KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY,

KUMASI - GHANA.

COLLEGE OF ENGINEERING

DEPARTMENT OF PETROLEUM ENGINEERING

A COMPARATIVE STUDY OF ANN, RF, DT, SVR AND GB ALGORITHMS WITH EMPIRICAL CORRELATIONS FOR PREDICTING OIL FORMATION VOLUME FACTOR.

PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE BSC. (ENG) DEGREE

By:

DABLAH URIEL SELIKEM

EWUSI EMMANUEL

AMOH SARPONG SAMUEL

ABELIMWINI SALOMEY

INCOOM ALBERT

EKORBOR AFIBA BADWO STEPHANIE

BAABA - YAKUUB ABDUL - MUUMIN YAKUUB YEMBONI

SUPERVISOR: CASPAR DANIEL ADENUTSI, PHD

SEPTEMBER, 2024

DECLARATION

We hereby declare that the project work entitled "A Comparative Study Of ANN, RF, DT, SVR And G-Boost Algorithms With Empirical Correlations For Predicting Oil Formation Volume Factor" submitted to the Department of Petroleum Engineering – Kwame Nkrumah University of Science and Technology is a record of an original work done by us under the supervision of Dr. Caspar Daniel Adenutsi, a lecturer of the Petroleum Programme, College of Engineering and this project work is submitted in partial fulfilment of the requirement for the award of Bachelor of Science Degree in Petroleum Engineering. The results embodied in this project report have not been submitted to any other university or institute for the award of any degree or diploma.

DABLAH URIEL SELIKEM ……..……

EWUSI EMMANUEL ……...

AMOH SARPONG SAMUEL ……...……

ABELIMWINI SALOMEY …….……..…….

INCOOM ALBERT …….

EKORBOR AFIBA BADWO STEPHANIE …….

BAABA - YAKUUB ABDUL - MUUMIN YAKUUB YEMBONI …….

SUPERVISOR'S DECLARATION

I certify that this project has been successfully undertaken and submitted under my supervision.

Date: 2nd September 2022.

Caspar Daniel Adenutsi, PhD ……

(Supervisor) (Signature)

DEDICATION

This thesis is first and foremost dedicated to the Almighty God, and to our supervisor, Dr. Caspar Daniel Adenutsi; whose expertise, guidance and patience have been invaluable . We also dedicate the work to our lovely family.

ACKNOWLEDGEMENT

We would like to express our gratitude to the Almighty God for his blessings during the study, which enabled us to successfully achieve our aims and objectives. We would like to sincerely thank our research supervisor, Dr. Caspar Daniel Adenutsi, for giving us the chance to carry out this study. Your mentorship and expertise have been the driving force behind the success of this project.

We also dedicate the work to our lovely families for their unwavering support and encouragements.

ABSTRACT

Currently, one of the most difficult challenges the oil and gas industry is faced with is accurately predicting oil formation volume factor, Bo. With applications ranging from data analysis to pattern recognition and target prediction, machine learning techniques have emerged as a promising tool to address this challenge. In contrast to conventional empirical correlations, the current study investigates the use of machine learning models, including Gradient Boosting, ANN, Support Vector Regression, Random Forest, and Decision Trees, in predicting this crucial parameter for oil formation volume factor. The results have demonstrated that these models are relatively efficient in the accurate prediction of the Bo, with Artificial Neural Network (ANN) model achieving the best performance than the rest of the techniques in minimizing error, with an APD% of 2.50, AAPD% of 4.20, and an R^2 value of 0.9850. ANN also had the lowest MSE (0.0011) and RMSE (0.0333).

The findings from this study provide valuable insight to industry practitioners in making more informed decisions in managing and developing hydrocarbon resources.

This is because Machine learning (ML) models have proven to predict Bo more accurately than traditional methods.

TABLE OF CONTENTS

DECLARATION I
SUPERVISOR'S DECLARATION II
DEDICATION III
ACKNOWLEDGEMENT IV
ABSTRACT V
TABLES IX
CHAPTER 1 1
INTRODUCTION 1
1.1 Background of Bo prediction 1
1.2 Problem statement 2
1.3 Justification 3
1.4 Aim and Objectives 3
1.5 LIMITATIONS 4
CHAPTER 2 5
LITERATURE REVIEW 5
2.1 EMPIRICAL CORRELATIONS FOR OIL FORMATION VOLUME FACTOR (Bo) PREDICTION 5
2.2 MACHINE LEARNING (ML) MODELS BEING USED FOR (Bo) PREDICTION. 15
2.3 ARCHITECTURE OF MODELS IMPLEMENTED. 16
2.3.1 Gradient Boosting 17
2.3.2 Support Vector Regression (SVR) 19
2.3.2 Random Forest 21
2.3.4 Artificial Neural Networks (ANN) 22
2.3.4 Decision Tree 24
CHAPTER 3 30
METHODOLOGY 30
3.1 INTRODUCTION TO THE WORKFLOW 30
3.1.0 Data Collection and Processing of Dataset 32
3.1.1 Data Characterization 32

3.1.2 Feature Scaling 34
3.1.3 Feature Selection 34
3.1.4 Pair plot of the data set 34
3.1.5 Pearson Correlation Analysis 37
3.1.6 Spearman Correlation Analysis 38
3.1.7 Analysis of Independent Variables 38
3.1.8 Ensuring Robust Model Performance through K-Fold Cross-Validation 40
3.1.9 Gradient Boosting 41
3.1.10 Support Vector Regression (SVR) 43
3.1.11 Random Forest 43
..... 44
3.1.12 Artificial Neural Network (ANN) 45
CHAPTER 4 47
RESULTS AND DISCUSSION 47
4.1 Results 47
4.1.1 Metrics Used to Assess the Models' Predictions 47
4.2 Discussion 49
4.2.1 Discussion of Empirical Correlations Results 49
4.2.2 Discussion of Machine Learning Models Results 51
CHAPTER 5 53
CONCLUSION AND RECOMMENDATION 53
5.1 CONCLUSION 53
5.2 RECOMMENDATIONS 54
REFERENCE 55

LIST OF FIGURES

Figure 1 Flowchart of Gradient Boosting Model Used 19
Figure 2 Flowchart of SVR Model Used. 21
Figure 3 . Biologic Nerve Cell (Neuron) 23
Figure 4 . Network with Hidden Layers and Output Layer 24
Figure 5 Workflow sequence used to contrast empirical and machine learning models 31
Figure 6 K-Fold Cross Validation 31
Figure 7 Pair plot of dataset 36
Figure 8 Pearson's Coefficient of Correlation Heatmap 37
Figure 9 Spearman Coefficients of Correlation Heatmap 38
Figure 10 Flowchart of Random Forest Model 44
Figure 11 Performance of Empirical Models 49
Figure 12 Performance of Machine Models 52

TABLES

Table 1 Empirical correlations used to predict oil formation volume factor (Bo) 9
Table 2 Bo published research on prediction using machine learning techniques. 25
Table 3 Statistical characterizations of the dataset 33
Table 4 Normalized Training Dataset 41
Table 5 Normalized Testing Dataset 41
Table 6 Model Evaluation of Empirical Models 50
Table 7 Model Evaluation of Machine learning models 52

LIST OF EQUATIONS

Eq. (1)..... 34
Eq. (**2**) 48
Eq. (**3**) 48
Eq. (**4**) 48
Eq. (**5**) 48
Eq. (**6**) 48

CHAPTER 1

INTRODUCTION

1.1 Background of Bo prediction

Within the oil and gas sector, the oil volume (including gas in solution) at reservoir conditions divided by oil volume at typical surface conditions is known as the Oil Formation Volume Factor (Bo). The discharge of dissolved gas causes the reservoir's volume to decline when oil is extracted.

Usually, it is larger than or equal to one. The computation of important petroleum engineering parameters, including in situ, reservoir simulation, well testing, and reservoir production calculation, heavily depends on this feature. Petroleum engineers typically depend on correlations that have been empirically formed, even though the majority of PVT data are generated in laboratories. This is brought on by the cost of obtaining a PVT analysis in addition to unavailable and faulty samples. Real-time decision-making requires quicker and less expensive techniques, and this quality is predicted using correlations that have been empirically developed. One of the key challenges in predicting (Bo) is the complex nature of reservoir fluids and their behavior under various pressures and temperatures. Traditionally, correlations and empirical models have been used to estimate Bo based on available data and reservoir characteristics.

Despite the existence of empirical correlations that serve as traditional tools for estimating the Bo, their limitations have prompted the exploration of advanced predictive methodologies.

This project delves into the historical evolution of these empirical correlations, their constraints, and the transformative potential of machine learning (ML) techniques as innovative solutions for enhancing the prediction of oil.

However, with the improvements in machine learning, there is a chance to increase the accuracy of these predictions by applying algorithms that can identify complex patterns and correlations in the data. Machine learning approaches, regression and artificial neural networks, have demonstrated promising results in estimating oil formation volume parameters. Numerous factors, such as reservoir pressure, temperature, fluid composition, and other pertinent variables, can be considered by these techniques.

1.2 Problem statement

Current methods for predicting (Bo), a vital property in oil production, rely on correlations derived from experiments. These correlations often have limitations. They might not capture the intricacies of real reservoirs, are restricted by specific data ranges, and don't account for unique reservoir characteristics. Additionally, obtaining the necessary data for these correlations can be difficult, especially in mature or unconventional oil fields. Because of these shortcomings, this research project investigates how machine learning approaches may be used to increase the precision and dependability of Bo forecasts.

1.3 Justification

Machine learning improves overall forecast accuracy by lowering bias and variation when compared to current empirical approaches.

With the predictions reducing the impact of individual model failures, the machine learning models can withstand outliers and noise in the data more effectively. By using many viewpoints and identifying a variety of patterns and correlations within the data, machine learning techniques improve their capacity for generalization while also lowering the possibility of model constraints.

1.4 Aim and Objectives

Aims

The aim is to create a novel and trustworthy machine learning model that will predict (Bo), providing an upgrade over the generally accepted traditional approaches currently in use.

Objectives

To achieve this aim, the following specific objectives will be addressed:

- Identify and extract key features (independent variables) from readily available reservoir fluid properties that significantly influence Bo. Dimensionality reduction techniques may be employed to optimize model performance and interpretability.
- Develop a robust Bo prediction model using machine learning algorithms. This involves training and evaluating various algorithms to select the one that achieves the highest accuracy and generalizability for predicting Bo.
- Utilizing pre-established metrics, perform a thorough assessment of the produced machine learning model's performance. This evaluation will compare its predictive accuracy

and efficiency against existing widely used conventional methods, such as empirical correlations.

- Analyze and compare the performance of the best-performing machine learning model with established empirical correlations for Bo prediction. This analysis will highlight the strengths and limitations of each approach**.**

1.5 LIMITATIONS

Although machine learning (ML) models have several drawbacks, they can greatly increase the accuracy of Formation Volume Factor (Bo) estimates. A major issue in ML is overfitting, where a model learns too much from the noise and unnecessary details in the training data, affecting its performance on new data. This is especially relevant to Bo prediction due to the diverse and complex geological and reservoir characteristics when used to forecast distinct reservoirs, an overfitted model may perform well on its training data but not generalize to other, unknown datasets, producing inaccurate results.

Furthermore, the effectiveness of ML models heavily depends on the quantity and quality of the data used in training. For Bo prediction, obtaining a large and accurate dataset that reflects subsurface conditions accurately is challenging. Data that is incomplete or biased can result in models that do not fully capture the complexity of reservoir behaviours, thus reducing prediction accuracy.

Selecting the right input features is crucial for ML model performance. Features that are not precise or relevant can lead to poor predictions. In the context of Bo prediction, it is essential to identify the key geological and fluid properties that affect the Bo.

