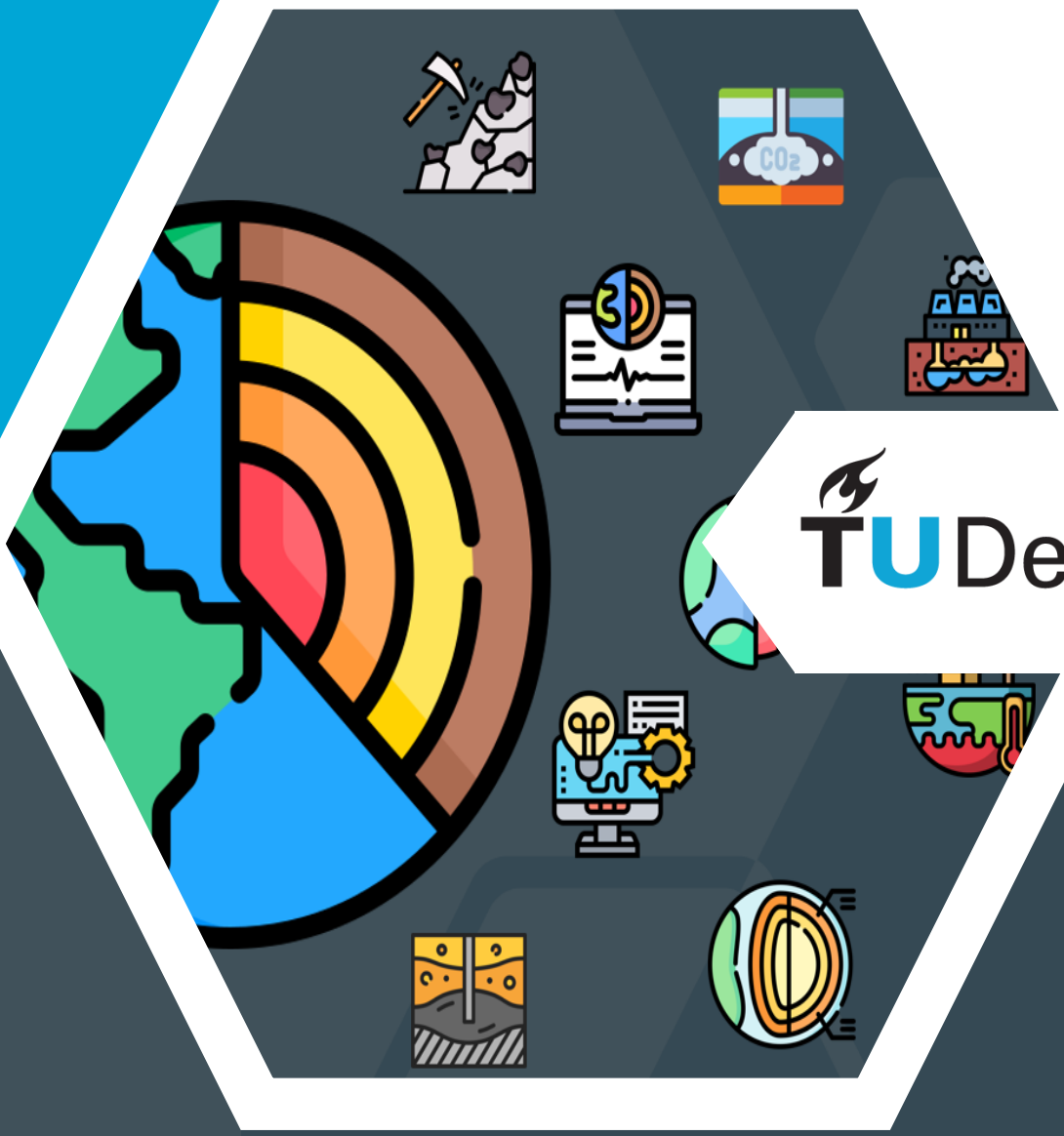
Evaluating the Performance of Multivariate Imputation by Chained Equations (MICE) when Predicting Missing Well-Log Data in Sedimentary Basins



TU Delft Delft University of Technology

GitHub

Luis Carlos Baez Lozada

MSc. Thesis Geo-Energy Engineering

August 2023

Evaluating the Performance of Multivariate Imputation by Chained Equations (MICE) when Predicting Missing Well-Log Data in Sedimentary Basins

by

Luis Carlos Baez Lozada

Student Number

5607426

to obtain the degree of
Master of Science
in Applied Earth Sciences, Geo-Energy Engineering
at the Delft University of Technology
to be defended publicly on Thursday August 24, 2023, at 9:00 AM.

Supervisors:	MSc. Parvin Kolah Kaj	P.KolahKaj@tudelft.nl	TU Delft
	Dr. V. Crombez	Vincent.Crombez@csiro.au	CSIRO
	Dr. G. Rongier (Committee Chair)	G.Rongier@tudelft.nl	TU Delft
	Dr. H.A. Abels (Committee Member)	H.A.Abels@tudelft.nl	TU Delft
	Dr. M. Soleymani Shishvan (Committee Member)	M.SoleymaniShishvan@tudelft.nl	TU Delft

TU Delft Delft University of Technology

PREFACE

This report is the result of my graduation research project “Evaluating the Performance of Multivariate Imputation by Chained Equations (MICE) when Predicting Missing Well-Log Data in Sedimentary Basins”. This is the last step towards fulfilling the requirements needed to obtain the title Master of Science in Applied Earth Sciences, Geo-Energy Engineering, at the Delft University of Technology.

First, I would like to express my deepest gratitude to my supervisor and committee member, Guillaume Rongier, for their support, guidance, and encouragement throughout my studies at this university and the course of this research. Their expertise, patience, and constructive feedback have been invaluable in shaping this thesis and my growth as a professional. I also want to especially thank my daily supervisor, Parvin Kolah Kaj, for her encouragement, interest, and support throughout the development of this research. I am also grateful to my supervisor, Vincent Crombez, and the members of my thesis committee, Masoud Soleymani Shishvan and Hemmo Abels, for their insightful comments and suggestions, which have significantly contributed to the quality of this work.

I also want to thank my family and friends for their unwavering love, understanding, and encouragement throughout my academic journey and for their belief in my abilities and their constant support have been a source of strength and motivation. Also, my most sincere thanks to my fellow students and study friends who gave me help and support during this process.

Finally, I dedicate this thesis to my wife, Juanita Goez, and my parents, Rodrigo Baez and Martha Lozada, whose support and unconditional love has been a light and hope that has guided me in my search for knowledge and personal growth.

SUMMARY

This research evaluates the applicability of Multivariate Imputation by Chained Equations (MICE) for estimating missing well-log data across different sedimentary basis. Utilizing various machine learning techniques including XGBoost (XGB), Random Forest (RF), K-Nearest Neighbors (KNR), and Bayesian Ridge (BR), the performance of MICE was tested on three different data sets from distinct geological contexts and preprocessing conditions with minimal user input.

The main results indicate that the performance of MICE varied across different data sets and well-logs, highlighting the complexity of imputing missing data in heterogeneous sedimentary basins. The number of iterations in MICE did not significantly impact the performance of the models, while data quality, pre-processing, and geological complexities played crucial roles. The Force-200 data set, which underwent extensive preprocessing, demonstrated better imputation performance compared to the Montney and Beetaloo data sets. Additionally, XGB often outperformed other algorithms, predicting missing values with different number of iterations.

The main conclusions drawn from this study emphasize the need for more research to minimize user input and to develop more robust and flexible approaches to imputing missing data in well-logs. The study highlights the challenge of determining a single set of hyperparameters optimal for all the well-logs, suggesting the need for more adaptable models or even advanced techniques like deep learning techniques. The research also suggests the importance of refining pre-processing techniques, exploring further combinations of well-logs, and developing cross-validation approaches that effectively replicates real-world scenarios to advance the application and reliability of MICE in data imputation of subsurface data with missing values.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 | INTRODUCTION

1.1 GENERAL CONTEXT AND PROBLEM STATEMENT

As the urgency to mitigate climate change intensifies, it becomes necessary to move away from high-carbon fossil energy sources, such as oil and gas, and towards more sustainable alternatives. Geothermal energy, due to its low carbon emissions, and energy storage, such as hydrogen storage, are emerging as viable options to meet the growing global energy demand and address climate challenges. To optimize the exploration, development, and production of these geological resources, well-logs are crucial to these processes by providing measurements of subsurface fluids and rocks. In geological and engineering fields, these measurements provide essential data for interpreting subsurface geology, characterizing reservoirs, evaluating wellbore stability, and designing engineering strategies (Darling, 2005).

A well-log refers to the recorded measurements and data obtained from various instruments or tools used in drilling and exploration activities. Well-logs consist of the determination of various geophysical and geological properties of the subsurface at different depths (Darling, 2005; Feng, Grana, & Balling, 2021), such as electrical resistivity, rock porosity, density, seismic wave velocity, among others. In addition, these measurements can also provide information about the characteristics of fluids present in the formations, such as water, oil, and gas content. By analyzing well-logs, geoscientists can develop a better understanding of the subsurface, improving their analysis and decision-making about resource exploration and production.

However, well-logs are often incomplete and certain data is partially or entirely missing. Missing data in well-logs can be attributed to operational issues, economic decisions, or the complexity of geological formations (Darling, 2005). During the drilling of a well, logging tools can fail, or some measurements are not carried out due to cost considerations, resulting in certain intervals with missing data or complete missing logs. This absence of data in well-logs presents significant problems which can impact the interpretation of subsurface conditions. For instance, missing data can lead to inaccurate reservoir characterization, making reservoir capacity prediction, resource estimation, and drilling decisions difficult. Additionally, well-logs are used to calibrate seismic data and help to build more accurate and detailed geological models of the subsurface. When well-

logs are not available, seismic interpretation becomes more challenging and less reliable, compromising understanding of the subsurface and its resources. Therefore, it is important to address the problem of missing data in well-logs to ensure better decision-making in this industry.

In recent years, the application of machine learning algorithms has gained popularity in geosciences, particularly in well-log analysis. These algorithms can extract hidden relationships in the available data and make predictions about properties of interest (Hallam, Mukherjee, & Chassagne, 2022; Feng, Grana, & Balling, 2021). This technology has proven its effectiveness in a variety of applications, including lithofacies identification, anomaly detection in well-logs, drilling parameters predictions, and estimation of rock properties such as porosity and permeability (Belyadi & Haghighat, 2021). However, the presence of missing data affects the potential of these algorithms, since these models mostly require complete data sets for optimal performance (Dixneuf, Errico, & Glaus, 2021).

Handling missing data is a common challenge in data science. Traditional techniques such as data deletion and mean mode substitution are common approaches used, but they have significant drawbacks dealing with missing values. Deleting data, for example, reduces the sample size in the data set and causes the loss of valuable information. Similarly, median mode substitution often produces biased estimates and does not effectively address the problem (Leke & Marwala, 2019). Additionally, these techniques do not account for possible relationships or patterns between missing values and other variables, which can result in inaccurate representations of the original data. These limitations make data deletion and mean mode substitution not recommended for predicting missing values in well-logs, especially in situations where prediction performance and maximizing data collection are critical. Instead, it is necessary to adopt more sophisticated and modern approaches that consider the complexity and interactions of the data to obtain more reliable and valuable results.

Previous studies have explored various strategies for addressing the issue of missing data in well-logs. One of these strategies relies on machine learning algorithms. For example, Lopes & Jorge (2018) and Feng, Grana, & Balling (2021) proposed the use of Random Forest and Gradient Boosted Trees for predicting missing values in a single well-log, demonstrating promising results in completing well-log data. However, in these studies, missing values were introduced artificially and randomly into well-logs for testing purposes that do not represent the nature of missing values in well-logs accurately. Additionally, Feng, Grana, & Balling (2021) employed a complete data set for training their model, which is not a realistic representation of the field situation where well-logs frequently have missing data.

Furthermore, Multivariate Imputation by Chained Equations (MICE) emerged as an innovative alternative for imputing missing well-log data (Hallam, Mukherjee, & Chassagne, 2022). In this research, MICE was applied using machine learning models such as K-Nearest Neighbors (KNR), Random Forest (RF), Gradient Boosted Trees (GBT), and Bayesian Ridge (BR) in two different

data sets from the Norwegian North Sea. Although the study proposed a methodology for simultaneous imputation of all input well-logs, the evaluation focused on three logs, limiting the understanding of effectiveness of this imputation method. Moreover, the method of introducing missing values in well-logs presented gaps that need further exploration.

