



KASDI MERBAH UNIVERSITY OUARGLA
ACADEMIC MASTER'S THESIS



Domain:earth and universe science

Sector:geology

Specialty:petroleum geology

Thème

**Prediction of Sonic Transit Time Using
Machine Learning Techniques: A Case Study
of the Oued Zine Field**

Presented by

BELDJEBLIA ANES

REBIA ABD ELKADER

Supported publicly on: 11/06/2026

The jury consisted of

BELKISER Mohamed Salah	PR	President	UKM Ouargla
KECHICHED Rabah	PR	Moderator/rapporteur	UKM Ouargla
AMEUR-ZAIMECHE Ouafi	MCA	Promotor	UKM Ouargla
BOUTHAGHANE Ayoub	PHD	Co- Promotor	UKM Ouargla

Academic year 2025-2026

Acknowledgment

At the outset, we thank Allah, the Almighty, for granting us the strength, patience, and perseverance necessary to complete this work.

We would like to express our sincere gratitude to our promoter, Mr. AMEUR-ZAIMECHE Ouafi (MCA), and our co-promoter, Mr. BOUTHAGHANE Ayoub (PhD), for their guidance, supervision, availability, valuable advice, relevant recommendations, and continuous support throughout the completion of this dissertation.

We express our appreciation to Mr. BELKISER Mohamed Salah (Professor), President of the Jury, and to Mr. KECHICHED Rabah (Professor), Examiner, for the honor they have bestowed upon us by accepting to evaluate this work.

We would also like to thank all the teachers and staff of the Department of Earth and Universe Sciences for the quality of their teaching, their availability, and the knowledge they have imparted to us throughout our academic journey.

Finally, we warmly thank our parents—our families, and all those who have supported and encouraged us throughout the completion of this dissertation.

Rebia Abdelkader & Beldjebbia Anes

Dedication

الحمد لله دائماً وإبدا الحمد لله الذي أعطاني القوة والصبر لإتمام هذه المذكره بعد
 بسم الله الرحمن الرحيم

بكل حب وامتنان، أهدي هذا العمل إلى أبي وأمي، أغلى ما أملك في هذه الحياة .

إلى أبي، الذي علّمني أن النجاح لا يأتي إلا بالصبر والاجتهاد، والذي لم يبخل عليّ يوماً بالنصيحة والدعم، وكان دائماً مصدر قوة وثقة في كل خطوة من خطواتي

وإلى أُمي، صاحبة القلب الكبير، التي سهرت وتعبت ودعت لي في كل وقت، والتي كانت سندي الحقيقي في لحظات التعب والضعف قبل لحظات الفرح والنجاح. مهما كتبت من كلمات فلن أوفيها حقها، ففضلها عليّ أكبر من أن يُوصف

إليهما أهدي ثمرة سنوات من الجهد والتعب، راجياً أن يكون هذا العمل سبباً في إسعادهما وفخرهما، وأن أكون قد رددت جزءاً يسيراً مما قدماه لي من حب وتضحية وعطاء

كما أهدي هذا العمل إلى عائلتي الكريمة، وإلى كل من وقف إلى جانبي وشجعني وساندني طوال مسيرتي الدراسية

اللهم احفظ والدي وأدم عليهما الصحة والعافية، واجعل هذا النجاح فرحةً لهما قبل أن يكون لي

الحمد لله كثيراً طيباً

بلجبلية أنس

Dedication

الحمد لله دائما وابدًا الحمد لله الذي اعطاني القوة والصبر لاتيتم هذه المذكرة بعد
بسم الله الرحمن الرحيم

I dedicate this work to my parents and sisters, who have always been the source of my strength and pride, and whose unwavering support has enabled me to reach this stage of my life.

To my father, who has always been my pillar of support and the guiding force that led me along the path of knowledge. You taught me the true meaning of determination and emphasized the importance of education and academic achievement, especially for young people in today's world.

To my mother, whose kindness, patience, and sacrifices accompanied me throughout my academic journey. She stood by my side with her constant prayers, love, and encouragement. Her support has been invaluable, and she has played a major role in the success I have achieved today.

To my dear sisters, who shared both the challenges and the joys of this journey and who have always been a source of encouragement and support.

To my friend Anes, who stood by my side throughout the preparation of this dissertation. We worked together hand by hand, and I hope that this achievement will bring us both success and happiness.

To everyone who contributed, directly or indirectly, to the completion of this research, I extend my sincere gratitude and appreciation.

This dissertation is but a humble token of gratitude and loyalty to all those who have supported me, feelings that will never fade regardless of the passage of time.

الحمد لله كثيرا طيبا

Rebia Abdelkader

Summary

Figures List.....	VI
Tables List	VI
Abstract.....	VII
GENERAL INTRODUCTION	X
CHAPTER I: Literature Review on Artificial Intelligence Methods	X
Introduction.....	2
1. Artificial Intelligence.....	3
1.1. Subfields of Artificial Intelligence.....	3
2. Machine Learning.....	5
2.1. Machine Learning Fundamentals.....	5
2.1.1. Historical Development of Machine Learning.....	5
2.1.2. Learning Process in Machine Learning.....	6
2.2. Types of Machine Learning	6
2.2.1. Supervised Learning	6
2.2.1.1. Core Supervised Learning Algorithms.....	7
2.2.2. Semi-Supervised Learning	12
2.2.3. Unsupervised Learning	13
2.2.3.1. Unsupervised Learning Algorithms	13
2.2.4. Reinforcement Learning	14
2.3. Applications of Machine Learning.....	15
3. Deep learning	15
3.1. Deep Learning Architectures.....	16
3.1.1. Artificial Neural Networks (ANNs):.....	16
3.2. Principles of Deep Learning.....	16
3.2.1. Convolutional Neural Networks (CNNs):.....	17
3.2.2. Recurrent Neural Networks (RNNs):	17
3.3. Applications of deep learning in Petroleum Geology and Engineering	18
Conclusion	19
CHAPTER II: Overview of the Study Area	2
Introduction.....	21
1. Geological and Geographical Context.....	21
1.1. The Sbaa Basin.....	21
1.2. The Oued Zine Field.....	21

2.	Stratigraphic Framework	22
2.1.	Precambrian Basement	22
2.2.	Cambrian Deposits	22
2.3.	Ordovician Reservoir (Units III and IV).....	23
2.4.	Silurian Formation.....	23
3.	Petroleum System	26
3.1.	Exploration History	26
3.2.	Source Rock	26
3.3.	Reservoir Rock.....	27
3.4.	Seal Rock.....	27
3.5.	Migration and Trapping Mechanisms	27
4.	Structural and Tectonic Framework	28
5.	Lithological Characteristics	28
	Conclusion	29
CHAPTER III: Methods and materials		21
	Introduction	31
1.	Conventional Method	32
1.1.	Direct Methods (Core Analysis).....	32
1.2.	Indirect Methods ((well loggings analyses).....	34
2.	Unconventional Method	39
2.1.	Data Collection	40
2.2.	Data Analysis.....	43
2.3.	The statistical study	43
2.4.	Data base separation	44
2.5.	Selection of the algorithms	45
2.6.	Model evaluation.....	46
3.	Materials use	48
3.1.	Excel.....	48
3.2.	Python.....	48
	Conclusion	48
CHAPTER IV: Application of Machine Learning for Sonic Log Estimation.....		31
	Introduction	50
	Data Collection and Analysis.....	50
1.	Results Using Drilling Data	51
1.1.	Methodology and results	51

1.2.	Prediction	54
1.3	Discussion.....	61
2.	Results Using Well Log Data	65
2.1.	Statistical study	65
2.2.	Prediction	68
2.3.	Discussion.....	73
2.4.	Future importance.....	74
3.	Comparison between the results of Drilling Data & Well Log Data.....	77
3.1.	Overview of the Scenarios	77
3.2.	Statistical Performance Comparison	77
3.3.	Feature Importance & Sensitivity Analysis	78
3.4.	Direct Discussion (Pros & Cons).....	78
	Conclusion	79
	General Conclusion	78
	References Liste	81

Figures List

Figure 1. Structural Map of Artificial Intelligence Fields	4
Figure 2. Machine learning types.....	5
Figure 3. Illustrate the structure of the Decision Tree Regressor.....	8
Figure 4. Structure of a Random Forest.....	9
Figure 5. Support Vector Machine Decision Boundary	11
Figure 6. Principal Features of the XGBoost Model.....	12
Figure 7. Training Procedure of Semi-Supervised Learning	13
Figure 8. Working Structure of Deep Learning Network	16
Figure 9. Principles of Deep Learning.....	17
Figure 10. Geological map of Algeria, scratched from the Paleozoic roof, with the location of the Sbaa basin (modified d'après WEC, 1995).	24
Figure 11. Synthetic Lithostratigraphic Column and Correlation of Stratigraphic Data of the Sbaa Basin with Associated Main Tectonic Events	25
Figure 12. Carte de maturité (actuelle) à la base du silurien radioactif dans le bassin de Sbaa (document GDF-SUEZ).	28
Figure 13. Logging during drilling and wireline log. (Zaki, 1994).	35
Figure 14. Electrical Logs Tools.....	36
Figure 15. Caliper Logging Tool.....	36
Figure 16. Radioactive Tools.....	37
Figure 17. Acoustic Tool.....	37
Figure 18. The solid tool body of the Sonic Scanner receiver section contrast with typical sonic logging tools.....	38
Figure 19. Machine Learning Workflow for compressional sonic log prediction.	40
Figure 20. https://scikit-learn.org/stable/_images/grid_search_workflow.png	45
Figure 21. Schematic illustration of the k-fold cross-validation procedure used for model evaluation.	47
Figure 22. Bivariate analysis (correlation)	53
Figure 23. Graphical results scenario 1	56
Figure 24. Graphical results scenario 2	57
Figure 25. Graphical results scenario 3	58
Figure 26. Graphical results scenario 4	59
Figure 27. Graphical results scenario 5	60
Figure 28. Random Forest feature importance	62
Figure 29. XGBoost feature importance	63
Figure 30. Support Vector Machine feature importance	64
Figure 31. Bivariate analysis (correlation)	67
Figure 32. Graphical results scenario 1	70
Figure 33. Graphical results scenario 2	71
Figure 34. Graphical results scenario 3	72
Figure 35. Random Forest feature importance.....	74
Figure 36. XGBoost feature importance	75
Figure 37. Support Vector Machine feature importance	76

Tables List

Table 1. Drilling parameters.	41
Table 2. well log parameters.	42
Table 3. Monovariate analysis Using Drilling Data.....	51
Table 4. Utilized Models.....	54
Table 5. Machine Learning model results Of Drilling Data.....	55
Table 6. Monovariate analysis Using Well Log Data.....	65
Table 7. Utilized Models.....	68
Table 8. Machine learning model results Of Drilling Data.	69

Abstract

Sonic transit time (DT) is considered one of the most important petrophysical parameters used for reservoir characterization and subsurface formation evaluation. This study aims to predict DT values in the Oued Zine Field, located within the Sbaa Basin in the western Algerian Sahara, using artificial intelligence techniques based on both drilling data and logging data.

Predictive models were developed using three machine learning algorithms: XGBoost, Random Forest, and Support Vector Machine to evaluate their capability in accurately estimating DT values. The results demonstrated the superiority of the XGBoost model compared to the other algorithms in both scenarios.

Using drilling data as input variables, the XGBoost model achieved an R^2 value of **0.9249** with RMSE and MAE values of 1.4324 and 1.0584, respectively, on the testing dataset. In contrast, when logging data were used as inputs, the model achieved an R^2 value of **0.9105** with RMSE = 1.27658 and MAE = 0.955662, indicating a high predictive capability for DT estimation using both datasets.

The findings confirm the effectiveness of machine learning techniques in improving reservoir characterization and reducing dependence on costly conventional measurements, while also highlighting the importance of drilling data for real-time prediction during drilling operations.

Résumé

Le temps de transit acoustique (DT) est considéré comme l'un des paramètres pétrophysiques les plus importants utilisés pour la caractérisation des réservoirs et l'évaluation des formations géologiques. Cette étude vise à prédire les valeurs de DT dans le Oued Zine Field, situé dans le Sbaa Basin au Sahara algérien occidental, en utilisant des techniques d'intelligence artificielle basées sur les données de forage (drilling data) et les données de diagraphies (logging data).

Des modèles prédictifs ont été développés à l'aide de trois algorithmes de machine Learning : XGBoost, Random Forest et Support Vector Machine afin d'évaluer leur capacité à estimer avec précision les valeurs de DT. Les résultats ont montré la supériorité du modèle XGBoost par rapport aux autres algorithmes dans les deux scénarios étudiés.

En utilisant les données de forage comme variables d'entrée, le modèle XGBoost a obtenu une valeur de **$R^2=0.9249$** , avec des valeurs de RMSE et MAE de 1.4324 et 1.0584 respectivement pour les données de test. En revanche, lorsque les données de logging ont été utilisées comme entrées, le modèle a atteint une valeur de **$R^2=0.9105$** , avec RMSE = 1.27658 et MAE = 0.955662, indiquant une forte capacité prédictive pour l'estimation du DT à partir des deux types de données.

Les résultats obtenus confirment l'efficacité des techniques de machine learning pour améliorer la caractérisation des réservoirs et réduire la dépendance aux mesures conventionnelles coûteuses, tout en mettant en évidence l'importance des données de forage pour la prédiction en temps réel durant les opérations de forage.

الملخص

يُعتبر زمن العبور الصوتي أحد أهم الخصائص البتروفيزيائية المستخدمة في توصيف المكامن وتقييم التكوينات تحت السطحية. تهدف هذه الدراسة إلى التنبؤ بقيم زمن العبور الصوتي في حقل وادي زين، الواقع ضمن حوض السبع في الصحراء الجزائرية الغربية، باستخدام تقنيات الذكاء الاصطناعي اعتماداً على كلٍ من بيانات الحفر وبيانات التسجيلات الجيوفيزيائية.

تم تطوير نماذج تنبؤية باستخدام ثلاث خوارزميات للتعليم الآلي: التعزيز التدريجي المُحسّن، والغابات العشوائية، وآلة المتجهات الداعمة، وذلك لتقييم قدرتها على تقدير قيم زمن العبور الصوتي بدقة. وقد أظهرت النتائج تفوق نموذج التعزيز التدريجي المُحسّن مقارنةً بالخوارزميات الأخرى في كلتا الحالتين.

عند استخدام بيانات الحفر كمتغيرات إدخال، حقق نموذج التعزيز التدريجي المُحسّن قيمة معامل التحديد بلغت 0.9249، مع قيم الجذر التربيعي لمتوسط مربع الخطأ ومتوسط الخطأ المطلق التي بلغت 1.4324 و 1.0584 على التوالي في مجموعة الاختبار. وفي المقابل، عند استخدام بيانات التسجيلات الجيوفيزيائية كمدخلات، حقق النموذج قيمة معامل التحديد بلغت 0.9105، مع الجذر التربيعي لمتوسط مربع الخطأ = 1.27658 ومتوسط الخطأ المطلق = 0.955662، مما يدل على قدرة تنبؤية عالية لتقدير زمن العبور الصوتي باستخدام كلتا مجموعتي البيانات.

تؤكد النتائج فعالية تقنيات التعلم الآلي في تحسين توصيف المكامن وتقليل الاعتماد على القياسات التقليدية المكلفة، مع إبراز أهمية بيانات الحفر في التنبؤ الآني أثناء عمليات الحفر.

GENERAL INTRODUCTION

The field of geosciences has witnessed significant advancements due to the integration of geology with artificial intelligence techniques, enabling more accurate and efficient analysis of subsurface data, particularly in the oil and gas sector. Machine learning algorithms have proven highly effective in handling large and complex geological datasets, extracting nonlinear relationships, and improving the prediction of petrophysical properties. This has contributed to enhancing reservoir characterization and reducing reliance on traditional, costly methods.

This study aims to apply machine learning models to predict sonic log (DT) in the Oued Zine Field, located within the Sbaa Basin, using both drilling data and well logging data. The objective is to improve reservoir characterization accuracy and support exploration and production decision-making processes.

This work is organized into four main chapters as follows:

Chapter

This chapter presents the theoretical background and recent developments in artificial intelligence methods, with a focus on machine learning algorithms and their applications in processing geological and petrophysical data. It also discusses their use in the oil and gas industry, as well as the challenges associated with reservoir characterization using AI techniques.

Chapter

This chapter provides a geological description of the study area, including the Sbaa Basin and the Oued Zine Field. It highlights the reservoir's geological characteristics, its economic importance, and the main exploration challenges in the region.

Chapter

This chapter describes the dataset and methodology used in this study, including data preprocessing and model development. Two different scenarios are considered: the first using drilling data, and the second using well logging data. Machine learning algorithms such as XGBoost, Random Forest, and Support Vector Machine are employed to compare predictive performance.

Chapter IV (Application of Machine Learning for Sonic Log Estimation):

This chapter presents the results of applying machine learning models for sonic log (DT) prediction. It analyzes and compares model performance under both input scenarios, highlighting the effectiveness of drilling and logging data in improving prediction accuracy and supporting decision-making in exploration and production activities.

CHAPTER I: Literature Review on Artificial Intelligence Methods

Introduction

The increasing interest in Artificial Intelligence (AI) in the oil and gas industry is primarily driven by the growing complexity of exploration and production operations. Traditional methods have become insufficient to effectively manage challenges such as resource depletion, high operational costs, and environmental constraints. This has led to a shift toward data-driven technologies capable of improving efficiency and decision-making processes. As highlighted by [Hyne \(2012\)](#), AI and Machine Learning (ML) have transitioned from optional innovations to essential tools in modern petroleum operations. In addition, [Gharbi et al. \(2005\)](#) emphasized that the rapid growth in successful application cases and the availability of domain expertise have significantly accelerated the integration of AI techniques in the industry.

AI has proven to be particularly valuable in handling large, complex, and real-time datasets, making it highly suitable for predictive applications in drilling and formation evaluation. Its ability to learn from data and adapt to new conditions enables more accurate and reliable predictions compared to conventional methods. According to [Shahab \(2011\)](#), AI systems can process multiple variables simultaneously and solve problems that were previously difficult or impossible using traditional approaches. This capability is directly applicable to Real-Time Sonic Log Prediction, where AI models analyze continuous drilling and formation data to estimate sonic properties in real time. Such applications reduce uncertainty, improve subsurface characterization, and support faster and more informed decision-making during drilling operations.

Beyond specific applications, AI offers several strategic advantages across the oil and gas value chain. It enhances operational efficiency by optimizing processes and minimizing downtime, while also reducing human errors through automation. Furthermore, AI contributes to improving safety by limiting human exposure to high-risk environments. As noted by [Oyedele et al. \(2021\)](#) and [Ologunde \(2024\)](#), AI enables faster data-driven decisions, continuous updating of reservoir models, and better overall system performance. Moreover, [Meadows et al. \(2004\)](#) demonstrated that AI-based solutions improve reliability and cost-effectiveness, making them essential for achieving long-term sustainability and competitiveness in the oil and gas industry.

1. Artificial Intelligence

Artificial Intelligence (AI) is a field of computer science focused on developing systems capable of performing tasks that typically require human intelligence, such as learning, reasoning, and decision-making (Russell and Norvig, 2021; IBM, 2024). It involves the simulation of human cognitive functions, enabling machines to perceive their environment, process data, and act in a goal-oriented manner (Manning, 2020). Various definitions emphasize AI as the automation of intelligent behavior and the computational ability to achieve specific objectives (Winston, 1992; Bellman, 1978; McCarthy, 2007;). Overall, AI can be considered an interdisciplinary domain that combines algorithms, data, and computational models to create adaptive and intelligent systems.

AI can be understood through four main perspectives: systems that think like humans, act like humans, think rationally, and act rationally (Russell and Norvig, 2021). Over time, AI has evolved from theoretical concepts to practical applications driven by advances in data availability and computational power.

In the oil and gas industry, AI plays an increasingly important role in analyzing complex datasets such as well logs, seismic data, and drilling information. It is widely used for reservoir characterization, lithology classification, and the prediction of key petrophysical properties, improving both efficiency and decision-making in exploration and production processes.

As AI continues to develop, it has led to several specialized subfields. Among them, Machine Learning has emerged as a key area for data-driven modeling and prediction, and will be discussed in the following section.

1.1. Subfields of Artificial Intelligence

Artificial Intelligence is a broad and multidisciplinary field that encompasses several subdomains, each addressing specific aspects of intelligent behavior (Nilsson, 2009; Russell and Norvig, 2021;). Among the most important subfields is Machine Learning (ML), which enables systems to learn from data and improve their performance without explicit programming (Mitchell, 1997). ML has become a key technology for leveraging data and improving efficiency in real-world applications [X].

Another important area is Computer Vision, which focuses on enabling machines to interpret and analyze visual data such as images and videos (Szeliski, 2010). This is particularly useful in domains such as seismic interpretation and subsurface imaging in the oil and gas industry.

Natural Language Processing (NLP) allows machines to understand and generate human language (Jurafsky and Martin, 2020), enabling applications such as automated reporting and data analysis.

Deep Learning, a subset of ML, relies on neural networks to model complex patterns in large datasets and has shown high performance in tasks such as image recognition and pattern detection (LeCun et al., 2015).

Reinforcement Learning enables systems to learn through interaction with their environment using reward-based mechanisms (Sutton and Barto, 2018), and is increasingly applied in optimization problems such as drilling and production strategies.

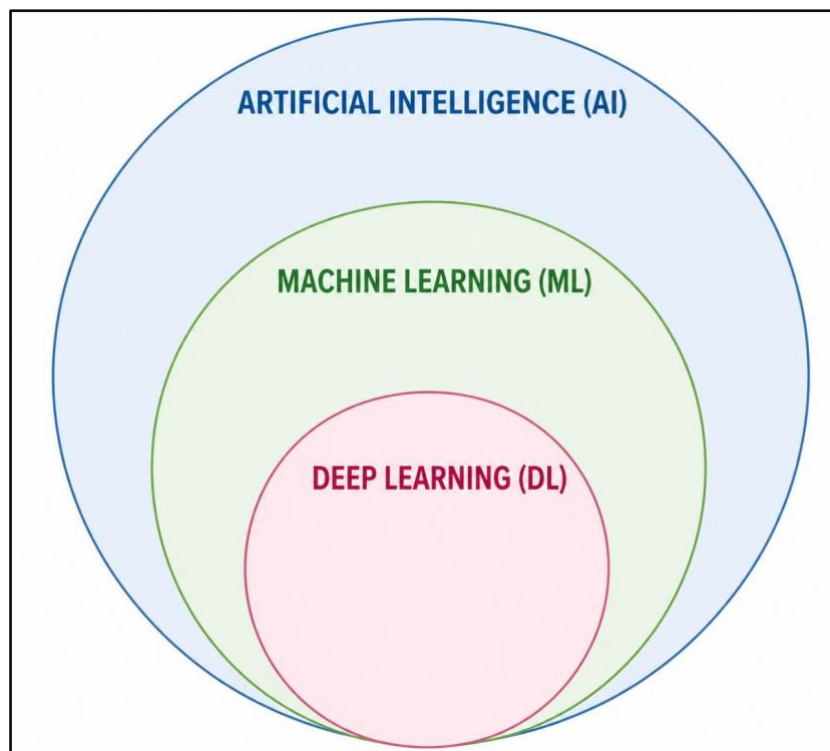


Figure 1. Structural Map of Artificial Intelligence Fields

2. Machine Learning

Machine Learning (ML) is a fundamental branch of artificial intelligence that focuses on the development of algorithms capable of learning from data and improving their performance without being explicitly programmed. Instead of relying on predefined rules or physical models, ML systems identify patterns, relationships, and structures within data to make predictions or decisions.

Formally, (Tom Mitchell 1997) defined machine learning as:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

Machine learning has become essential in modern data-driven industries, enabling automation, predictive analytics, and intelligent decision-making across various domains.