CHAPTER 2

LITERATURE REVIEW

2.1 EMPIRICAL CORRELATIONS FOR OIL FORMATION VOLUME FACTOR (Bo) PREDICTION.

Predicting the Bo is essential in the area of petroleum engineering for several applications, including production optimization, reservoir characterization, and figuring out an oil field's recoverable reserves(Honarpour et al., 2006).

Over time, several empirical correlations have been constructed to determine the volume factor of oil formation. The relations between the formation volume factor and other reservoir fluid parameters are based on the examination of experimental data and empirical relationships(Salim Basaleh, n.d.) such as oil composition, temperature, and other parameters.

The earliest mathematical model for Bo estimation was developed by Katz in 1942(Ahmed, 2018; Standing & Katz, 1942). Katz's study was based on 16 saturated hydrocarbon samples at certain pressure and temperature ranges. Petrophysical parameters were employed in the graphical correlation. The correlation was only documented graphically, and because it included graphs and mathematics, it was challenging to apply.

Standing further refined Katz's correlation by incorporating additional variables such as oil gravity and reservoir temperature into the equation. To create his graph, the association was based on tests conducted on samples from 105 experimental data sets that were gathered

from various Californian oil fields (Standing, 1947). Standing introduced an empirical correlation that included oil-specific gravity and reservoir temperature as variables in the estimation of the Bo. Standing's correlation, developed in 1947, became widely used and is still utilized today.

Another notable correlation was developed by Glaso in 1980 based on Standing's work (Glaso, 1980). Glaso's correlation, 1980 published a reservoir fluid properties correlation for hydrocarbon mixes in the North Sea. It was based, however somewhat altered, on Standing's model. The way Glaso's model corrects the oil gravity in correlation corrected value based on reservoir temperature and oil viscosity is a key component that helps explain the paraffinic effect. Correction variables for the occurrence of non-hydrocarbons are used to adjust the correlation also takes into account the influence of nonhydrocarbons on saturation pressure. He utilized 45 different oil samples in all, the majority of which were from the North Sea oil fields and ranged from volatile oils to heavy oils and tars.

Additionally, Vasquez and Beggs in 1977 developed correlations based on data obtained from fields worldwide by applying the concept of regression analysis on 6000 data sets collected from fields worldwide, generally applicable for all oil types and covering a wide range of pressures, temperatures, and oil properties (Vazquez & Beggs, 1977). In their correlation, Vazquez and Beggs added reservoir pressure and isothermal oil compressibility to the input parameters

Al-Marhoun (1988) developed a mathematical relation for Bo at bubble point pressure via nonlinear multiple regression analysis for 160 experimental data sets gathered from 69

Middle Eastern reservoirs. The correlation was shown to depend on gas-specific gravity, gas solubility, oil-specific gravity, and reservoir conditions. The Al-Marhoun correlation could not offer reliable predictions beyond the conditions for which it was established. The correlation could not accurately predict oil formation volume factors under extreme conditions, such as very high or very low temperatures and pressures, and in reservoirs with highly viscous oils or gas condensates.

Dokla & Osman (1992) employed 51 data sets from the UAE crudes to provide an additional correlation for the prediction of BO. Once more, they built his connection using the Al-Marhoun (1988a) correlation as a foundation.

Omar & Todd (1993) used both linear and non-linear regression analysis (Omar & Todd, 1993) to develope a PVT correlation for Malaysian crude oil based on 93 data sets gathered from different reservoirs in the country. Since Standing's (1947) study served as the foundation for their correlation, identical input parameters were applied. With a R^2 of 99% and an APD of 1.44%, the correlation created to forecast Bo at the bubble point pressure is said to have significantly improved accuracy.

Petrosky and Farshad proposed correlations using samples from Gulf of Mexico oils, Petrosky & Farshad (1993). The Standing's Bo correlation is the foundation of this correlation. Nonlinear multiple regression analysis was utilized in order to get the best outcomes.

Almehaideb (1997) suggested a novel correlation to forecast the PVT characteristics of UAE crude oil. For his investigation, 62 data samples were gathered from several oil fields

in the United Arab Emirates. Regression analysis was then carried out to create correlations that might forecast the PVT characteristics of crude oil in the United Arab Emirates. The development of Bo correlation requires three input parameters: reservoir temperature, gas-to-oil ratio, and oil-specific gravity. The correlation value is 0.9985, and the stated average absolute percentage error for calculating Bo at the bubble point pressure is 1.35%.

Al-Shammasi (1999, 2001) offered a new correlation. This analysis was done with 1709 global data sets that were gathered from 13 published journals. In his initial connection, he used four metrics: gas solubility, temperature of the reservoir, gas-specific gravity, and GOR. The gas-specific gravity is the second association he found. For the first and second equations, the reported average absolute percentage errors are 1.806% and 3.033%, respectively.

These correlations have been continuously refined and expanded to accommodate diverse reservoir characteristics and fluid properties, contributing to advancements in petroleum engineering practices. However, it is essential to acknowledge the ongoing development of correlations and the potential of emerging technologies.

While empirical correlations have been widely used in the prediction of Bo, they have some limitations.

One limitation is based on regional data which may not accurately represent the characteristics of different oil reservoirs. For example, the correlations developed by Katz and Standing were based on specific regions such as California, and may not apply to oil reservoirs in other areas. Likewise, Glaso's correlation was developed for North Sea crudes and may not be suitable for reservoirs in other regions.

Another limitation of empirical correlations is their reliance on simplified assumptions and linear relationships between variables. These assumptions may not hold for all oil reservoirs, particularly those with complex fluid systems or non-linear relationships between variables. In addition, traditional correlations are based on a limited range of experimental data and may not capture the full complexity of all oil reservoir behavior

Table 1 Empirical correlations used to predict oil formation volume factor (Bo)

Author	**Origin**	**Dataset**	**Applicable Range**	**Limitations**
Standing & Katz (1942)	California	105	Bo:1.024-2.15(bbl/stb) T: 100 - 258 (F˚) Rs:201-425(scf/stb) API: 16.5 - 63.8 γg: 0.9 - 0.955	- Ignores non-hydrocarbon components
Vazquez & Beggs (1977b)	Worldwide	5008	Bo:1.028-2.226(bbl/stb) T: 75 - 294 (F˚) Rs: 0 - 2199 (scf/stb) API: 15.3 - 59.3 γg : 0.65 - 1.28 Pb: 15 - 6055 (psia)	- May not represent certain oil types well

Author	Origin	Dataset	Applicable Range	Limitations
Glaso (1980b)	North Sea	45	Bo:1.032-2.588(bbl/stb) T: 80 - 280 (F˚) Rs: 90 - 2637 (scf/stb) API: 22.3 - 48.1 γg : 0.65 - 1.276 Pb:165 - 7142 (psia)	- Limited applicability, struggles with generalization
Al-Marhoun (1988)	Middle East	160	Bo:1.032-1.997(bbl/stb) T: 74 - 240 (F˚) Rs: 26 - 1602 (scf/stb) API: 19.4 - 44.6 γg:0.75 - 1.673 Pb:130 - 357 (psia)	- Inaccurate for oils unlike Middle Eastern crudes
Kartoatmodjo & Schmidt (1994)	Worldwide	5392	Bo:1.007-2.747(bbl/stb) T:75 - 320 (F˚) Rs: 0 - 2890 (scf/stb) API: 14.4 - 59 γg: 0.4824 - 1.1.668 Pb:24.7 - 4746.7(psia)	- Potential bias towards specific regions/oil types

Author	Origin	Dataset	Applicable Range	Limitations
Dokla & Osman (1992b)	UAE	51	Bo:1.216-2.493(bbl/stb) T: 190 - 275 (F˚) Rs: 181 - 2266 (scf/stb) API:28.2 - 40.3 γg: 0.798 - 1.29 Pb: 590 - 4640 (psia)	- Limited capture of crude oil property diversity (even within UAE)
Macary & El-Batanoney (1993)	Gulf of Suez	90	Bo: 1.2 - 2 (bbl/stb) T: 130 - 290 (F˚) Rs: 200 - 1200 (scf/stb) API: 25 - 40 γg: 0.7 - 1 Pb: 1200 - 4600 (psia)	- Application to other regions with significantly different compositions could result in less accurate predictions of

Author	Origin	Dataset	Applicable Range	Limitations
Petrosky Jr. & Farshad (1998b)	Gulf of Mexico (Texas, Louisiana)	81	Bo:1.118-1 .623(bbl/stb) T: 114 - 288 (F˚) Rs: 217 - 1406 (scf/stb) API: 16.3 - 45.0 γg: 0.58 - 0.85 Pb: 1574 - 6528(psia)	- Significant variations within Gulf of Mexico crudes and other locations might not be fully captured.
Omar & Todd (1993b)	Malaysia	93	Bo:1.085-1.954(bbl/stb) T: 125 - 280 (F˚) Rs: 142 - 1440 (scf/stb) API: 26.6 - 53.2 γg: 0.612 - 1.32 Pb: 790 - 3851 (psia)	- Limited applicability, might not be accurate for crudes with significantly different compositions.
Hanafy et al. (1997)	Egypt	324	Bo: 1.032- 4.35(bbl/stb) T: 107 - 327 (F˚) Rs: 7 - 4272 (scf/stb) API: 17.8 - 48.8 γg: 0.623 - 1.627 Pb: 36 - 5003 (psia)	- It might not fully capture the diversity of crude oil systems globally.

Author	Origin	Dataset	Applicable Range	Limitations
Almehaideb (1997b)	UAE	62	Bo:1.14 -3.562(bbl/stb) T: 190 - 306 (F°) API: 30.9 - 48.6 Rs: 128 - 3871 (scf/stb) γg: 0.746 - 1.116 Pb: 501 - 4822 (psia)	- Restricted applicability to a wider variety of UAE crudes with diversity in oil compositions.
Hemmati & Kharrat (2007)	Iran	287	Bo: 1.091 - 2.54(bbl/stb) T: 77.5 - 290 (F°) Rs: 125 - 2189.25 (scf/stb) API: 18.8 - 48.34 γg : 0.523 - 1.415 Pb: 348 - 5156 (psia)	- It might be more accurate for Iranian oil than others because of compositional similarities.

Author	Origin	Dataset	Applicable Range	Limitations
Al-Shammasi (1999, 2001)	Worldwide	1243	Bo: 1.02 - 2.916(bbl/stb) T: 74 - 341.6 (F˚) Rs: 6 - 32.98.6 (scf/stb) API: 6 - 63.7 γg: 0.51 - 3.44 Pb: 31.70 - 7127 (psia)	Not all-encompassing.
Elsharkawy & Alikhan (1997)	Middle East	254	Bo: 1.057 - 1.770 (bbl/stb) T:130 - 250 (°F) Rs: 34 - 1400 (scf/stb) API: 20 - 45 γg: 0.663 - 1.064 Pb: 317 - 4375 (psia)	When employed for oils from other geographical areas with significantly different geochemical characteristics, the predictions might hold less precision

2.2 MACHINE LEARNING (ML) MODELS BEING USED FOR (Bo) PREDICTION.

To overcome these limitations, researchers have turned to machine learning techniques. Machine learning techniques offer the potential to overcome the limitations of traditional correlations in predicting (Bo). These machine learning models can capture non-linear relationships and handle complex data sets, allowing for more reliable estimations of oil formation volume factor.

Moreover, machine learning models can be trained on a larger and more diverse dataset, encompassing various reservoir characteristics and fluid properties. This provides for a more thorough knowledge of the factors influencing Bo and increases the accuracy of forecasts. Overall, while empirical correlations have been widely used in the prediction of oil formation volume factor, they have their limitations.