Despite the valuable contributions of these studies in addressing missing data in well-logs, they have limitations in predicting these values. Some common limitations include that the missing values are often added randomly for testing, which does not accurately represent real-world scenarios where entire well-logs might be missing. Moreover, most studies typically focus on a single or few missing logs, whereas in practical settings, missing values occur in almost all well-logs. Additionally, they often used one or two benchmark data sets with unclear pre-processing steps, making it difficult to apply their findings to different geological contexts or data sets. Therefore, this research aims to conduct a comprehensive evaluation of the performance of MICE with minimal user input, where all well-logs have missing values and entire well-logs could be missing. The goal is to determine the applicability of MICE to predict missing values in well-logs in different geological contexts, reflecting real-world scenarios as closely as possible.

1.2 RESEARCH QUESTION

This project aims to address some of the limitations identified in previous studies and literature. The objective is to develop a framework that allows simultaneous imputation of various well-logs using the MICE approach and incorporates cross-validation techniques to assess performance realistically with minimal user input. The main research question guiding this study is the following:

How does MICE perform in settings where all the types of well-logs have missing values and entire well-logs are missing?

To answer this question, we use three data sets from different sedimentary basins and apply MICE without extensive pre-processing or pre-selection of inputs to compare the performance of the various machine learning models in predicting missing values in well-logs. Additionally, we define sub-questions to help answer the main research question evaluating the performance of MICE:

- How to define a validation study that thoroughly tests MICE's ability to impute well-logs?
- What is the computational efficiency and performance of MICE for different datasets with missing well-log data?
- Which data sets to use to determine MICE's domain of applicability?
- Which machine learning methods to focus on in MICE?

- How to efficiently perform hyper-parameter tuning for those methods?
- Is there a generic configuration for MICE that provides reasonable results for all types of well logs and all datasets?

1.3 THESIS STRUCTURE

This thesis is divided into five chapters, each of them contributes to answering the research question and provides insights into the application of MICE for predicting well-log data with missing values.

Literature Review	Provides a critical appraisal and evaluation of existing concepts, methods, and applications related to well-logging, machine learning, missing values, and imputation techniques.
Data Analysis	Presents an overview of the data sets used, including geological context, data description and missing value analysis.
Methodology	Describes the proposed framework for predicting missing values in well-logs using multivariate imputation by chained equations (MICE).
Results and Discussion	Presents the findings of the project, discussing the computational efficiency and performance of the implemented approaches.
Conclusions and Recommendations	Summarizes the key findings and evaluates the prediction of missing data, including recommendations for future research on implementing multiple imputation techniques for well-log data.

2 | LITERATURE REVIEW

This section provides a detailed review of the literature to contextualize and validate the research carried out. First, the concept and importance of well-logs is examined before discussing the problem of missing values, including its definition, causes, impacts, and the various patterns and mechanisms of missing data in well-logs. Various methods for handling missing data are then reviewed, from traditional techniques to more robust strategies such as multivariate imputation by chained equations (MICE).

Subsequently, machine learning is explored, investigating its common models and the potential of more sophisticated ensemble models. These approaches are critically evaluated for their effectiveness and suitability in addressing the problem of missing well-log data. Finally, the machine learning workflow is explained, from data preprocessing to implementation, detailing techniques for developing a robust and reliable method for imputing missing values in well-logs.

2.1 WELL-LOGS

2.1.1 Defining Well-Logs

Well-logs are records of measurements taken during the drilling and exploration of oil and gas wells. The process of acquiring these measurements, called well-logging, uses various tools and instruments to gather data about subsurface geology, rock formations, fluid content and other properties of the well (Darling, 2005; Feng, Grana, & Balling, 2021). These measurements are usually taken at regular intervals at different depths in the wellbore, as can be seen in the Figure 1.

Currently, different types of well-logs are used for lithology identification, formation evaluation, hydrocarbon detection, and rock mechanical property analysis in the oil and gas industry (Evenick, 2018). Each type of log provides specific information about subsurface characteristics such as density, gamma ray, resistivity, and porosity. Table 1 presents a summary of some of the most used well-log types for this project, including their abbreviations.

Table 1. Types of well-logs, including abbreviations and their applications (Darling, 2005; Evenick, 2018).

Well-Log	Abbreviation	Description	Application
Density	RHOB	Measures bulk density of the formation.	Lithology, Porosity.
Gamma-ray	GR	Measures the natural radioactivity of a formation. Helps to identify shales and organic content.	Lithology.
Sonic	DT	Measures the time of compressional sound waves in the formation. It is a good indicator of density and the presence of gas.	Lithology, Mechanical properties.
Resistivity	RES	Measures the flow of electricity through a formation. Helps to detect hydrocarbon from water.	Lithology, Porosity, Hydrocarbons.
Spontaneous Potential	SP	Measures the electrical that arises due to salinity differences between the borehole fluid and the fluid formation. This is a good indicator of formation permeability and can distinguish shale from carbonates and sandstones.	Permeability, Fluid content.
Porosity (e.g., Neutron Porosity)	NPHI	Measures the porosity of the formation based on the quantity of hydrogen content.	Porosity, Fluid storage capacity.

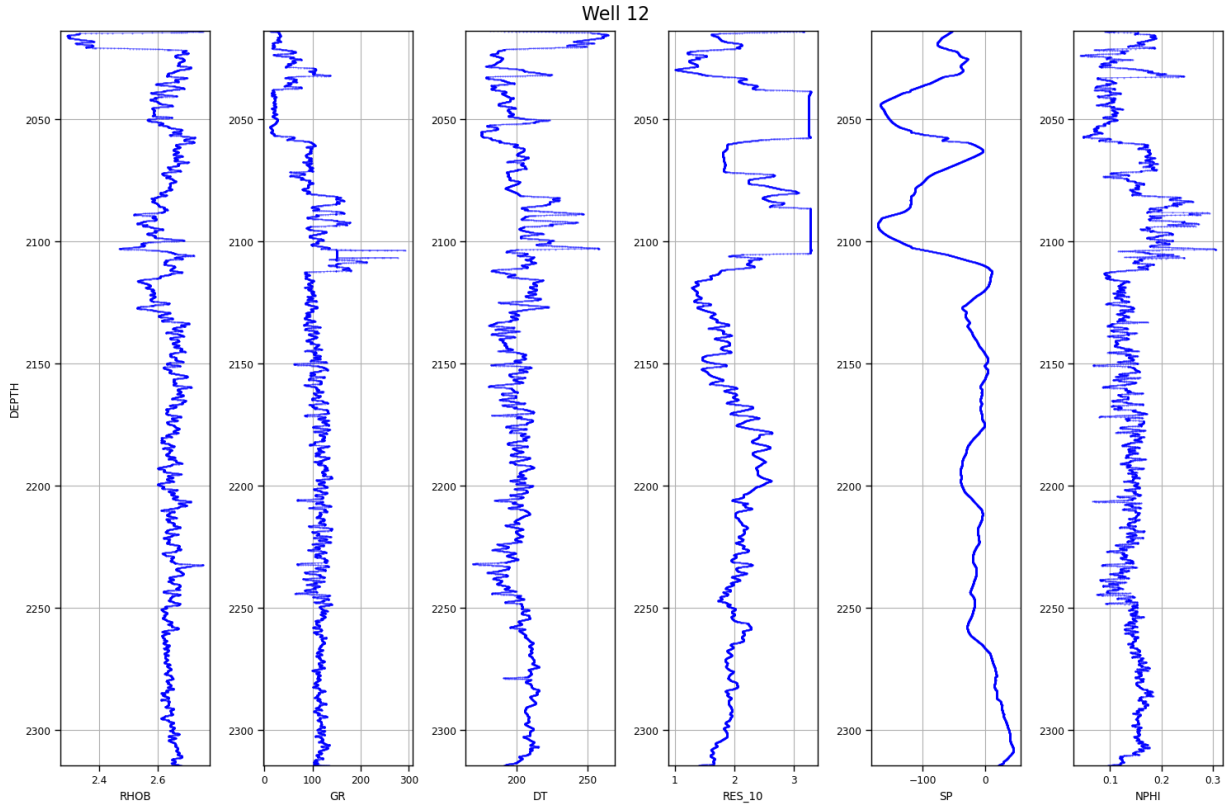


Figure 1. Well-logs of the Montney data set Well 12 from section Data Analysis, including Density (RHOB), Gamma-ray (GR), Sonic (DT), Resistivity (RES), Spontaneous Potential (SP) and Neutron Porosity.

2.1.2 Importance of Well-logs

Well-logs play a fundamental role in geosciences, providing essential information throughout the life cycle of a wellbore, from exploration to subsurface optimization. They have a significant impact

on sectors such as oil and gas, improving the evaluation and classification of underground formations, which makes it possible to determine the best use of resources for the production of hydrocarbons, geothermal energy and gas storage (Darling, 2005; Liu, 2015).

The importance of well logging is reflected in its wide application in various disciplines. Geologists, for example, use well-logging as a mapping technique to investigate the subsurface, allowing them to understand geological properties and formations. Petro-physicists employ well-logs to evaluate the hydrocarbon production potential of a reservoir, allowing them to estimate properties such as porosity and permeability. Geophysicists use well-logs as complementary information for seismic analysis, which allows them to integrate multiple sources of subsurface information. Reservoir engineers, on the other hand, rely on well-logs to obtain values for use in simulations, which help them in optimizing drilling and production operations (Liu, 2015).

Besides these applications, well-logs can contribute to achieving economic and environmental sustainability goals. They serve as a source of information to identify the most productive zones for drilling, reducing costs, and increasing productivity. Moreover, well-logs can be employed in real time to identify drilling risks, detection of unstable formations and possible leaks. As a result, well-logs allow prompt decision-making and operations adjustments to guarantee the integrity, stability, and safety of well operations, while reducing environmental risks (Darling, 2005; Liu, 2015; Lopes & Jorge, 2018).

Despite the benefits and wide uses of well-logs, the main drawback is that these measurements are rarely complete. In the following sections, we will explain in more detail the topic of missing values in well-logs. We will start by defining a missing value, explore its possible causes, and understand why it is a problem in well-logs. Additionally, we will explore the different missingness mechanisms and patterns for missing values to select the best approach to handle missing data. In this context, we will investigate how these missing values can be addressed by showing different techniques, with a particular focus on multivariate imputation by chained equations (MICE).

2.2 MISSING DATA

2.2.1 Defining Missing Data

Missing data refers to the absence of values in a data set (Galli, 2022; Little & Rubin, 2020). In other words, if an observation that was intended to be measured, collected, and recorded is not present in the data set, it is considered a missing value. For example, if gamma ray (GR) measurements in a well are supposed to be recorded at every meter of depth, but at certain depths this measurement is not available, these points are considered as missing values. Furthermore, missing data can extend beyond certain points or individual observations. For

instance, if spontaneous potential (SP) log has not been performed on a well, the entire log could be missing in the data set. Figure 2 visually shows missing values in a well and serves as a representation of the examples mentioned above. It is relevant to note that missing values can be represented as not a number (NaN) or simply remain blank in the data set. For this project, we use the NaN representation.

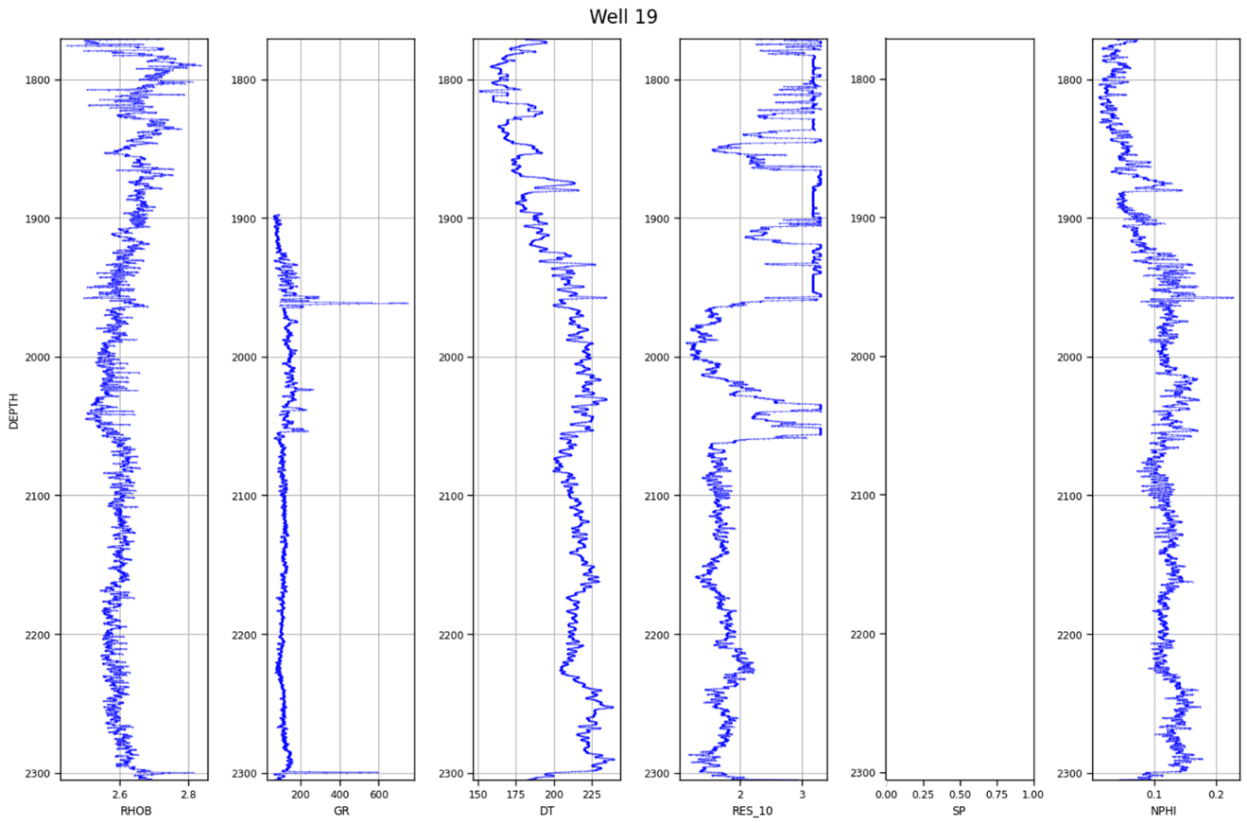


Figure 2. Visualization of missing values in well-logs, using the Well 19 of the Montney data set from section Data Analysis.