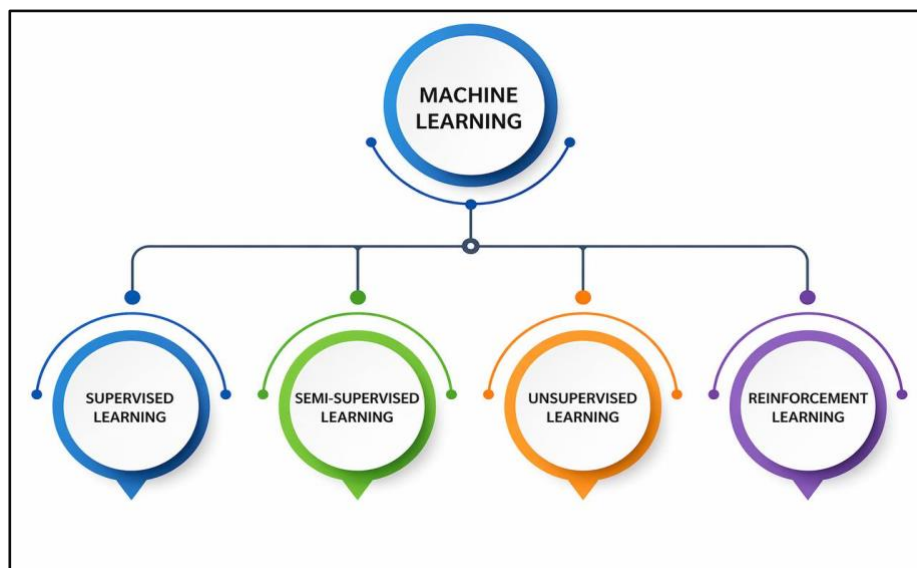


Figure 2. Machine learning types.

2.1. Machine Learning Fundamentals

2.1.1. Historical Development of Machine Learning

The term Machine Learning was first introduced by Arthur Samuel in 1952. Early developments include the perceptron model by Frank Rosenblatt in 1957, which laid the foundation for neural networks.

In 1967, the nearest neighbor algorithm marked a significant step in pattern recognition. However, progress slowed during the 1970s due to limitations in computational power and dominance of traditional computing paradigms.

A breakthrough occurred in 1982 with Hopfield networks, followed by the introduction of backpropagation in 1986, which significantly improved neural network training. In 1997, IBM's Deep Blue defeated world chess champion Garry Kasparov, demonstrating the potential of machine learning.

In the 21st century, the availability of big data and computational power has accelerated ML applications across various industries.

2.1.2. Learning Process in Machine Learning

The machine learning process generally consists of two main phases:

- Training phase: The model learns patterns from a dataset by adjusting its internal parameters based on input features and corresponding outputs.
- Testing phase: The trained model is evaluated on unseen data to measure its performance and generalization capability.

This process relies on four fundamental components:

- Data storage: Efficient handling and retrieval of large datasets.
- Model training (abstraction): Extracting meaningful patterns from data.
- Generalization: Applying learned knowledge to new data.
- Evaluation: Measuring model performance using metrics such as accuracy or error rate.

2.2. Types of Machine Learning

Machine learning algorithms are broadly classified into four main categories:

2.2.1. Supervised Learning

Supervised learning is a technique where models are trained using labeled data. The goal is to learn a mapping function from input variables X to output variables Y . The objective is to learn a mapping function that can predict outputs for new, unseen inputs.

Supervised learning is mainly divided into:

⇒ **Classification**

Classification (involves predicting discrete labels).

Examples include:

Spam detection (spam / not spam) .

Medical diagnosis (disease / no disease).

⇒ **Regression**

Regression focuses on (predicting continuous values).

Examples include:

Predicting oil production rates

Estimating reservoir properties such as porosity and permeability

Classification:

Classification assigns input data into predefined categories.

Types of Classification:

- a) **Binary Classification**
- b) **Multiclass Classification**
- c) **Multi-label Classification**

2.2.1.1. Core Supervised Learning Algorithms

Several algorithms are widely used in supervised learning

A. Decision Tree:

Decision trees can be used to optimize exploration and production processes by identifying the most critical factors influencing these processes. For example, decision trees can help determine the likelihood of discovering hydrocarbons in a particular region based on geological and geophysical data.

$$Y = f(x) = \sum_{i=1}^n w_i \cdot I(x \in R_i)$$

Y : is the predicted outcome,

x : is the input feature vector,

w_i : is the weight associated with the i th region, $I(x \in R_i)$ is an indicator function that returns 1 if x belongs to the i th region R_i and 0 otherwise.

Application of Decision Trees in Predicting Well Performance:

A case study was conducted to evaluate the effectiveness of decision trees in predicting well performance in a mature oil field 1. The study used a dataset of 100 wells, with features including well design, operational conditions, and production data. The results showed that decision trees were able to accurately predict well performance, with an average error of 5%.

Interpreting and implementing decision trees involve understanding the outputs and insights provided by the tree, integrating the tree with other analytical tools, and following best practices for implementation

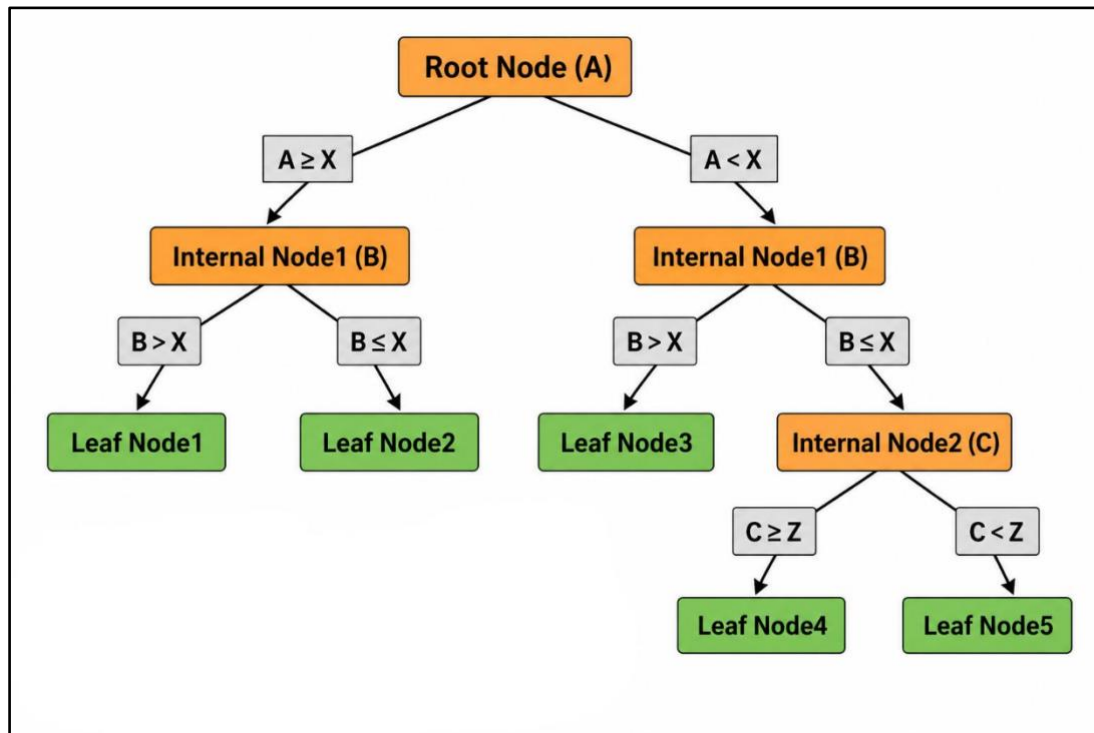


Figure 3. Illustrate the structure of the Decision Tree Regressor.

B. Random Forest:

An ensemble learning method introduced by [Leo Breiman \(2001\)](#), which combines multiple decision trees to improve accuracy and robustness. Its improve classification accuracy by building many decision trees instead of a single one. Each tree is trained using randomness, such as random samples of data, random feature selection, or random split choices. After growing a large number of independent trees, the final prediction is made by majority voting among all trees. This approach increases accuracy and reduces errors compared to using only one decision tree.

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\}$$

➤ Principle of Random Forest:

A Random Forest consists of a collection of decision trees, where each tree is trained using a random subset of the data and features. Formally, it can be defined as a set of classifiers $\{h(x, \Theta_k)\}$, where each tree depends on a random vector Θ

k , independently drawn from the same distribution (Breiman, 2001). Each tree in the forest gives a prediction, and the final output is determined by:

Majority voting for classification problems. Averaging for regression problems. This ensemble approach reduces variance and improves generalization compared to a single decision tree.

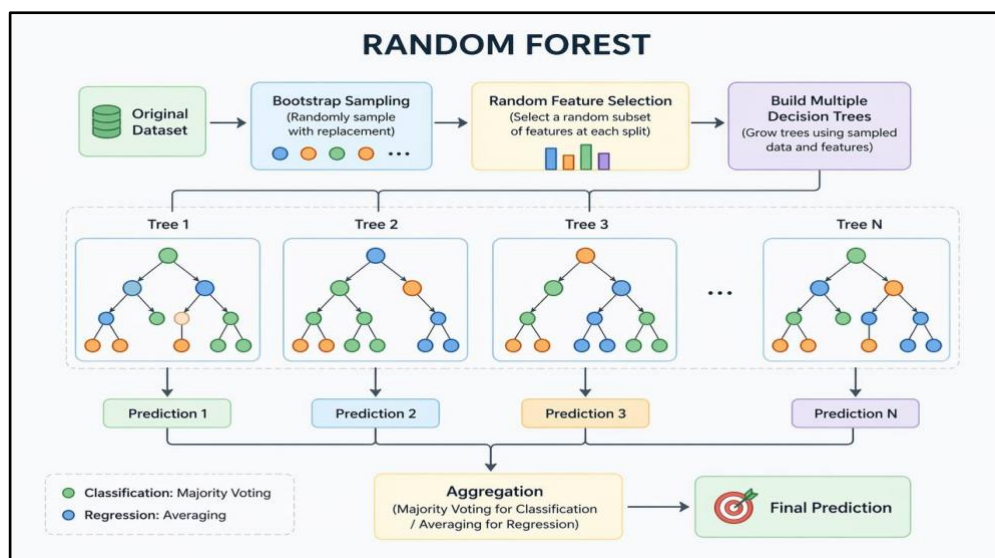


Figure 4. Structure of a Random Forest.

C. Support Vector Machine (SVM):

Developed by [Corinna Cortes and Vladimir Vapnik \(1995\)](#), SVM finds the optimal hyperplane that separates data into classes with maximum margin. They are powerful methods of data classification. They are based on the principle of structural risk minimization (SRM), which gives them a good generalization capacity. In their simple form, SVMs seek an optimal hyperplane separating classes with the greatest possible margin while reducing errors. Their main strength is to project the data into a higher dimensional space in order to make linear separation easier, without being penalized by the scarcity of data in this space.

$$y(x) = \text{sign} \left[\sum_{i=1}^N \alpha_i y_i K(x_i, x) + b \right]$$

$y(x)$: decision function (decision hypersurface) used to classify sample x .

N : number of data points.

i : index of samples from 1 to N .

α_i : positive real constants.

y_i : label of the data point (class -1 or +1).

$K(x_i, x)$: kernel function.

x_i : the i -th input data.

x : Vector belonging to the sample space.

b : real constant.

➤ Principle of support vector machine:

SVM works by learning from labeled data and determining the optimal decision boundary (hyperplane) that separates different classes. The algorithm focuses on maximizing the margin between classes, which improves classification accuracy and generalization. When the data are not linearly separable, SVM uses a kernel function

to project the data into a higher-dimensional space, where linear separation becomes possible. The points closest to the hyperplane, called support vectors, play a crucial role in defining the decision boundary. In practical applications such as petroleum drilling, SVM can analyze sensor data to automatically classify different drilling operations. Its ability to handle complex and high-dimensional data makes it a reliable and efficient classification method.

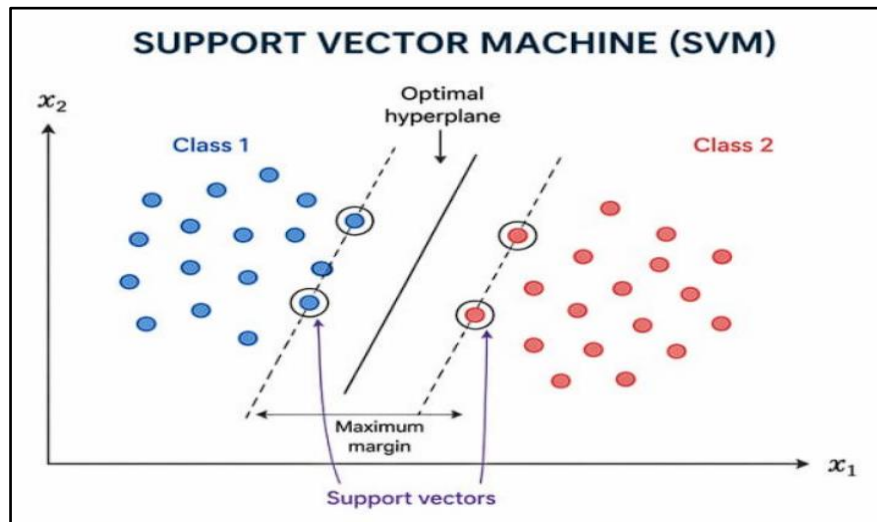


Figure 5. Support Vector Machine Decision Boundary

D. Extreme Gradient Boosting (XGBoost):

XGBoost (Extreme Gradient Boosting), developed by [Chen and Guestrin \(2016\)](#), is a powerful supervised machine learning algorithm used for both classification and regression tasks. It is an optimized implementation of gradient boosting that builds predictive models based on decision trees. XGBoost is widely used due to its high performance, scalability, and efficiency in handling large and complex datasets.

XGBoost belongs to the family of ensemble learning methods, specifically boosting techniques, where multiple weak learners (decision trees) are combined to form a strong predictive model.

$$\hat{y}_i = \alpha \sum_{k=1}^k f_k(x_i)$$

y_i : the predicted value corresponding to the input x_i

x_i : the input data (i-th sample)

α : learning rate of the model

f_k : total number of regression trees

K : output of the k -th regression tree

➤ **Principle of Extreme Gradient Boosting:**

The fundamental idea of XGBoost is to construct models sequentially, where each new tree is trained to correct the errors made by the previous ones. This is achieved using the concept of gradient boosting, where the model minimizes the difference between predicted and actual values.



Figure 6. Principal Features of the XGBoost Model

2.2.2. Semi-Supervised Learning

Semi-supervised learning is a machine learning approach that combines both labeled and unlabeled data during the training process. It is particularly useful in situations where labeled data are limited or expensive to obtain, while a large amount of unlabeled data is available. This approach lies between supervised learning, which uses only labeled data, and unsupervised learning, which relies entirely on unlabeled data.

➤ Principle of Semi-Supervised Learning

The main idea of semi-supervised learning is to improve model performance by leveraging the structure of unlabeled data in addition to a small set of labeled examples. The algorithm assumes that data points with similar characteristics are likely to belong to the same class.

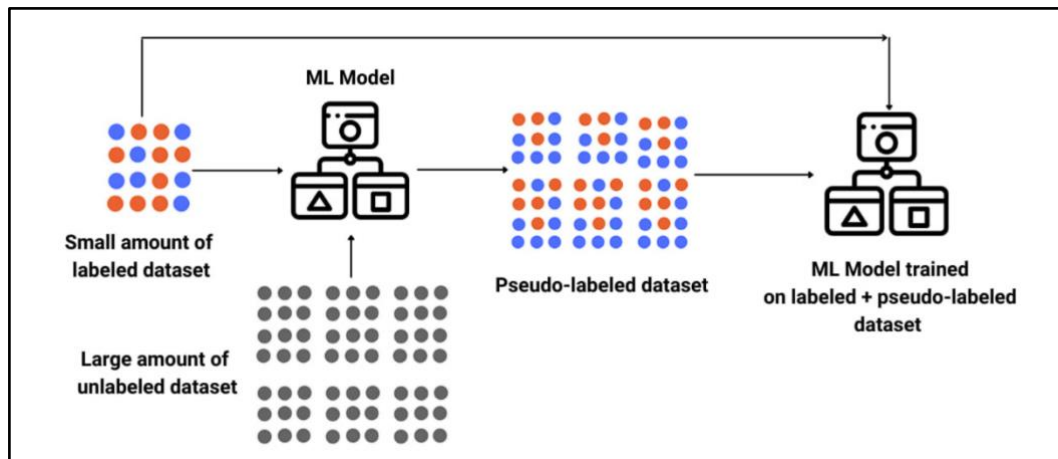


Figure 7. Training Procedure of Semi-Supervised Learning

2.2.3. Unsupervised Learning

Unlike supervised learning, where each training example is associated with a labeled output, unsupervised learning focuses on discovering hidden structures, patterns, or relationships within the data itself. In this approach, the model analyzes input data without any prior knowledge of the output, making it entirely data-driven.

Unsupervised learning aims to explore the inherent organization of the dataset by identifying similarities, differences, and underlying features. It is commonly used to group data into clusters, reduce dimensionality, or extract meaningful representations from complex datasets. While reducing the cost of data labeling.

2.2.3.1. Unsupervised Learning Algorithms

A. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used statistical technique for dimensionality reduction. Its objective is to transform the original data into a new set of variables, called principal components, which capture the maximum variance in the data. The first principal component represents the direction of the highest variance,

followed by the second, and so on. By selecting only the most significant components, PCA reduces the number of variables while preserving essential information.

➤ **Principal Component Analysis uses:**

Reducing dimensionality and simplifying datasets.

Data visualization in lower-dimensional spaces.

Noise reduction by eliminating less important components.

Preprocessing data for improving the performance of other algorithms.

B. Clustering

Clustering is the process of grouping data points into clusters such that elements within the same group are more similar to each other than to those in other groups. It is one of the most important tasks in unsupervised learning.

➤ **Clustering uses:**

- a) **K-means:** partitions data into K clusters based on distance to cluster centroids. It is simple and efficient but works best with well-separated clusters.
- b) **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** groups data based on density and is effective in detecting outliers or noise.
- c) **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):** designed for large datasets, it builds a hierarchical structure to reduce computational complexity.

2.2.4. Reinforcement Learning

Reinforcement Learning is a branch of machine learning where an agent learns to make decisions by interacting with an environment. The agent receives rewards or penalties for its actions and aims to maximize the cumulative reward over time through trial-and-error learning.

➤ **The principle of Reinforcement Learning**

The principle of Reinforcement Learning is based on continuous interaction between an agent and its environment. At each step, the agent observes a state, takes an action, and receives feedback in the form of a reward or penalty. Using this feedback, the agent updates its strategy (policy) through trial and error in order to improve future decisions and maximize cumulative reward.

2.3. Applications of Machine Learning

Machine learning is widely applied across multiple industries:

Healthcare: Disease detection and medical diagnosis

Finance: Fraud detection and risk assessment

Manufacturing: Process optimization and quality control

Transportation: Autonomous vehicles and route optimization

Oil and Gas Industry:

- Reservoir characterization
- Drilling optimization
- Production forecasting
- Predictive maintenance

In petroleum engineering, ML plays a critical role in improving efficiency, reducing operational risks, and enabling data-driven decision-making.

3. Deep learning

Deep Learning (DL) is an advanced subfield of machine learning that relies on artificial neural networks composed of multiple hidden layers. It is inspired by the structure and functioning of the human brain and is designed to automatically learn hierarchical representations from data. Unlike traditional machine learning techniques, which often require manual feature extraction, deep learning models can directly process raw data and extract relevant features autonomously ([LeCun et al., 2015](#)).

Deep learning has demonstrated remarkable success in various domains such as computer vision, natural language processing, and scientific computing due to its ability to model complex and nonlinear relationships. In recent years, it has also gained increasing importance in engineering fields, particularly in petroleum geology and petroleum engineering, where large and complex datasets are common.

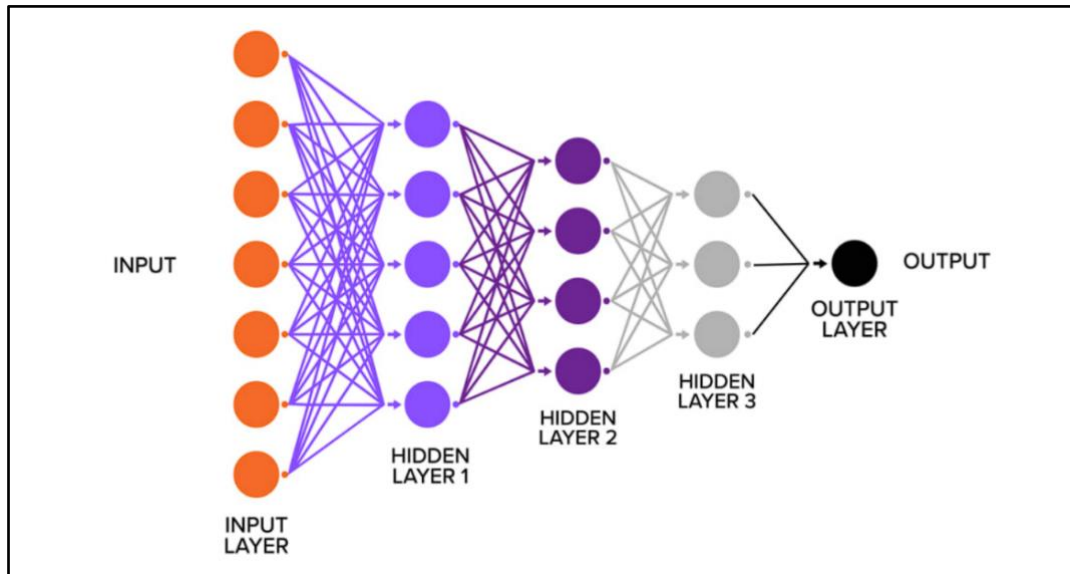


Figure 8. Working Structure of Deep Learning Network

3.1. Deep Learning Architectures

3.1.1. Artificial Neural Networks (ANNs):

Artificial Neural Networks (ANNs) are the fundamental building blocks of deep learning. They consist of interconnected neurons organized into three main layers: input layer, hidden layers, and output layer. Each neuron processes input data through weighted connections and applies an activation function to produce an output. During training, the model adjusts its weights using optimization algorithms such as gradient descent to minimize prediction error.

3.2. Principles of Deep Learning

Deep learning models are based on artificial neural networks (ANNs), which consist of layers of interconnected neurons. These layers include:

Input Layer: Receives raw data such as seismic signals, well logs, or reservoir properties.

Hidden Layers: Perform feature extraction and transformation through nonlinear operations.

Output Layer: Produces the final prediction or classification result.

Each neuron applies a weighted sum of its inputs followed by an activation function (such as ReLU, sigmoid, or tanh) to introduce nonlinearity into the model. The learning process involves adjusting the weights and biases using optimization algorithms such as gradient descent and back propagation.

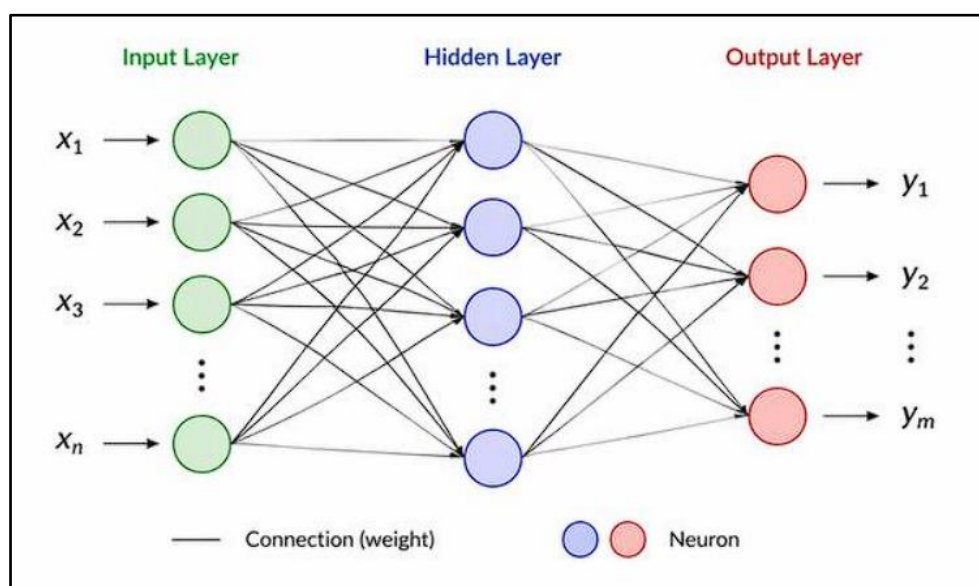


Figure 9. Principles of Deep Learning

3.2.1. Convolutional Neural Networks (CNNs):

Convolutional Neural Networks (CNNs) are specialized deep learning models designed for processing structured data such as images. They use convolutional layers to automatically detect features like edges, textures, and shapes. Pooling layers are used to reduce dimensionality and improve computational efficiency. CNNs have achieved remarkable success in computer vision tasks, including object detection and image classification (LeCun et al., 2015).

