. A., . . . Razouki, A. (2020). Machine Learning Assisted Petrophysical Logs Quality Control, Editing and Reconstruction . *OnePetro*.
- Thanh, T. T. (2023). Estimate Petrophysical Properties by Using Machine Learning Methods. *ResearchGate*.

- Tufail S, R. H. (2023). *Tufail S, RiggAdvancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. .* Retrieved from mpdi.com: https://www.mdpi.com/2079-9292/12/8/1789
- Wakeel Hussain, M. L. (2023). Machine learning - a novel approach to predict the porosity curve using geophysical logs data: An example from the Lower Goru sand reservoir in the Southern Indus Basin, Pakistan. *Journal of Applied Geophysics*.