2.2.2 Causes and Problems of Missing Data in Well-Logs

Missing values in well-logs can be attributed to various factors, including commercial considerations, geological conditions, unavailability of logging tools, mechanical or operational failures, and data quality control procedures (Darling, 2005). Commercial considerations play a critical role in the integrity of well-log data. Due to financial constraints, companies can be forced to make decisions that impact the acquisition of this information. These decisions could include limiting measurements in certain depths and formations, and in some cases, not logging some wells. Consequently, these considerations lead to reduced data collections and a lack of well-log data.

Mechanical and operational failures also contribute to missing data. During the drilling process, well-logging equipment often has technical problems and failures that can lead to incorrect readings and incomplete data. Additionally, complex geological conditions, such as unstable

formations, can interrupt the logging process resulting in gaps in the data. Data quality control is another factor that affects the presence of missing data in well-logs. During data processing, outliers, errors, and certain data are identified for later removal to ensure the final record is accurate and reliable. However, this process can result in missing values and a subsequent reduction in observations in the final well-log.

Missing values in well-logs directly affect subsurface analysis and interpretation, as incomplete data can complicate the understanding of subsurface geology and the evaluation of reservoir quality (Darling, 2005; Liu, 2015). Since well-logs are fundamental for defining rock properties and fluid content, incomplete well-logs lead to inaccurate predictions of formations characteristics and reservoir performance. For instance, when resistivity and porosity logs are incomplete, identifying hydrocarbons and estimating their storage capacity becomes challenging, resulting in unreliable estimates for determining reservoir potential.

Furthermore, missing well-log data can compromise the integration of well-logs with other data sources, such as seismic data. While seismic data provides a broad and general view of the subsurface, it does not provide precise details about rock or fluid properties at specific points such as well-logs. Therefore, the combination of these two types of data gives a more complete and accurate picture of the subsurface. However, when well-log data is incomplete or missing, seismic data calibration becomes less reliable, leading to greater uncertainties in seismic interpretation and understanding of the subsurface.

2.2.3 Missingness Mechanisms and Patterns

The identification of the mechanism and pattern of missing data is crucial to determine the most appropriate strategy to handle missing data in well-logs. Each approach has its own strengths and limitations that depend on the characteristics of the data and the research objective. To effectively address the problem of missing data, it is necessary to understand the reasons for the missing values, called missing mechanisms, and the structure or pattern that these missing values present in the data (Little & Rubin, 2020).

The missing mechanism refers to the process by which data is missing, specifically about the relationship between missing data and the values of variables in a data set. Understanding the missing mechanism is crucial because it can influence the validity of the methods used to handle missing data and the potential impact on the results of statistical analyzes or machine learning models (Dixneuf, Errico, & Glaus, 2021; Little & Rubin, 2020). There are three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

- **Missing Completely at Random (MCAR):** MCAR occurs when the missing values are unrelated to any systematic reason (White, Royston, & Wood, 2011; Huyen, 2022).

- **Missing at Random (MAR):** MAR refers to missing values that can be explained by other observed variables, introducing a pattern to the missing data (White, Royston, & Wood, 2011; Huyen, 2022).
- **Missing not at Random (MNAR):** MNAR is the most problematic mechanism and occurs when the missingness is related to both observed and unobserved variables, including the variable itself. MNAR indicates that there is a reason why the values are missing from the data set (White, Royston, & Wood, 2011; Huyen, 2022).

In the context of well-logs, missing values could be considered as missing not at random due to business decisions, geological conditions, and tool limitations. For instance, to save costs, a company might record only certain intervals, which are the most promising for hydrocarbon or geothermal energy production, using other available data or geological models as a reference. Another example could be the failure of logging tools in specific formations due to collapses or complexity in data collection, resulting in gaps in the data for this interval. In these cases, the missing data is directly related to unobserved measurements; therefore, it would be classified as missing not at random. However, it is important to note that the reasons for missingness vary by data set.

Missing patterns, on the other hand, refer to the structure of the missing values in the data. It describes how missing data is distributed in a data set, providing information about the reason for the missing values. There are different types of missingness pattern which includes univariate non-response, multivariate, monotone, general, and file matching (Berglund & Heeringa, 2014; Little & Rubin, 2020). In the Figure 3 shows the examples of missingness patterns.

- **Univariate Non-response:** It occurs when data is missing for a particular variable, often because respondents chose not to provide that information (Berglund & Heeringa, 2014; Little & Rubin, 2020).
- **Multivariate:** This pattern occurs when there is missing data in multiple variables (Berglund & Heeringa, 2014; Little & Rubin, 2020).
- **Monotone:** It occurs in different variables due to participants dropping out or leaving the study for unknown reasons (Berglund & Heeringa, 2014; Little & Rubin, 2020).
- **General:** This occurs in multiple variables throughout the data set, without a specific order.
- **File matching:** It occurs when data sets are merged from different sources (Berglund & Heeringa, 2014; Little & Rubin, 2020).

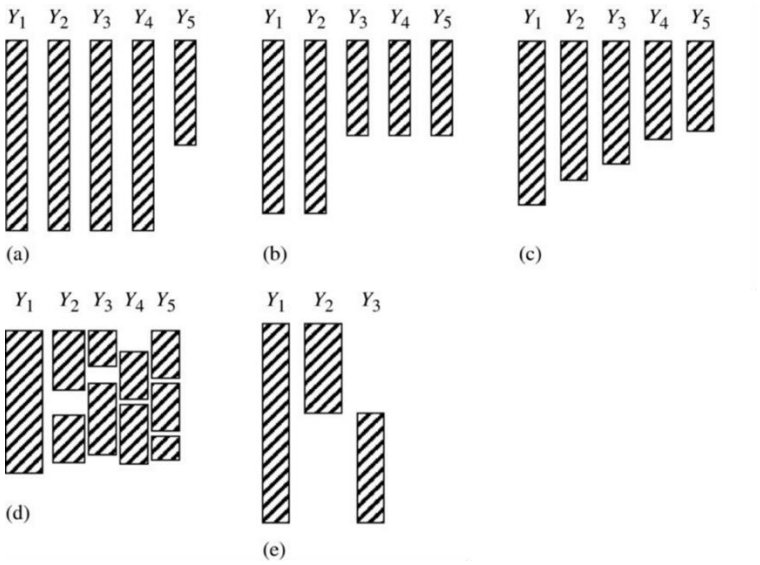


Figure 3. Examples of missingness patterns modified from (Little & Rubin, 2020): (a) Univariate non-response, (b) Multivariate, (c) Monotone, (d) General, and (e) File matching.

Well-log data often includes information from multiple wells, each well providing specific geological insights and different logs based on depth. Consequently, each row within these data sets represents a measurement at a particular depth within a given wellbore, while each column indicates a different property or log type. Regarding the missingness pattern, missing values in well-logs present a generalized structure. Where the missing data is scattered throughout the data set, distributed across multiple wells, depths, and variables such as geological properties or logs, as can be seen in the following Figure 4.

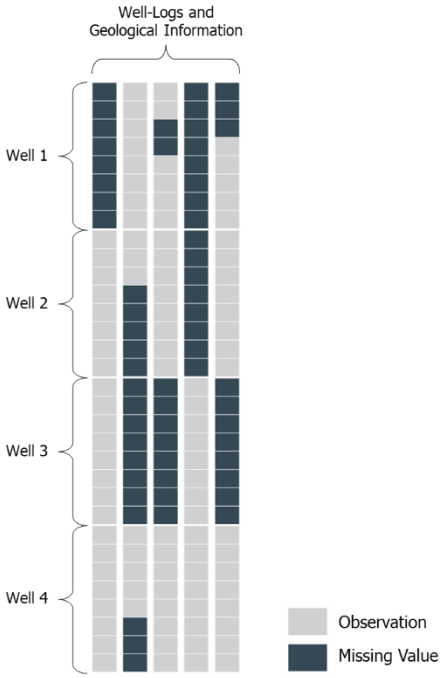


Figure 4. Generalized pattern of missing data in well-logs.

2.3 HANDLING MISSING DATA

Handling missing data is not a simple task, as it requires an understanding of the mechanism of missingness and the use of appropriate techniques to deal with it. In this section, we will discuss some of the traditional approaches used to handle missing data, each with their strengths and limitations. Among these methods, we will explain more in detail the multivariate imputation by chained equations (MICE) method due to its promising results in predicting missing values in well-logs.

2.3.1 Traditional Methods for Handling Missing Values

I. Data Deletion

Data deletion consists of removing observations or features with missing values. Although this method is quick and easy to implement, it can result in the removal of a significant proportion of the data set if there are missing values in multiple features (Galli, 2022). It is important to note that when missing values follow a missing not at random (MNAR) mechanism, data deletion can introduce bias and loss of information, which affects the reliability of the subsequent analysis (Gallatin & Albon, 2023).

II. Single Imputation

Imputation is the process of replacing missing data with substituted values, with the objective of producing a complete data set (Leke & Marwala, 2019). These methods can be univariate or multivariate based if the approach considers relationships with other variables in the data set during the imputation process. For instance, multivariate imputation predicts missing values of a variable using information from other variables in the data set. Additionally, imputation methods can also be categorized into single and multiple imputation.

Single imputation replaces the missing values with a single imputed value to produce a complete data set. Although this method is simple to implement with minimal computational resources, single imputation can introduce bias into the data set, especially for missing not at random (MNAR) mechanisms. This method does not capture the variability and uncertainty associated with the missing values (Little & Rubin, 2020; Berglund & Heeringa, 2014). Therefore, single imputation may underestimate uncertainty and lead to potentially invalid inferences. Two common single imputation methods are mean median substitution and regression methods.

Mean-Median Substitution

This method replaces missing values with the mean or median variable. However, this approach can change the distribution of the original variables if there is a high percentage of missing data, and lead to biased estimates. Mean median substitution is more effective when the data is missing completely at random, MCAR (Galli, 2022; Leke & Marwala, 2019).

Regression Methods

Regression methods have been used to estimate missing values by setting the target variable as the dependent variable and the remaining features as independent variables or predictors. However, regression methods are not an ideal solution to deal with MNAR data (Jafari, 2022). These methods assume that the missing values are missing at random (MAR), considering that the probability of missingness is only related to the observed variables. Consequently, regression methods may not capture the relationship between the missing values and the unobserved variables in MNAR, leading to biased estimates and misleading results (Leke & Marwala, 2019).

III. Multiple Imputation

Multiple imputation creates various data sets with different imputed values for the missing entries. These data sets are then combined to obtain a single set of estimates, resulting in a final completed set (Little & Rubin, 2020). This method considers the variations between the imputed data sets, effectively incorporating the uncertainty arising from the imputation process into the final results (Dixneuf, Errico, & Glaus, 2021).

Multiple imputation is multivariate. It preserves the distribution of each variable and maintains associations among variables, making it more robust and precise, reducing bias in subsequent analysis (Berglund & Heeringa, 2014). A popular and flexible technique within multiple imputation is multivariate imputation by chained equations (MICE), which is known for its applicability to various types of missing data (Dixneuf, Errico, & Glaus, 2021; Hallam, Mukherjee, & Chassagne, 2022).