In recent years, the use of machine learning techniques for predicting Bo in the oil and gas industry has gained significant attention. Various studies have explored the application of machine learning models such as the transparent open Box (TOB), ANNs, SVM, genetic programming, and random forest in predicting the formation volume factor of oil. Rashidi et al (2021), Saghafi et al (2019), Wood & Choubineh (2019) and Gouda & Attia (2024) conducted a comprehensive study on the use of artificial ANNs for predicting formation volume factor based on a combination of input variables such as reservoir pressure, temperature, oil gravity, and gas-oil ratio. Their research demonstrated the effectiveness of ANNs in capturing complex non-linear relationships between reservoir conditions and oil volume, leading to accurate predictions.

Similarly, Rashidi et al (2021) investigated the application of support vector machines for formation volume factor prediction. Their study showcased the capability of SVM in handling high-dimensional data and identifying patterns that contribute to accurate estimation of oil volume factors, E. A. El-Sebakhy (2009).

Karimnezhad et al (2014)) in his paper, suggested a novel connection to forecast the Bo for crudes from the Middle East. There were 429 distinct crude oil data sets from Middle Eastern reservoirs that were utilized. Of these, 286 data sets were chosen at random to serve as training data and test data to build the correlation and validate the correlation, respectively. These results are more accurate for Middle East crudes than any previous empirical correlations at the time and demonstrate strong agreement with data.

These studies collectively underscore the potential of machine learning approaches in predicting formation volume factor, offering promising avenues for improving the efficiency and accuracy of oil extraction processes.

2.3 ARCHITECTURE OF MODELS IMPLEMENTED.

Traditional methods, relying on empirical correlations and laboratory measurements, fall short in accuracy and efficiency (Standing, 1947). By comparing Support Vector Regression (SVR), Random Forest, Gradient Boosting, Decision Trees, and Artificial Neural Network (ANN) we aim to identify the most effective model for predicting oil formation volume factor (OFVF).

2.3.1 Gradient Boosting

Gradient Boosting is a powerful machine learning model that has received a lot of attention lately because of its outstanding performance, across a wide range of applications, from predictive modeling to decision-making tasks. This algorithm is particularly adept at handling heterogeneous features, noisy data, and complex dependencies, making it a popular choice for challenges such as web search, recommendation systems, and weather forecasting.

Fundamentally, by combining the advantages of several weak models, the Gradient Boosting method is a method of collective learning that creates a powerful predictive model. Iteratively training a series of weak models to fix the mistakes produced by the prior model is how the method operates. This process of sequential model building, known as boosting, allows the algorithm to gradually improve its performance and achieve state-of-the-art results.

Errors are an essential part of every machine-learning system. Bias error and variance error are the two basic categories of error. We can reduce the bias error of the model by using the gradient boosting approach. This algorithm's primary concept is to develop models successively, with each new model aiming to fix the flaws in the preceding one.

Gradient Boosting is used in regression issues when our target variable is continuous, as in the case of Bo prediction. The goal of adding weak learners to reduce the loss function as Mean Squared Error (MSE), use gradient descent. This procedure includes:

- Basic Model: To get early forecasts, start with a basic model.
- Determine Residuals: Determine the differences (residuals) between the expected and actual numbers.

- Fit New Model: To forecast these residuals, create a new model.
- Update Predictions: To obtain updated predictions, add the new model's predictions to the earlier ones.
- Repeat: Keep going until the residuals are as little as possible, or for a predetermined number of iterations.

Gradient Boosting efficiently lowers the total error by repeatedly concentrating on the errors of the prior models, improving predictive accuracy.

The main ideas of gradient boosting are explained in the context of regression, with an emphasis on reducing mistakes and gradually enhancing the model's performance.

The Gradient Boosting algorithm's adaptability and customization are two of its main benefits. The optimization aim of the technique may be modified to suit particular applications by modifying the loss function. This allows the algorithm to be optimized for different types of problems, such as classification, regression, or ranking tasks.

The mathematical framework of the Gradient Boosting algorithm can be broken down into several core components.

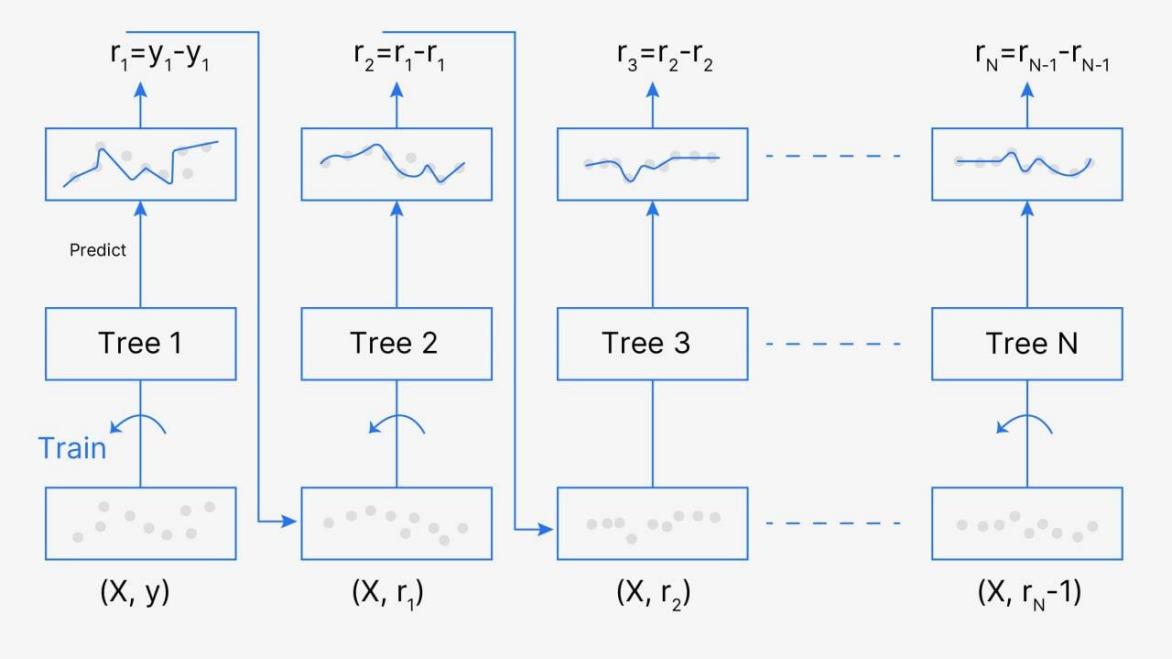

Figure 1 Flowchart of Gradient Boosting Model Used

2.3.2 Support Vector Regression (SVR)

SVR a type of SVMs specifically designed for predicting continuous values, offers a powerful learning approach for predicting OFVF. This allows reservoir engineers to optimize oil production and reservoir management strategies.

Unlike SVMs used for classification, which identify decision boundaries to separate data points into distinct classes, SVR focuses on finding a function that best fits the training data for continuous variables like OFVF. This function aims to reduce the prediction error between the actual and estimated Bo (Vapnik, 1995).

A key strength of SVR for OFVF prediction lies in its ability to handle high-dimensional data. Reservoir properties influencing OFVF can be numerous, leading to datasets with many input variables like pressure, temperature, gas-specific gravity, oil-specific gravity, gas

solubility and API gravity. SVR excels at handling these complex scenarios, similar to SVMs (Vapnik, 1995).

Another significant advantage of SVR is its capability to model non-linear relationships. The relationship between various reservoir properties and OFVF can be intricate and non-linear. Similar to SVMs, SVR employs kernel functions to transform the data into a higher-dimensional feature space. This allows the model to capture these non-linear relationships, potentially leading to more accurate OFVF predictions compared to algorithms limited to linear models (Schölkopf & Smola, 2002).

Furthermore, SVR can perform well even with limited training data, a common challenge when collecting large datasets for specific oil reservoirs. Additionally, SVR shares the memory efficiency of SVMs, which becomes beneficial when dealing with vast amounts of data from multiple wells or formations (Vapnik, 1995).

Beyond these core advantages, SVR offers additional benefits for OFVF prediction. Like SVMs, SVR exhibits robustness to outliers, which can be present in reservoir data due to measurement errors or specific reservoir characteristics (Cortes & Vapnik, 1995). While not as interpretable as simpler models, SVR can still provide some insights into the relationship between reservoir properties and OFVF through analysis of the support vectors (Guyon & Vapnik, 2002). The figure below illustrates the flowchart of the SVR model used in predicting OFVF.

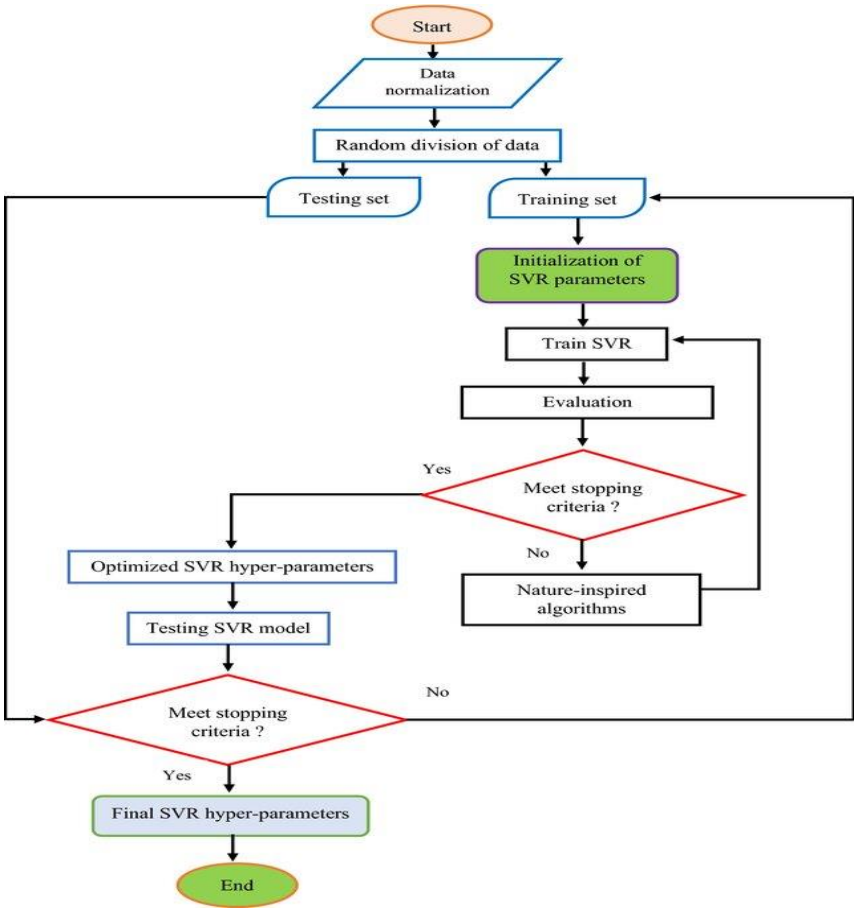

Figure 2 Flowchart of SVR Model Used.

2.3.2 Random Forest

Random Forest, a robust ensemble learning technique, emerges as a promising tool for predicting Oil Formation Volume Factor (OFVF). This methodology constructs multiple decision trees and aggregates their predictions to enhance predictive accuracy while mitigating overfitting (Breiman, 2001). Each tree is independently trained on a randomly selected subset of the data (bootstrap sampling) and a subset of features (feature bagging) (Hastie, Tibshirani, & Friedman, 2009). The final prediction is derived by averaging the outputs of all trees in the forest.

The strength of Random Forest lies in its ability to handle complex datasets characterized by high dimensionality, non-linear relationships, and missing values (Liaw and Wiener, 2002). By merging the forecasts from many different trees, the algorithm effectively reduces the risk of overfitting, a common challenge in predictive modeling. Moreover, Random Forest offers a balance between bias and variance, resulting in models that generalize well to unseen data.