3.2.2. Recurrent Neural Networks (RNNs):

Recurrent Neural Networks (RNNs) are designed to handle sequential data such as time series or text. They maintain a memory of previous inputs, allowing them to capture temporal dependencies. However, traditional RNNs may struggle with long-

term dependencies, which led to the development of advanced architectures such as Long Short-Term Memory (LSTM) networks. These models are widely used in applications like speech recognition and language translation ([LeCun et al., 2015](#)).

3.3. Applications of deep learning in Petroleum Geology and Engineering

➤ Reservoir Characterization

Deep learning models are used to analyze seismic data and well logs to estimate reservoir properties such as porosity, permeability, and fluid saturation. This improves the accuracy of subsurface modeling.

➤ Seismic Interpretation

CNN-based models are widely used for automatic fault detection, horizon tracking, and lithology classification, significantly reducing manual interpretation time.

➤ Drilling Optimization

Deep learning helps in predicting drilling parameters, detecting anomalies, and optimizing drilling operations to reduce cost and improve efficiency.

➤ Production Forecasting

Time-series models such as RNN and LSTM are used to predict oil and gas production rates based on historical data, enabling better decision-making.

➤ Molecular and Microscopic Analysis

At the microscopic level, deep learning is integrated with molecular dynamics simulations to study fluid behavior, rock-fluid interactions, and chemical processes within reservoirs. It enables accurate modeling of phenomena such as methane hydrate formation, CO₂ sequestration, and hydrocarbon behavior under extreme conditions.

Conclusion

Machine learning has emerged as a powerful modern tool that enables the analysis of complex and large-scale datasets by capturing nonlinear relationships without relying solely on explicit physical models. In this chapter, the main learning paradigms supervised, unsupervised, semi-supervised, and reinforcement learning were presented, along with key algorithms such as decision trees, Random Forests, and artificial neural networks, highlighting their effectiveness in predictive modelling.

These techniques are characterized by strong generalization capabilities, allowing them to handle noisy, incomplete, and heterogeneous data. The importance of data pre-processing, model training, and evaluation was also emphasized to ensure reliable performance. When applied to subsurface and geophysical data, these methods significantly improve prediction accuracy and pattern recognition.

With the rapid growth of computational resources and data availability, machine learning plays a crucial role in enhancing decision-making, reducing uncertainty, and improving operational efficiency, supporting the development of intelligent and data-driven solutions.

CHAPTER II: Overview of the Study Area

Introduction

The study area is located within the Saharan platform, one of the most stable cratonic domains in North Africa, which has remained relatively undeformed since the Paleozoic era ([WEC, 1995](#)). This platform extends over a vast region and is characterized by a series of sedimentary basins that differ in their geological evolution and hydrocarbon potential. In Algeria, the Saharan platform is clearly distinguished from the northern Atlas domain by the South Atlas flexure, which represents a major structural boundary separating two contrasting tectonic regimes.

The Saharan domain is subdivided into several basins that developed under different sedimentary and tectonic conditions ([Beuf et al., 1971](#)). Among these basins, the Sbaa basin occupies a strategic position in the western part of the Algerian Sahara and represents an important area for hydrocarbon exploration. Its geological configuration, combined with the presence of thick Paleozoic sedimentary sequences, makes it a favorable environment for the development of petroleum systems.

Within this basin, the Oued Zine field constitutes a key area of interest due to its structural configuration and reservoir potential ([Wazir, 2014](#)). The field is associated with complex tectonic features inherited from multiple geological events, which have influenced both sedimentation and hydrocarbon accumulation.

1. Geological and Geographical Context

1.1. The Sbaa Basin

The Sbaa basin is located in the western Algerian Sahara and forms a narrow synclinal depression extending over approximately 32,000 km². It is bounded by the Saoura range to the north and northeast, which separates it from the Timimoun basin, and by the Kahal Tabelbala structure to the southwest. To the south, it is limited by the Reggane basin ([WEC, 1995](#)).

1.2. The Oued Zine Field

The Oued Zine field is situated in the central-eastern part of the basin and is characterized by a monoclinial structure dipping southward. This structure is associated with a major east-west trending fault system, which plays a key role in controlling the

structural configuration and hydrocarbon accumulation. This fault is believed to have originated during the Pan-African orogeny and was later reactivated during the Hercynian phase ([Wazir, 2014](#)). The fault system exhibits a complex geometry, splitting into two main branches with different orientations and dip angles. This structural complexity reflects the influence of multiple tectonic phases, particularly those related to the Pan-African and Hercynian events ([Allègre & Caby, 1972](#); [Coward & Ries, 2003](#)).

2. Stratigraphic Framework

2.1. Precambrian Basement

The Precambrian basement forms the structural foundation of the Saharan platform and consists mainly of crystalline rocks such as granites, schists, and metamorphic formations belonging to the West African craton and the Tuareg shield ([Allègre & Caby, 1972](#)).

This basement was shaped during the Pan-African orogeny, which ended around 620 Ma. Following this tectonic phase, erosion processes led to the formation of an irregular surface that served as a base for Paleozoic sedimentation ([Beuf et al., 1971](#)).

2.2. Cambrian Deposits

The Cambrian marks the beginning of widespread sedimentation across the Saharan platform. These deposits rest unconformably on the Precambrian basement, forming a major regional discontinuity known as the infra-Cambrian unconformity ([Beuf et al., 1971](#)).

Lithologically, Cambrian formations are dominated by coarse sandstones and quartzites deposited in fluvial environments. These deposits progressively evolve into marine conditions, reflecting a transition to tidal and estuarine environments.

2.3. Ordovician Reservoir (Units III and IV)

The Ordovician formations represent the most important reservoir interval in the study area. These deposits are subdivided into Units III and IV, reflecting distinct depositional environments ([Beuf et al., 1971](#)).

Unit III consists of alternating sandstones and clays deposited under marine conditions, while Unit IV corresponds to glaciogenic deposits formed during the Hirnantian glaciation. These deposits are highly heterogeneous and include sandstones, microconglomerates, and glaciomarine sediments ([Ghienne et al., 2007](#)).

The reservoir properties of these formations are strongly influenced by depositional and diagenetic processes, leading to variability in porosity and permeability.

2.4. Silurian Formation

The Silurian is characterized by the deposition of organic-rich shales under anoxic marine conditions. These “hot shales” are considered the primary source rocks for hydrocarbons in North Africa ([Lüning et al., 2000](#)).

These formations are characterized by high total organic carbon content and strong hydrocarbon generation potential, making them essential for the petroleum system.

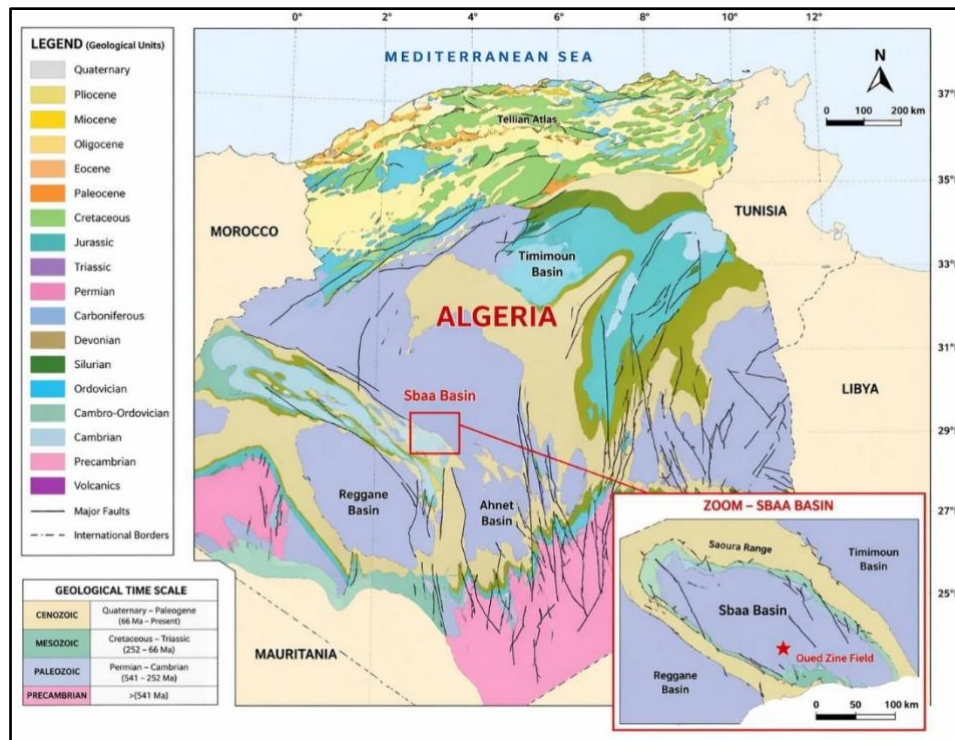


Figure 10. Geological map of Algeria, scratched from the Paleozoic roof, with the location of the Sbâa basin (modified d'après WEC, 1995).

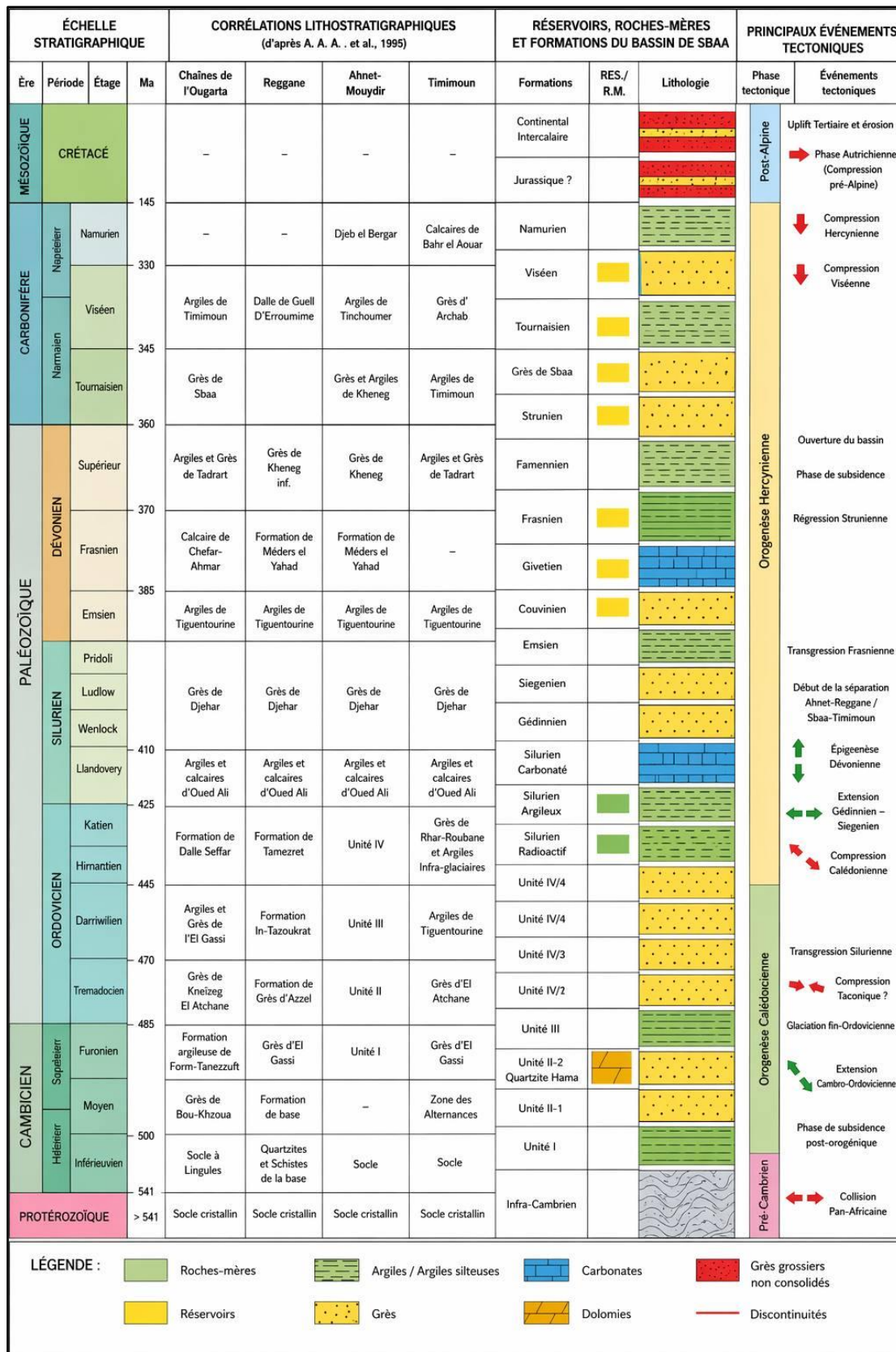


Figure 11. Synthetic Lithostratigraphic Column and Correlation of Stratigraphic Data of the Sbaa Basin with Associated Main Tectonic Events.

3. Petroleum System

3.1. Exploration History

Exploration activities in the study area began in the early 1960s with initial geological and geophysical surveys aimed at identifying hydrocarbon potential ([WEC, 1995](#)). Early work relied mainly on basic seismic data and regional mapping, which provided a general understanding of the subsurface structure.

With time, exploration techniques improved significantly, particularly during the 1970s and 1980s, through the use of enhanced seismic acquisition and drilling methods ([Beuf et al., 1971](#)). In recent years, the introduction of three-dimensional (3D) seismic imaging and digital geological modeling has greatly enhanced subsurface interpretation, allowing for more accurate identification of reservoirs and reducing exploration uncertainty ([Akkouche, 2007](#)).

3.2. Source Rock

The primary source rocks in the study area correspond to Silurian shales, which are characterized by a high content of organic matter and are widely recognized as the main hydrocarbon source in North Africa ([Lüning et al., 2000](#)). These formations, commonly referred to as “hot shales,” contain predominantly Type II kerogen, which is particularly favorable for the generation of hydrocarbons under suitable thermal conditions([Boote.et.al,1998](#)).

The richness and quality of these source rocks are controlled by depositional conditions, particularly anoxic marine environments that favored the preservation of organic material. As a result, these Silurian shales play a fundamental role in the petroleum system and contribute significantly to hydrocarbon generation within the basin

3.3. Reservoir Rock

The main reservoir rocks of the Oued Zine Field are the Ordovician sandstones. These formations exhibit variable porosity and permeability due to depositional and diagenetic processes. Despite the heterogeneity in reservoir quality, they represent the principal hydrocarbon-bearing intervals within the Sbaa Basin and constitute the main target for hydrocarbon exploration and production.

3.4. Seal Rock

Silurian clay formations act as an effective seal within the petroleum system due to their very low permeability and significant thickness ([Lüning et al., 2000](#)). These fine-grained deposits form impermeable barriers that prevent the upward migration of hydrocarbons, allowing their accumulation within underlying reservoir units.

The sealing capacity of these formations is further enhanced by their lateral continuity across the basin, making them a key element in preserving hydrocarbons over geological timescales ([Boote et al., 1998](#)).

3.5. Migration and Trapping Mechanisms

Hydrocarbon migration within the study area is primarily controlled by fault systems and structural discontinuities, which act as pathways facilitating the movement of hydrocarbons from source rocks to reservoir units ([Beicip-Franlab, 2000](#)). These migration processes are influenced by pressure gradients, permeability variations, and the structural configuration of the basin.

Trapping mechanisms are mainly structural in nature, with anticlines representing the dominant type of trap. These structures, often associated with faulting, create favorable conditions for hydrocarbon accumulation by forming closed geometries that prevent further migration ([Coward & Ries, 2003](#)).

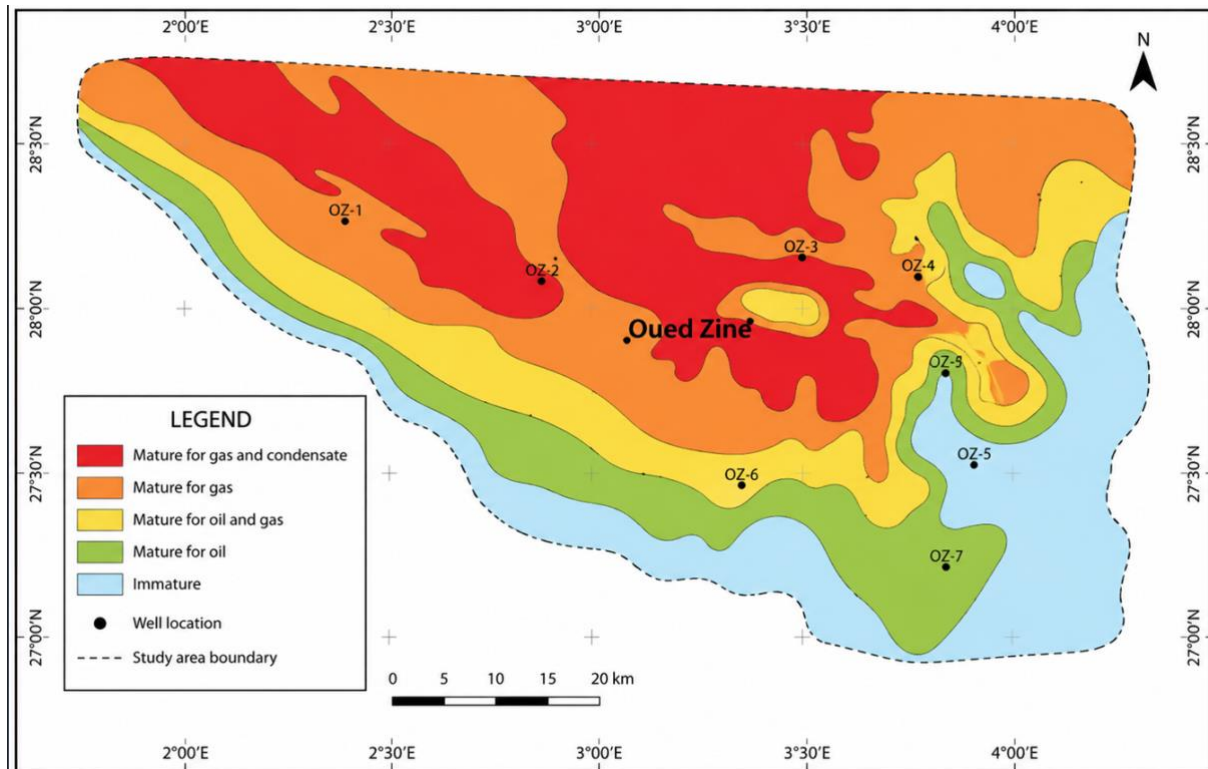


Figure 12 *Hydrocarbon Generation Maturity Map of the Oued Zine Field, Sbaa Basin (document GDF-SUEZ).*

4. Structural and Tectonic Framework

The structural evolution of the study area is controlled by multiple tectonic phases, including the Pan-African, Hercynian, and Alpine events. The Hercynian phase played a dominant role in shaping the present structure through folding and faulting (Coward & Ries, 2003).

These structures control reservoir distribution and hydrocarbon accumulation, making them essential elements in exploration studies.

5. Lithological Characteristics

The lithology of the study area is dominated by sedimentary rocks, including sandstones, shales, and carbonate formations. Sandstones represent the main reservoir rocks, while shales act as both source and seal formations.

The presence of volcanic rocks such as basalt indicates localized magmatic activity, which may influence the thermal evolution of the basin.

Conclusion

The geological framework of the study area reflects a complex interaction between stratigraphy, tectonics, and petroleum system development. The Cambro-Ordovician reservoirs and Silurian source rocks constitute the main elements controlling hydrocarbon potential.

Structural features such as folds and faults govern hydrocarbon migration and trapping, while lithological variations influence reservoir quality. This understanding provides a solid basis for further reservoir modeling and predictive studies.

CHAPTER III: Methods and materials

Introduction

The determination of petrophysical parameters such as porosity, permeability, density, and acoustic properties is essential in petroleum engineering because these parameters control reservoir quality, fluid storage capacity, and hydrocarbon production performance. They are commonly obtained from core analysis and well logging techniques, which remain the principal sources of subsurface reservoir evaluation data ([Torsæter et al., 2000](#)). Accurate estimation of these properties is therefore fundamental for reservoir characterization and field development decisions.

Among these parameters, sonic properties are of particular importance because they provide valuable information about rock and fluid behavior in the reservoir. Sonic logging is based on measuring the travel time or velocity of acoustic waves through formations surrounding the borehole. These measurements are used to estimate porosity, lithology, geomechanical characteristics, and seismic correlations. Sonic and ultrasonic tools are also applied for casing inspection, cement evaluation, and well integrity monitoring ([Liu et al., 2016](#)).

Conventional methods for sonic determination mainly include core analysis, sonic logging, and seismic surveys. These methods provide direct measurements but are often costly and time-consuming. Recently, non-conventional approaches based on artificial intelligence can predict sonic logs from other well logs such as gamma ray, density, neutron, and resistivity. Techniques such as Neural Networks and Random Forest offer faster, lower-cost, and accurate results, making them increasingly attractive for modern reservoir evaluation ([Alobaidi et al., 2015](#); [MDPI, 2025](#)).

Recently, non-conventional approaches have gained increasing attention in petroleum reservoir evaluation due to their ability to reduce costs, save time, and provide reliable predictions where direct sonic measurements are unavailable. In particular, the integration of machine learning techniques with conventional petrophysical data has emerged as a promising tool for improving sonic log estimation and supporting modern reservoir characterization.

1. Conventional Method

1.1. Direct Methods (Core Analysis)

1.1.1. Sonic Core Analysis

In sonic core analysis, the velocity of acoustic waves is measured on rock samples (core plugs) collected from the borehole under controlled laboratory conditions. Ultrasonic transducers are used to transmit acoustic pulses through the sample, and the travel time between the transmitter and receiver is recorded.

Two main types of seismic waves are involved:

P-waves (Primary waves): These are compressional waves that propagate through both solids and fluids. They are the fastest seismic waves and are widely used in petrophysical measurements because their velocity is highly sensitive to rock properties such as porosity and lithology ([Wyllie et al., 1958](#); [Tiab, 2015](#)).

S-waves (Secondary waves): These are shear waves that propagate perpendicular to the direction of wave motion. They only travel through solid materials and do not propagate in fluids, making them particularly useful for evaluating the rigidity and mechanical strength of rocks ([Guéguen, 1994](#)).

By measuring the travel time of these waves through the core sample, their velocities can be calculated. These velocities reflect the elastic and physical properties of the rock, making sonic core analysis a fundamental tool in rock physics for understanding the mechanical behavior of geological formations under controlled laboratory conditions ([Gassmann, 1951](#); [Tiab, 2015](#)).

1.1.2. Porosity from Sonic Core Analysis

Porosity in sonic core analysis is determined indirectly by relating the propagation of acoustic waves to the internal structure of the rock. The fundamental idea is that acoustic waves travel faster through dense, compact rocks with few voids, while their velocity decreases in rocks with higher pore content.

When an acoustic wave passes through a core sample, its propagation is influenced by both the solid rock matrix and the fluids occupying the pore spaces. As porosity increases, the wave spends more time traveling through the less rigid fluid-

filled pores, which results in a slower overall propagation ([Wyllie et al., 1958](#); [Guéguen, 1994](#)).

1.1.3. Lithology from Sonic Core Analysis

Lithology identification using sonic core data relies on analyzing the elastic response patterns of different rock types rather than simply measuring wave velocity. Each lithology exhibits a distinct acoustic behavior due to differences in mineral composition, grain contacts, and internal fabric.