However, it is important to note that the idea of imputation can create a false belief that the data is complete. Even though it offers a convenient solution to handle missing data, it can introduce biases in subsequent analyses (Little & Rubin, 2020). The imputation model, whether single or multiple imputation, may not perfectly match the true underlying distribution assumptions or the assumed missing data mechanism, which can bias the results. Moreover, multiple imputation can also introduce biases if important variables are omitted, incorrect relationships are assumed, or crucial interactions are omitted in the imputation model (Berglund & Heeringa, 2014).

2.3.2 Multivariate Imputation by Chained Equations (MICE)

Multivariate imputation by chained equations (MICE) is used to impute incomplete multivariate data (Buuren & Groothuis-Oudshoorn, 2011). This method employs an iterative procedure in which each variable with missing values is imputed based on the other variables, resulting in multiple imputations that are combined to create a final complete data set (Berglund & Heeringa, 2014; Dixneuf, Errico, & Glaus, 2021).

The MICE steps are as follows:

1. **Initialization:** Initial imputations for missing values are created using a simple method like mean imputation from observed values.
2. **Iteration:** In each iteration, an incomplete variable is selected, and its missing values are imputed based on the observed values and imputed values of other variables. This process is repeated for all incomplete variables.
3. **Convergence:** The stability of the imputations across iterations is evaluated, and the process continues until convergence is achieved.
4. **Combination:** Multiple imputations are combined to obtain a single imputed data set, on which statistical analysis can be performed.

MICE offers several advantages for imputing missing data. It can handle various types of variables and complex data structures, including categorical, continuous, and ordinal variables (White, Royston, & Wood, 2011). By generating multiple imputations, MICE captures the uncertainty associated with missing data and enables proper estimation of standard errors and valid statistical inference. Furthermore, MICE conserves the relationships between variables by imputing missing values based on observed values and their associations with other variables, regardless of the missingness mechanism (Buuren & Groothuis-Oudshoorn, 2011).

However, it is important to acknowledge certain limitations and assumptions associated with MICE. The method assumes that the missing data follows missing at random (MAR) or missing completely at random (MCAR) mechanisms. For this reason, MICE may introduce bias and inaccurate imputation if the missing data is missing not at random (MNAR) (Buuren & Groothuis-Oudshoorn, 2011). Additionally, it is crucial to determine the appropriate model and the number of iterations to find a balance between computational efficiency and imputation quality. Otherwise, unreliable results could be obtained. For this reason, it is important to monitor converge, and empirical evidence presented by Buuren & Groothuis-Oudshoorn (2011) suggest that convergence is often achieved with a relatively small number of iterations, ranging from 10 to 20.

MICE allows to select the number of iterations. When the number of iterations is set to 1, the algorithm performs a single imputation, although it is designed for multiple imputation. In this configuration, the algorithm could estimate the missing values using to some extent the relationships between variables. However, the result may not include all the complexities and characteristics of the data that can be better captured through multiple iterations.

In the context of well-logs, the multivariate imputation by chained equations (MICE) is considered a promising method for handling missing data. The MICE approach provides a more robust prediction by generating multiple imputed data sets and it has the advantage of simultaneously estimating all well-logs, yielding a complete final data set. A previous study has shown promising results using MICE in a limited number of well-logs (Hallam, Mukherjee, & Chassagne, 2022).

2.4 MACHINE LEARNING

Machine learning is a powerful tool that allows us to develop models based on empirical data, without the need for physical laws. The objective of this technique is to determine the dependencies between variables, which allows predictions and supports decision-making on new and unseen data (Bangert, 2021).

In this project, we focus our attention on supervised learning regression models, a type of machine learning model. Supervised learning is used to predict an outcome based on other input features, and when it is applied in regression, it aims to predict continuous numerical values (Müller & Guido, 2016). This approach is widely used in geosciences using well-log data, for tasks such as lithofacies identification, well-log anomaly detection, production estimation, and rock property prediction (Belyadi & Haghighat, 2021; Bangert, 2021).

Furthermore, machine learning can be used to estimate missing data. This is carried out by considering the feature with missing values as a target vector and using the remaining features for prediction (Gallatin & Albon, 2023). Previous studies by Feng, Grana, & Balling (2021) and Lopes & Jorge (2018) have demonstrated effective use of ensemble methods, including Random Forest and Gradient Boosted Trees, to predict missing values in a single well-log.

However, machine learning models require complete features and sufficient data for training and prediction. When substantial data is missing, it reduces the amount of information available to the model, potentially leading to less accurate and less robust predictions. Since well-log data sets are often incomplete, this limitation prevents the optimal performance of these models. Moreover, most machine learning algorithms cannot deal with missing values in the target and feature arrays by default, which requires addressing them beforehand during the pre-processing step (Gallatin & Albon, 2023).

It is important to highlight that diverse machine learning algorithms can be employed for the imputation process. Some of these models can be combined with certain imputation techniques to obtain a more robust and generalizable prediction (Gallatin & Albon, 2023). For example, the multivariate imputation by chained equations (MICE) approach can be implemented in combination with ensemble methods and other machine learning algorithms. Using this approach allows the prediction of all variables with missing data without extensive pre-processing, considering the relationships between the different variables in the data (Dixneuf, Errico, & Glaus, 2021; Hallam, Mukherjee, & Chassagne, 2022).

In this section, we will explore different supervised learning regression models used in this project, categorized into common baseline models and ensemble models. We will highlight their capabilities, strengths, and effectiveness in well-log data.

2.4.1 Common Baseline Models

I. K-Nearest Neighbors (KNN)

The KNN regressor is an algorithm that predicts the value of a data point considering the values of its k nearest neighbors. The algorithm finds the closest data points in the training data set to make a prediction for a new data point. The choice of the value of the number of neighbors “k” is an important parameter to obtain reliable results (Bangert, 2021). This model is simple to implement, fast and effective in capturing local patterns in the data. The main advantage of this model is that it can give reasonable performance without much tuning, and it is a good reference method before implementing more advanced algorithms (Müller & Guido, 2016).

II. Bayesian Ridge (BR)

The Bayesian Ridge regression model is a linear model that incorporates Bayesian inference for regression analysis. This model introduces regularization that prevents overfitting and improves generalization. One of the primary advantages of the Bayesian ridge regression is its ability to incorporate prior information about parameters and construct good prior distributions (Michimae & Emura, 2022). By assigning probability distributions to the regression coefficients, this algorithm allows a comprehensive representation of uncertainty in the estimated parameter and predictions (Belyadi & Haghighat, 2021). Moreover, Buuren & Groothuis-Oudshoorn (2011) suggest the use of this model for handling missing data in multivariate imputation by chained equations (MICE).

III. Decision Tree

A decision tree can be used for various tasks, not limited to regression. It implements a hierarchy of if/else questions guiding the decision-making process, where each tree node represents a question or a terminal node, also called a leaf. This terminal node provides the final outcome or prediction (Belyadi & Haghighat, 2021).

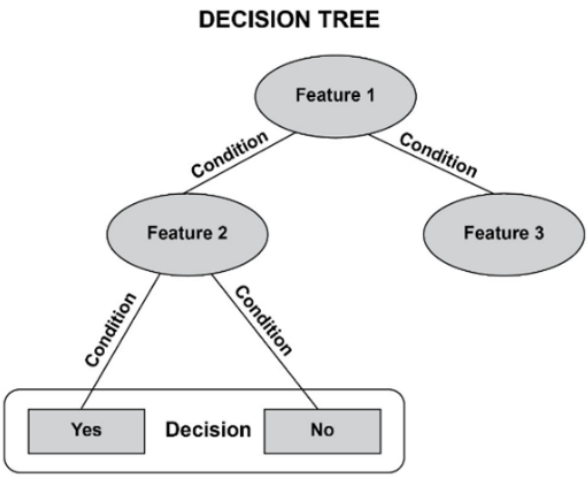


Figure 5. Structure of a decision tree from (Jeyaraman, Olsen, & Wambugu, 2019).

Decision trees have advantages that make them easy for non-experts to visualize and understand such as their interpretability. Furthermore, algorithms based on decision trees do not need pre-processing such as normalization or feature standardization, these algorithms work well with data at different scales. However, the main drawback of this model is that decision trees tend to overfit and offer poor generalization performance (Müller & Guido, 2016). As a solution, ensemble methods are used instead of a single decision tree.

2.4.2 Ensemble Models

Ensemble methods combine multiple machine learning models to create more robust models and improve prediction performance. The most common ensemble methods are based on Decision Trees, such as Random Forest and Gradient Boosting (Müller & Guido, 2016).

I. Random Forest (RF)

Random Forest is an ensemble approach that combines predictions from multiple decision trees to make more accurate predictions (Bangert, 2021). This algorithm addresses the problem of overfitting and sensitivity to outliers in individual decision trees by introducing randomness and diversity into the model. Random Forest builds several decision trees, each randomly trained on different subsets of data and features, using a technique called bagging or bootstrap aggregation (Müller & Guido, 2016).

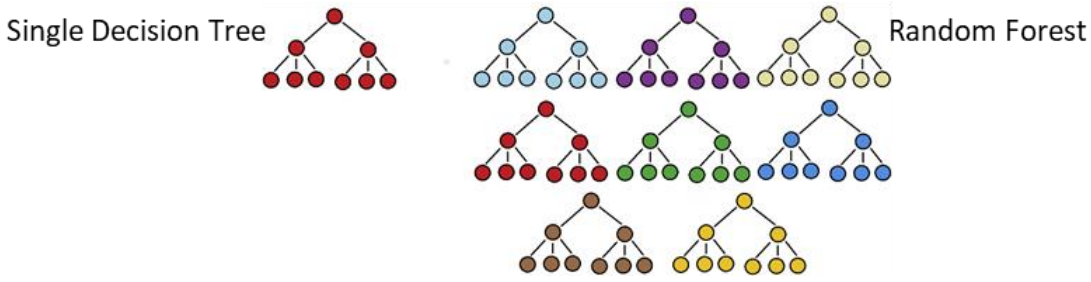


Figure 6. Decision Tree and Random Forest from (Belyadi & Haghighat, 2021).

Bagging is a technique for generating multiple subsets of data by randomly sampling using replacement from the original data set. Each subset, defined as bootstrap sample, is the same size as the original data set, but may contain duplicate instances and exclude some original instances. By creating various bootstrap samples, Random Forest builds different decision trees, introducing model diversity and reducing variance. The result is the average of the predictions of multiple trees (Belyadi & Haghighat, 2021; Müller & Guido, 2016).

As mentioned above, Random Forest can reduce overfitting, improving the stability of the ensemble and improving its generalization capacity. However, this model can be computationally expensive and time consuming, especially for large data sets. Furthermore, tuning hyperparameters for optimal performance can be time consuming and computationally intensive (Belyadi & Haghighat, 2021; Müller & Guido, 2016).

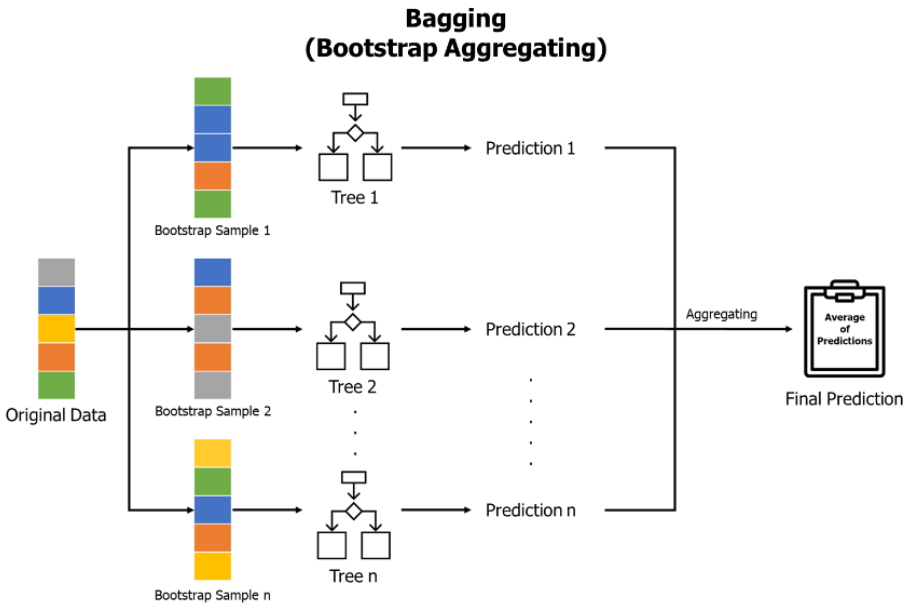


Figure 7. Bagging or Bootstrap aggregation for regression tasks.