Random Forest can successfully handle the difficulties brought on by the complexity and variety of reservoir data in the context of Bo prediction. The algorithm may create prediction models with higher accuracy and reliability than conventional techniques by combining several variables, including reservoir parameters, fluid composition, temperature, pressure, and composition.

2.3.4 Artificial Neural Networks (ANN)

ANNs learn through an algorithm known as backpropagation that modifies the weights of connections between neurons (Fig. 5) to minimize prediction errors (Rumelhart, Hinton, & Williams, 1986). This iterative process allows the network to gradually improve its predictive accuracy. In the context of OFVF, ANNs can learn to identify intricate relationships between various input parameters (pressure, temperature, gas specific gravity, oil specific gravity, gas solubility and API gravity) and the corresponding OFVF values.

The architecture of an ANN is a critical determinant of its performance. It involves specifying the number of hidden layers, neurons per layer, and the type of activation functions. Multi-layer perceptrons (MLPs) are commonly employed for OFVF prediction.

These networks are made up of an output layer, an input layer, and one or more hidden layers.(Fig. 6). The hidden layers extract relevant features from the input data, while the output layer produces the predicted Bo values.

ANNs may learn intricate patterns thanks to the nonlinearity that activation functions provide to the system. Tanh, sigmoid, and Rectified Linear Units (ReLU) are examples of common activation functions. The performance of the network can be greatly affected by the activation function selection. ReLU's ability to lessen the effects of the vanishing gradient problem and its processing efficiency have made it more and more popular. However, depending on the particulars of the OFVF data, different activation functions could be more appropriate.

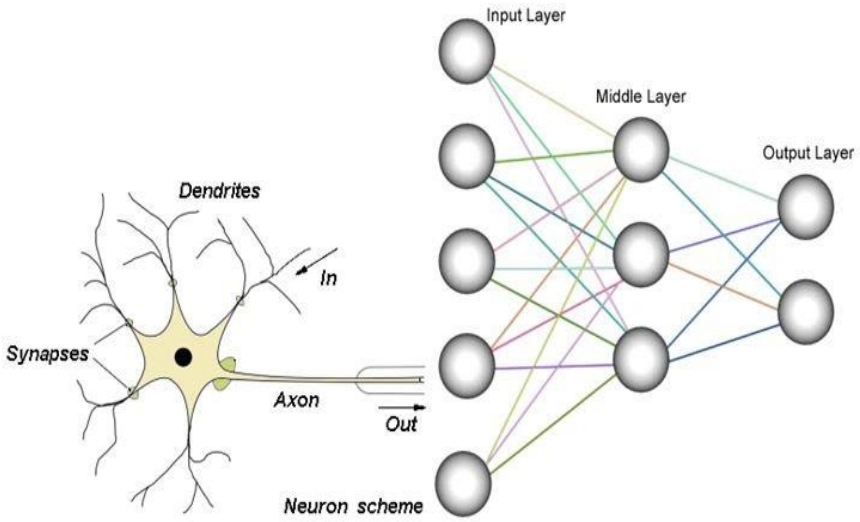

Figure 3 . Biologic Nerve Cell (Neuron)

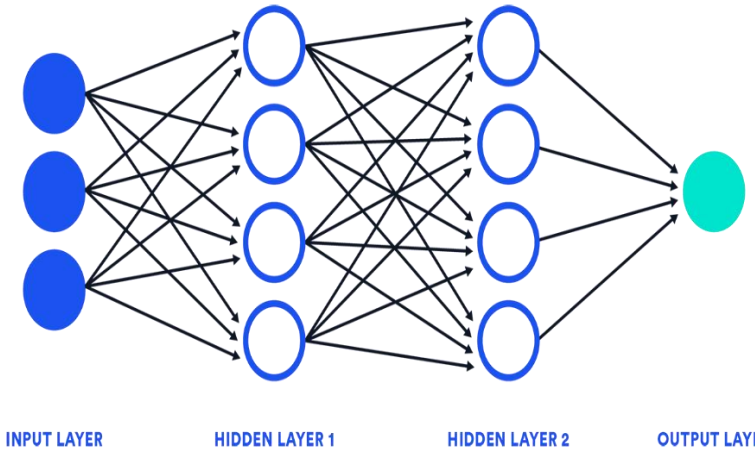

Figure 4 . Network with Hidden Layers and Output Layer

2.3.4 Decision Tree

Decision trees divide data according to feature properties using a hierarchical framework. The tree is made up of nodes, whose branches indicate the potential results of the test and each internal node represents a test on an attribute. Leaf nodes at the terminals of the tree represent projected values or class labels. Recursive partitioning creates the tree by having an algorithm choose the best feature to divide the data into segments that maximize information gain and reduce impurity. To classify or predict a new data instance, Beginning at the root node, one must travel the tree. and moving through the branches that correspond to the attribute values of the instance until they reach a leaf node, which provides the predicted value or assigned class. (Stone, 1984).

Table 2 Bo published research on prediction using machine learning techniques.

Author	Formula Displayed	Origin	Data Number	Method	Errors
Gharbi et al (1999)	No	World Wide	5200	ANN	R^2 = 0.9875 RMSC = - SD = 2.43 AAPD% = 1.97 APD% = 1.53
Boukadi et al (1999)	No	Oman	92	ANN	R^2 = 0.95 RMSC = - SD = - AAPD% = 1.6 APD% = 0.02
Al-Marhoun & Osman (2002)	No	Saudi	282	ANN	R^2 =- RMSC = - SD = 0.6626 AAPD%=0.5116 APD% = 0.2173
Goda et al (2003)	No	Middle East	160	ANN	R^2 = 0.9981 RMSC = - SD = - AAPD% = 0.0307 APD% = -0.0078

Author	Formula Displayed	Origin	Data Number	Method	Errors
Malallah et al (2006)	No	California	105	ACE	R^2 = 0.9854 RMSC = - SD = 0.04 AAPD% = 1.94 APD% = 0.002
El-Sebakhy et al (2007)	No	Globally (782 dataset)	Dataset1 Dataset2 Dataset3	ANN SVM	R^2 = 0.9769 RMSC = 1.4625 SD = - AAPD% = 1.3718 APD% = 0.1808 R^2 = - RMSC = 0.62 SD = 0.4743 AAPD%=0.3527 APD% = -0.006 R^2 = 0.9992 RMSC = 0.65 SD = - AAPD%=0.3856

Author	Formula Displayed	Origin	Data Number	Method	Errors
					APD% = 0.0651
E. A. El-Sebakhy (2009)	No	Globally	1246	ANN SVM	R^2 =0.997 RMSC = - SD = 0.4743 AAPD%=0.353 APD% = -0.006
Dutta & Gupta (2010)	No	India	1852	ANN	R^2 =0.9890 RMSC = - SD = 1.848 AAPD%=1.779 APD% = -
Moghadam et al (2011)	No	Iran	218	ANN	R^2 =0.997 RMSC = - SD = 0.65 AAPD%=0.53 APD% = -3×10^{-5}
Khoukhi (2012)	No	Globally (1225 dataset)	Dataset1	(ANN, GA-ANN ,	R^2 = 0.9998 RMSC = 0.0111 SD = -

Author	Formula Displayed	Origin	Data Number	Method	Errors
			Dataset2	GA - ANFIS	AAPD%= - APD% = - R^2 =0.9979 RMSC = 43.42 SD = 6.523 AAPD%=4.241 APD% = -0.103
Rafiee-Taghanaki et al (2013)	No	Globally	569	ANN & LSSVM	R^2 =0.95 RMSC = 0.64 SD = - AAPD%=1.45 APD% = -0.07
Shokrollahi et al (2015)	No	Globally	756	MLP, RBF, LSSVM and CMIS	R^2 =0.9760 RMSC = 0.294 SD = 0.0295 AAPD%=1.4611 APD% = -
Salehinia et al (2016)	No	Iran	755	NARX-HW	R^2 =0.999 RMSC = - SD = -

Author	Formula Displayed	Origin	Data Number	Method	Errors
				,ANFIS-GP , ANFIS-FCM)	AAPD%= 0.0137 APD% =$7.1\times`10^{-3}$
Seyyedattar et al (2020)	No	World Wide	569	LSSVM-CSA & ANFIS	R^2 = 0.9994 RMSC = - SD = - AAPD% = 0.099 APD% = 0.012

CHAPTER 3

METHODOLOGY

Accurate prediction of Oil Formation Volume Factor (OFVF) as an essential fluid property is crucial for optimizing reservoir management and production. Traditional methods, relying on empirical correlations and laboratory measurements, often fall short in terms of efficiency (Standing, 1947). To address these limitations, this research investigates machine learning algorithms to develop Bo prediction models. By comparing Support Vector Machines (SVM), Gradient Boosting, Random Forests, Decision Trees, and Artificial Neural Networks (ANN), we aim to identify the most effective model for predicting OFVF.

3.1 INTRODUCTION TO THE WORKFLOW

The workflow (Fig. 4) for the proposed methods is illustrated in this section. To get accurate and trustworthy findings, the method employed is k-fold cross-validation (Fig. 5). To be more precise, the selected characteristics are used to divide the dataset. This entails splitting the dataset into k groups based on a random arrangement. The remaining dataset is utilized to validate the trained models, while $k-1$ batches were employed as the model's training set of data. Every portion of data is utilized for validation just once thanks to the k number of times this procedure is done. At last, the model with the best validation performance is designated as the recommended model. For this investigation, a k value of 10 was determined to be suitable.

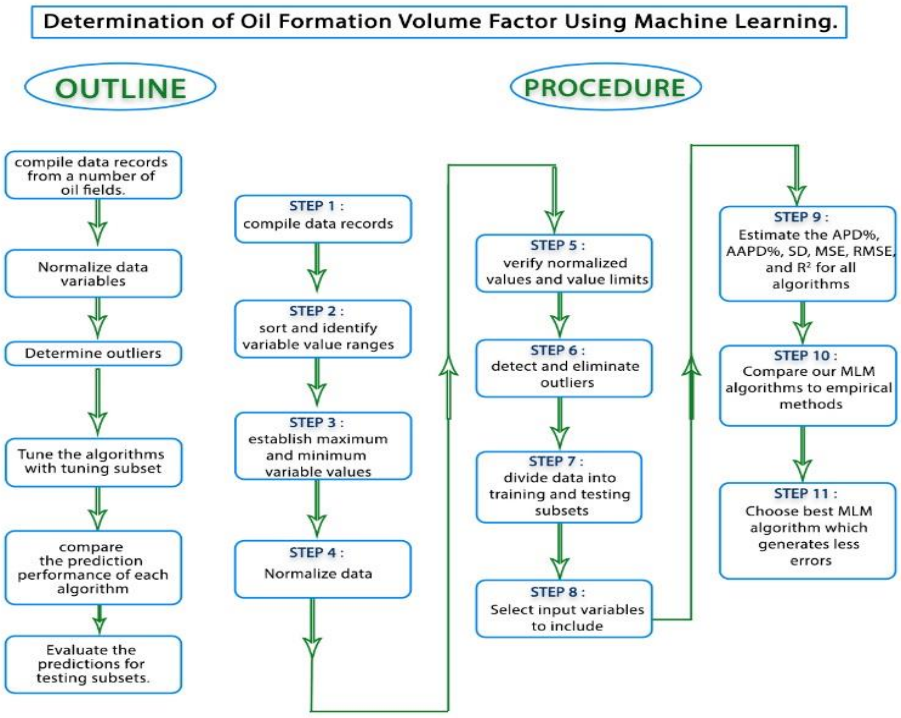

Figure 5 Workflow sequence used to contrast empirical and machine learning models