Instead of focusing on travel time alone, interpretation is often based on cross-analysis between compressional and shear wave responses, where the relationship between these parameters helps distinguish between clastic and carbonate rocks. The rigidity of the grain framework and the degree of cementation strongly influence these acoustic signatures, allowing a more reliable differentiation between lithologies with overlapping physical properties.

This approach is commonly integrated with rock physics models to improve classification accuracy, where experimental core measurements serve as calibration data for identifying lithological variations in reservoir rocks ([Guéguen, 1994](#); [Tiab, 2015](#)).

1.1.4. Saturation estimation from Sonic Core Analysis

Saturation of fluids within rocks (water, oil, or gas) is estimated by analyzing their effect on the elastic response of the rock. The presence of fluids in the pore spaces modifies the stiffness of the medium, which in turn influences the behavior of acoustic waves within the sample.

The rock response is compared between dry and saturated states, where differences in fluid type lead to changes in elastic properties. This effect is particularly significant in the presence of gas, which causes a noticeable reduction in rock stiffness and wave propagation velocity.

This variation are interpreted using rock physics models such as the Gassmann equation. then used indirectly to estimate the degree of saturation and relate it to reservoir properties in petroleum systems ([Gassmann, 1951](#)).

1.2. Indirect Methods ((well loggings analyses)**1.2.1. Well Logging**

Well logging is a geophysical technique used to measure and record the physical properties of subsurface formations along a borehole using specialized downhole tools. It is used to characterize lithology, evaluate reservoir quality, and identify fluid content within geological formations. The method includes different types of logs such as gamma ray, resistivity, density, and sonic logs, each based on a specific physical principle. These measurements provide continuous information with depth that is essential for formation evaluation and hydrocarbon exploration (Pang et al., 2022; Springer Nature, 2021). Well logging operations are performed under different well conditions depending on the stage of drilling and completion we have :

a) Open Hole Logging :

Open hole logging is performed immediately after drilling, before casing installation. It provides direct and detailed measurements of formation properties such as lithology, porosity, and fluid content for accurate formation evaluation.

b) Cased Hole Logging :

Cased hole logging is carried out after casing and cementing the well. It focuses on reservoir monitoring and production evaluation, including fluid movement, cement integrity, and well performance.

c) Bottom Hole :

Bottom hole logging (Logging While Drilling) is done near the drill bit during drilling. It provides real-time formation data, allowing immediate interpretation and quick drilling decisions such as well trajectory adjustment and reservoir identification.

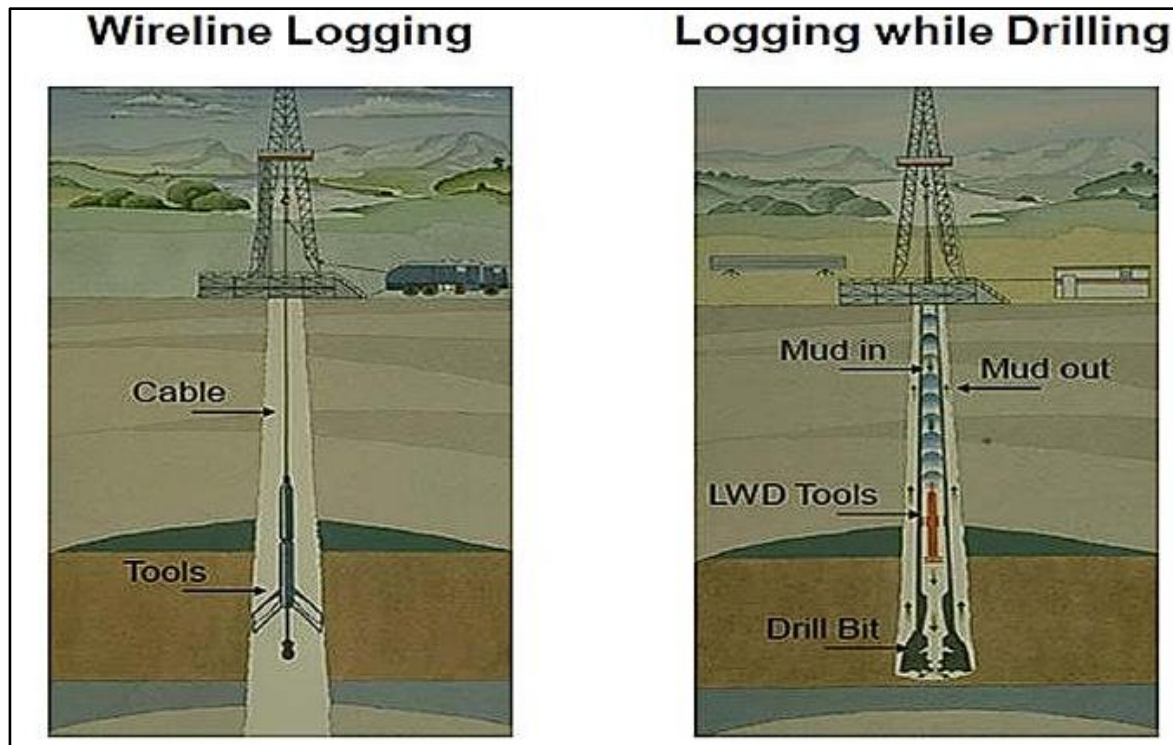


Figure 13. Logging during drilling and wireline log. (Zaki, 1994).

1.2.1.1. Well Logging Tools

The logging tools collectively provide a complete evaluation of subsurface formations by measuring electrical, mechanical, radioactive, and acoustic responses of rocks, which are essential for reservoir characterization and hydrocarbon exploration.

a) Electrical Logs:

Spontaneous Potential (SP) Log : which records natural electrical potentials generated by ion movement between formation water and drilling mud.

Resistivity Logs : which measure the ability of rocks to conduct electric current and are mainly used to detect fluid types and hydrocarbon zones.

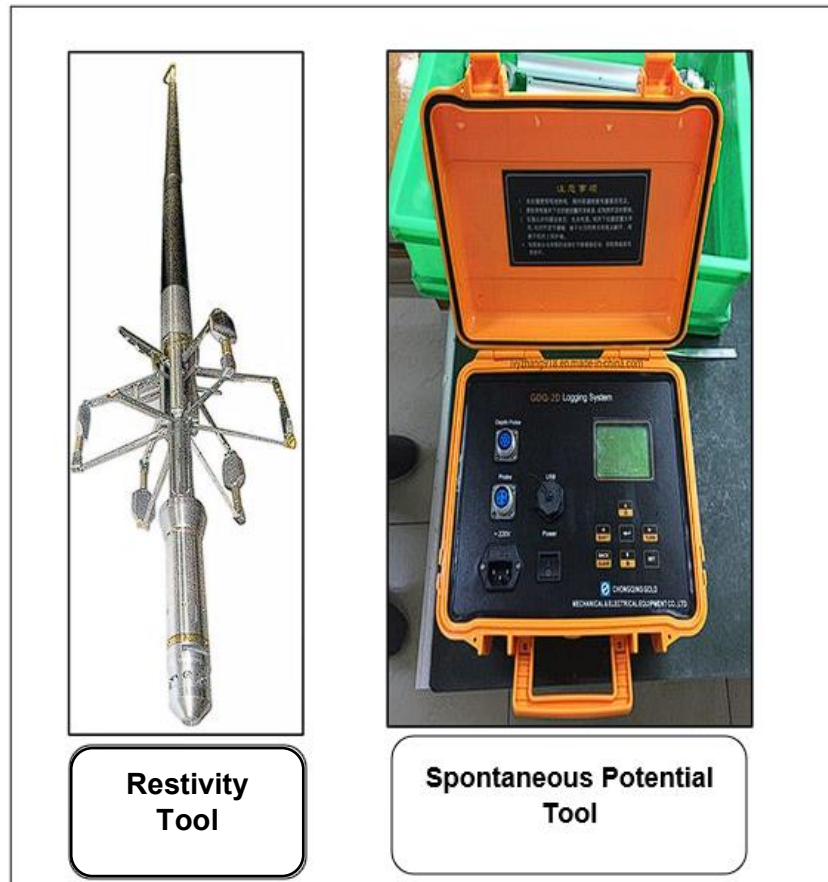


Figure 14. Electrical Logs Tools.

b) Mechanical Logs :

Caliper Log: which measures the diameter of the borehole and helps identify borehole conditions such as washouts or constrictions.



Figure 15. Caliper Logging Tool

c) Radioactive (Nuclear) Logs :

Gamma Ray Log: used to identify shale content and lithology.

Neutron Log: used to estimate formation porosity by detecting hydrogen content.

Density Log: which measures formation bulk density using gamma-ray scattering to evaluate porosity and lithology.

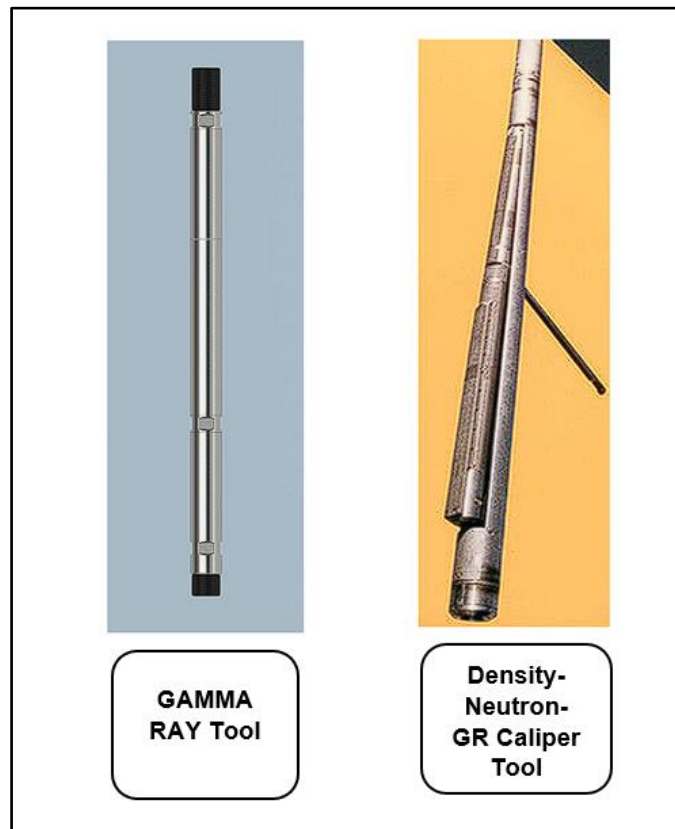


Figure 16. Radioactive Tools

d) Acoustic Logs :

Sonic Log : which measures the travel time of acoustic waves through the formation and is used to determine porosity and mechanical rock properties.



Figure 17. Acoustic Tool

1.2.2. Sonic Log

The idea of these observations depends on the transmission and reception of sound waves, the most important of which is the monitoring of the cement bond and the monitoring of the places where the packaging tubes are fetched Casing Collar Locator (Antwan, 1988).

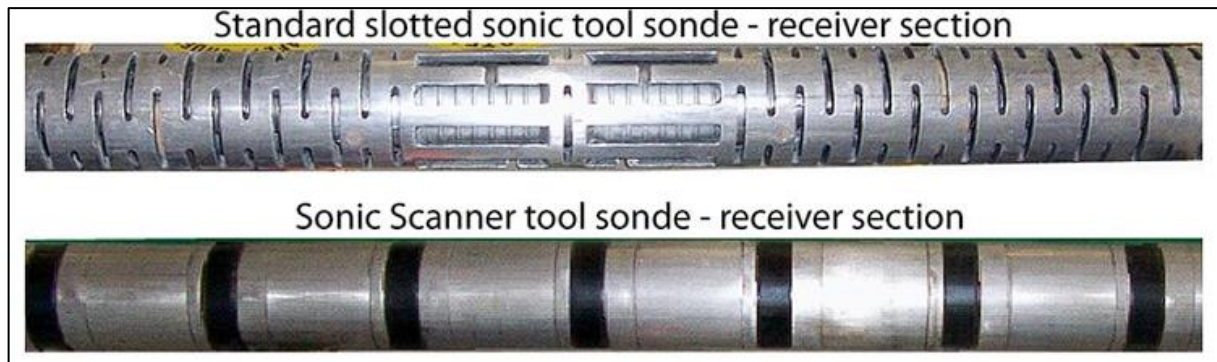


Figure 18. The solid tool body of the Sonic Scanner receiver section contrast with typical sonic logging tools (<https://csegrecorder.com/assets/images/articles/2009-05-sonic-fig11.jpg>)

a) Sonic uses

Provides a continuous record of seismic velocity and travel time along the borehole, which is useful for calibrating seismic data

- Supplies data for generating synthetic seismograms.
- Allows estimation of porosity, especially when combined with FDC and CNL logs.
- Supports stratigraphic correlation between formations.
- Helps identify lithology and facies.
- Detects fractures.
- Indicates compaction levels.
- Assists in recognizing overpressure zones.
- Contributes to the identification of source rocks.

b) Work Method

- transmitter: emits compressional sound waves into the formation.
- Receivers: detect these waves and measure the time taken to travel through the formation.
- travel time (Δt): is the difference in arrival time between receivers.

This travel time depends on:

- Lithology
- Porosity
- Pore fluid type

Consists of :

- 2 transmitters
- 4 receivers

Operation:

The first transmitter emits waves detected by two receivers.

The process is repeated with the second transmitter and the other two receivers.

The tool is positioned in the center of the borehole without contacting the walls to ensure accurate measurements.

2. Unconventional Method

Machine learning prediction is the process of applying trained models to infer unknown values by capturing patterns and relationships embedded in historical data. These models are capable of learning complex and non-linear interactions between variables, which allows them to generalize their knowledge to new, unseen data ([Springer, 2021](#)). In the case of sonic logs, this approach is particularly useful for estimating missing or unavailable measurements by exploiting the relationship between drilling parameters and acoustic properties, leading to more continuous and reliable subsurface characterization ([Maldonado-Cruz et al., 2023](#); [Santos et al., 2022](#)).

To predict sonic log values from drilling parameters, a structured methodology was adopted which are as follows :

- Data collection.
- Data analysis.
- Statistical study.
- Database separation.
- Selection of the algorithms.
- Model evaluation.

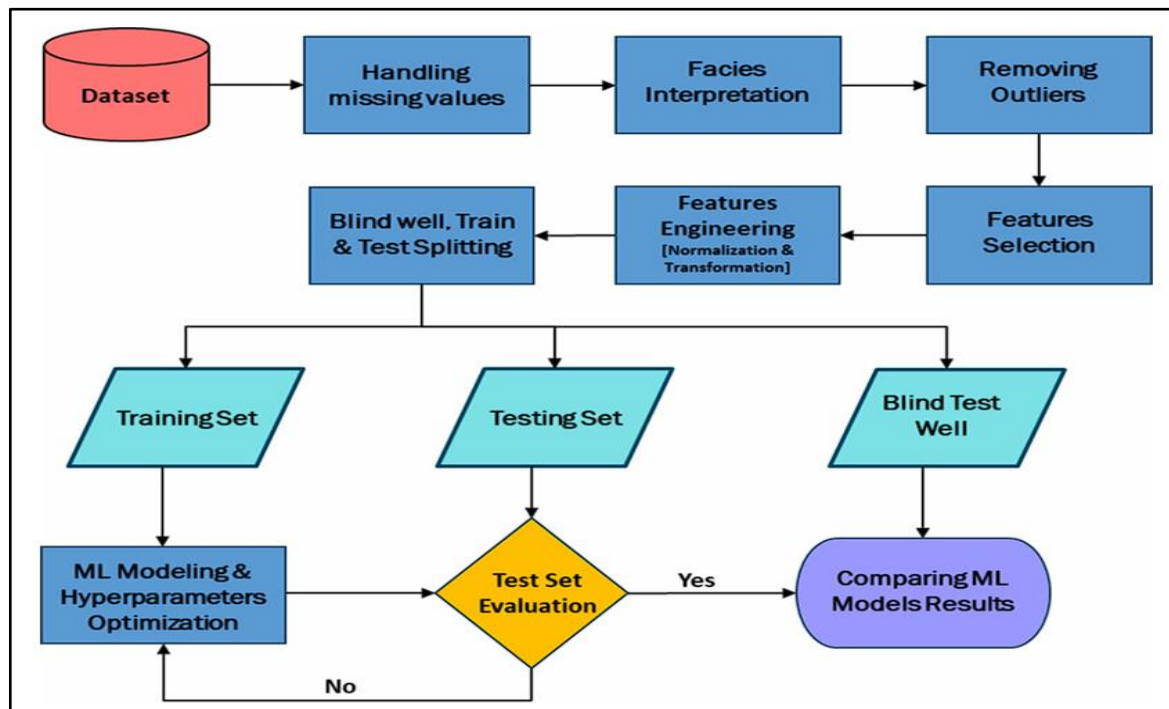


Figure 19. Machine Learning Workflow for compressional sonic log prediction.

2.1. Data Collection

The data used in this study were collected from the Oued Zine field located in the Sbaa Basin, Algeria. Two categories of data were used: drilling data and well log data. The drilling dataset consisted of operational parameters recorded continuously during drilling operations, while the well log dataset included petrophysical measurements obtained from wireline logging tools.

Table 1. Drilling parameters.

Parameter	Unit	What it Means
Depth	ft or m	Represents the measured depth of the well during drilling operations. It is used to correlate drilling parameters with formation properties.
RPM T	rpm	Represents the rotational speed generated by the top drive or rotary table during drilling operations. It controls how fast the drill string rotates inside the wellbore.
Weight on Bit (WOB)	kN	The downward force applied on the drill bit to improve rock cutting efficiency during drilling.
Rotation per Minute (RPM)	rpm	Indicates how fast the drill string and drill bit are rotating while drilling.
Torque	kN·m	Represents the rotational force required to rotate the drill bit through the formation.
Rate of Penetration (ROP)	m/h or ft/h	Describes how quickly the drill bit advances through the rock formation.
Standpipe Pressure (SPP)	psi or bar	The pressure measured inside the drilling circulation system while pumping drilling fluid.
Flow In	L/min	The amount of drilling fluid entering the well during circulation operations.

Table 2.well log parameters.

Parameter	Unit	What it Means
Depth	ft or m	Represents the measured depth of the well during drilling operations. It is used to correlate drilling parameters with formation properties.
GR (Gamma Ray)	API	Measures the natural radioactivity of formations and is mainly used to identify shale content and lithology.
NPHI (Neutron Porosity)	% or v/v	Estimates formation porosity by measuring the hydrogen concentration within the formation fluids.
RHOB (Bulk Density)	g/cm ³	Measures the bulk density of the formation and is commonly used to evaluate porosity and lithology.
RT (True Resistivity)	ohm·m	Measures the true electrical resistivity of the formation and helps identify hydrocarbon-bearing zones.
CAL (Caliper Log)	in (inches)	Measures the diameter of the borehole and helps detect washouts or irregular wellbore conditions.

2.2. Data Analysis

Before applying machine learning algorithms, the collected drilling and logging data must first go through a pre-processing stage to improve its quality and reliability. In many cases, raw data may contain:

- missing values
- noisy measurements
- inconsistent records

That can reduce the accuracy of the prediction model. Therefore, data cleaning and normalization are commonly performed to make the different parameters more consistent and easier for the algorithm to interpret. It is also important to match the drilling parameters with the correct depth intervals so they accurately correspond to the available wireline log data. In addition, correlation analysis and feature selection are often used to identify the most relevant variables and remove unnecessary information. These steps help improve the efficiency and overall performance of machine learning models in petroleum engineering applications ([Wang, H., 2023](#); [Al-Mudhafar, W. J., 2020](#)).

2.3. The statistical study

The statistical study is an essential first step in analysing drilling and well log data before applying machine learning models. It helps to better understand how drilling parameters are related to the target response by exploring the data in a simple and structured way. Through basic statistical analysis such as correlations and data distribution, it becomes easier to identify important trends, detect inconsistencies, and select the most relevant variables that influence prediction accuracy. This step ensures that the dataset is reliable and well-prepared for building an effective sonic log or drilling rate prediction model ([Anemangely et al., 2018](#); [Sabah et al., 2019](#)).

➤ Monovariate Analysis :

Univariate (monovariate) analysis studies each variable separately to describe its distribution, central tendency, and variability. It is used to detect outliers and assess data quality before modeling. In drilling and machine learning applications, it helps

understand input uncertainty and prepare reliable datasets ([Wang et al., 2023](#); [Towards Data Science, 2020](#); [Filzmoser et al., 2009](#)).

For input features: Univariate analysis is used to examine each drilling parameter individually through statistical tools such as histograms, boxplots, and density plots in order to identify the distribution shape, detect outliers, and highlight possible data inconsistencies before modeling.

For output (sonic log): The same approach is applied to the sonic log values to understand their distribution behavior, which is essential for selecting appropriate scaling, normalization, or transformation techniques prior to machine learning implementation ([Filzmoser et al., 2009](#); [Towards Data Science, 2020](#))

➤ **Bivariate Analysis:**

Bivariate analysis studies the relationship between two variables—typically an input feature and an output response—to understand how they vary together in a dataset. In sonic log and petroleum data analysis, it is used to identify the most influential drilling and petrophysical parameters affecting the sonic response before modelling ([Khushaba et al., 2022](#); [William Paterson University, 2020](#); [Bahadori et al., 2018](#)).

Scatter plots are commonly used to visualize linear or non-linear relationships.

while correlation coefficients such as [Pearson](#) and [Spearman](#) help quantify the strength and direction of these relationships.

2.4. Data base separation

Database separation is a key step in machine learning where the dataset is divided into different subsets to ensure a reliable evaluation of the model. Typically, the data is split into a **training set**, which is used to fit the model, and a **testing set**, which is used to evaluate its ability to generalize to unseen data. A common practice is to use an **80/20 split**, although other strategies such as cross-validation or leave-one-out approaches are also applied in multi-well or complex datasets to improve

robustness and reduce bias in performance estimation (Alpaydin, 2021; DIVA Portal, 2020; BSR, 2022).

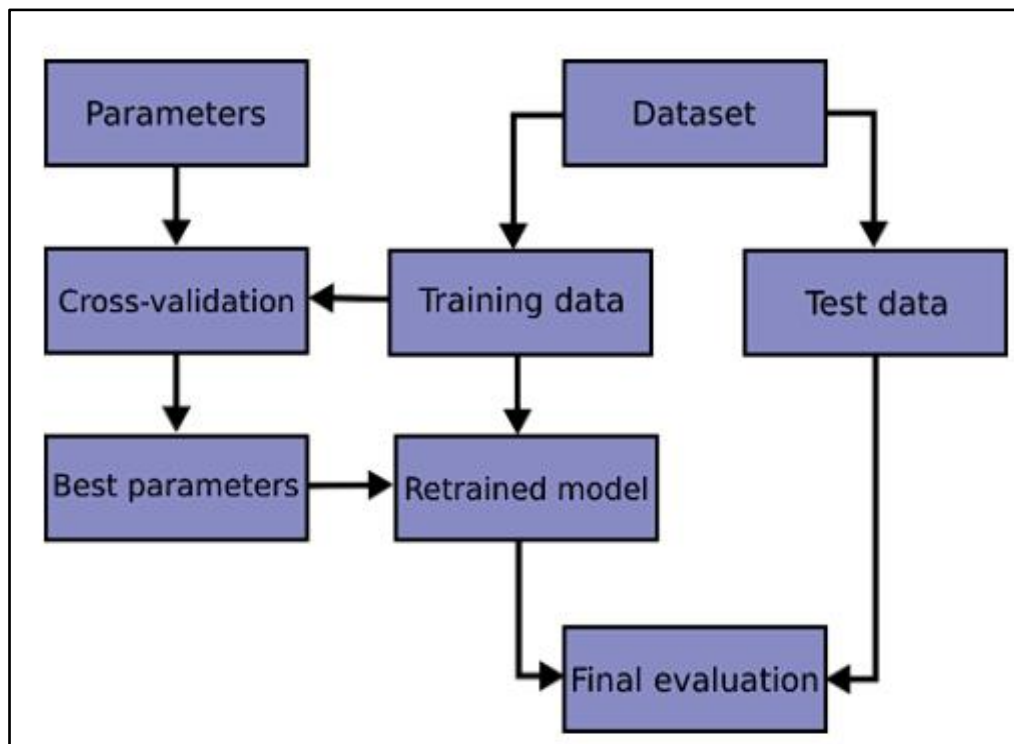


Figure 20. https://scikit-learn.org/stable/_images/grid_search_workflow.png

2.5. Selection of the algorithms

algorithm selection is an important step in machine learning because no single algorithm performs best across all datasets or problem types. Different algorithms have varying strengths depending on data characteristics and problem complexity, making selection necessary for optimal performance. (Rice, 1976; Kerschke et al., 2018).