II. Gradient Boosting Trees

Gradient Boosting Trees is an ensemble method that sequentially trains multiple decision trees using boosting. Unlike the Random Forest, it focuses on reducing bias by correcting for errors made by previous models (Müller & Guido, 2016). Each new tree is built to improve ensemble performance by assigning more weight to instances that were incorrectly predicted. The predictions from each tree are combined, often using a weighted average or other techniques, to obtain the final prediction (Bangert, 2021).

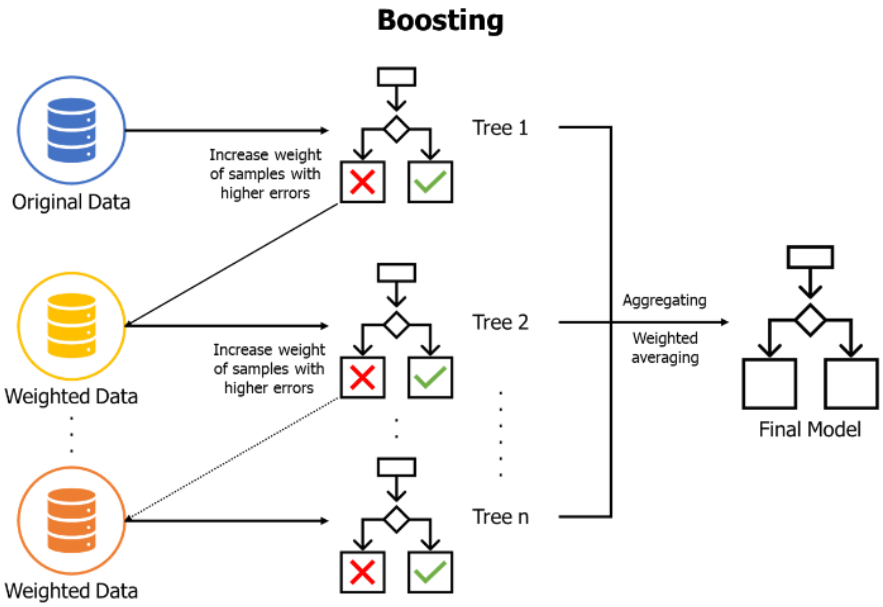


Figure 8. Boosting for Regression tasks.

This technique is frequently employed in machine learning competitions due to its high predictive accuracy. One of the most implemented models is XGBoost (XGB), which not only returns excellent performance, but also it has the capability to handle missing values within a dataset, eliminating the need for the implementation of an imputation strategy. However, its main drawback is that requires careful tuning of the parameter and may take a long time to train (Bangert, 2021).

2.5 A CRITICAL APPRAISAL OF APPROACHES TO ADDRESS MISSING WELL-LOG DATA

Previous studies have addressed the challenge of missing data in well-logs, proposing a variety of techniques that take advantage of machine learning algorithms. Random Forest and Gradient Boosted Trees have shown promising results in predicting and imputing missing values in well-log data (Lopes & Jorge, 2018; Feng, Grana, & Balling, 2021). However, these approaches artificially and randomly introduced missing values, which can misrepresent real-world scenarios, where some well-logs may be completely missing.

An innovative alternative that addresses this problem is the implementation of multivariate imputation by chained equations (MICE) (Hallam, Mukherjee, & Chassagne, 2022). This method applies machine learning models such as K-Nearest Neighbors (KNR), Random Forest (RF), Gradient Boosted Trees (GBT), and Bayesian Ridge (BR), to impute missing values in well-log data sets. The effectiveness of this approach has shown its successful application in Norwegian North Sea data sets, such as the Volve data set and the Force-2020 well-logging machine learning data set, where the Gradient Boosted Trees (XGB) algorithm obtained the best performance.

Nevertheless, the evaluation of MICE focused on three logs: Sonic Sheer Slowness (DTS), Sonic Compressional Slowness (DTC) and Density (RHOB). This limits the generalizability of this imputation method. Different well-logs, such as gamma-ray (GR), resistivity (RES) and neutron porosity (NPHI), measure various physical properties with unique patterns, dependencies, and ranges of values. Consequently, models trained on a specific set of well-logs may not generalize effectively to others that behave differently, respond to different geological conditions, and have variable value ranges. For this reason, it is critical to test these methods in a wider variety of well-logs and geological settings.

Additionally, the study by Lopes & Jorge (2018) has revealed how machine learning models can fill gaps in Neutron Porosity (NPHI) logs, using a data set from offshore Dutch wells in the North Sea. However, the study does not present a clear methodology for dealing with missing values and lacks a detailed explanation of how they used the wells for training and testing, making it challenging to verify their claims and adapt the results to other contexts.

Similarly, the study of Feng, Grana, & Ballin (2021) addressed the prediction of missing measures in the travel-time of the shear velocity (DTS), using the Volve data set. Nevertheless, the paper shows certain limitations. First, the authors did not explicitly mention how missing values were created. It seems that missing values were randomly introduced into the target log to create gaps. Second, they used a complete database for training the model, which is not realistic since well-logs are often incomplete. Additionally, this research was limited to a single data set and used hold-out validation, which may lead to poor generalization and inaccurate performance evaluation due to the temporal or spatial dependencies of the data.

Despite the valuable contributions of these studies in predicting missing well-log data, common limitations indicate the need for more comprehensive investigation. Most of these studies introduce random missing values for the tests, which fail to represent real-world situations where complete well-logs might be missing. They also focus on single or limited number of logs although missing values are found in almost all well-logs. Furthermore, these approaches are often tested on one or two reference data sets, with unclear pre-processing steps Although the Dutch and Norwegian data sets are known for their high-quality records due to rigorous protocols the absence of clear pre-processing guidelines can make it difficult to transfer findings to different geological contexts or data sets.

Based on these limitations, it is necessary to implement a robust and complete machine learning workflow. This methodology should reflect realistic missing well-log data scenarios, accommodate a variety of well-log types, and ensure its generalizability when tested across different data sets and geological contexts. It is essential to develop a framework that not only accurately imputes missing data, but also evaluates its effectiveness and applicability to different data sets. Therefore, the following section provides an overview of a standard machine learning workflow and the steps required to design a robust and reliable framework for missing well-log data imputation.

2.6 MACHINE LEARNING WORKFLOW

Machine learning workflow is a sequence of activities or processes crucial for deploying a successful machine learning model for predictions (Jeyaraman, Olsen, & Wambugu, 2019). Each step plays an important role to ensure effective model training, evaluation, and deployment. The understanding of these processes is essential since it allows developing a more robust proposal to predict the missing values in well-logs. The workflow typically consists of data pre-preprocessing, data splitting, model training, evaluation, and deployment.

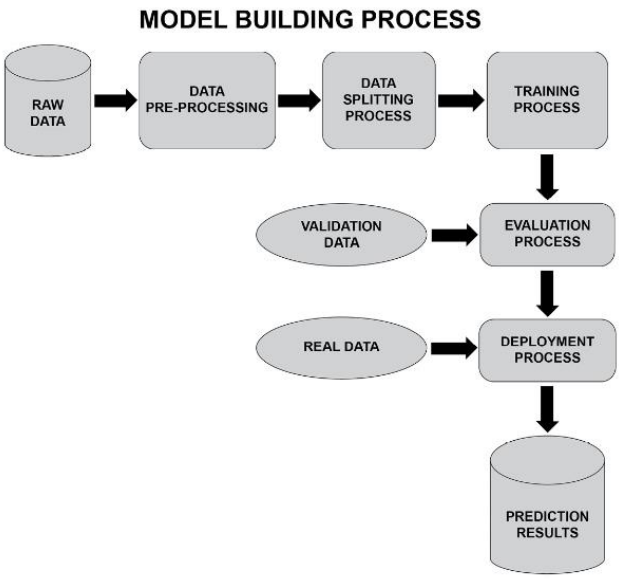


Figure 9. Standard Machine Learning Workflow (Jeyaraman, Olsen, & Wambugu, 2019)

2.6.1 Data Pre-processing

Data pre-processing is the initial step in the machine learning framework to prepare data for training a model. This includes processes such as data cleaning, feature selection, and feature engineering (Jeyaraman, Olsen, & Wambugu, 2019).

Data cleaning is essential to address missing values, outliers, or incorrect values that can affect the performance of a prediction model. Traditionally, these values are handled before training the model by implementing strategies such as data deletion, mean-mode substitution, and imputation methods. In the context of this study, we focus specifically on addressing missing values. However, it is important to note that the evaluation of missing value predictions is not commonly performed directly in current practices. For this reason, the proposed framework not only predicts the missing values, but it also evaluates those predictions directly, filling a gap in the current methodology and providing a more comprehensive analysis of the imputation process.

Following the data cleaning step, the most important features are selected to train the model, which usually are the variable highly correlated with the output variable. Additionally, feature engineering techniques are applied to obtain new features that allow a better performance of the model (Jeyaraman, Olsen, & Wambugu, 2019).

Data scaling is required in certain situations to ensure that certain algorithms are not biased by the magnitude of the data. Feature scaling guarantees that each input variable contributes equally to the learning process, regardless of its range or unit. Even though normalization or standardization of the data is recommended for imputation algorithms like MICE, the suitability of scaling depends on the specific model used within the MICE frameworks. Different models may

have different requirements regarding data scaling. For instance, algorithms such as Random Forest and XGBoost do not require and are not affected by feature scaling. Additionally, data scaling can speed up the training process and improve model performance (Belyadi & Haghighat, 2021). Normalization is a common technique for scaling data in an interval between 0 and 1.

$$Feature\ Normalization\ (X') = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2.6.2 Data Splitting

Data splitting is the partitioning of data between subsets for training, validation, and testing. The goal of this process is to test and evaluate the performance of machine learning models. It is common that 80% of the data is allocated to train the model and the remaining 20% is kept for testing and evaluation (Bangert, 2021; Jeyaraman, Olsen, & Wambugu, 2019).

The training set is crucial for model training, as it allows the model to learn patterns and relationships. For this reason, it is essential to ensure that the training data is diverse and representative of different scenarios to ensure robust generalization. On the other hand, testing sets evaluate the performance of the model on unseen data (Jeyaraman, Olsen, & Wambugu, 2019).

In machine learning applications, it is common to split data randomly between training and test sets. However, this practice may introduce uncertainty to the model since it is possible that the training set is not representative of the overall data distribution, potentially leading to a biased model. Another drawback of this technique is that it only provides a single evaluation of the model, which may not be indicative of its true performance on unseen data (Belyadi & Haghighat, 2021).

To address the limitation of a single evaluation, validation sets and cross-validation techniques are employed. The validation set is a separate fraction of the data that is used as additional indicator to evaluate the model, commonly used to hyperparameter tuning. Cross validation, on the other hand, asses the generalization performance of supervised machine learning models by generating multiple validations sets. This method provides a more robust evaluation compared to using a single split into a training and test sets (Belyadi & Haghighat, 2021).

Cross validation is a technique for evaluating how the trained model will perform on unseen data. This technique takes full advantage of the data set for robust model evaluation, providing a more realistic estimate of model performance when applied to unseen data. In the context of well-logs, the evaluation process is typically performed randomly and does not consider the characteristics of missing values in well-logs. To address this limitation and achieve a more robust evaluation, it is necessary to implement an approach that uses cross-validation in a realistic manner. This approach should replicate how the missing values are present in well-logs, providing a more precise assessment of the performance of a model.

The most popular cross-validation technique is k-fold, where the data set is divided into K subsets or folds. Subsequently, the model is trained and tested K times, with each fold as test set once and the remaining folds as training sets. This approach allows us to evaluate the model on different subsets of data, reducing bias and variance in the evaluation. In k-fold cross-validation, the value of k is typically 5 or 10, but it can be user-specified. It is important to note that as the value of k increases, the computation time also increases (Belyadi & Haghighat, 2021; Müller & Guido, 2016).