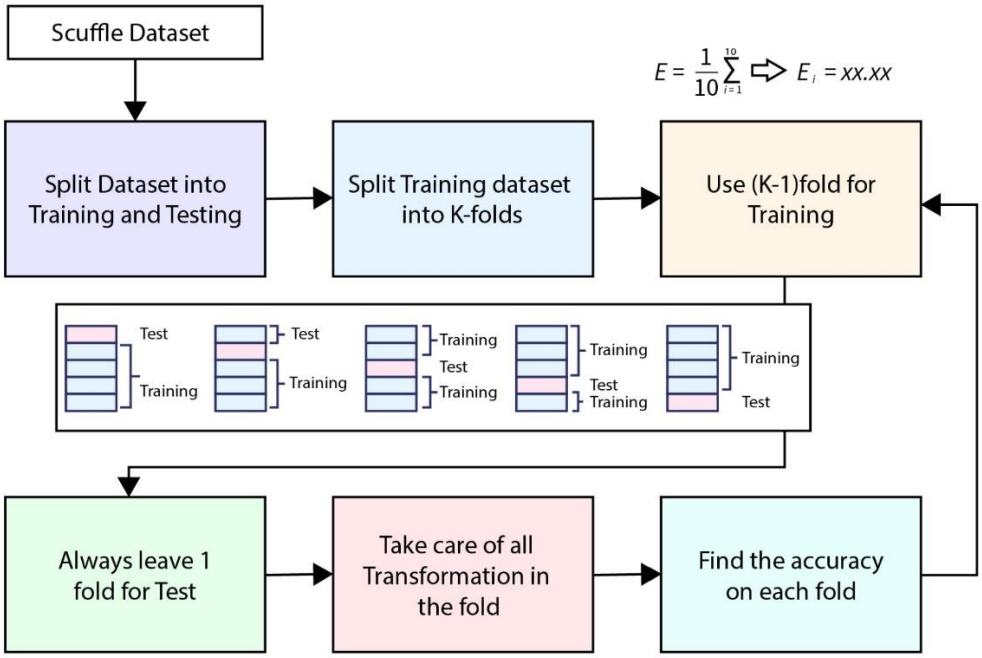

Figure 6 K-Fold Cross Validation

3.1.0 Data Collection and Processing of Dataset

The first step in this study is to collect a comprehensive dataset of relevant variables for predicting Bo. This may include data on well characteristics, reservoir properties, and production history.

Once the data is collected, it must be preprocessed to address any issues such as missing values, outliers, or feature redundancy. Feature selection will be a crucial subtopic of the data preprocessing stage, as identifying the most informative variables improve performance.

The dataset used consists of 221 observations with the following features: Rs (solution gas-oil ratio), T (temperature), γg (gas specific gravity), γo (oil specific gravity), P (pressure), and API, along with the target variable Bo (oil formation volume factor). The data sources and initial exploration steps are described in detail, highlighting the importance of understanding the reservoir's heterogeneity and the potential challenges in obtaining accurate in-situ measurements. Feature engineering is then performed to create additional relevant variables, such as derived ratios or combinations of the existing features. The methodology employed in this study focuses on developing robust predictive models for the (Bo).

3.1.1 Data Characterization

Dataset provided for this research paper consists of several key petrophysical variables. After establishing a comprehensive dataset, the next step was to perform detailed data exploration and preprocessing. This included analyzing the statistical properties of each variable, such

as the mean, standard deviation, minimum, and maximum values, as well as the 25th, 50th, and 75th percentiles. The data cleaning process involved identifying and addressing any outliers or erroneous data points to ensure the reliability of the subsequent analysis.

There were 221 input data linked to for this study. Using these data, innovative machine learning models are assessed to calculate the (Bo) based on easily accessible input factors.

Table 3 Statistical characterizations of the dataset

	Count	**Mean**	**Standard Deviation**	**Min**	**25%**	**50%**	**75%**	**Max**
Rs	221	563.95475	365.42092	24.000	330.000	468.0000	694.0000	2496.0000
T	221	225.30769	36.062136	125.000	198.000	232.0000	248.0000	296.0000
γg	221	1.366964	0.564725	0.612	0.811	1.3539	1.7167	3.4445
γo	221	0.829068	0.024172	0.766	0.815	0.8280	0.8370	0.8950
P	221	1452.6877	916.22016	55.000	715.000	1370.0000	2058.0000	4975.0000
API	221	39.274661	4.950105	26.600	37.500	39.4000	42.0000	53.2000
Bo	221	1.450834	0.252608	1.085	1.297	1.3990	1.5200	2.7130

3.1.2 Feature Scaling

To accommodate the highly variable units and values of the data, feature scaling is done during the initial processing of the data. The dataset's input variables are standardized using this method within a predetermined range. All data records' variable values must match in order to prevent computations from being biased by data ranges. are standardized to have a mean of 0 and a standard deviation of 1, as shown in Eq.(1).

$$Xstd = \frac{X - \mu}{\sigma} \qquad \text{Eq. (1)}$$

Where Xstd is the standardized value of X, X is the original value ,μ is the mean of the attribute, and σ is the standard deviation of the attribute.

3.1.3 Feature Selection

A critical preprocessing stage in machine learning is feature selection, which entails locating and picking the most pertinent characteristics from a dataset to create prediction models. By focusing on the most informative attributes, feature selection helps to create more accurate, efficient, and interpretable models.

3.1.4 Pair plot of the data set

A pair plot is a visualization technique used in data exploration to graphically summarize the relationships between multiple variables in a dataset. In essence, it produces a scatter plot matrix in which every variable is displayed against every other variable. This allows for a quick overview of the distribution of individual variables and the potential relationships between them.

Pair plots are invaluable tools in data analysis for several reasons. They help identify potential correlations between variables, which can be crucial for feature selection in machine learning models. Additionally, they can reveal patterns, outliers, and the overall

distribution of data points, providing insights into the data's characteristics. When analyzing the correlations between input and target variables in regression situations, pair plots are especially helpful.

When employing machine learning to anticipate Bo, a pair plot is an effective exploratory data analysis (EDA) tool. It facilitates the visualization of the correlations between the target variable (Bo) and the predictor variables (Rs, T, γg, γo, Pb, API). By examining these relationships, we can gain insights into which variables might be important predictors of Bo. Additionally, pair plots can help identify potential multicollinearity issues among the predictor variables, which can affect model performance.

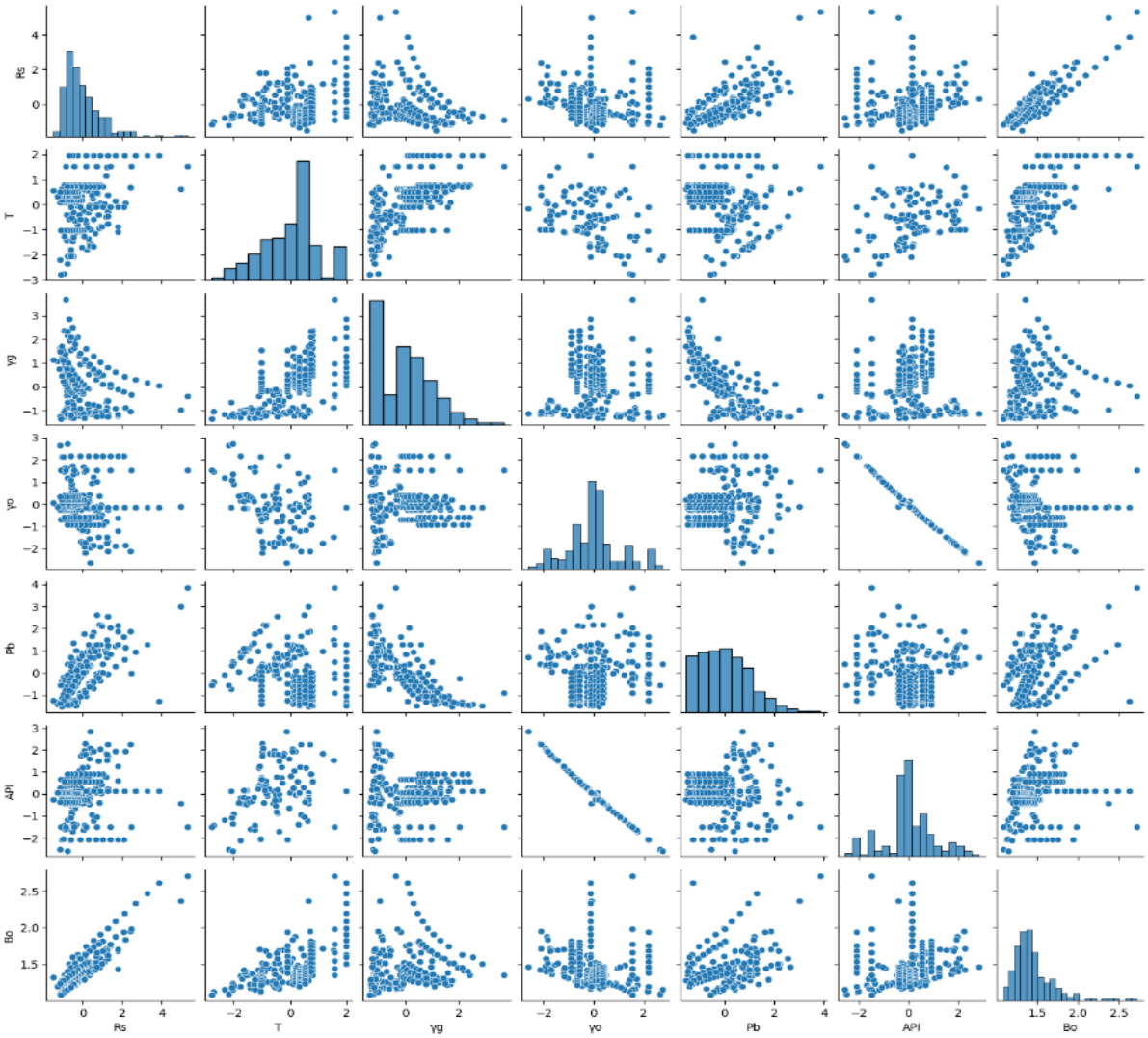

Figure 7 Pair plot of dataset

3.1.5 Pearson Correlation Analysis

The linear correlations between the input characteristics and Bo were found using Pearson correlation analysis. The outcomes show that the most relevant features are Rs, T, γo, API and P, consistent with the findings from the Spearman correlation analysis.