The selection process is usually as follows:

- A set of candidate algorithms is defined according to the problem type.
- Each algorithm is trained and tested on the same dataset using a consistent evaluation protocol.
- Their performance is compared using validation methods such as cross-validation or holdout testing.
- The algorithm with the best generalization performance is selected for final use.

2.6. Model evaluation

a) Training

Common machine learning algorithms such as Random Forest, XGBoost, and Deep Neural Networks are widely used because they can capture complex relationships between drilling parameters and formation properties. During the training phase, the model learns iteratively from the training dataset by adjusting its internal parameters to minimize prediction errors and improve performance. Proper tuning and validation are essential to ensure stable convergence, reduce overfitting, and achieve reliable generalization on unseen data (Babaei et al., 2020; Džafić et al., 2022; Zhang et al., 2025).

b) Test

To know and measure performance we use the following metrics

RMSE (Root Mean Square Error): is a performance metric used in machine learning and regression analysis to measure the difference between predicted and actual values. It gives higher importance to large prediction errors, making it useful for evaluating the overall accuracy and reliability of predictive models (Chai and Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - Y_i)^2}, (0 \leq RMSE < +\infty)$$

MAE (Mean Absolute Error): is a commonly used evaluation metric in machine learning that measures the average absolute difference between predicted values and actual observations. Lower MAE values indicate better model accuracy and smaller prediction errors (Chai and Draxler, 2014).

$$MAE = \frac{1}{N} \sum_{i=1}^N |Y_i - X_i|, (0 \leq MAE < +\infty)$$

R² (Coefficient of Determination): is a statistical metric used to evaluate how well a machine learning or regression model explains the variability of the target data.

Higher R^2 values indicate that the model provides a better fit and can explain a larger proportion of the observed variation in the dataset (Chicco et al., 2021).

$$R^2 = \left[\frac{\frac{1}{N} \sum_{i=1}^N (XI) \cdot (\overline{XI})}{\sqrt{(XI - \overline{XI})^2} \cdot \sqrt{(YI - \overline{YI})^2}} \right]^2, (-1 < R < +1)$$

NSE (Nash–Sutcliffe Efficiency): is a statistical metric used to evaluate the predictive performance of regression and hydrological models by comparing predicted values with observed data. An NSE value closer to 1 indicates better model accuracy and stronger agreement between predictions and actual observations (Nash and Sutcliffe, 1970).

$$NSE = 1 - \left[\frac{\sum_{i=1}^N [XI - YI]^2}{\sum_{i=1}^N [XI - \overline{YI}]^2} \right], (-\infty < NSE \leq 1)$$

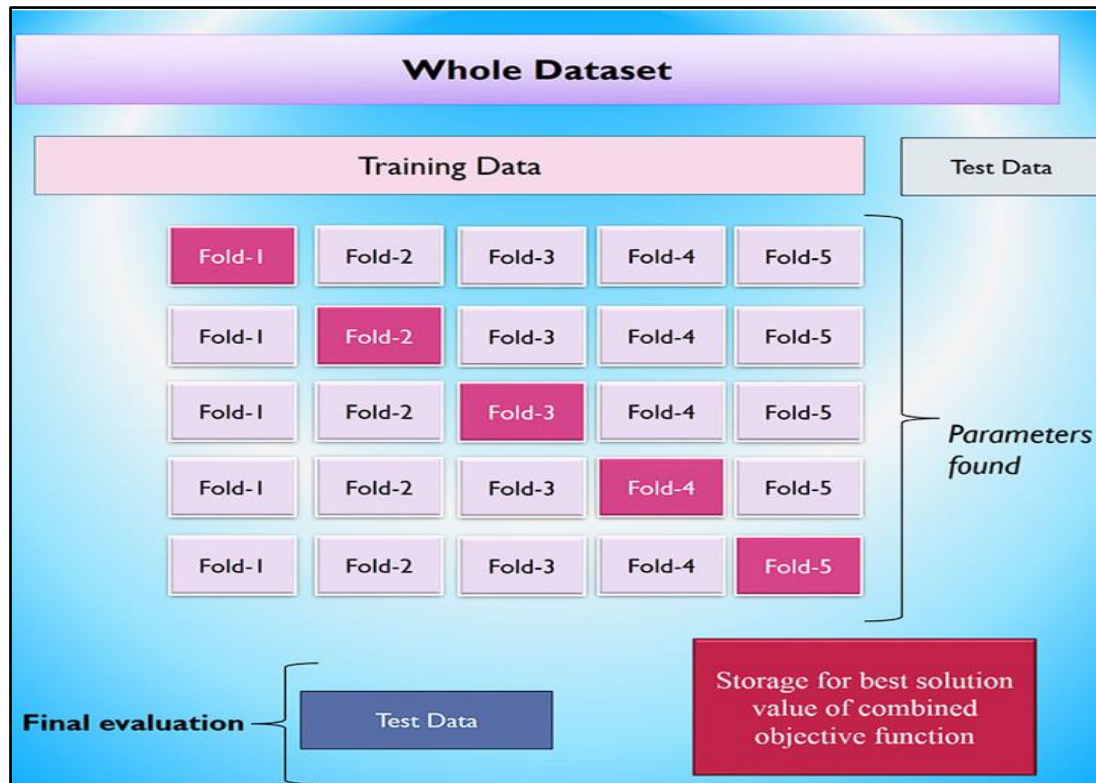


Figure 21. Schematic illustration of the k-fold cross-validation procedure used for model evaluation.

3. Materials use

3.1. Excel

In order to obtain data in an organized way, it was necessary to use Excel .We have corrected the missing, incomplete and illogical data of both drilling data and log data. And that in a very short time. These actions are necessary at the data processing stage before moving on to more advanced processing by Python using filters, formulas and simple schemes used.

3.2. Python

Python was used as the main programming language in this work because of its simplicity and efficiency in data analysis and machine learning applications. It provided an effective environment for processing drilling and well log data, cleaning the dataset, and developing predictive models for sonic log estimation. In addition, Python offers several powerful libraries such as Pandas for data handling and organization, NumPy for numerical operations, and Scikit-learn for implementing and evaluating machine learning models ([Adewumi et al., 2021](#)).

Conclusion

In this chapter, the available methods for measuring and determining sonic transit time (DT) were reviewed, ranging from conventional direct and indirect approaches to modern machine learning-based prediction techniques. The tools associated with each method were also presented and described, highlighting their role in obtaining DT values and supporting formation characterization and evaluation.

With regard to the present study, this chapter focused on the methodology adopted, as well as the tools and software used for data processing and analysis. In addition, the main steps involved in data preparation, predictive model development, and performance evaluation were presented. These elements provided the methodological foundation for the DT prediction tasks carried out in this work and serve as a basis for the presentation and discussion of the results in the following chapter.

CHAPTER IV: Application of Machine Learning for Sonic Log Estimation

Introduction

This part focuses on the application of machine learning for sonic log estimation using both drilling data and well log data together. Unlike many previous studies that rely on a single type of dataset, this work combines multiple sources of field data to improve the quality and reliability of the prediction.

The study follows a structured workflow where the data is first prepared and analysed, then used to build and train different machine learning models. These models are tested and their results are compared to identify the most accurate approach for predicting sonic logs.

This integrated approach provides a more realistic representation of subsurface conditions and improves the accuracy of the estimated results for petroleum reservoir analysis.

Data Collection and Analysis

To begin the process of predicting the values of the sound recording using machine learning, we collected drilling parameter data from the wad Zein area. These continuous measurements were recorded both during the drilling process as well as the well log measurements during and after drilling. After obtaining the measurements data, we evaluated the dataset for completeness and accuracy with the aim of removing any missing or Invalid records. This was done using Excel. After completion and confirmation of the validity of the data, it was transferred to an Excel file for simple processing of the following processing steps. The integration of drilling data with well logs will allow for better accuracy in the correlation of drilling parameters with depth because it provides a more reliable Geological reference to help build subsequent machine learning models.

1. Results Using Drilling Data

1.1. Methodology and results

1.1.1. Statistical study

A. Monovariate Analysis

Table 3. Monovariate analysis Using Drilling Data.

	DT	DEPTH	ROP	WOB	WOH	RPM	RPM T	Flow IN	SPP	Torque
Count	187.000	187.000	187.00000	187.00000	187.00000	187.000000	187.000000	187.000000	187.000000	187.000000
Mean	61.531620	2383.000	4.534225	3.97861	3.97861	72.689840	381.914439	865.459893	2819.320856	2774.82352
Std	5.542409	54.126	2.981359	1.94261	1.94261	12.832672	269.409396	75.981479	176.559910	914.81580
Min	55.668914	2290.000	0.400000	0.00000	0.00000	37.000000	49.000000	609.000000	1793.000000	1749.00000
25%	58.364083	2336.5000	2.400000	3.00000	3.00000	63.000000	88.000000	866.000000	2716.000000	1953.50000
50%	60.008567	2383.000	3.500000	4.00000	4.00000	66.000000	610.000000	866.000000	2804.000000	2220.00000
75%	63.616967	2429.500	5.750000	5.00000	5.00000	86.500000	628.000000	867.000000	2935.500000	3684.00000
Max	89.549700	2476.000	13.500000	8.00000	8.00000	101.000000	663.000000	1846.000000	3243.000000	4408.00000

The dataset contains 187 observations, which is a reasonable number for training a machine learning model. All key drilling parameters such as WOB, RPM, Torque, and ROP are included and show some variability in their values.

Torque has the most variable values (1749 to 4408), indicating a significant variability of drilling conditions at varying depths. The ranges for ROP and WOB also indicate non-constant values for these parameters during fluid operations. SPP has a value range of 1793 to 3243, which indicates a significant amount of variability in pressure conditions occurring at different depth intervals. Variability caused by drilling operations is expected in real-world drilling operations because it provides data for the learning of the model.

The range of depth for this dataset is from 2290 to 2476 feet, representing the section of drilling operations for which data has been recorded. Continuous recording of drilling parameters has resulted in the data set being able to closely represent the actual formation conditions. Data parameters such as Flow IN have little variation and appear to have mostly consistent hydraulic conditions during drilling.

B. Bivariate Analysis

DEPTH & DT: A highly negative correlation is present between these variables (CC = -0.62). As DEPTH increases, DT will decrease. Within the remaining two variables, SPP (CC = -0.57) and RPM T (CC = -0.54), moderate negative correlations are also shown. For both ROP (CC = 0.33) and Flow IN (CC = 0.30), DT has a weak positive correlation.

WOB & Torque: Moderate to weak correlations are exhibited between WOB and most of the other variables. The two variables showing the strongest correlation (moderate to weak) were Torque and ROP.

SPP: A moderate negative correlation exists between SPP and DT (CC = -0.57) indicating that increased fluid flow is associated with increased surface pressure. All other correlations are relatively weak.

Torque: A moderate positive correlation exists between Torque and WOB indicating that an increase in WOB is associated with higher values of Torque. A similar positive correlation exists for WOB and ROP; all other correlations were relatively weak.

ROP: A weak positive correlation exists between ROP and DT (CC = 0.33), while a weak negative correlation exists between ROP and DEPTH. Correlation values were within the weak to moderate correlation range for all other variables.

RPM: Moderate positive correlations exist between RPM and WOB and Torque, while all other correlations are in the weak to moderate correlation category.

the correlation matrix reveals weak to moderate correlations among most variables. The most notable relationships are the strong negative correlation between DT and DEPTH (CC = -0.62), the moderate negative correlations between DT and both SPP

and RPM T, and the positive correlation between SPP and Flow IN which reflects a logical physical relationship. In contrast, ROP shows weak correlations with all other parameters, suggesting that it is governed by combined or non-linear effects rather than simple linear relationships.

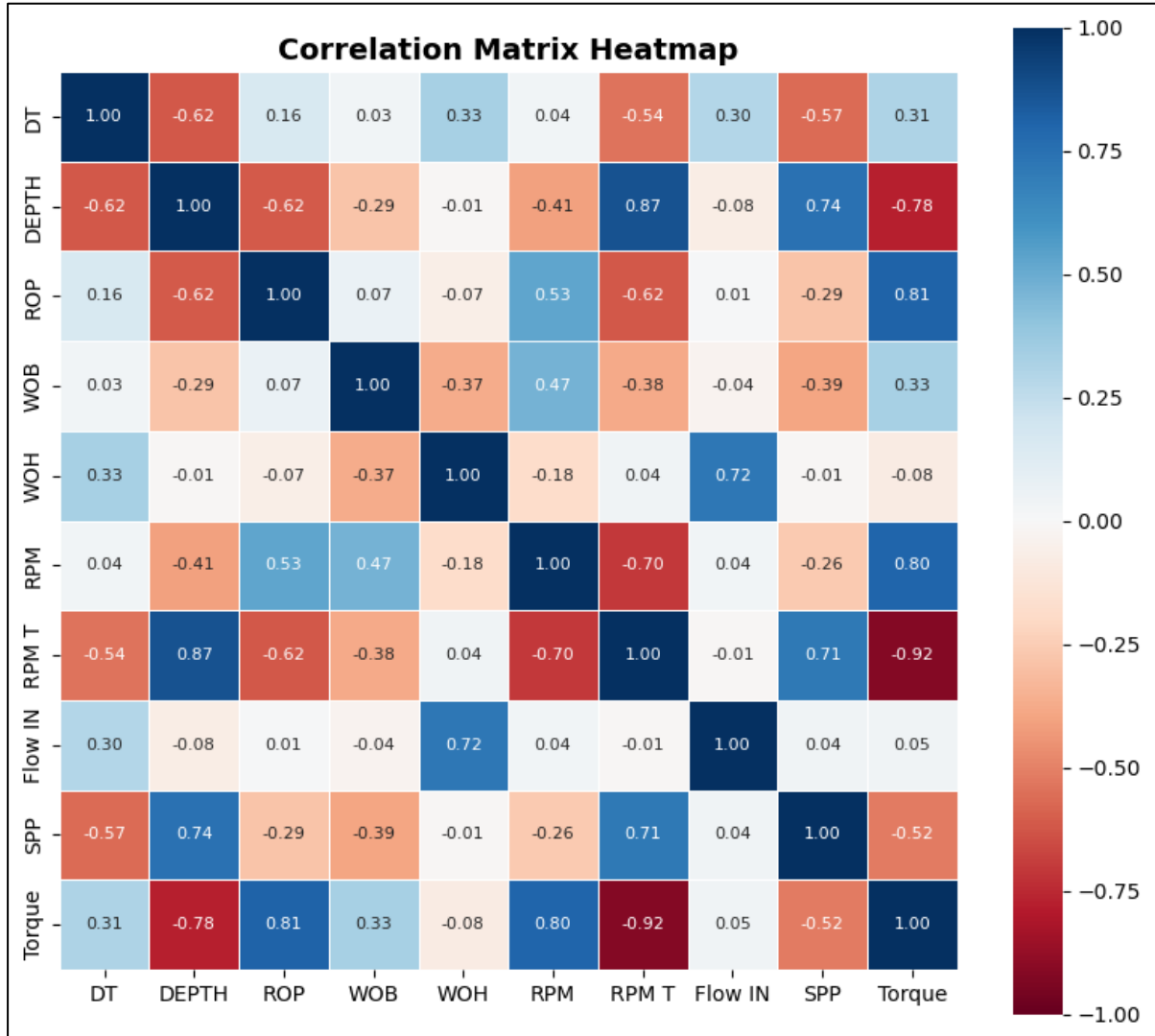


Figure 22. Bivariate analysis (correlation)

1.2. Prediction**1.2.1. Machine Learning model's implementation****Table 4.** Utilized Models.

Combination	Input Variables	Output
	Drilling Data	
1	DEPTH,ROP,WOB,WOH,RPM,RPM T,Flow IN,SPP,Torque	Sonic
2	DEPTH,ROP,WOH,RPM,RPM T,Flow IN,SPP,Torque	Sonic
3	DEPTH,ROP,WOH,RPM T,Flow IN,SPP,Torque	Sonic
4	DEPTH,WOH,RPM T,Flow IN,SPP,Torque	Sonic
5	DEPTH,WOH,RPM T ,SPP,Torque	Sonic

This table shows the groups of variables used in acoustic prediction, where a set of cases from 1 to 5 was obtained, and one or more variables were omitted at each step. The main or first group consists of: Depth, ROP, WOB, WOH, RPM, Flow In, and Torque.

This approach was adopted to evaluate the influence of each variable on Sonic prediction results and to identify the variables with limited or no impact on prediction performance. The comparison between the different cases also helped simplify the prediction process while maintaining reliable results.

1.2.2. Performance evaluation

Table 5. Machine Learning model results Of Drilling Data.

ML Models		Training set				Testing Set	
	R ²	RMSE	MAE		R ²	RMSE	MAE
				Drilling Data			
RF 1	0.9813	0.7657	0.5371		0.8959	1.6857	1.1408
RF 2	0.9798	0.7954	0.5470		0.8925	1.7131	1.1524
RF 3	0.9796	0.7994	0.5509		0.8933	1.7066	1.1617
RF 4	0.9787	0.8167	0.5577		0.8996	1.6554	1.1063
RF 5	0.9804	0.7850	0.5480		0.9048	1.6121	1.1062
XGB 1	1.0000	0.0031	0.0022		0.9250	1.4311	1.0677
XGB 2	1.0000	0.0037	0.0026		0.9249	1.4324	1.0584
XGB 3	1.0000	0.0042	0.0029		0.9242	1.4386	1.0481
XGB 4	1.0000	0.0093	0.0066		0.9259	1.4220	0.9968
XGB 5	1.0000	0.0092	0.0065		0.9238	1.4428	1.0134
SVM 1	0.0814	5.3680	2.5087		0.1994	4.6754	2.2304
SVM 2	0.0845	5.3591	2.5042		0.2023	4.6668	2.2235
SVM 3	0.0897	5.3438	2.4964		0.2076	4.6512	2.2109
SVM 4	0.0992	5.3158	2.4811		0.2179	4.6211	2.1812
SVM 5	0.1013	5.3095	2.4778		0.2198	4.6155	2.1773