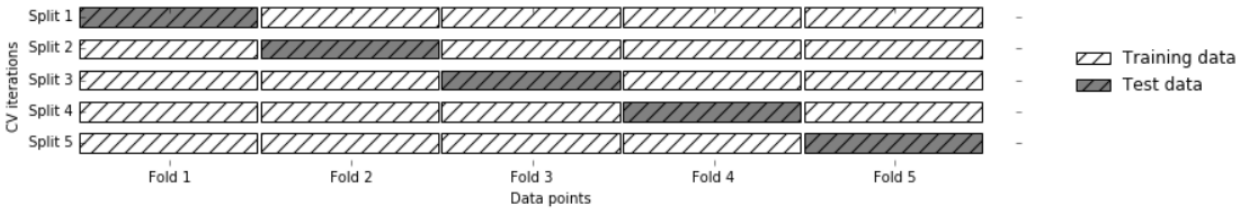


Figure 10. Data splitting in five-fold cross-validation from (Müller & Guido, 2016)

As can be seen in the Figure 10, five-fold cross validation is performed, the data is split into five parts of approximately equal size called folds. Subsequently, a sequence of models is trained, where each model is trained using one-fold as the test set and the remaining folds as the training set. This process is repeated until each fold has been used as the test set. Another variation of cross-validation is GroupKFold, which considers groups in the data that should not be split when creating the training and test sets (Müller & Guido, 2016). This approach is particularly useful in geoscience application, where multiple observations from the same well need to be generalized to other wells. In the following Figure 11, it is shown that in each split in GroupKFold ensures that each group is entirely in the training or test set.

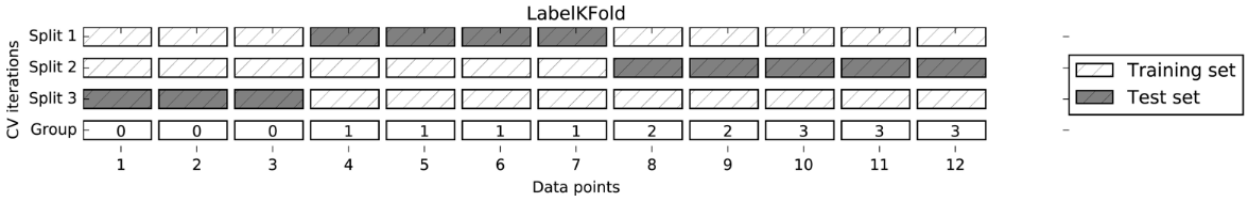


Figure 11. Data splitting with GroupKFold from (Müller & Guido, 2016)

2.6.3 Training Process

The training process in machine learning involves the selection and training of suitable models to learn the underlying patterns and relationships within the labeled training data set. During this step, models are trained using iterative optimization algorithms to minimize errors or maximize likelihood. By adjusting parameters, the model learns to make predictions (Jeyaraman, Olsen, & Wambugu, 2019). Therefore, it is crucial to balance the complexity and generalization of the model, considering bias-variance trade-off.

The bias-variance trade-off is used in machine learning to deal with the relationship between bias and variance in predictive models. Bias in a model refers to the tendency to oversimplify or make certain assumptions that may or may not reflect with the true patterns or relationships present in the data, while variance relates to the variation in predictions for a specific data point (Belyadi & Haghighat, 2021). For instance, high bias refers to a model that oversimplifies the data, leading to underfitting and poor performance on both training and testing data. High variance, on the other hand, refers to a model that fits the training data too closely, including noise and random fluctuations, resulting in overfitting. Overfitting leads to high performance on the training data but poor performance on the testing data as the model fails to generalize to new, unseen data (Bangert, 2021).

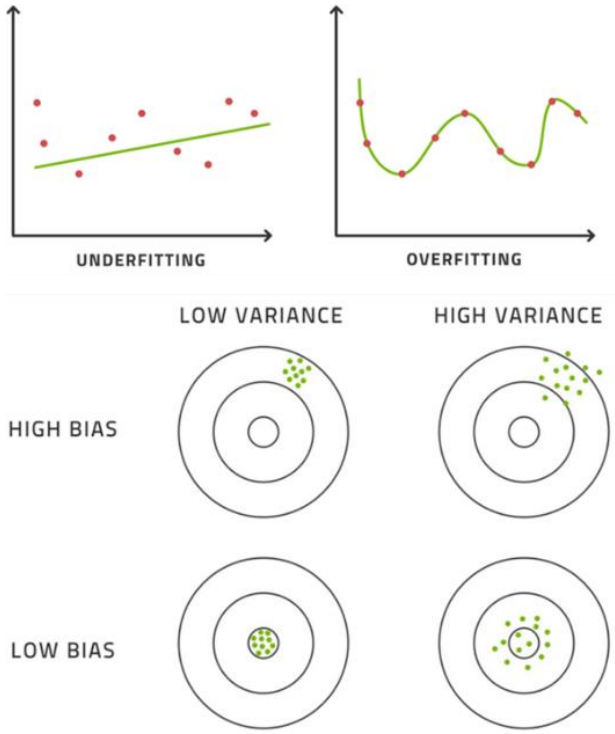


Figure 12. Illustration of underfitting and overfitting by darts example from (Bangert, 2021). High bias – Underfitting and High variance - Overfitting

The optimal model aims to achieve a balance between bias and variance by finding an appropriate level of complexity. To optimize the bias-variance trade-off, techniques such as regularization, cross-validation, and ensemble methods are employed. Regularization prevents overfitting by adding constraint to the model. Cross-validation evaluates performance on unseen data and guides the selection of the best model complexity. Ensemble methods combine multiple models to reduce variance and improve the performance of the predictions (Belyadi & Haghighat, 2021).

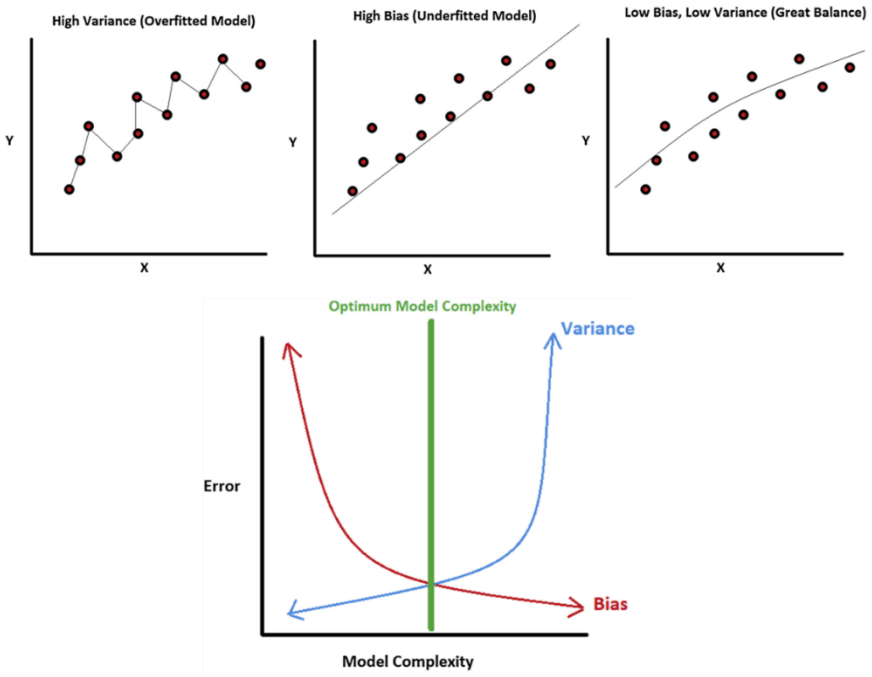


Figure 13. Balance of the model from (Belyadi & Haghighat, 2021)

2.6.4 Evaluation Process

The evaluation process is crucial for assessing the performance and effectiveness of a supervised machine learning model (Jeyaraman, Olsen, & Wambugu, 2019). To evaluate the performance of a trained model, various evaluation metrics are implemented. The following metrics are the most frequent used:

I. Mean Squared Error (MSE)

It measures the average squared difference between the predicted values and the actual values. A lower MSE shows better performance, indicating that the predictions of the model are closer to the true values (Belyadi & Haghighat, 2021).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{\_pred\ i})^2$$

II. Mean Absolute Error (MAE)

It calculates the average absolute difference between the predicted values and the actual values. It provides a measure of the average prediction error. Similarly, a lower MAE implies better performance since it represents a smaller average prediction error (Belyadi & Haghighat, 2021).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_{\_pred\ i}|$$

III. The Coefficient of Determination or R-Squared Scored (R2)

It measures the goodness of fit for a regression model (Müller & Guido, 2016). R2 indicates the proportion of the variance in the output variable that can be explained by the input variables. Higher R2 values, closer to 1, implies better model performance, as it indicates that a larger proportion of the variation in the data is captured by the model (Bangert, 2021).

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_{\_pred\ i})^2}{\sum_{i=1}^{n}(y_i - y_{\_avg\ i})^2}$$

The evaluation metrics allow us to compare the performance of each model or approach implemented in the project. Additionally, it is important to visualize the results to observe trends and patterns that metric evaluations do not provide. The visualizations could make it possible to identify specific regions or intervals in which the model performs well or poorly, highlighting areas for improvement.

2.6.5 Deployment Process

The focus of this process is on selecting the model with optimal performance and integrating it for real-world application after following a machine learning framework. In my opinion, the evaluation of model performance is a critical aspect. By comparing the performance of different models, we can gain valuable insights into their strengths and weaknesses. This allows us to select the most suitable model or to determine if the approach is reliable to use in real world applications.

3 | DATA ANALYSIS

This section investigates and analyzes the three data sets used for this project: Montney, Beetaloo and Force-200. Each of these data sets represents a unique geologic formation with distinct data preprocessing techniques, providing a rich source of information for well-log data exploration. The following sections are divided into three main subsections, each of which focuses on one of the chosen data sets. Within each subsection, the geological context is explained, and exploratory data analysis (EDA) is performed. The geological context provides essential background on formation, sedimentology, stratigraphy, and other relevant geological aspects. The EDA section, on the other hand, describes how these data sets were selected and pre-processed, and provides statistical information about the data sets, including analysis of missing values.

3.1 MONTNEY

3.1.1 Geological Context

The Montney formation is located in the Western Canadian Sedimentary Basin (WCSB), as can be seen in Figure 14. This formation spans parts of Alberta to northeastern British Columbia and exhibits a variety of unique sedimentological and stratigraphic features. The Montney formation has been used for oil and gas production since it contains conventional and unconventional petroleum accumulations (Crombez, Rohais, Baudin, & Euzen, 2016; Ducros, Sassi, Vially, Euzen, & Crombez, 2017).

The WCSB is a foreland basin where the Montney formation lies. This foreland basin gets thicker from East to West with maximum thickness of 5 km at the deformation limit. The sediments started to accumulate in the basin from the Early Paleozoic to the Cenozoic. The formation and subsequent evolution of this region were significantly influenced by the orogeny of the Canadian Cordillera, which began in the Middle Jurassic. This event transformed the WCSB from the western margin of Pangea into a foreland basin (Crombez, Rohais, Baudin, & Euzen, 2016; Ducros, Sassi, Vially, Euzen, & Crombez, 2017).

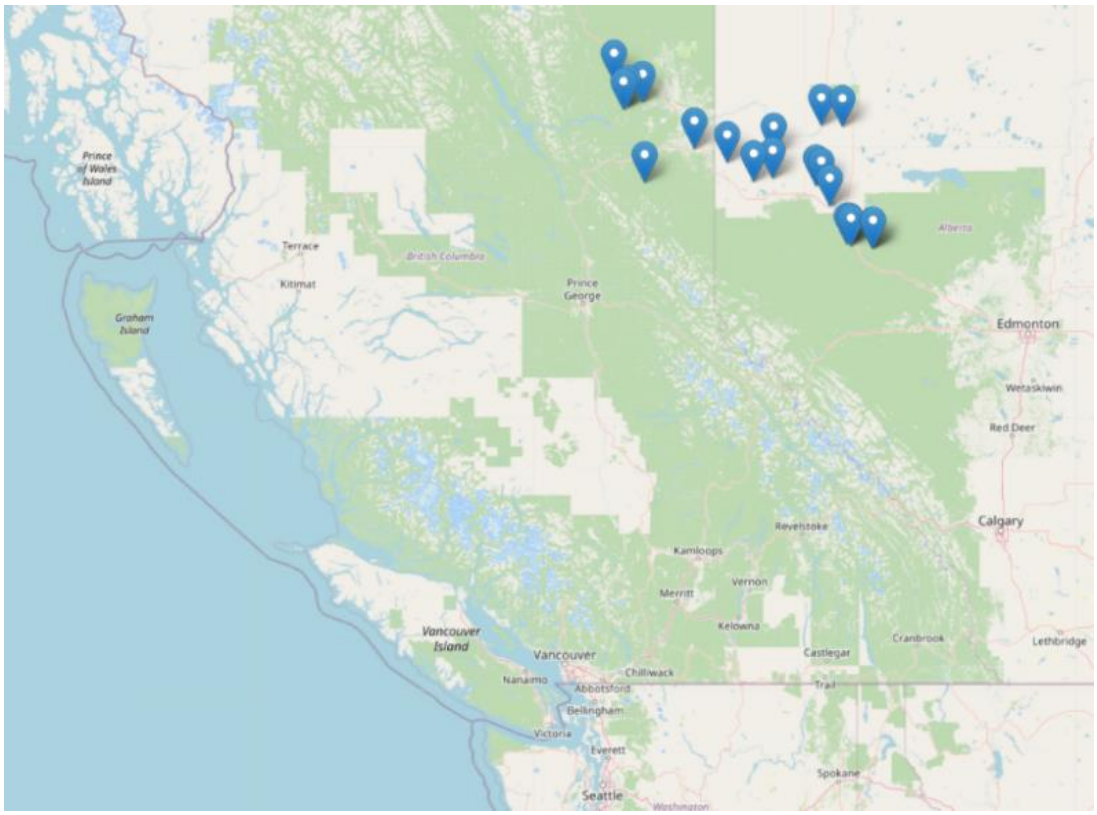


Figure 14. Geographical location of Montney Wells.