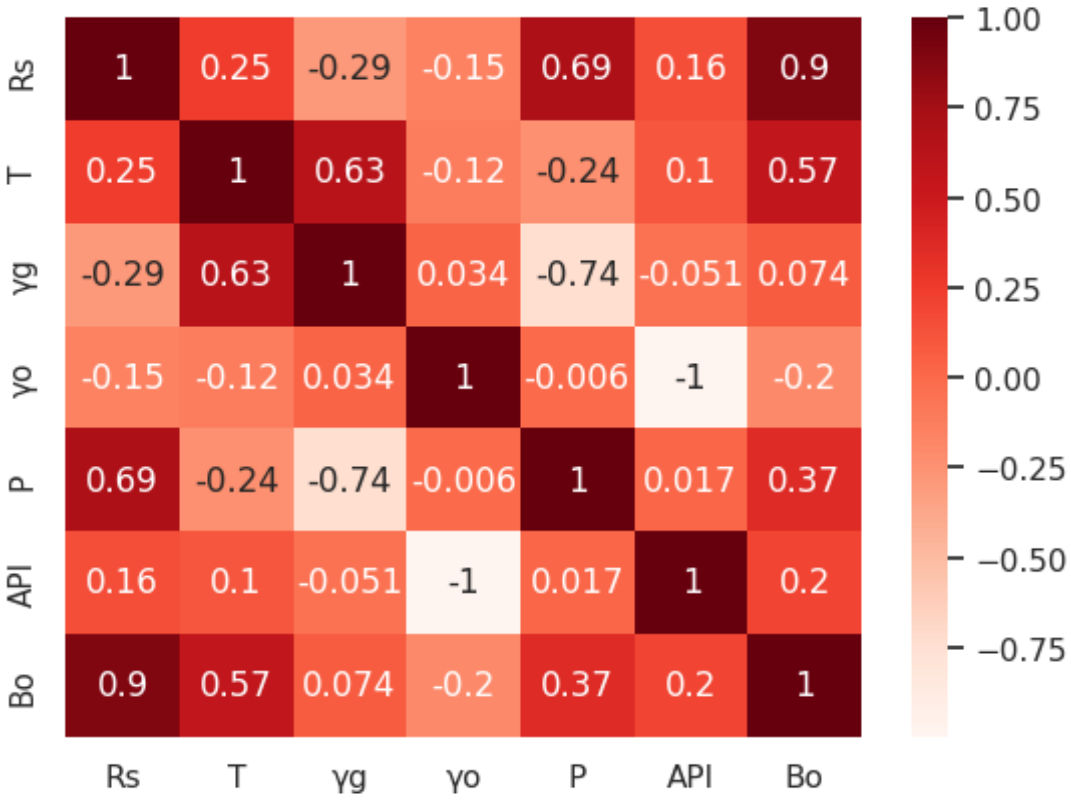

Figure 8 Pearson's Coefficient of Correlation Heatmap

3.1.6 Spearman Correlation Analysis

Spearman correlation was performed to identify monotonic relationships between input features and the Bo. The results indicate that the most relevant features are Rs, T, γo,API, and P, as they exhibit strong correlations with the target variable.

Figure 9 Spearman Coefficients of Correlation Heatmap

3.1.7 Analysis of Independent Variables

The following four criteria largely determine the (Bo) values of crude oils, as has been well-established via past empirical connections and diverse applications: Resolution Temperature, Oil Specific Gravity (γo), Gas Specific Gravity (γg), Gas Oil Ratio (Rs), and API Gravity (API) are the values that are obtained from this data.

Determining the relative importance of these variables in influencing the (Bo) is useful. The linear link between normally distributed data may be quantified using the Pearson correlation coefficient. Assuming that the relevant influencing factors are regularly distributed or linearly connected is not practical. Utilizing the Spearman rank correlation approach or other non-parametric statistical tests to assess the non-linear relationships affecting the dependent variables of the Oil Formation Volume Factor (Bo) is therefore more relevant.

The non-parametric Spearman's correlation coefficient is represented throughout the range of -1 (perfect negative correlation) to 1 (perfect positive correlation), similar to Pearson's correlation coefficient. A zero value indicates a complete lack of association.

The p-values for the (Bo) are shown in Figures 7 and 8, along with an identification of the major impacting input factors. These findings indicate the following:

Resolution Ratio of Gas to Oil (Rs): Given its significant positive connections with the Oil Formation Volume Factor (Bo), this variable has the most influence over the latter. The second most significant factor, temperature (T), has a substantial positive association with bo. Bo and API Gravity (API) have a positive association. The formulaic link between API and oil-specific gravity (γo) explains this.

Oil Specific Gravity (γo): Shows negative correlations with Bo and has a similar magnitude of correlation as API, but in the opposite direction due to their formulaic relationship. Gas Specific Gravity (γg : This is the least influential variable with Spearman's correlation coefficient close to zero.

In the gathered dataset of 221 records, the factors' relative effect on the Oil Formation Volume Factor (Bo) is summarized as follows: (Rs) is greater than (T) is greater than API or Oil Specific Gravity is greater thanGas Specific Gravity (γg).These correlations are shown

in Figures 5 and 6, which also show how each variable affects the Oil Formation Volume Factor (Bo) in respect to the others. Strong correlations between predictors and the response variable suggest that selecting these predictors would enhance the predictive performance of the models, which is consistent with the findings from the literature. Achieving a balance between predictive performance and model complexity requires careful consideration of the threshold selection A cutoff of 0.35 finds an appropriate medium between dimensionality reduction and model accuracy preservation. Those that exceed the cutoff are deemed pertinent and kept for further examination or model building. The approach takes into consideration both linear and non-linear connections between characteristics and the target variable by integrating both Pearson and Spearman correlations. The acquired results are Rs, T, P, and γo. To confirm that there was no association between the chosen characteristics, further analysis was carried out. P and Rs have a significant association. Since Rs and Bo showed a better link, P was discarded.

3.1.8 Ensuring Robust Model Performance through K-Fold Cross-Validation

To ensure robust model evaluation and prevent overfitting, we employed a K-Fold cross-validation strategy. Using this strategy, our dataset was split into five folds of the same size. Four folds were merged to create the training set in each iteration, while the remaining fold was used as the validation set. Five times through, this procedure was carried out, with a turn-taking fold serving as the validation set.

Using the "iloc" indexing technique and the K-Fold function from the scikit-learn library, we effectively divided the dataset into training and validation sets for every iteration. The training and validation features were stored in the X_train and X_test variables, respectively,

while the target variable values were stored in the y_train and y_test variables. This methodical methodology made it possible to evaluate performance in-depth across several data subsets.

Table 4 Normalized Training Dataset

	Count	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Rs	177	0.00708	0.993	-1.4810	-0.6554	-0.3071	0.4088	4.9563
T	177	0.05708	0.985	-2.7322	-0.5922	0.1860	0.6307	1.9647
γ_o	177	-0.04498	1.005	-2.6150	-0.5833	-0.0857	0.3289	2.7338

Table 5 Normalized Testing Dataset

	Count	**Mean**	**STD**	**Min**	**25%**	**50%**	**75%**	**Max**
Rs	44.0	-0.0285	1.0496	-1.2945	-0.5964	-0.1932	0.3203	5.2992
T	44.0	-0.2296	1.0462	-2.7878	-1.0299	0.0748	0.5195	1.9647
γ_o	44.0	0.1809	0.9811	-1.8273	-0.0961	0.0594	0.4948	2.1948

3.1.9 Gradient Boosting

In this study, we employed the Gradient Boosting from the scikit-learn library to model the relationship between the input features and the target variable. Gradient Boosting is an ensemble learning strategy that produces a sequence of decision trees, with each new tree being trained to fix the mistakes of its predecessors. By merging the best features of several weak learners into one robust predictive model, this iterative technique improves predictive performance.

The Gradient Boosting Regressor was configured using the default parameters provided by the scikit-learn library. The model utilizes “n_estimators” set to 100, indicating that it will run 100 boosting stages to iteratively refine its predictions. “learning_rate” is assigned to 0.1, which controls the step size at each iteration, influencing how much the model is updated in response to the estimated error. Additionally, the “max_depth” is configured to 3, establishing the maximum depth of the individual regression estimators to prevent overfitting while still capturing essential patterns in the data. The “min_samples_split” parameter is set to 2, meaning that "min samples leaf" is set to 1, guaranteeing that every leaf node has at least one observation, and defining the minimal number of samples needed to divide an internal node. The “subsample” parameter is configured at 1.0, meaning that the full dataset is used for fitting the individual base learners, which can enhance the model’s stability. Lastly, the “loss” function is set to 'ls', which stands for least squares regression, the default loss function optimized in regression tasks. These default settings provide a robust foundation for the model, enabling it to learn effectively from the training data while balancing complexity and generalization.

Utilizing the training data ({X_train), we fitted the model.} and {y_train}), where {X_train} indicates the feature set and `y_train} indicates the matching target values, after setting up the regressor. By means of this fitting process, the regressor is able to ascertain the patterns present in the data, therefore producing a predictive equation that can be used to the prediction of yet-to-be-observed data.

3.1.10 Support Vector Regression (SVR)

In order to capture complicated interactions within the data, we combined the usage of a radial basis function (RBF) kernel with SVR model for this research. With the kernel parameter set to 'rbf' when the SVR model was constructed, the method was able to translate the input features into a higher-dimensional space, improving its capacity to match nonlinear patterns. Following model configuration, the SVR was set to the training dataset, where the goal values are indicated by ‘y_train’ and the input characteristics are represented by ‘x_train’. The model may discover the underlying correlations in the training data through this fitting procedure.

Subsequently, we generated predictions for training data using the fitted SVR model. The predicted values were stored in the variable ‘y_predtrainsvr’, which provides the model's performance on the training set. By comparing these predictions to the actual target values, we can evaluate the model's accuracy and assess its capacity to generalize to unseen data. Overall, the use of SVR with an RBF kernel presents a powerful approach for regression tasks, particularly when dealing with complex, nonlinear relationships.

3.1.11 Random Forest

In this study, we employed a Random Forest Regressor to model the relationship between the features in our dataset and the target variable. The model was instantiated with a specified number of estimators ($n\_estimators = 10$) and a fixed random seed ($random\_state = 42$) to ensure reproducibility of the results.

The Random Forest algorithm operates by constructing multiple decision trees during training and outputting the average prediction of these individual trees for regression tasks. By setting the number of estimators to 10, we aimed to balance computational efficiency with predictive performance. The choice of a random state of 42 ensures that the results can be replicated in future experiments by providing a consistent starting point for the random number generation involved in the training.

After instantiation, the model was fitted to the training data using the `fit` method, which involves training the ensemble of decision trees on the features (X_train) and the target values (y_train). This process enables the model to learn the underlying patterns in the data, preparing it for subsequent predictions on unseen data. Overall, the implementation of the Random Forest Regressor represents a robust approach to tackling the regression problem posed by our dataset.

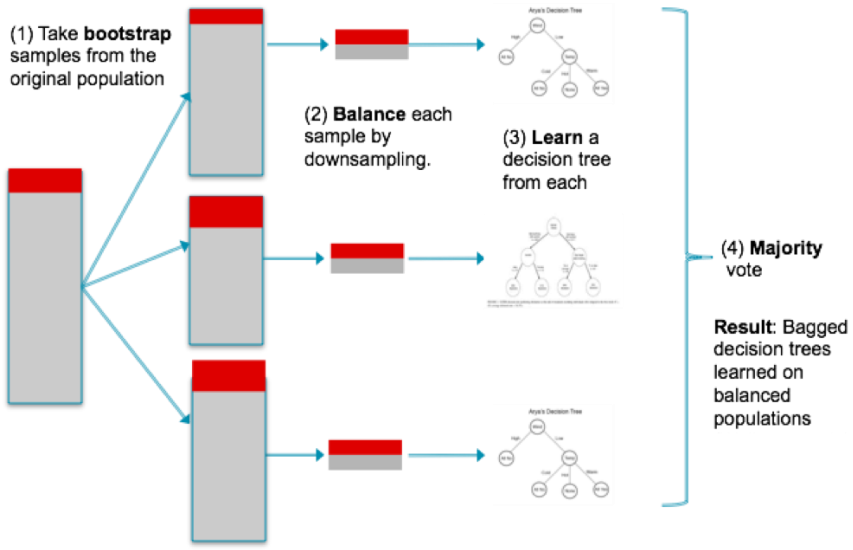

Figure 10 Flowchart of Random Forest Model

.

3.1.12 Artificial Neural Network (ANN)

Using TensorFlow's Keras API, we built a neural network model for this study in order to predict the target variable based on the given characteristics. The architecture of the model was specified as sequential, enabling simple layer stacking.

The input layer (X_train.shape[1]) is the initial layer and defines the form of the input data. It is configured to match the of features in the training dataset. By doing this, the model is guaranteed to obtain accurate input dimensions. After the input layer, we added a dense hidden layer of 128 321 neurons that used is Relu. A common option for hidden layers is the ReLU activation, which adds non-linearity while retaining effective training, allowing the model to pick up intricate patterns.