Cross Plots

Cross Plots - Case 1

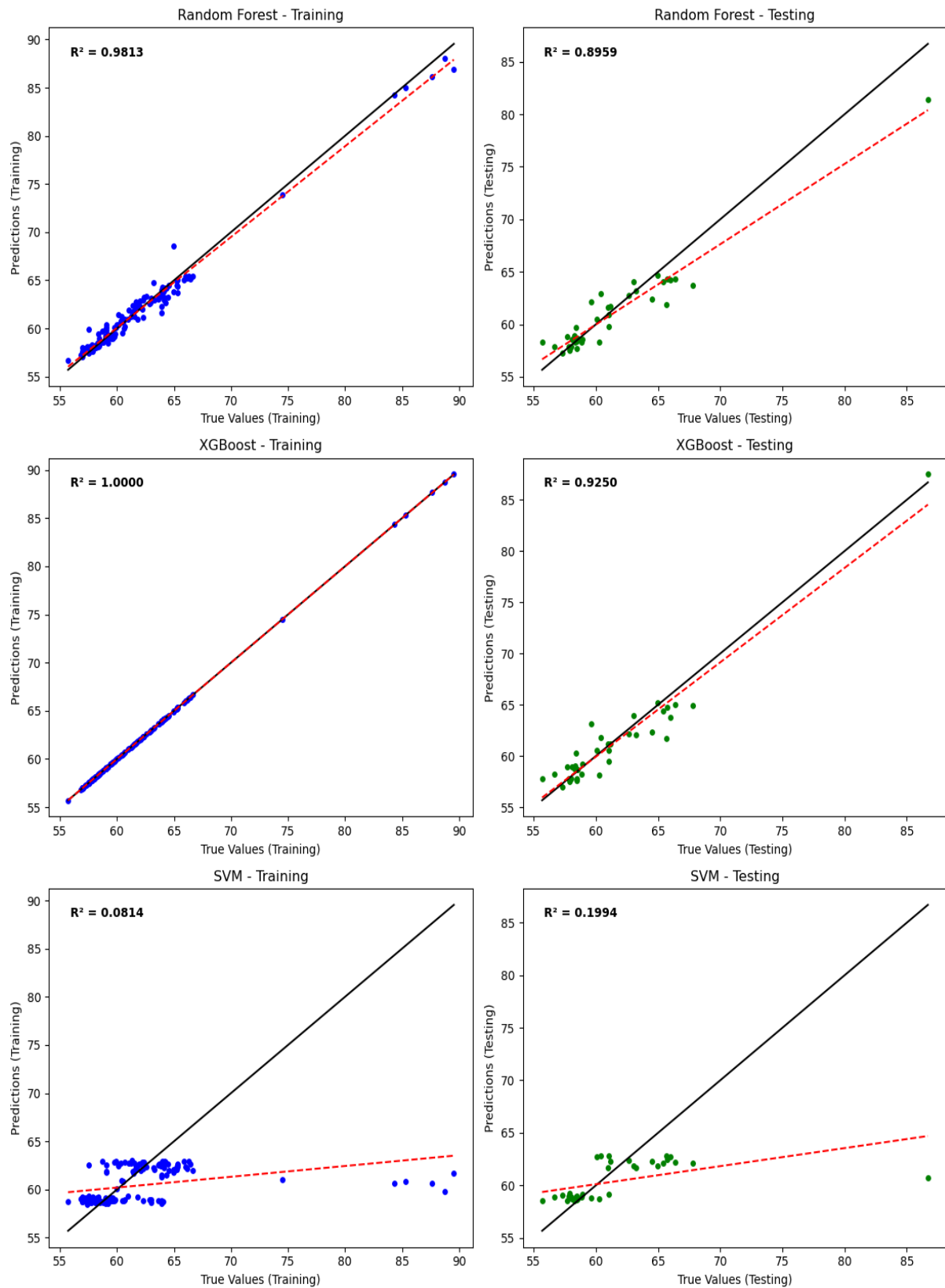


Figure 23. Graphical results scenario 1

Cross Plots - Case 2

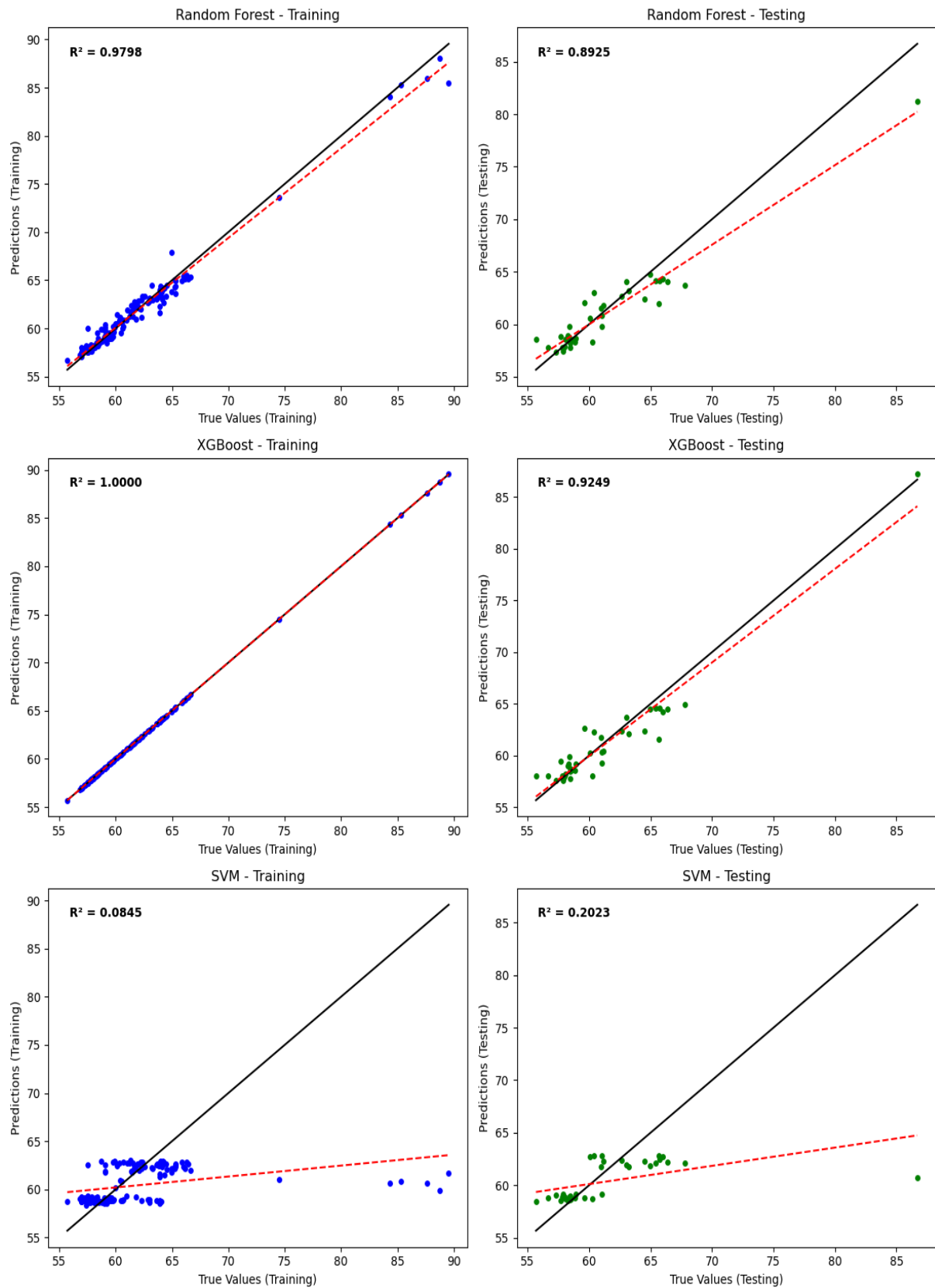


Figure 24. Graphical results scenario 2

Cross Plots - Case 3

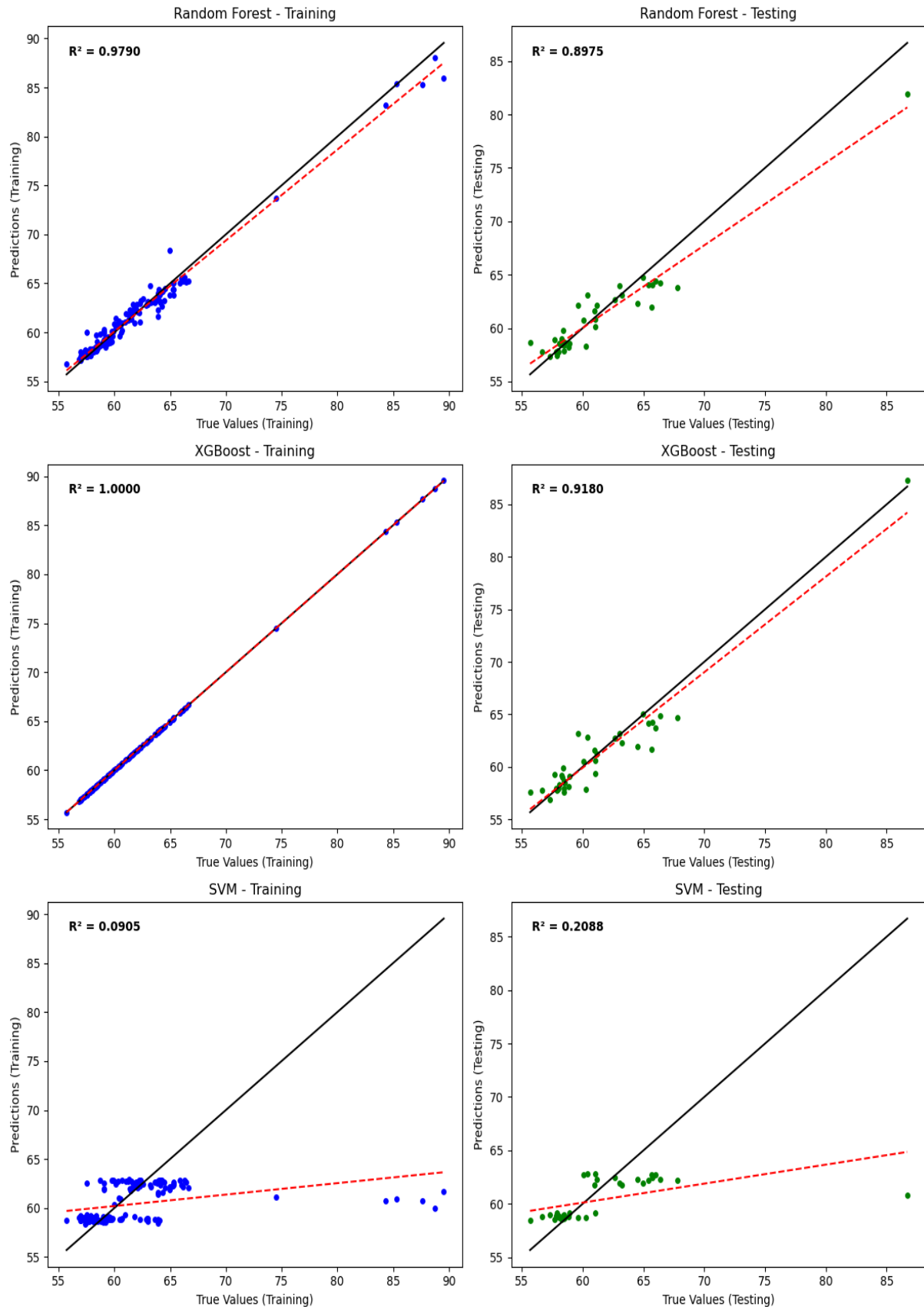


Figure 25. Graphical results scenario 3

Cross Plots - Case 4

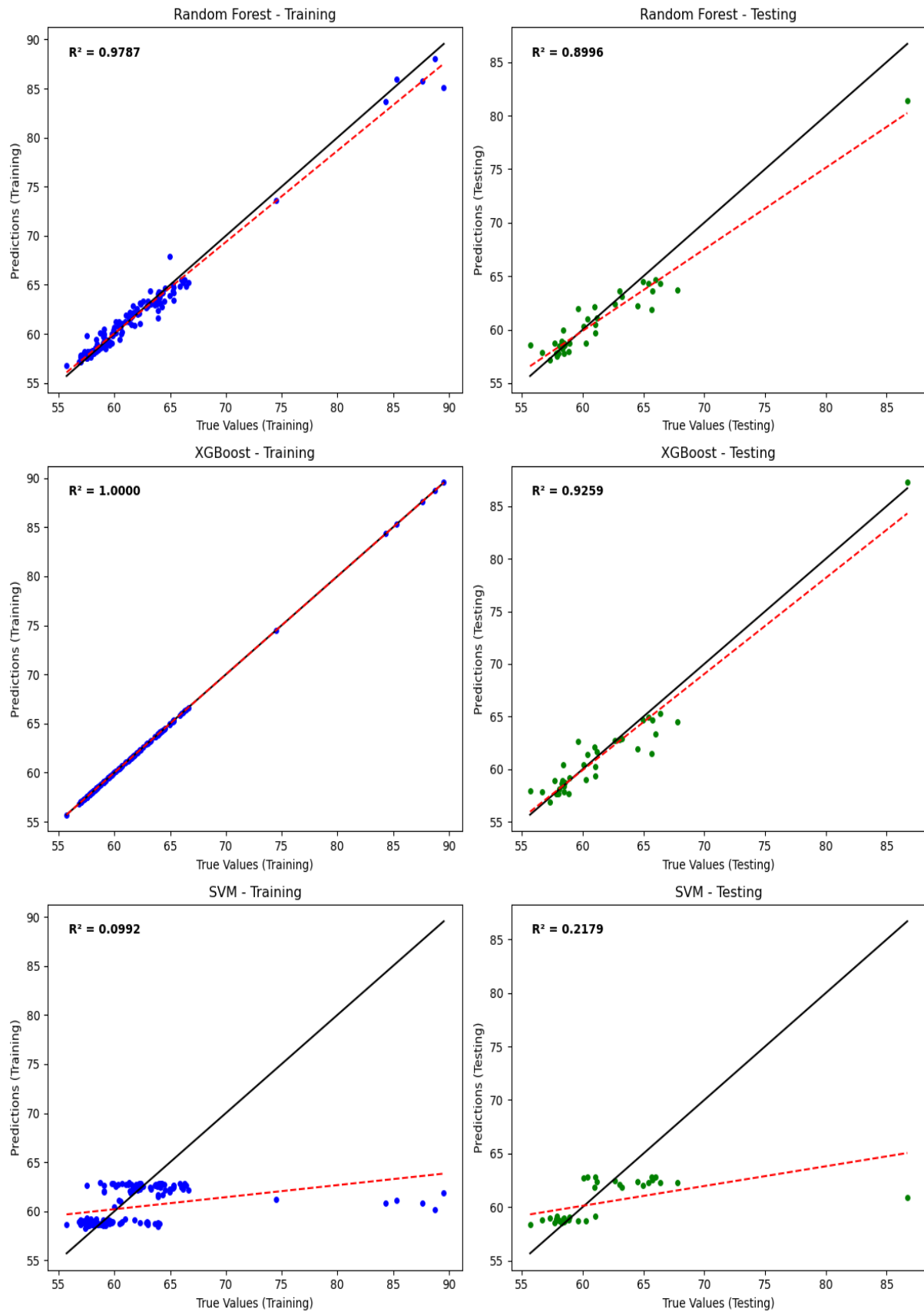


Figure 26. Graphical results scenario 4

Cross Plots - Case 5

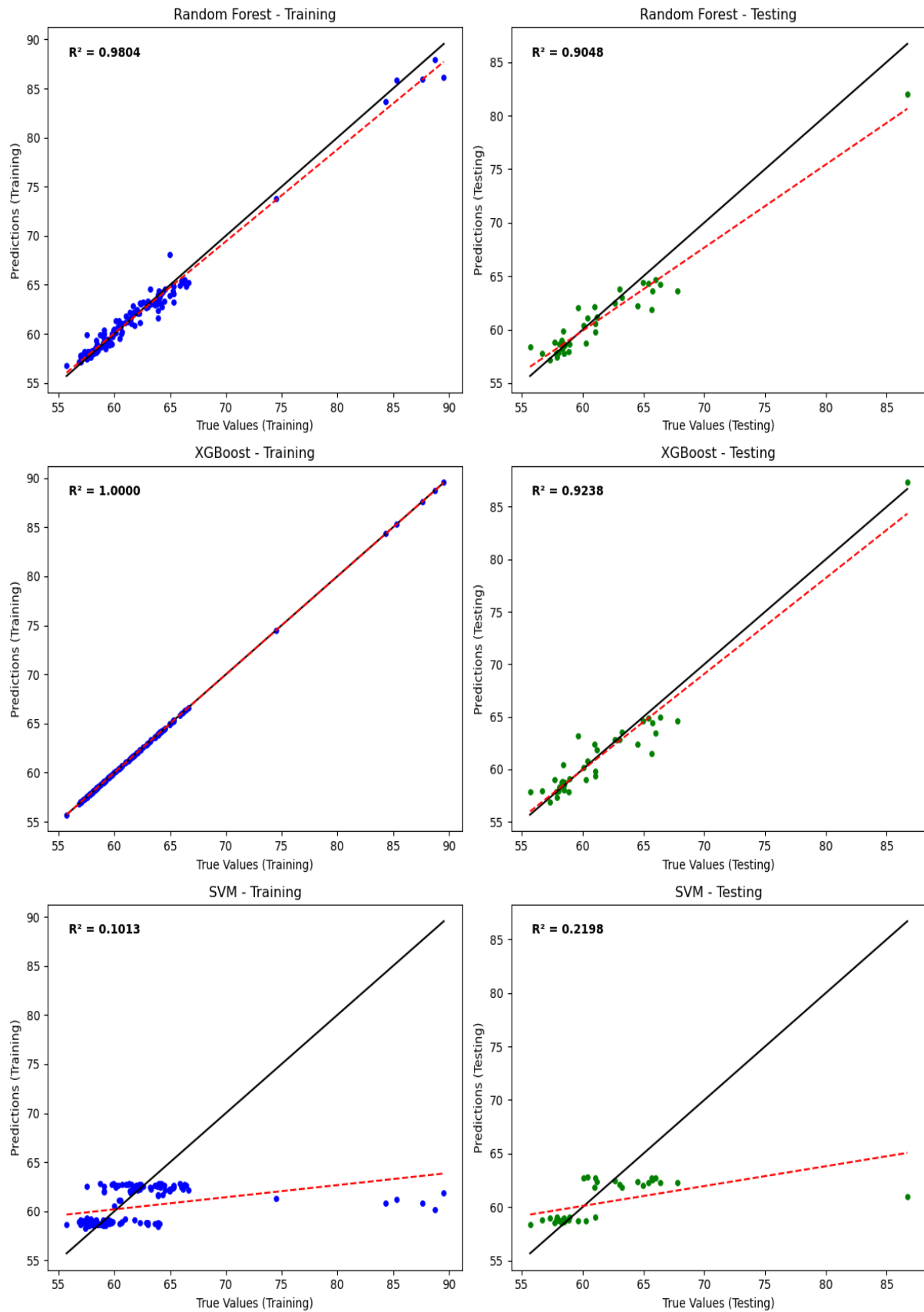


Figure 27. Graphical results scenario 5

1.3. Discussion

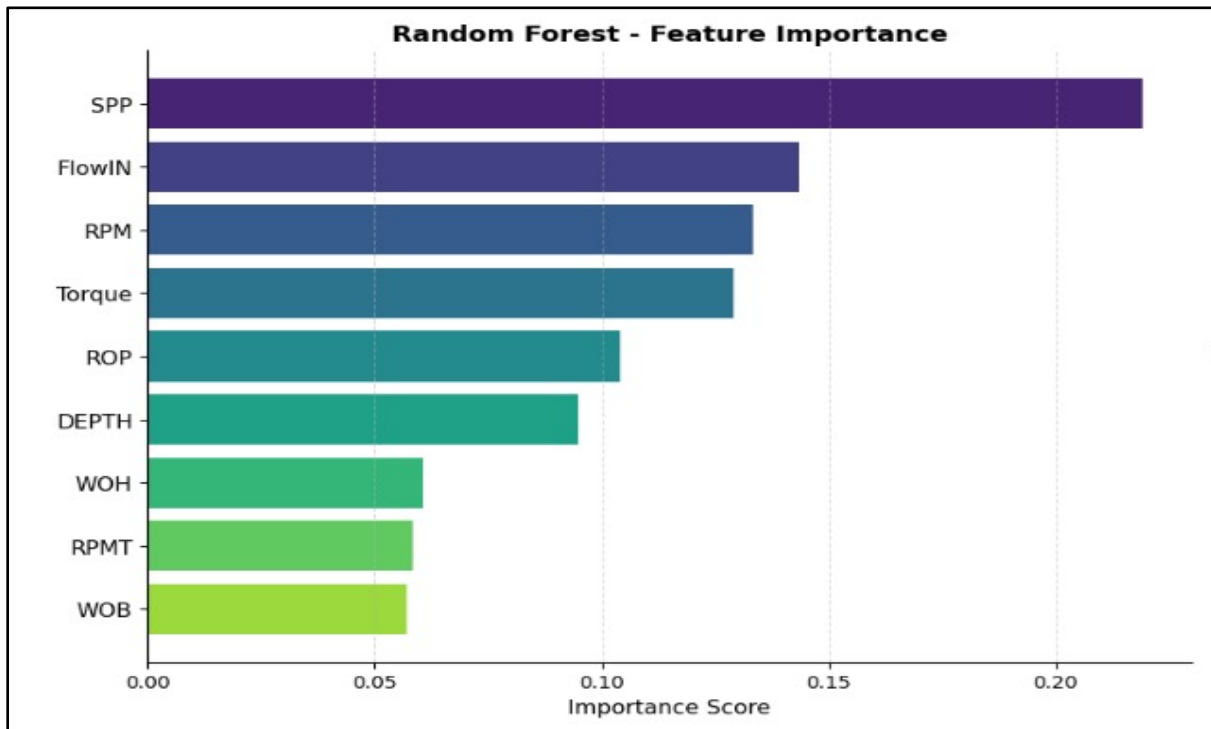
The previous results show a recent study in which we compared three machine learning the previously mentioned drilling data to predict the values of the acoustic signal. The models were as follows: Random Forest, XGBoost, Support Vector Machine.

Random Forest (RF): The test results show a high similarity with the training results for R^2 , where the differences were noticeable between 0.08 and 0.09. For both RMSE and MAE, there was a clear increase in error; the closest RMSE results were around 1.61, and for MAE around 1.10. The best and closest results were obtained by RF5. Overall, these results reflect the stability of the Random Forest model and its good ability to generalize in data prediction.

XGBoost (XGB): we observe perfect performance in training results, where R^2 is equal to 1, indicating an excellent fit, with slight variations compared to testing results, where R^2 remains stable around 0.923–0.926. As for RMSE and MAE, there is a noticeable increase in error between training and testing, indicating a degradation in performance and a decrease in the model's ability to generalize. The RMSE increases from values as low as 0.0031–0.0093 in training to around 1.42–1.44 in testing, while MAE shows a similar trend around 0.996–1.067. Overall, the best results among the models were obtained by XGB4, showing the best balance between accuracy and generalization.

Support Vector Machine: SVM has a low performance level based on the training results obtained with an R^2 value around 0.08–0.10, indicating a weak fit of the model to the data. A slight improvement is observed in the test results, where R^2 increases to approximately 0.20–0.22. In both training and testing, RMSE and MAE show a moderate decrease from training to testing, indicating a small improvement in generalization, although the overall error remains relatively high. The RMSE decreases from approximately 5.31–5.36 in training to around 4.61–4.67 in testing, while MAE also shows a similar decreasing trend from about 2.47–2.50 to around 2.17–2.23. SVM shows lower predictive performance compared to the other models, despite a slight improvement in the testing phase.

feature importance

**Figure 28.** Random Forest feature importance

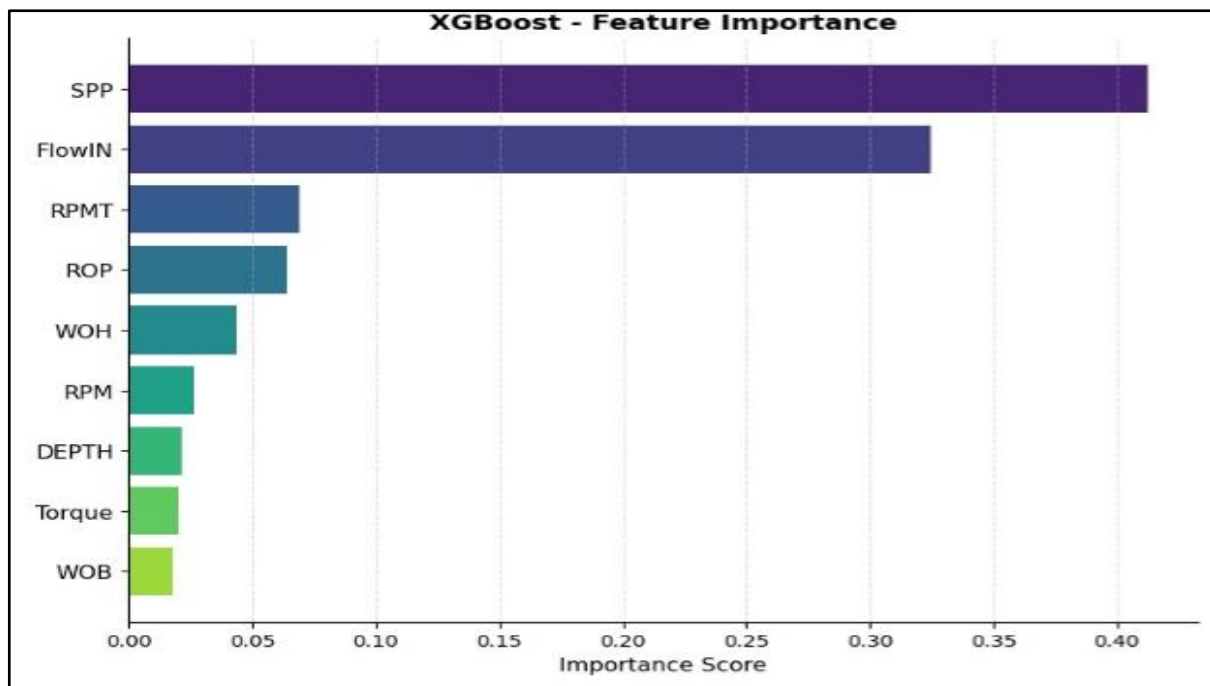


Figure 29. XGBoost feature importance

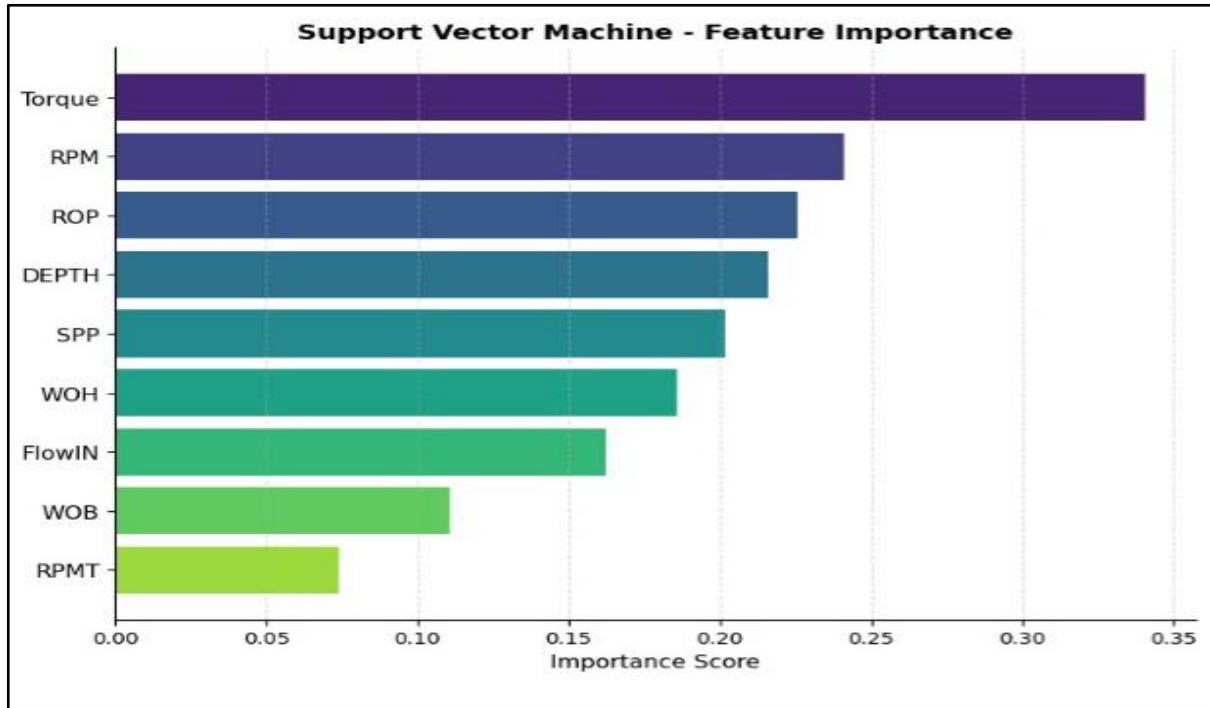


Figure 30. Support Vector Machine feature importance

From the feature importance chart of the Random Forest model, it is evident that Standpipe Pressure (SPP) and Flow In represent the most influential variables in predicting the sonic log, as they contribute significantly to improving the model's predictive accuracy and overall performance. In contrast, the remaining parameters such as Hook Load (WOH), Rotations Per Minute at the top drive (RPM T), and Weight on Bit (WOB) exhibit lower levels of importance, with WOB showing the weakest contribution among all the considered variables.