The sedimentology of the Montney formation is complex and diverse. The Montney formation initially formed in the Peace River area, and it is characterized by dark gray shale to argillaceous siltstone interbedded with shale. In Alberta, this formation changes from dolomitic bioclastic sandstone to fine to very fine-grained grey sandstone, siltstone, and shale towards British Columbia. The most proximal deposits recorded in Alberta and British Columbia are tidal deposits. Therefore, the formation contains delta deposits dominated by waves, tides, and rivers along with turbiditic deposits (Crombez, Rohais, Baudin, & Euzen, 2016). The Montney formation transitions from foreshore, tidal, and shoreface sandstone, siltstone, and coquina bed deposits to offshore-transition and offshore organic-rich siltstones and turbiditic deposits (Ducros, Sassi, Vially, Euzen, & Crombez, 2017).

The Montney formation was deposited during a second-order sequence, approximately 5 million years, which is subdivided into several third-order sequences composed of multiple para-sequences (Crombez, Rohais, Baudin, & Euzen, 2016). The Triassic succession, which includes the Montney formation, corresponds to a mix of siliciclastic, evaporitic, and carbonate sedimentation (Ducros, Sassi, Vially, Euzen, & Crombez, 2017). In British Columbia, the formation is divided into three units: Lower, Middle, and Upper Montney, each reflecting specific periods in the geologic timescale (Crombez, Rohais, Baudin, & Euzen, 2016). The Montney and Doig formations together form a pro-grading clastic ramp deposited during the Early to Middle Triassic,

highlighting the transition from a passive to an active margin setting (Ducros, Sassi, Vially, Euzen, & Crombez, 2017).

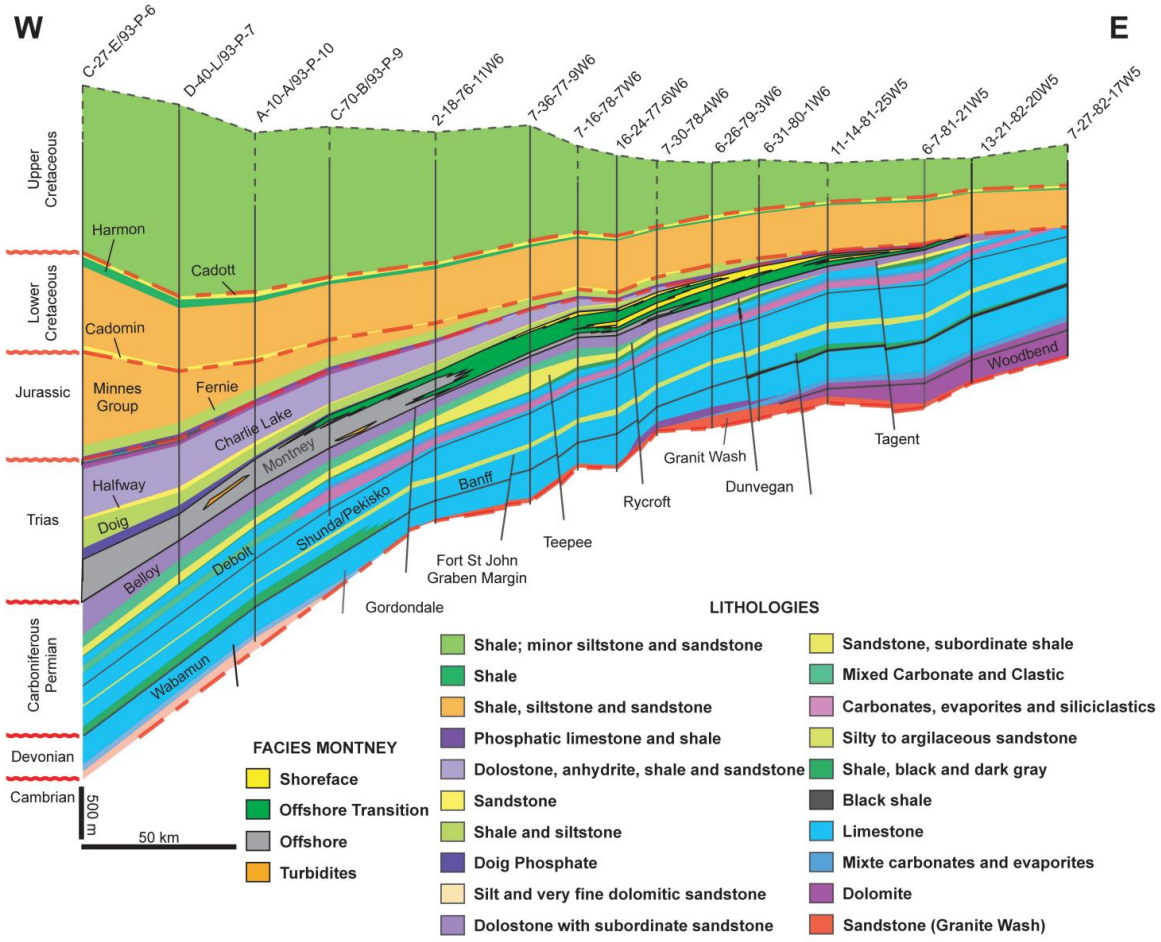


Figure 15. Cross-section with the main stratigraphic intervals and lithologies of the Montney formation from (Ducros, Sassi, Vially, Euzen, & Crombez, 2017).

3.1.2 Exploratory Data Analysis (EDA)

I. Selection and Description of the Data Set

The Montney data set contains wellbore information from the year 2001, including lithography and stratigraphy. This dataset was selected for its diverse geologic complexity, characterized by diverse sedimentological and stratigraphic features. Its sedimentary composition includes shales, siltstones, and various sandstones, and the formation exhibits tidal, wave-dominated deltas and river deposits with turbiditic deposits. Due to the high variability of depositional environments, from shallow to deep marine, different wells can have completely different patterns.

Furthermore, the presence of missing values in the data set reflects the common challenge in well log data, aligning with the objective of this thesis to assess the ability of the proposed framework to estimate these missing values. For these reasons, the Montney data set represents

a challenge for machine learning applications, which is interesting for evaluating the generalization and performance of the model in predicting missing values in different wells.

Unlike other data sets, the Montney data set preserves the integrity of the original data. Although a selection of wells was performed, other preprocessing steps, such as depth alignment, were not performed. Therefore, machine learning models can be evaluated in a more realistic scenario, which conserves how the data is originally obtained.

II. Data Analysis

This data set initially comprises 122,374 samples, with 19 features extracted from 20 wells. However, for the purpose of this study, we focus only on well-logs; therefore, we reduce the number of features to 12. These features include identification of the well, depth, coordinates, well-logs, lithography, and stratigraphy.

Table 2. Montney features used for the project.

Feature	Description	Data Type
WELL_ID	Identification of the well	Categorical
X	X-coordinate of the well	Continuous
Y	Y-coordinate of the well	Continuous
DEPTH	Depth of the well	Continuous
RHOB	Bulk Density	Continuous
GR	Gamma Ray	Continuous
DT	Sonic Travel Time	Continuous
RES_10	Log 10 of Resistivity	Continuous
SP	Spontaneous Potential	Continuous
NPHI	Neutron Porosity	Continuous
LITHO	Lithology	Categorical
STRAT	Stratigraphic information	Categorical

As can be seen in the Table 2, the data present 3 categorical characteristics and 9 continuous ones. The categorical variables are well identification, lithography, and sequence stratigraphy. Regarding the lithography, the data set presents 9 different formations, of which 60% of the data is made up of the Montney formation. This formation is followed by the Doig and Halfway formations with 28% and 10% of the data respectively. On the other hand, the most frequent lithostratigraphic groups are HST4, HST1, LST3 and HST2 with percentages of 32%, 16%, 11% and 11% respectively. Therefore, the high amount of data from certain formation or lithostratigraphy groups may lead to biased estimates, impacting the ability of the model ability to accurately predict properties in other formations or groups. To address this issue, it is possible to improve performance on imbalanced data by implementing MICE with robust models like Random Forest and XGBoost. These machine learning models can handle imbalanced data better than other algorithms, capturing complex relationships between the well-logs.

The resistivity log was transformed to log base 10, which is a common practice that helps to reduce data variability and skewness. It is important to mention that we transform all the resistivity logs for the 3 data sets. The frequency distribution of the original resistivity and the transformed resistivity are shown in the following figure:

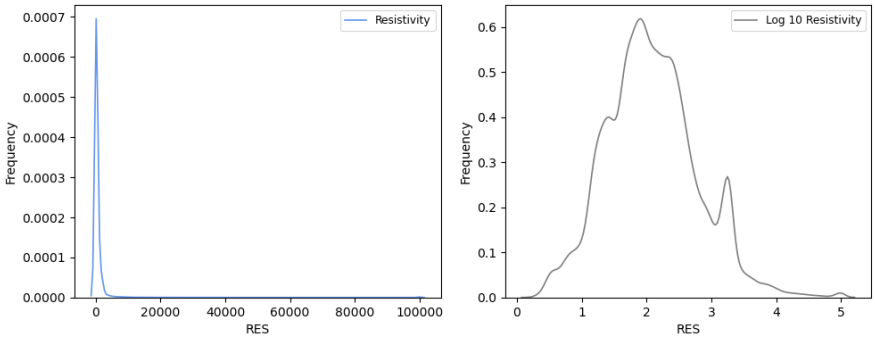


Figure 16. Histogram of Resistivity vs Log 10 Resistivity for Montney.

Table 3. Descriptive Statistics for numerical features in Montney data set.

	count	mean	std	min	25%	50%	75%	max
DEPTH	122374	2236.37	423.63	825.62	2043.20	2214.14	2367.30	4670.80
RHOB	97195	2.63	0.06	1.89	2.60	2.63	2.66	3.13
GR	78812	100.99	38.46	0.00	79.25	104.19	122.16	746.48
DT	81112	200.58	23.54	3.36	185.05	199.09	209.80	505.83
RES_10	98066	2.09	0.70	0.28	1.62	2.04	2.51	5.00
SP	92701	68.89	347.96	-362.23	-203.41	-6.16	148.54	964.69
NPHI	93366	0.12	0.05	0.00	0.09	0.12	0.15	0.64

Furthermore, descriptive statistics for numerical variables for well-logs, including depth, is performed. The data set presents a wide range of depth in the wells, ranging from approximately 800 m to 4,600 m. It was observed in the Figure 15, the variability of depth may be due to that the wellbores are in a foreland basin, which gets thicker from East to West. For this reason, wells located in the eastern part of the formation tend to be shallower than those located in the western part.

Furthermore, the complexity and diversity of the sedimentology of this region is reflected in the well-log data. Different rock compositions in the basin, such as shales, siltstones, and different types of sandstones, from dolomitic bioclastic sandstones to fine-grained to very fine-grained gray sandstones, could be associated with variability in well-log measurements. For instance, sandstones present higher porosity and lower density than shales, resulting in variations in the readings. In addition, different types of depositional environments such as tidal, wave-dominated deltas and rivers with turbiditic deposits along the basin can introduce variations in recorded readings since they can result in sequences of different lithologies, grain sizes, and sedimentary structures. For instance, wave-dominated delta present coarser sediments than tidal

environments, resulting in higher porosity readings. Therefore, we consider that implementing MICE may predict missing values considering the observed variability and heterogeneity in the well-log data, which are influenced by the different rock types and depositional environments.