When attempting to predict a continuous value in regression tasks, the output layer's single neuron with no activation function is used. The model was created using the Adam optimizer, which is renowned for its flexible learning rate capabilities, and the loss function was set to mean squared error (MSE),

The model was trained using the ‘fit’ method, specifying a total of 100 epochs and a batch size of 10. With this setup, the model may handle the training data in tiny batches and loop over it several times, which can aid in convergence and enhance generalization. Furthermore, we included validation data (X_test, y_test) to track the model's effectiveness on unseen data during training, helping to detect overfitting and ensure robust predictive performance. Overall, this approach leverages deep learning to model the complex relationships in the dataset effectively.

3.1.11 Decision Tree

To verify the repeatability of findings, the model was instantiated with a predefined random seed (random_state=42). For regression problems, the Decision Tree algorithm is a non-parametric supervised learning technique that divides the feature space recursively into discrete areas according to the input data. This enables the model to accurately represent the underlying structure of the data by generating predictions based on the average target value in each zone.

We fitted f to train the model after constructing the regressor, which required supplying the training dataset (X_train and y_train). In order to create a tree-like model that most accurately depicts the links between the characteristics and the target variable, the algorithm examined the training data to find patterns and linkages. As a result of the training process, decision rules are produced, which serve as the foundation for the regression predictions and enable the model to generate predictions on data that hasn't yet been seen. All things considered, this analysis's usage of a Decision Tree Regressor provides a simple yet effective method for capturing intricate connections within the dataset.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Results

This chapter presents the outcomes of the machine learning models we created to predict Bo and talks about the ramifications of what we discovered. We were successful in capturing intricate, non-linear correlations between reservoir parameters and Bo by utilizing a variety of machine learning approaches. We demonstrate the gains made by contrasting our machine learning models' performance with conventional empirical correlations.

4.1.1 Metrics Used to Assess the Models' Predictions

To evaluate the efficacy of the machine learning models responsible for forecasting Bo, a collection of assessment measures is necessary. These metrics offer numerical assessments of the correctness and dependability of the model.

The coefficient of determination R^2, is a statistical indicator of a regression model's quality of fit. R^2 has a range of 0 to 1, where 0 represents a bad fit and 1 represents a perfect match. R^2 is not a comprehensive method to adjust for overfitting. When a model performs well on the training set but badly on the evaluation set, this is known as overfitting. Therefore, additional criteria were employed to assess the models' performance. Mean Squared Error (MSE) is a crucial parameter for evaluating the predictive power of a machine learning model. The average squared difference between the actual goal values and the predicted target values of a dataset is computed. The primary objective of the Mean Squared Error (MSE) is to quantify the extent to which a model's predictions correspond with reality. The values of good models are closer to zero.

One often-used statistic that determines the average size of the errors between expected and real Bo values is Mean Absolute Error (MAE). Although useful, MAE might be distorted by outliers, which can affect the evaluation as a whole. To address this, Average Percentage Deviation (APD) was considered. APD measures the average percentage difference between predicted and actual values, offering a relative perspective on prediction errors. However, APD is susceptible to the direction of errors, as positive and negative deviations can offset each other. To mitigate this, Absolute Average Percentage Deviation (AAPD) calculates the absolute value of percentage deviations before averaging, providing a more robust measure of prediction accuracy.

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Bo_{predicted\ i} - Bo_{actual\ i})^2}{\sum_{i=1}^{n}(Bo_{predicted\ i} - \frac{\sum_{i=1}^{n} Bo_{actual\ i}}{n})^2} \qquad \text{Eq. (2)}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Bo_{actual\ i} - Bo_{predicted\ i})^2 \qquad \text{Eq. (3)}$$

$$MAE = \frac{\sum_{i=1}^{n}|Bo_{actual\ i} - Bo_{predicted\ i}|}{n} \qquad \text{Eq. (4)}$$

$$APD = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Bo_{actual\ i} - Bo_{predicted\ i}}{Bo_{actual\ i}}\right| \times 100 \qquad \text{Eq. (5)}$$

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(B_i - \overline{B})^2} \qquad \text{Eq. (6)}$$

4.2 Discussion

4.2.1 Discussion of Empirical Correlations Results

The Al-Marhoun correlation has the greatest R-squared value of 0.96116 among the statistical measures, showing a rather significant connection between the actual and expected Bo values. The Standing correlation has the lowest AAPD and APD values, indicating that, on average, its predictions are more in line with the actual readings.

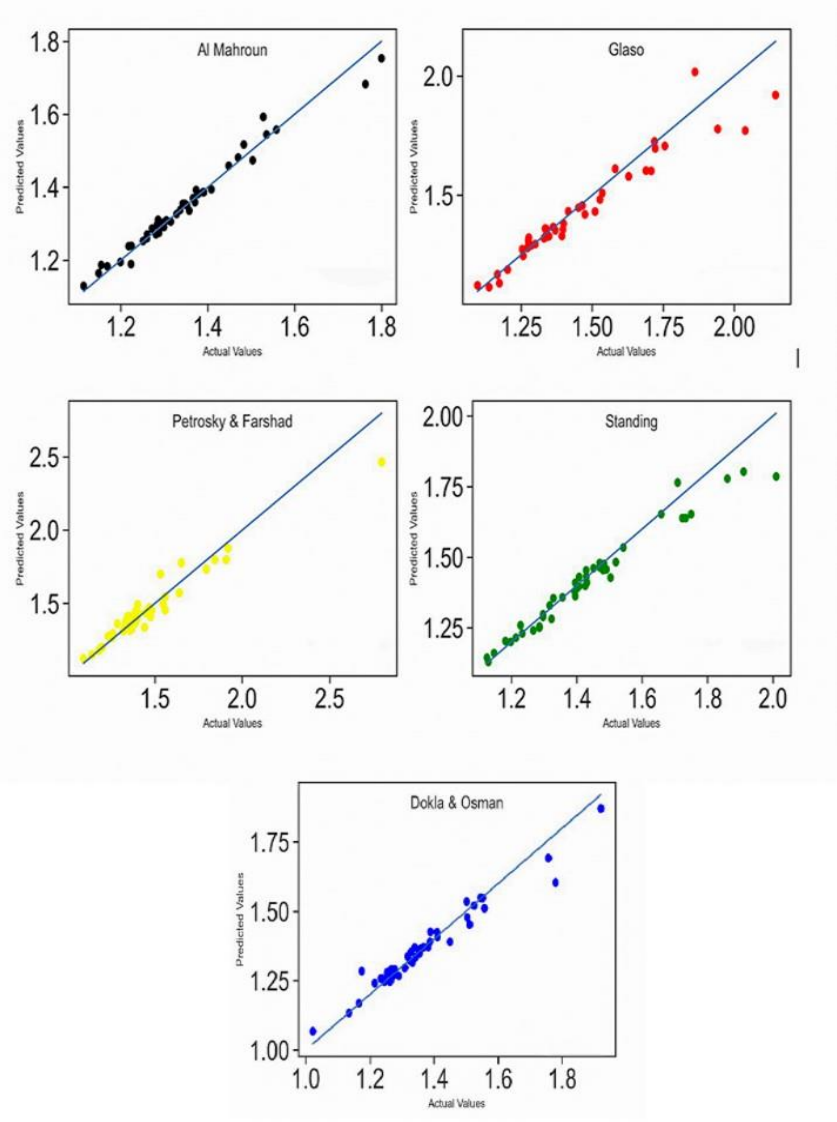

Figure 11 Performance of Empirical Models

Table 6 Model Evaluation of Empirical Models

Models	APD%	AAPD%	SD	MAE	MSE	RMSE	R^2
(Standing & Katz, 1942)	4.93	5.80	0.107	0.1078	0.0183	0.1351	0.9250
(Glaso, 1980)	12.50	21.91	0.2211	0.2543	0.1596	0.3995	0.9388
(Al-Marhoun, 1988)	**12.50**	**13.31**	**0.2034**	**0.2284**	**0.0820**	**0.2863**	**0.96116**
(Dokla & Osman, 1992)	9.67	11.52	0.1809	0.1695	0.0575	0.2397	0.9146
(Macary & El-Batanoney, 1993)	-7.31	8.01	0.1110	0.1221	0.0234	0.1531	0.9119
(Petrosky & Farshad, 1993)	13.55	13.59	0.1981	0.2336	0.0852	0.2920	0.9278
(Kartoatmodjo & Schmidt, 1994)	14.34	15.01	0.2159	0.2512	0.0990	0.3147	0.5158
(Almehaideb, 1997)	13.70	14.64	0.2227	0.25136	0.0988	0.3142	0.6577

4.2.2 Discussion of Machine Learning Models Results

The best R-squared value attained by the ANN, 0.985, indicates that machine-learning methods may still be improved upon to further enhance these Bo forecasts. The Bo ranges across which each correlation and method performs well and poorly may be determined with the use of this investigation.

With a mean around zero, the Bo prediction error models are less than the experimental models.

Random Forest follows closely behind ANN, showing competitive performance in terms of APD%, SD, MAE, and RMSE. While its R-squared value is slightly lower than ANN, it still demonstrates a strong predictive capability.

Decision Tree, Gradient Boosting, and SVM exhibit progressively worse performance as we move down the list. Their error metrics increase steadily, and their R-squared values decrease, indicating a declining ability to capture patterns in the data.

The machine learning models demonstrated superior accuracy and reliability in predicting Bo compared to traditional empirical correlations. This improvement can be attributed to the model's ability to capture, non-linear relationships between reservoir properties that are often missed by empirical approaches.

Unlike empirical correlations, which are often region-specific, the machine learning model generalized well across different reservoir conditions. This broad applicability makes it a valuable tool for petroleum engineers working in diverse geographical locations.

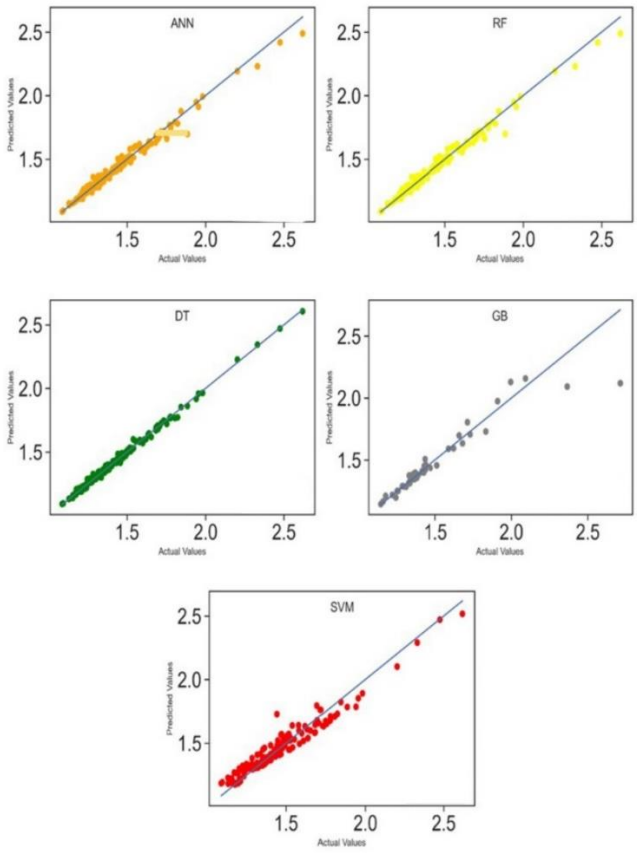

Figure 12 Performance of Machine Models

Table 7 Model Evaluation of Machine learning models

Models	APD %	AAPD %	SD	MAE	MSE	RMSE	R^2
ANN	**2.50**	**4.20**	**0.028**	**0.0376**	**0.0011**	**0.0333**	**0.9850**
RF	3.99	6.11	0.02810	0.0512	0.0016	0.0399	0.9800
DT	4.50	6.67	0.04970	0.0519	0.0027	0.0519	0.9725
GBM	5.20	8.73	0.0513	0.0657	0.0027	0.0524	0.96259
SVM	6.80	9.41	0.0643	0.0714	0.0036	0.0799	0.9496

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

The ability of machine learning algorithms to precisely forecast the (Bo) from reservoir parameters has been shown by this study. We have demonstrated the advantage of machine learning models in capturing intricate, non-linear connections within the data by contrasting the results of several machine learning methods with conventional empirical correlations.