On the other hand, the results of the XGBoost model clearly and explicitly highlight the most influential variables, with significant differences in their levels of importance. Standpipe Pressure (SPP) emerges as the most dominant feature, followed by Flow In with a noticeable margin. Beyond these two variables, the remaining parameters exhibit relatively weak influence on the model's predictions, particularly Torque and Weight on Bit (WOB), which show the lowest levels of contribution among all features.

In contrast, the results of the SVM model show that all variables contribute more equally compared to the other two models. However, there are still some differences between the features. Torque appears as the most influential variable, followed by

RPM. WOB comes before last in terms of influence, while ROP represents the least influential feature in the model, although it still shows a noticeable level of contribution.

We can observe some common findings among the results, as WOB was one of the least influential variables in all three models. Therefore, it can be assumed that, with further development and improvements of the models, it may be possible to exclude this variable while maintaining the accuracy of the results, or even achieving more accurate predictions.

2. Results Using Well Log Data

2.1. Statistical study

A. Monovariate Analysis

Table 6. Monovariate analysis Using Well Log Data.

	DEPTH	GR	NPHI	RHOB	RT	CAL	DT
Count	1208	1208	1208	1208	1208	1208	1208
Mean	2383.349789	35.535814	0.03091	2.547443	526.428314	5.957328	61.337639
Std	53.725181	30.134775	0.044792	0.064972	606.977299	0.081218	5.408623
Min	2289.9624	14.5637	0.0002	2.3637	4.325	5.7591	52.1974
25%	2337.7779	22.479325	0.0058	2.49575	83.3556	5.9075	58.153675
50%	2383.7646	27.1875	0.01735	2.5635	190.98285	5.9399	59.90375
75%	2429.7513	35.847675	0.042425	2.5958	882.1808	5.972575	63.001825
Max	2475.738	325.6337	0.3101	2.7654	1950	6.3024	92.3668

There are 1208 observations in the dataset, which can be considered as a reasonable sample size for training an ML model. It includes all key logging and drilling related parameters such as DEPTH, GR, NPHI, RHOB, RT, CAL, DT, etc. that each exhibit some level of variability.

DEPTH varies between approximately 2289 and 2476 which represents the drilled length interval that is contained in the dataset. GR varies significantly from about 14.56 to 325.63. This indicates changes in lithological characteristics as you move through the wellbore. RT also exhibits a high degree of variability from 4.32 to 1950, which represents a significant change in the resistivity of formations.

Parameters such as RHOB and CAL demonstrate more moderate levels of variability than GR and RT. NPHI shows low levels of variability in general, with limited dispersion. DT, which is the target variable that will be used to predict future outcomes, ranges approximately from 52 to 92 reflecting a varying amount of acoustic response across the formations. Overall, the dataset adequately reflects realistic variations that exist in the subsurface therefore making it a valuable resource for building predictive ML models.

B. Bivariate Analysis

DEPTH and RT: There is a strong negative correlation between DEPTH and RT (CC = -0.69). As DEPTH becomes larger, RT becomes smaller. The other variables also display relatively moderate to strong correlations with DEPTH; the exceptions are NPHI (CC = -0.62) and DT (CC = -0.60), both exhibiting moderate negative correlation, and RHOB (CC = 0.58), exhibiting moderate positive correlation. Both GR (CC = -0.19) and CAL (CC = 0.23) exhibit weak correlations with DEPTH.

GR and NPHI: Moderate positive correlations are observed between GR and both NPHI and DT, while the correlations of the remaining variables with GR are considered weak.

RT: RT has a moderate negative correlation with RHOB, indicating that the intervals of maximum density values also have a minimum resistivity. Weak positive correlations are observed between RT and both NPHI and DT, while the remaining correlations are weak.

RHOB: A moderate positive correlation exists between RHOB and DEPTH, indicating that increasing depth is associated with higher bulk density values. In contrast, RHOB shows a moderate negative correlation with RT, while all remaining correlations are weak to moderate.

DT: STRONG positive correlation between DT and NPHI, and moderate negative correlation between DT and DEPTH. Correlations between DT and all other variables range from weak to moderate.

NPHI: Moderate positive correlations exist between NPHI and both GR and RT, while a strong positive correlation is observed with DT. In contrast, weak negative correlations are shown with RHOB and CAL.

In general, the correlation matrix shows weak to moderate correlations between the majority of the variables investigated. Of interest is the strong negative correlation between RT and DEPTH (CC = -0.69), and the moderate negative correlation between DEPTH and both NPHI and DT, along with the strong positive correlation between NPHI and DT. Thus, this represents a logical petrophysical correlation. In comparison, CAL has weak correlations with the majority of the parameters, indicating that the effects of borehole enlargement on the other logging variables is essentially independent from those of borehole enlargement.

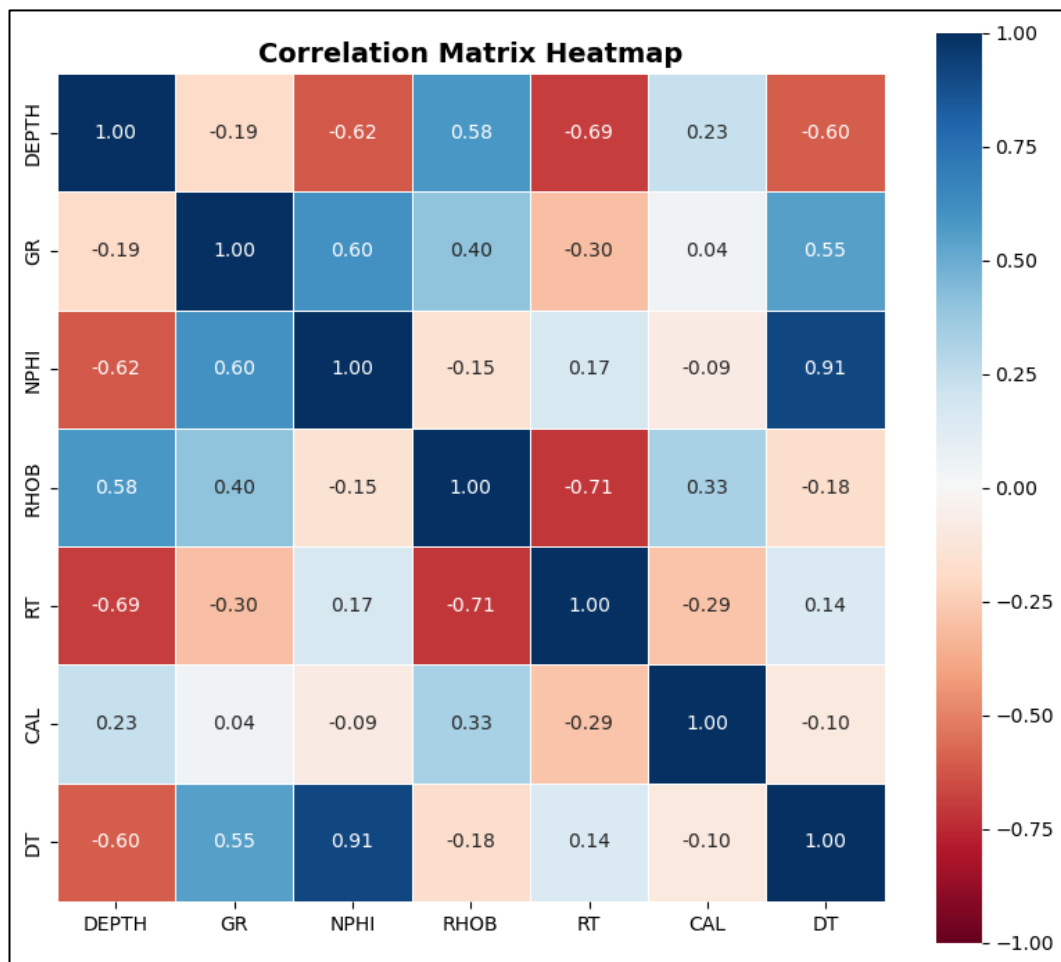


Figure 31. Bivariate analysis (correlation)

2.2. Prediction**2.2.1. Machine Learning model's implementation****Table 7. Utilized Models**

Combination	Input Variables	Output
	Well Log Data	
1	DEPTH,GR,NPHI,RHOB,RT,CAL	Sonic
2	DEPTH,GR,NPHI,RHOB,RT	Sonic
3	DEPTH,GR,RHOB,RT	Sonic

This table shows the groups of variables used in sonic log prediction, where a set of cases from 1 to 3 was obtained, and one or more variables were omitted at each step. The main or first group consists of: DEPTH, GR, NPHI, RHOB, RT, and CAL.

This approach was adopted for the same reasons as the previous time. Identification of variables with limited or no effect on forecasting performance. The comparison of different cases also helped to simplify the forecasting process while maintaining reliable results.

2.2.2. Performance evaluation

Table 8. Machine learning model results Of Drilling Data.

ML Models		Training set				Testing Set	
	R ²	RMSE	MAE		R ²	RMSE	MAE
				Well Log Data			
RF 1	0.989285	0.585369	0.398118		0.89223	1.400828	1.071473
RF 2	0.989035	0.592161	0.402341		0.886194	1.43952	1.106611
RF 3	0.989586	0.577086	0.394171		0.895098	1.382061	1.055886
XGB 1	0.992787	0.48027	0.329467		0.9105	1.27658	0.955662
XGB 2	0.999354	0.143775	0.103593		0.888659	1.423848	1.060801
XGB 3	0.99929	0.150649	0.107321		0.89328	1.39399	1.068638
SVM 1	0.997322	0.29266	0.191949		0.874754	1.510145	1.141908
SVM 2	0.992274	0.497062	0.323853		0.878548	1.487094	1.115697
SVM 3	0.007945	5.632504	2.82886		0.109352	4.027068	2.28115

Cross Plots

Cross Plots - Case 1

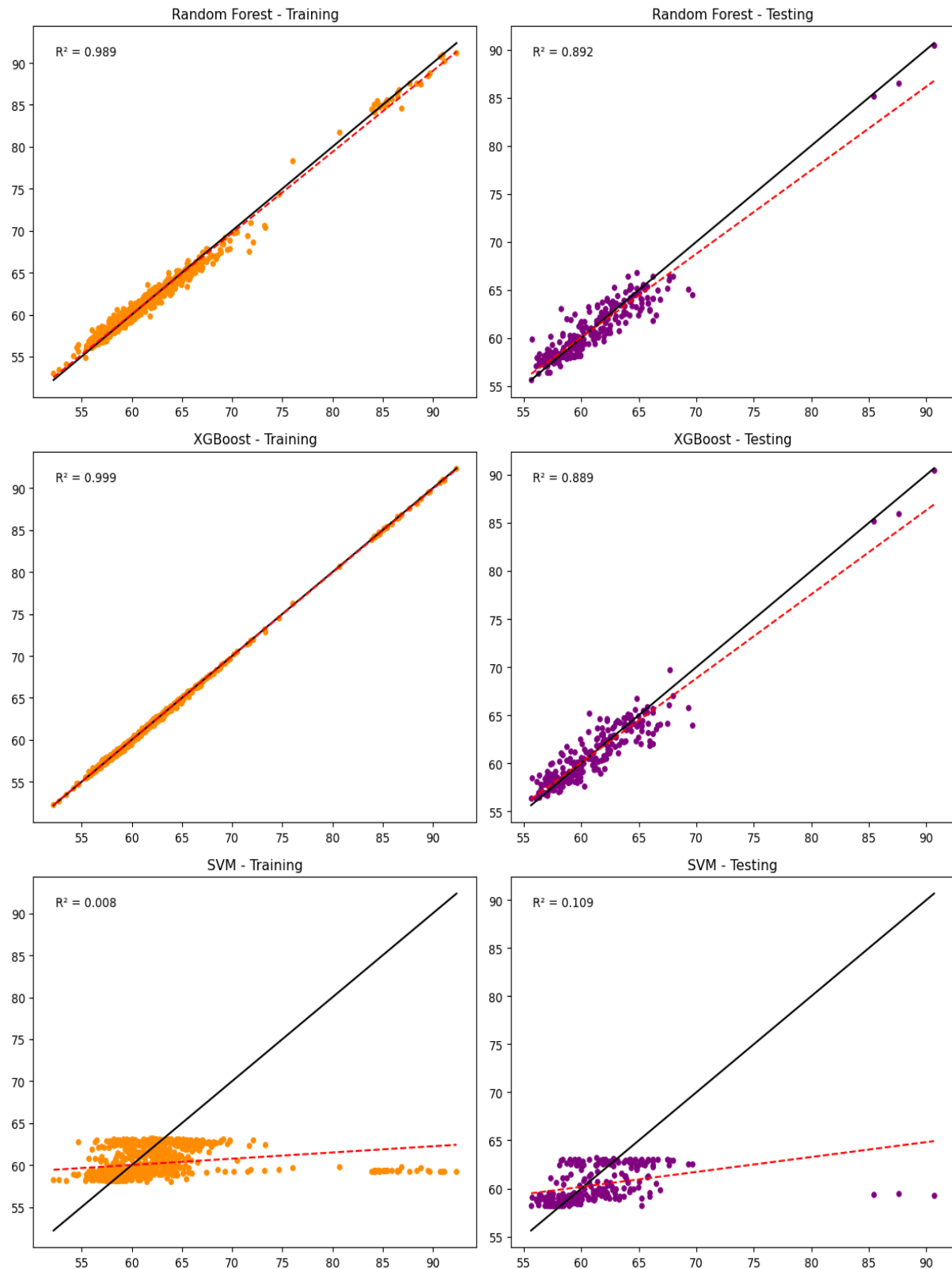


Figure 32. Graphical results scenario 1

Cross Plots - Case 2

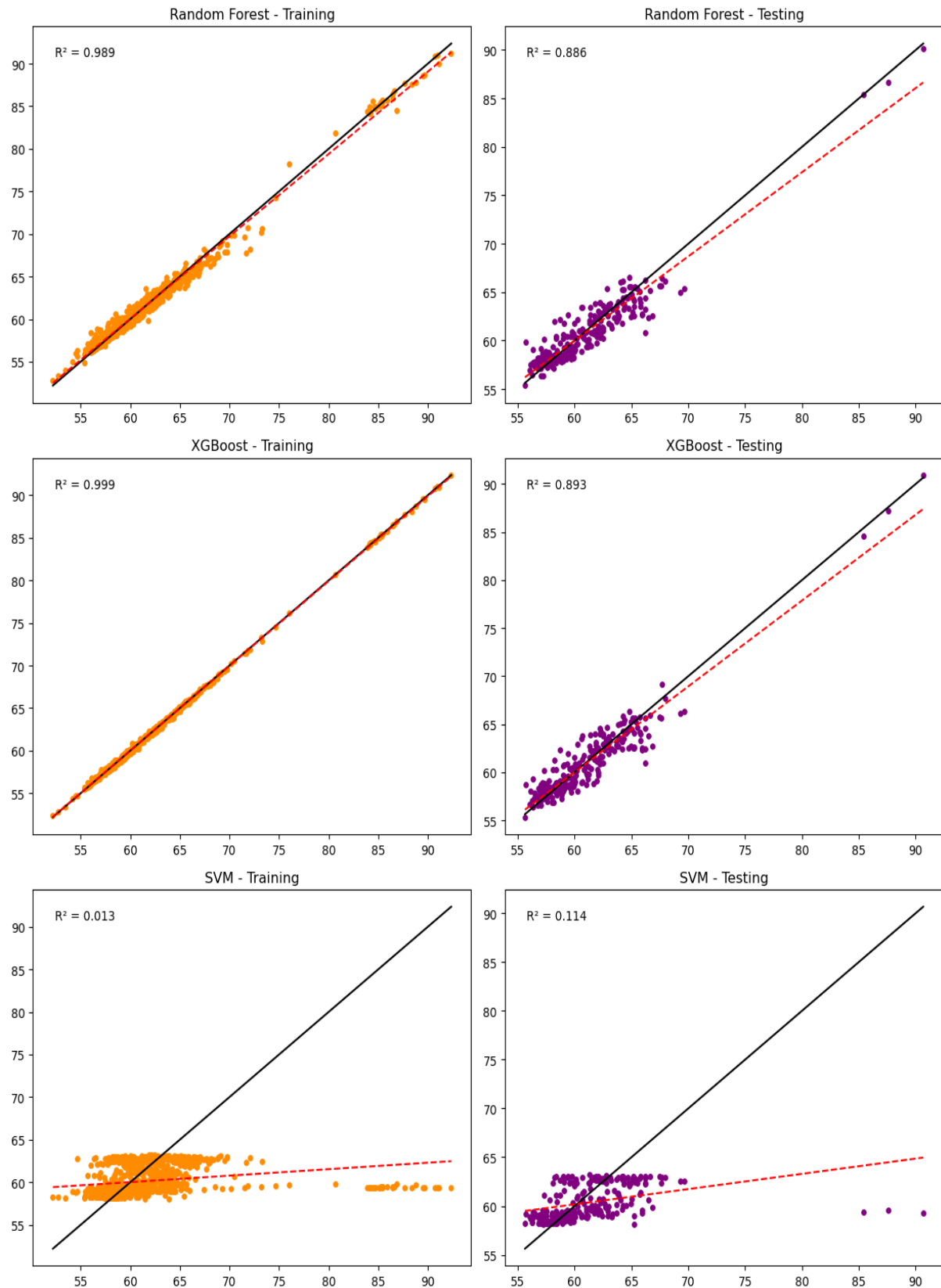


Figure 33. Graphical results scenario 2

Cross Plots - Case 3

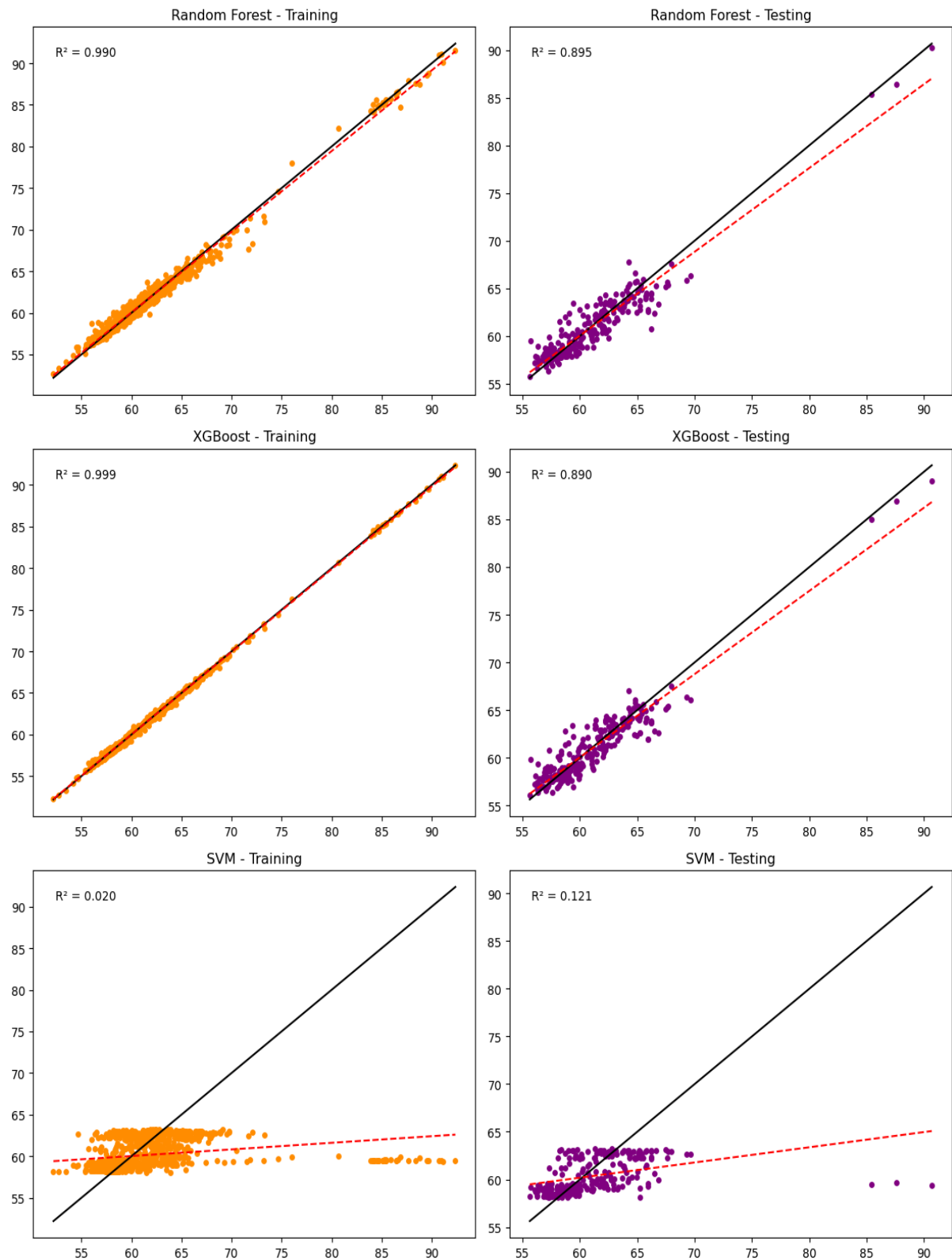


Figure 34. Graphical results scenario 3

2.3. Discussion

After completing the experiment by drilling the data and we moved to the same experiment with the same approach and steps that recorded by the Well Log data, the same machine learning machinery was used to predict which was represented by: Random Forest, Xgboost, Vector Machine Support.

Random Forest: In the training data, the model performed extremely well with the R^2 consistently around the 0.989 marking indicating a very strong fit and the model is well-trained from one iteration to the next. However, there was a marked decline in test data performance with the tested R^2 value range being around 0.886 to 0.895 – exhibiting an acceptable level of generalisation . There was also a noted increase in both RMSE and MAE metrics from training to testing, with RMSE increasing from about 0.57-0.59 to somewhere in the range of 1.38-1.44, while MAE increased from roughly 0.39-0.40 to around 1.05-1.10. Overall RF3 consistently achieves superior performance; providing the best balance between prediction accuracy and error levels on unseen data.

XGBoost (XGB): The model, had the capability to fit quickly with high R^2 values between about 0.992–0.999 to the training dataset, as shown through their high accuracy associated with pattern learning. There was also a clear drop in R^2 values from the training to testing phase (testing values of approx 0.889–0.910) indicating a generalising fall-off between training and testing performance. In addition, the RMSE and MAE metrics produce much higher values from the training error compared with the testing dataset (RMSE values of approx 0.14–0.48 to approx 1.27–1.42 on unseen data; MAE values of approx 0.10–0.33 to approx 0.95–1.07). Therefore, based on performance, the XGB configuration demonstrated the best/balanced performance for achieving the greatest overall accuracy and generalisation.

Support Vector Machine: The models have produced inconsistent results across the runs during training, resulting in strong results (R^2 close to 0.99, SVM1 and SVM2) or worse results (SVM3 with $R^2 < 0$), showing the models are not properly interpreting the underlying patterns. The models' performance stabilizes during testing, although with lower values ($R^2 = 0.10$ to 0.88), indicating weak to moderate predictive ability depending upon configuration. Error metrics (e.g., RMSE and MAE) increase greatly

when moving from training to testing. For example, RMSE increases from approximately 0.29–0.49 (up to 5.63 in the worst case) during training to 1.48–1.51 during testing, while MAE increases from approximately 0.19–0.32 (up to 2.82) during training to 1.11–1.14 during testing. Overall, SVM2 represents the best configuration among configurations tested, exhibiting the most stable performance on previously unobserved data.