A key finding of this study is the ability of machine learning models to generalize well across diverse reservoir conditions. The results have demonstrated that these models are relatively efficient in the accurate prediction of the Bo, with Artificial Neural Network (ANN) model achieving the best performance than the rest of the techniques in minimizing error, with an APD% of 2.50, AAPD% of 4.20, and an R^2 value of 0.9850. ANN also had the lowest MSE (0.0011) and RMSE (0.0333). The broader applicability of machine learning models makes them a valuable tool for petroleum engineers working in various geological settings.

Despite the significant advancements achieved, there is still room for improvement in Bo prediction. Future research could focus on exploring additional machine learning algorithms, incorporating more comprehensive datasets, and delving deeper into feature engineering techniques to further enhance predictive capabilities.

In conclusion, this study has provided compelling evidence of the effectiveness of machine learning in predicting Bo. By leveraging these advanced techniques, the petroleum industry can potentially optimize reservoir management, improve production forecasting, and make more informed decisions regarding field development strategies.

5.2 RECOMMENDATIONS

To fully harness machine learning potential, the industry should prioritize acquiring extensive, high-quality data from various reservoir environments. Acquiring a larger and more diverse dataset is crucial for training robust machine learning models. Incorporating data from various reservoirs with different characteristics can improve model generalization.

Cutting-edge feature engineering methods can reveal hidden patterns in the data, resulting in more precise forecasts. Combining the strengths of machine learning and empirical models could create hybrid approaches that offer enhanced performance by leveraging the strengths of both approaches.

The successful application of machine learning in real-time reservoir management requires efficient algorithms and robust model deployment strategies. Continuous improvement through feedback loops will be essential for maximizing the benefits of this technology. Therefore, investigating the feasibility of implementing machine learning models for real-time Bo prediction in reservoir management systems can enable proactive decision-making

REFERENCE

Al-Marhoun, M. (1988). PVT Correlations for Middle East Crude Oils. *Journal of Petroleum Technology*, *40*, 650–666. https://doi.org/10.2118/13718-PA

Al-Marhoun, M. A., & Osman, E. A. (2002). *Using Artificial Neural Networks to Develop New PVT Correlations for Saudi Crude Oils*. https://doi.org/10.2118/78592-MS

Almehaideb, R. A. (1997). Improved PVT correlations for UAE crude oils. *Proceedings of the Middle East Oil Show*, *1*, 109–120. https://doi.org/10.2118/37691-MS

Al-Shammasi, A. A. (1999). Bubble Point Pressure and Oil Formation Volume Factor Correlations. *All Days*. https://doi.org/10.2118/53185-MS

Al-Shammasi, A. A. (2001). A Review of Bubble point Pressure and Oil Formation Volume Factor Correlations. *SPE Reservoir Evaluation & Engineering*, *4*(02), 146–160. https://doi.org/10.2118/71302-PA

Boukadi, F., Al-Alawi, S., Al-Bemani, A., & Al-Qassabi, S. (1999). Establishing PVT correlations for Omani oils. *Petroleum Science and Technology*, *17*(5), 637–662. https://doi.org/10.1080/10916469908949738

Dokla, M. E., & Osman, M. E. (1992). Correlation of PVT properties for UAE crudes. *SPE Formation Evaluation*, *7*(1), 41–46. https://doi.org/10.2118/20989-PA

Dutta, S., & Gupta, J. P. (2010). PVT correlations for Indian crude using artificial neural networks. *Journal of Petroleum Science and Engineering*, *72*(1–2), 93–109. https://doi.org/10.1016/J.PETROL.2010.03.007

El-Sebakhy, E. A. (2009). Forecasting PVT properties of crude oil systems based on support vector machines modeling scheme. *Journal of Petroleum Science and Engineering*, *64*(1–4), 25–34. https://doi.org/10.1016/J.PETROL.2008.12.006

El-Sebakhy, E., Sheltami, T., Al-Bokhitan, S., Shaaban, Y., Raharja, I., & Khaeruzzaman, Y. (2007). Support vector machines framework for predicting the PVT properties of crude-oil systems. *SPE Middle East Oil and Gas Show and Conference, MEOS, Proceedings*, *3*, 1416–1429. https://doi.org/10.2118/105698-MS

Elsharkawy, A. M., & Alikhan, A. A. (1997). Correlations for predicting solution gas/oil ratio, oil formation volume factor, and undersaturated oil compressibility. *Journal of Petroleum Science and Engineering*, *17*(3–4), 291–302.

Gharbi, R. B., Elsharkawy, A. M., & Karkoub, M. (1999). Universal neural-network-based model for estimating the PVT properties of crude oil systems. *Energy and Fuels*, *13*(2), 454–458. https://doi.org/10.1021/EF980143V

Glaso, O. (1980). Generalized Pressure-Volume-Temperature Correlations. *Journal of Petroleum Technology*, *32*(05), 785–795. https://doi.org/10.2118/8016-PA

Goda, H. M., Shokir, E. M. E. M., Fattah, K. A., & Sayyouh, M. H. (2003). Prediction of the PVT data using neural network computing theory. *Society of Petroleum Engineers - Nigeria Annual International Conference and Exhibition 2003, NAICE 2003*. https://doi.org/10.2118/85650-MS

Gouda, A., & Attia, A. M. (2024). Development of a new approach using an artificial neural network for estimating oil formation volume factor at bubble point pressure of Egyptian crude oil. *Journal of King Saud University - Engineering Sciences*, *36*(1), 72–80. https://doi.org/10.1016/J.JKSUES.2022.08.001

Hanafy, H. H., Macary, S. M., ElNady, Y. M., Bayomi, A. A., & El Batanony, M. H. (1997). Empirical PVT correlations applied to Egyptian crude oils exemplify significance of using regional correlations. *Proceedings - SPE International Symposium on Oilfield Chemistry*, 733–737. https://doi.org/10.2118/37295-MS

Hemmati, M. N., & Kharrat, R. (2007). *A Correlation Approach for Prediction of Crude Oil PVT Properties*. https://doi.org/10.2118/104543-MS

Honarpour, M. M., Nagarajan, N. R., & Sampath, K. (2006). Rock/Fluid Characterization and Their Integration—Implications on Reservoir Management. *Journal of Petroleum Technology*, *58*(09), 120–130. https://doi.org/10.2118/103358-JPT

Karimnezhad, M., Heidarian, M., Kamari, M., & Jalalifar, H. (2014). A new empirical correlation for estimating bubble point oil formation volume factor. *Journal of Natural Gas Science and Engineering*, *18*, 329–335. https://doi.org/10.1016/J.JNGSE.2014.03.010

Kartoatmodjo, T., & Schmidt, Z. (1994). Large data bank improves crude physical property correlations. *Oil and Gas Journal;(United States)*, *92*(27).

Khoukhi, A. (2012). Hybrid soft computing systems for reservoir PVT properties prediction. *Computers and Geosciences*, *44*, 109–119. https://doi.org/10.1016/J.CAGEO.2012.03.016

Macary, S. M., & El-Batanoney, M. H. (1993). Derivation of PVT Correlations for the Gulf of Suez Crude Oils. *Journal of The Japan Petroleum Institute*, *36*(6), 472–478. https://doi.org/10.1627/JPI1958.36.472

Malallah, A., Gharbi, R., & Algharaib, M. (2006). Accurate estimation of the world crude oil PVT properties using graphical alternating conditional expectation. *Energy and Fuels*, *20*(2), 688–698. https://doi.org/10.1021/EF0501750

Moghadam, J. N., Salahshoor, K., & Kharrat, R. (2011). Introducing a new method for predicting PVT properties of Iranian crude oils by applying artificial neural networks. *Petroleum Science and Technology*, *29*(10), 1066–1079. https://doi.org/10.1080/10916460903551040

Omar, M. I., & Todd, A. C. (1993). *Development of New Modified Black Oil Correlations for Malaysian Crudes*. https://doi.org/10.2118/25338-MS

Petrosky, G. E., & Farshad, F. F. (1993). Pressure-volume-temperature correlations for Gulf of Mexico crude oils. *Proceedings - SPE Annual Technical Conference and Exhibition*, *Sigma*, 395–406. https://doi.org/10.2118/26644-MS

Rafiee-Taghanaki, S., Arabloo, M., Chamkalani, A., Amani, M., Zargari, M. H., & Adelzadeh, M. R. (2013). Implementation of SVM framework to estimate PVT properties of reservoir oil. *Fluid Phase Equilibria*, *346*, 25–32. https://doi.org/10.1016/J.FLUID.2013.02.012

Rashidi, S., Mehrad, M., Ghorbani, H., Wood, D. A., Mohamadian, N., Moghadasi, J., & Davoodi, S. (2021). Determination of bubble point pressure & oil formation volume factor of crude oils applying multiple hidden layers extreme learning machine algorithms. *Journal of Petroleum Science and Engineering*, *202*, 108425. https://doi.org/10.1016/J.PETROL.2021.108425

Saghafi, H. R., Rostami, A., & Arabloo, M. (2019). Evolving new strategies to estimate reservoir oil formation volume factor: Smart modeling and correlation development.

Journal of Petroleum Science and Engineering, *181*, 106180. https://doi.org/10.1016/J.PETROL.2019.06.044

Salehinia, S., Salehinia, Y., Alimadadi, F., & Sadati, S. H. (2016). Forecasting density, oil formation volume factor and bubble point pressure of crude oil systems based on nonlinear system identification approach. *Journal of Petroleum Science and Engineering*, *147*, 47–55. https://doi.org/10.1016/J.PETROL.2016.05.008

Salim Basaleh, S. (n.d.). *Bin-Gadeem Salem Mubarak Saleh (1) Bin-Gadeem Ali Salem Mubarak (3)*.

Seyyedattar, M., Ghiasi, M. M., Zendehboudi, S., & Butt, S. (2020). Determination of bubble point pressure and oil formation volume factor: Extra trees compared with LSSVM-CSA hybrid and ANFIS models. *Fuel*, *269*. https://doi.org/10.1016/J.FUEL.2019.116834

Shokrollahi, A., Tatar, A., & Safari, H. (2015). On accurate determination of PVT properties in crude oil systems: Committee machine intelligent system modeling approach. *Journal of the Taiwan Institute of Chemical Engineers*, *55*, 17–26. https://doi.org/10.1016/J.JTICE.2015.04.009

Standing, M. B. (1947). A pressure-volume-temperature correlation for mixtures of California oils and gases. *Drilling and Production Practice*.

Standing, M. B., & Katz, D. L. (1942). Density of Natural Gases. *Transactions of the AIME*, *146*(01), 140–149. https://doi.org/10.2118/942140-G

Vazquez, M., & Beggs, H. D. (1977). Correlations for fluid physical property prediction. *Proceedings - SPE Annual Technical Conference and Exhibition*, *1977-October*. https://doi.org/10.2118/6719-MS

Wood, D. A., & Choubineh, A. (2019). Reliable predictions of oil formation volume factor based on transparent and auditable machine learning approaches. *Advances in Geo-Energy Research*, *3*(3), 225–241. https://doi.org/10.26804/AGER.2019.03.01