2.4. Future importance

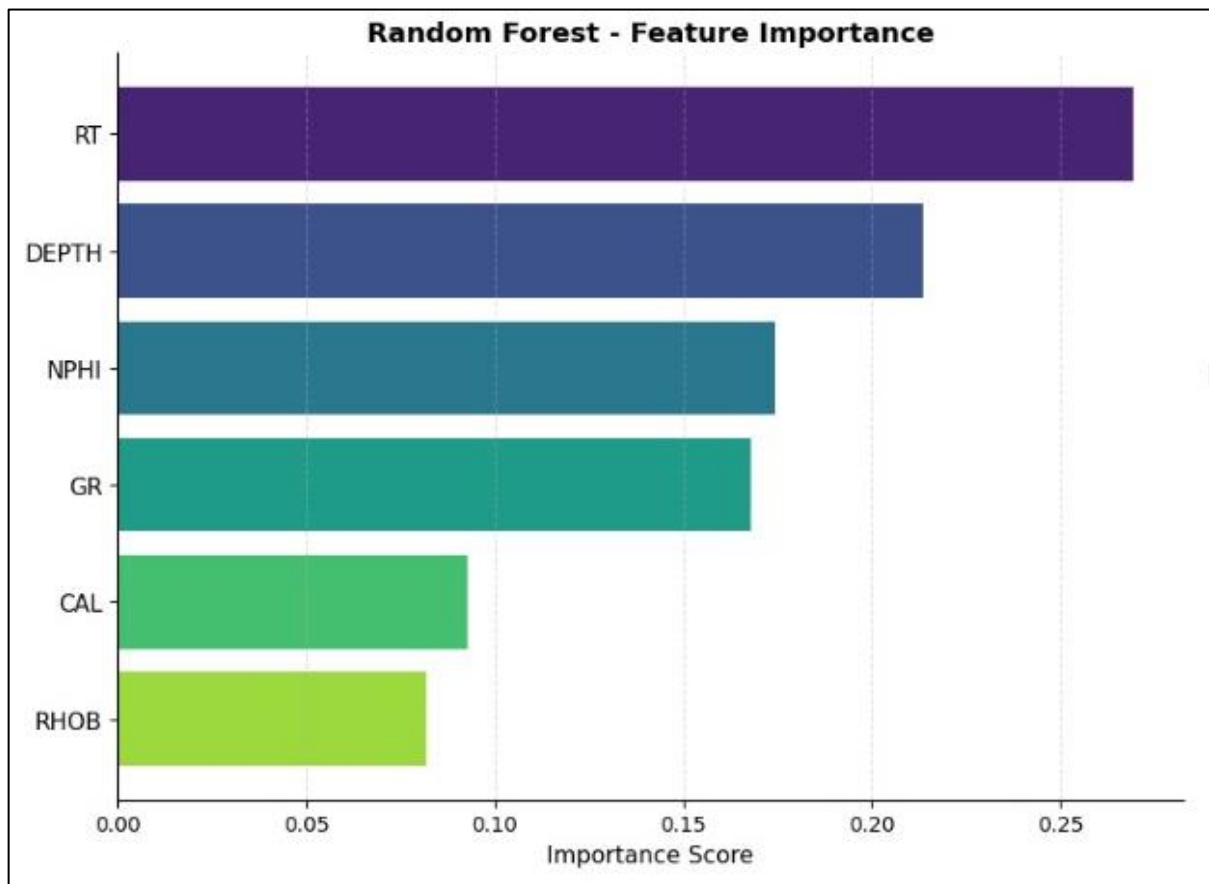


Figure 35. Random Forest feature importance

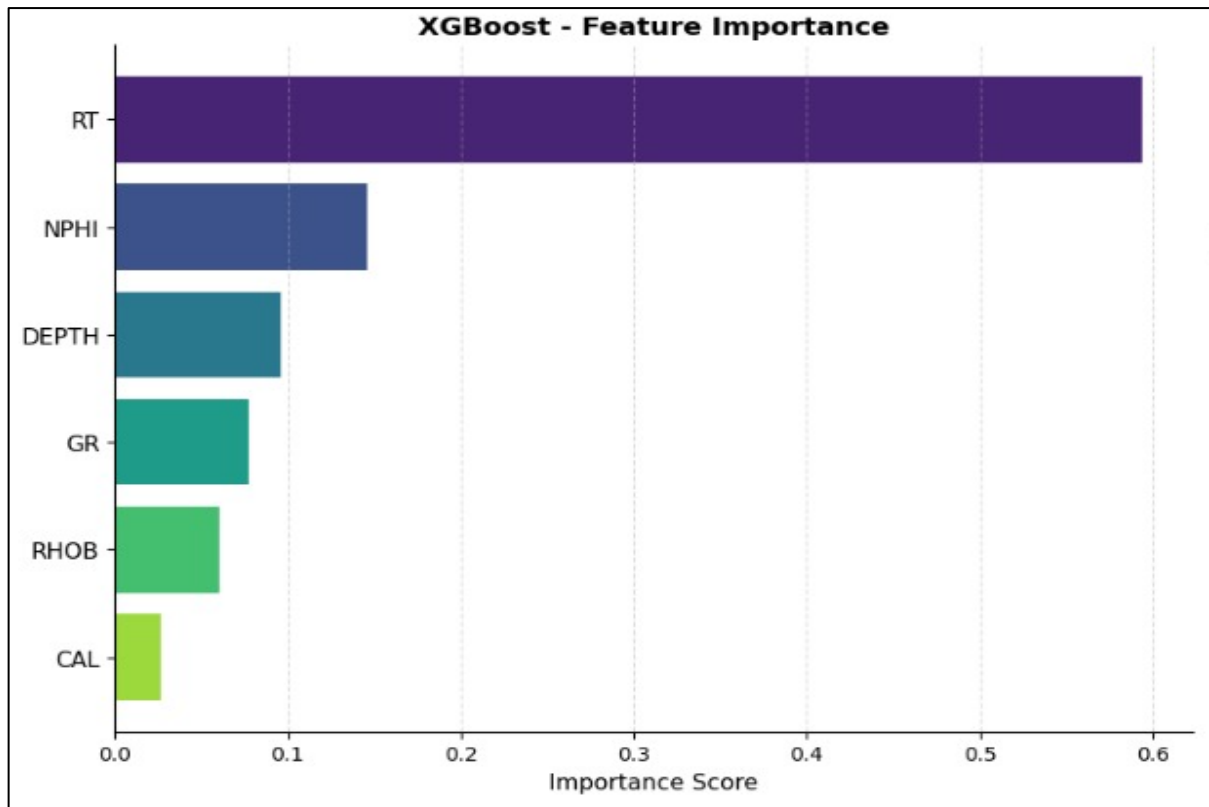


Figure 36. XGBoost feature importance

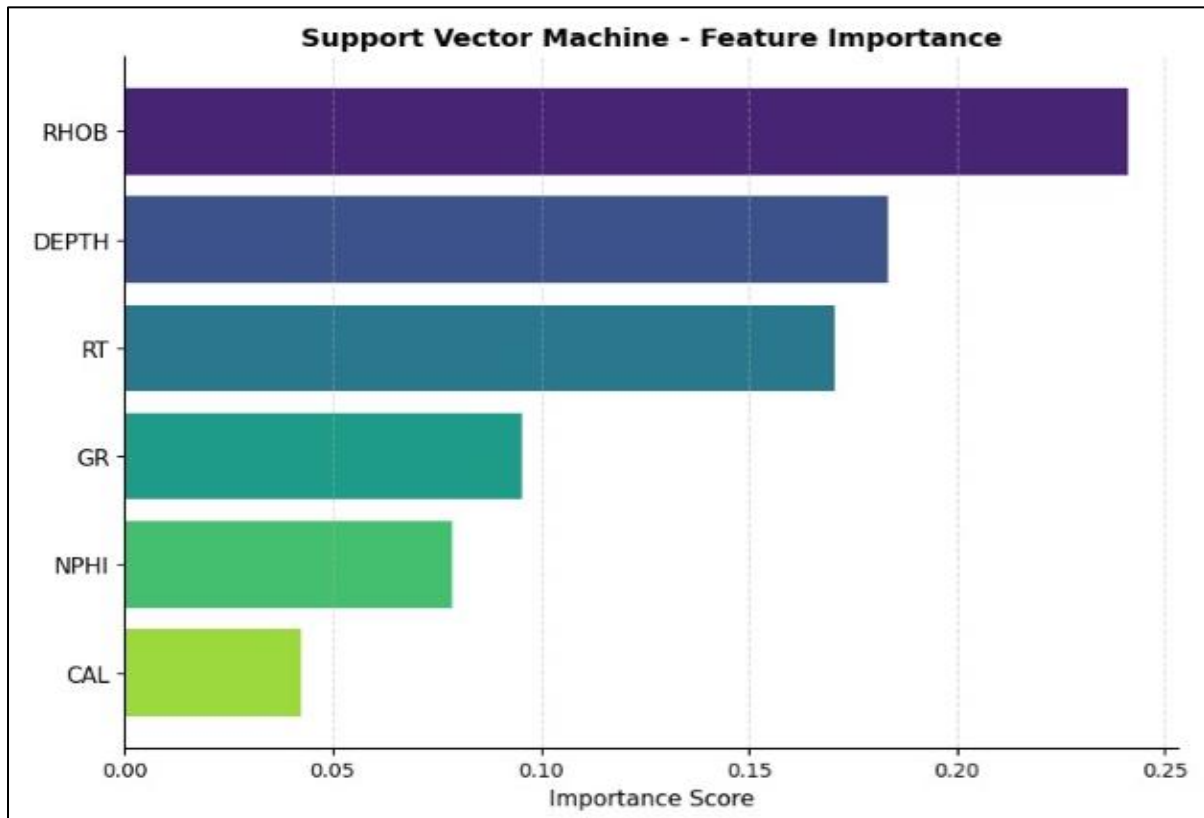


Figure 37. Support Vector Machine feature importance

Analysis

The feature importance results of the Random Forest model indicate that Resistivity (RT) has the highest impact on predicting the target variable, contributing most strongly to the model's overall performance. DEPTH follows as the second most important predictor, while Neutron Porosity & Gamma Ray (NPHI and GR) also demonstrate meaningful contributions to the prediction process. On the other hand, Caliper & Bulk Density (CAL and RHOB) show relatively low influence within the model, with RHOB being the least significant variable among all inputs.

Similarly, as observed in the drilling data, a comparable pattern is also seen in the results of the XGBoost model, where there is a significant variation in the contribution of the variables to the predictions. The most influential and primary contributor to the model's output is resistivity (RT). In contrast, the other variables show relatively weak contributions, particularly the Caliper (CAL), which exhibits the lowest level of importance among all the features.

In the SVM model, we observe a clear shift in variable importance, where RHOB, which was among the least contributing variables in the previous two models, has now become the most influential one, followed by depth. However, CAL remains the least contributing variable, consistently showing poor contribution across all three models.

3. Comparison between the results of Drilling Data & Well Log Data

3.1. Overview of the Scenarios

To accurately estimate the sonic transit time (Δt) within the reservoir, the predictive capability of artificial intelligence was evaluated under two distinct data scenarios:

Scenario 1 (Drilling Data): Utilizes real-time operational parameters recorded continuously during drilling operations. The core input features include Depth, Rotational Speed (RPM), Weight on Bit (WOB), Torque, Rate of Penetration (ROP), Standpipe Pressure (SPP), and Flow In.

Scenario 2 (Well Log Data): Utilizes standard geophysical petrophysical measurements obtained through wireline open-hole logging tools. The input features include Depth, Gamma Ray (GR), Neutron Porosity (NPHI), Bulk Density (RHOB), True Resistivity (RT), and Caliper Log (CAL).

Both scenarios were modeled and tested using three machine learning algorithms: XGBoost, Random Forest (RF), and Support Vector Machine (SVM).

3.2. Statistical Performance Comparison

The quantitative performance of the models across both datasets was evaluated using the Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) on the testing datasets.

Key Statistical Insights:

Algorithm Superiority: Across both distinct inputs, the XGBoost model demonstrated clear superiority in capturing complex, non-linear subsurface relationships compared to Random Forest and Support Vector Machines.

Fit Accuracy (R^2): Drilling data yielded a slightly higher (R^2) value (0.9249) than Well Log data (0.9105). This implies that the combined mechanical variations of the drill string and hydraulics adequately map the structural transitions of the formations.

Error Margins (RMSE / MAE): Well log data produced slightly lower absolute error values ($\backslash$ RMSE = 1.2766 vs 1.4324). This marginal advantage is mathematically expected since wireline log inputs (like bulk density and neutron porosity) directly measure physics-based properties of the rock matrix and pore fluids, which share a tight petrophysical connection to sonic wave propagation.

3.3. Feature Importance & Sensitivity Analysis

The comparative mechanism of how these models arrive at their predictions highlights a fundamental difference in application focus:

Drilling Parameters Sensitivity: Sensitivity analysis shows that parameters tracking rotational and hydraulic configurations hold high influence. Notably, Weight on Bit (WOB) was systematically found to be one of the least influential variables across all three machine learning architectures. This is a valuable finding for your mémoire, indicating that future iterations of the model could potentially exclude WOB entirely while safely maintaining or even improving prediction accuracy.

Well Log Parameters Sensitivity: Bivariate correlation studies show strong petrophysical harmony in the dataset. There is a strong positive correlation between (DT) and Neutron Porosity (NPHI), alongside moderate negative correlations between (DT) and Depth. This conforms directly to classic rock physics formulations (e.g., the Wyllie time-average equation), confirming that the AI models are effectively respecting real subsurface physics. Furthermore, the Caliper Log (CAL) exhibits weak correlations with other variables, proving that borehole enlargements or washouts did not introduce damaging artifacts into the logging dataset.

3.4. Direct Discussion (Pros & Cons)

When compiling your final discussion, you can frame the choice between these two data types through the lens of operational value and financial economy:

A. Drilling Data Advantages (Real-Time Optimization)

Real-Time Execution: Predictive modeling via drilling data allows for the continuous generation of "pseudo-sonic logs" while drilling. This acts as an invaluable real-time asset for engineers to optimize drilling parameters instantly, detect overpressure zones safely, and adjust well trajectories without awaiting subsequent logging runs.

Extreme Cost Efficiency: Collecting mechanical drilling parameters requires no additional downhole tools or operational pauses. Utilizing AI here completely bypasses the premium costs and high risks associated with deploying expensive conventional sonic logging strings into unstable hole sections.

B. Well Log Data Advantages (Direct Physical Calibration)

Direct Reservoir Interpretation: Well logs measure in-situ physical traits (such as electrical resistivity, natural gamma radiation, and bulk density). These properties yield a slightly narrower error profile ($\text{MAE} = 0.955$) because they provide structural data explicitly linked to porosity and lithology changes across formations.

Seismic and Volumetric Reference: The predictions made via well logs are highly optimal for establishing precise depth-to-time seismic calibrations and generating synthetic seismograms needed by reservoir geologists for broader asset evaluation.

Conclusion

In conclusion, this chapter demonstrated that machine learning workflows can effectively predict sonic transit time within the Oued Zine Field (Sbaa Basin), offering a reliable alternative to costly and high-risk conventional wireline logging runs. Across both evaluated data environments, the XGBoost algorithm consistently outperformed the other models, proving highly adept at mapping complex, non-linear subsurface relationships. The drilling data scenario achieved a remarkably strong statistical fit, while the well log scenario provided slightly tighter error margins due to the direct, physics-based connection between petrophysical rock properties and acoustic wave velocity. Sensitivity analyses validated these models, revealing expected geological alignments with neutron porosity in the logging scenario, while showing that Weight on Bit was the least influential parameter in the drilling scenario—suggesting future models can be simplified by omitting it entirely.

From an operational perspective, both data frameworks serve distinct, high-value purposes within asset development. Utilizing surface drilling parameters allows operators to generate continuous "pseudo-sonic logs" in real time, enabling immediate parameter optimization, pore pressure detection, and drilling risk mitigation without incurring extra tool costs or borehole risks. Conversely, utilizing standard well logs delivers the exact, physics-calibrated data required by geologists for precise depth-to-time seismic correlations and static reservoir modeling. Ultimately, this study confirms that implementing artificial intelligence in the Sbaa Basin successfully bridges the gap between mechanical surface operations and deep petrophysical properties, providing a cost-effective, safe, and highly accurate alternative to traditional data acquisition.

General Conclusion

General Conclusion

This work was carried out within the framework of improving prediction methods for sonic log (DT), which is considered one of the most important petrophysical parameters used in subsurface reservoir characterization and rock property evaluation. With the increasing demand for accurate, fast, and cost-effective solutions in exploration and production activities, the use of artificial intelligence techniques combined with drilling and well logging data has become an effective approach to enhance geological interpretation and support decision-making processes.

This study was applied to the Oued Zine Field, located within the Sbaa Basin. This area is characterized by a relatively complex geological setting, which makes the acquisition and interpretation of petrophysical data challenging from both technical and economic perspectives. This complexity highlights the need for intelligent predictive models capable of handling nonlinear data and extracting hidden relationships within subsurface parameters.

In the data preparation stage, a comprehensive database was constructed based on drilling data and well logging data. This stage involved data cleaning and consistency correction to ensure high-quality inputs. A set of well logs, including gamma ray, density, porosity, and resistivity, was used as input features for predicting sonic log (DT), which reflects the elastic and physical properties of geological formations.

Predictive models were developed using three machine learning algorithms: XGBoost, Random Forest, and Support Vector Machine. Two different scenarios were considered: the first using drilling data and the second using well logging data. This approach allowed a comparative evaluation of model performance based on input data type and its impact on prediction accuracy.

The results showed that all models achieved good predictive performance for DT estimation, with XGBoost demonstrating superior accuracy and lower error values in both scenarios. Furthermore, logging data-based models provided more stable predictions, while drilling data proved highly effective for real-time prediction during drilling operations, making it particularly valuable for operational decision-making.

Overall, the findings confirm the effectiveness of machine learning techniques in improving sonic log prediction and enhancing reservoir characterization, while reducing reliance on costly conventional measurements.

BIBLIOGRAPHY

References Liste

1. Ajona, A., Kurniawan, A., & Putra, D. (2023). Application of Multiple Linear and Polynomial Regression in the Sustainable Biodegradation Process of Crude Oil. ResearchGate.
2. Akkouche, R. (2007). Petroleum Systems and Basin Evolution in the Saharan Platform.
3. Allègre, C. J., & Caby, R. (1972). Pan-African Orogeny.
4. Alobaidi, W., Tian, G. Y., & Sophian, A. (2015). Applications of Ultrasonic Techniques in Oil and Gas Pipeline Industries: A Review. *American Journal of Intelligent Systems*, 5(3A), 64–71.
5. Bantan, R. A. R., Almutairi, A. F., Alharbi, N. K., & Aljohani, M. S. (2020). Unsupervised Learning Techniques in Data Analysis. *IEEE Access*.
6. Beicip-Franlab. (2000). Petroleum System Studies.
7. Beuf, S., et al. (1971). Les Grès du Paléozoïque Inférieur au Sahara.
8. Beuf, S., et al. (1971). Lower Paleozoic Sandstones of the Sahara.
9. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
10. Boote, D. R. D., Clark-Lowes, D. D., & Traut, M. W. (1998). Palaeozoic Petroleum Systems of North Africa. In *Petroleum Geology of North Africa*. Geological Society, London.
11. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
12. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Wadsworth.
13. Chapelle, O., Schölkopf, B., & Zien, A. (2006). Semi-Supervised Learning. MIT Press.

14. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
15. Classification of Petroleum Well Drilling Operations Using Support Vector Machine (SVM). Adriane B. S. Serapião, Rogério M. Tavares, José Ricardo P. Mendes, Ivan R. Guilherme.
16. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273–297.
17. Coward, M. P., & Ries, A. C. (2003). Tectonic Evolution of Sedimentary Basins.
18. Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: A Review. Zeravan Arif Ali, Ziyad H. Abduljabbar, Hanan A. Taher, Amira Bibo Sallow, Saman M. Almufti.
19. Foote, K. D. (2019). A Brief History of Machine Learning. Dataversity.
20. Gassmann, F. (1951). Über die Elastizität poröser Medien.
21. Ghienne, J. F., et al. (2007). Ordovician Glaciation.
22. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
23. Guéguen, Y., & Palciauskas, V. (1994). Introduction to the Physics of Rocks. Princeton University Press.
24. Habeeb, A. (2023). Introduction and Investigation into Oil Well Logging Operations. Journal of Engineering, 29(2), 72–91.
25. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques (3rd ed.). Elsevier.

26. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
27. Hebb, D. O. (1949). *The Organization of Behavior*. Wiley.
28. Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554–2558.
29. Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31(8), 651–666.
30. Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.
31. Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science*, 349(6245), 255–260.
32. Kabir, M. M., & Luo, X. (2020). Clustering Analysis Methods in Machine Learning. *Journal of Data Science*.
33. Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4, 237–285.
34. Lai, J., et al. (2022). Geophysical Well-Log Evaluation in the Era of Unconventional Hydrocarbon Resources: A Review. *Surveys in Geophysics*, 43(3), 913–957.
35. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 521(7553), 436–444.
36. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4), 541–551.

37. Liu, Y., Li, X., Zheng, Y., & He, J. (2016). Sonic and Ultrasonic Measurement Applications for Cased Oil Wells. *Journal of Petroleum Exploration and Production Technology*, 6, 647–664.
38. Lüning, S., Craig, J., Fitches, W. R., Mayouf, J., & Busrewil, M. T. (2000). Lower Silurian “Hot Shales” in North Africa and Arabia: Regional Distribution and Depositional Model. *Earth-Science Reviews*, 49(1–4), 121–200.
39. Lüning, S., et al. (2000). Silurian Source Rocks.
40. MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.
41. Maldonado-Cruz, E., et al. (2023). Sonic Well-Log Imputation Through Machine-Learning-Based Uncertainty Models.
42. MDPI. (2025). Prediction of Sonic Well Logs Using Deep Neural Network: Application to Petroleum Reservoir Characterization. *Geosciences*.
43. Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
44. Mondol, N. H. (2015). Rock Physics and Ultrasonic Measurements on Core Samples.
45. Nasteski, V. (2017). An Overview of the Supervised Machine Learning Methods.
46. Natue. (2025). Machine Learning Model Optimization for Compressional Sonic Log Prediction. *Scientific Reports*.
47. Oil and Gas Production Forecasting Using Decision Trees, Random Forest, and XGBoost. Mays A. Al Shabaan & Zainab N. Nemer.

48. Pang, X., et al. (2022). Geophysical Well-Log Evaluation in the Era of Unconventional Hydrocarbon Resources.
49. Pedregosa, F., et al. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
50. Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1), 81–106.
51. Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
52. Rosenblatt, F. (1957). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6), 386–408.
53. Saba, T., & Ngepah, N. (2022). Overlapping Clustering Techniques: A Review. *Applied Soft Computing*.
54. Sajed, O. K. M. (n.d.). Well Logging Lecture Notes: Sonic Logging.
55. Saleh, K., Mabrouk, W. M., & Metwally, A. (2025). Machine Learning Model Optimization for Compressional Sonic Log Prediction Using Well Logs in Shahd SE Field, Western Desert, Egypt. *Scientific Reports*, 15(1), 14957.
56. Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229.
57. Santhanam, T., Padmavathi, S., & Venkatesh, P. (2016). Performance Analysis of XGBoost Algorithm.
58. Sarker, I. H. (2020). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*.
59. Santos, J. E., et al. (2022). Synthetic Sonic Log Generation with Machine Learning: A Contest Summary from Five Methods.

60. Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel Networks That Learn to Pronounce English Text.
61. Springer Nature. (2021). Well Logging and Formation Evaluation Studies.
62. Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.
63. Tavallaei, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A Detailed Analysis of the KDD CUP 99 Dataset.
64. Tiab, D. (2015). Petrophysics: Theory and Practice of Measuring Reservoir Rock Properties.
65. Torsæter, O., & Abtahi, M. (2000). Experimental Reservoir Engineering Laboratory Workbook.
66. Van Engelen, J. E., & Hoos, H. H. (2020). A Survey on Semi-Supervised Learning. *Machine Learning*, 109(2), 373–440.
67. Wazir, I. (2014). Structural Analysis of Oued Zine Field.
68. WEC. (1995). Geological Map of Algeria.
69. Wyllie, M. R. J., Gregory, A. R., & Gardner, L. W. (1958). Elastic Wave Velocities in Heterogeneous and Porous Media. *Geophysics*, 23(3), 459–493.
70. Zhang, D., Wang, Y., & Chen, Y. (2014). Machine Learning in Geosciences: Applications and Challenges. *Geoscience Frontiers*, 5(3), 349–361.
71. Zhu, X. (2005). Semi-Supervised Learning Literature Survey. University of Wisconsin-Madison