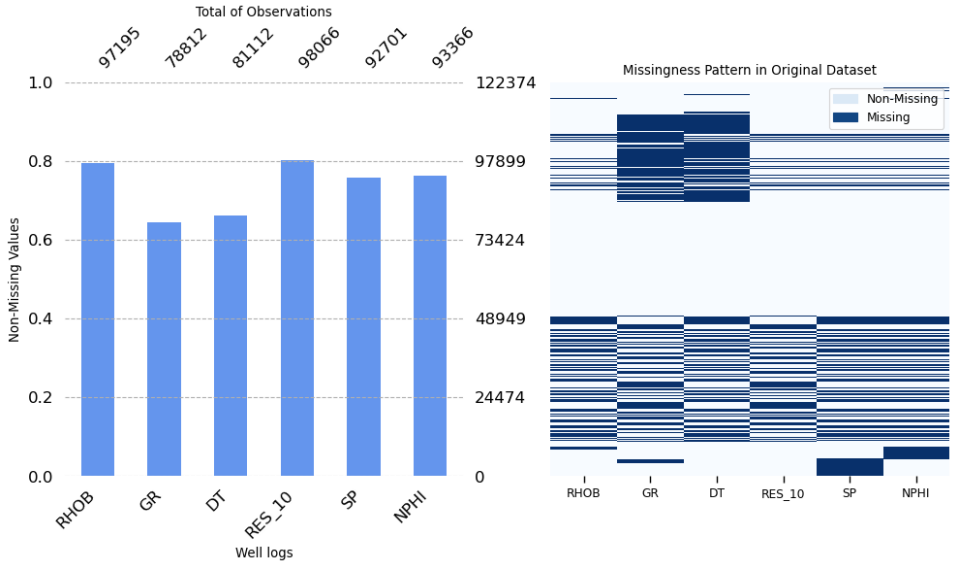


Figure 17. Missingness Analysis of well-logs in Montney data set.

The data contains missing values in each of the well-logs, as illustrated in Figure 13. Missing values occur in a range of 20% to 36%, with percentages of missing values for RHOB, GR, DT, RES_10, SP, and NPHI of 20%, 36%, 34%, 20%, 24% and 24%, respectively. We observe a generalized missingness pattern with a missing not at random mechanism. Therefore, methods such as data deletion, mean-mode substitution and simple imputation are not suitable approaches to deal with missing values in well-logs due to their limitations and potential biases.

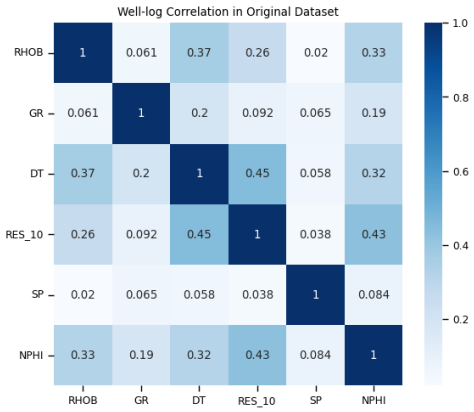


Figure 18. Correlation Matrix of well-logs in Montney data set.

The well-log correlation matrix shows that DT and RES_10 have the highest correlation of 0.45. Similarly, NPHI and RES_10 logs are moderately correlated, 0.43. It is also seen that DT and

RHOB are moderately correlated. Therefore, we consider that MICE can take advantage of all these correlations to impute missing values, including the low correlated well-logs.

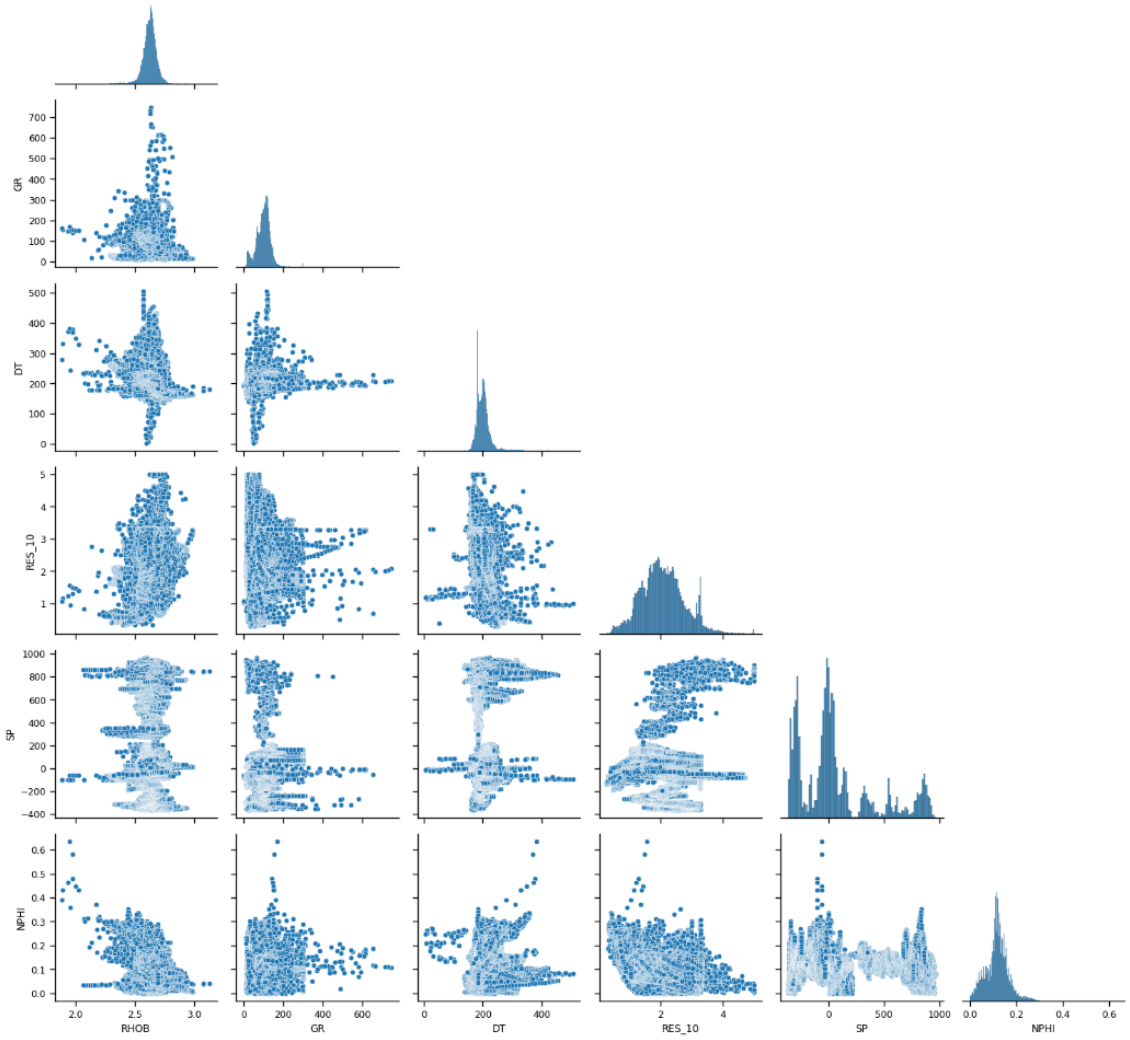


Figure 19. Montney pair plot of well-logs.

As can be seen in the scatter plots and distributions, the data is dispersed with no clear pattern or trend. The well-logs relationship is not clearly linear, which makes it difficult to interpret visually. Although pair plots provide insights about how features are related, they are limited in capturing other relationships within the data. However, MICE may capture these ignored relationships in the data set since this method incorporates the relationships and interactions between variables when inputting missing values. This could allow us to capture associations and interactions between well-logs that are not identified from simple visual inspection.

3.2 BEETALOO

3.2.1 Geological Context

Located in the Northen Territory of Australia, the Beetaloo sub-basin is part of the McArthur Basin, and it is one of the oldest petroleum systems in the world. This intracratonic basin contains sandstones and unconventional shale reservoirs such as the middle Velkerri and lower Kyalla shales, which were deposited approximately between 1400 - 1280 Ma and 1250 - 1190 Ma, respectively (Crombez, et al., 2022; Faiz, et al., 2021).

Figure 20. Geographical location of Beetaloo Wells.

The Beetaloo sub-basin is defined by several fault zones and geological features. The basin is bounded by the Batchelor and Urapunga fault zones in the north, Batten Fault Zone and the Murphy high mark to the east, the Helen Springs high in the south, and the Birrindudu Basin to the west (Crombez, et al., 2022). Moreover, the sub-basin has experienced multiple burials and uplifts, recording the formation and breakup of different supercontinents and tectonic activities (Faiz, et al., 2021).

The Beetaloo sub-basin includes formations such as Velkerri and Kyalla, which are part of the Roger group that includes rocks from the Mesoproterozoic age. These formations are mainly composed of siltstone and claystone, with occasional sandstone and limestone. The Velkerri formation is known for its organic rich mudstones and siltstones, indicative of a high-energy, wave-dominated nearshore depositional environment. The Amungee member, situated in the middle Velkerri, is considered the most attractive member for unconventional resources (Piane, et al., 2021). The Moroak sandstone, another important formation of Beetaloo, is characterized by fine

to medium grained sandstone interlayered with minor coarse-grained sandstone, conglomerate, and siltstone. These formations have been shaped by a variety of depositional environments from shallow marine deltaic to offshore, including wave and fluvial environments (Crombez, et al., 2022).

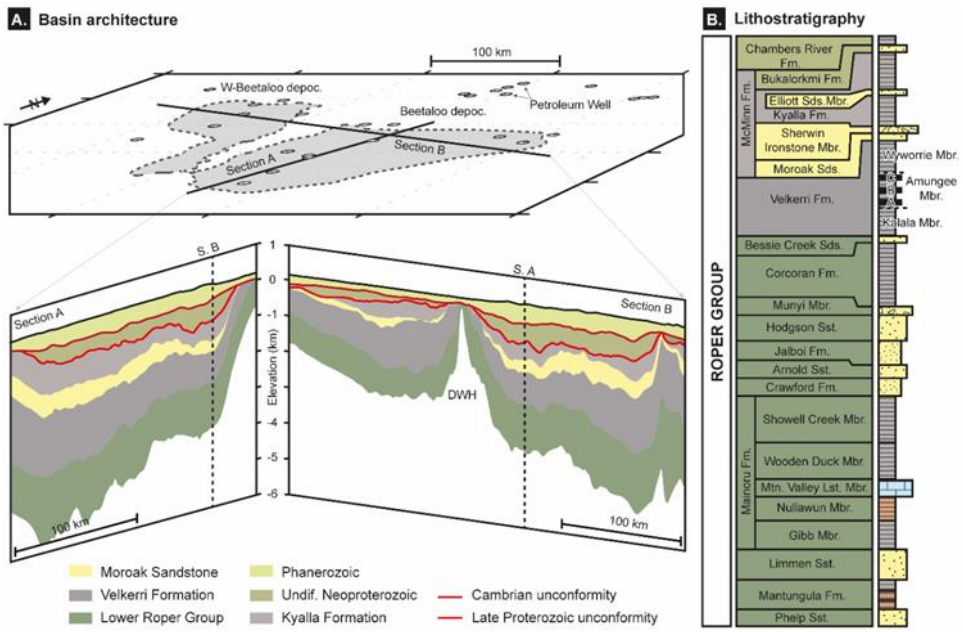


Figure 21. 2D cross-section from the subsurface of the Beetaloo sub-basin with the simplified lithostratigraphy of the Roger group from (Crombez, et al., 2022).

3.2.2 Exploratory Data Analysis (EDA)

I. Selection and Description of the Data Set

The Beetaloo data set contains well-log information, including lithography and stratigraphy from the North of Australia. This data set was chosen for its unique geological context, characterized by a variety of depositional environments such as shallow marine deltaic to offshore, along with wave and fluvial environments. Compared to Montney, Beetaloo may present less heterogeneity, making it an easier case study and providing a contrasting environment in which to assess the proposed framework.

Furthermore, the lack of pre-processing allows for a more realistic evaluation of the proposed framework, and missing values in the data reflects a common challenge in well-logs. These factors, including the geological context of Beetaloo, align with the objectives of this thesis and present another different scenario to rigorously test the performance of the model to estimate missing values. Like Montney, the Beetaloo data set was subjected to well selection and resistivity transformation using the base 10 logarithm.

II. Data Analysis

This data set contains 352,143 samples with 12 features extracted from 32 wells after removing unnecessary features. These features include well identification, depth, coordinates, well-logs, lithography, and stratigraphy:

Table 4. Beetaloo features used for the project.

Feature	Description	Data Type
WELL_ID	Identification of the well	Categorical
X	X-coordinate of the well	Continuous
Y	Y-coordinate of the well	Continuous
DEPTH	Depth of the well	Continuous
RHOB	Bulk Density	Continuous
GR	Gamma Ray	Continuous
DT	Sonic Travel Time	Continuous
RES_10	Log 10 of Resistivity	Continuous
SP	Spontaneous Potential	Continuous
NPHI	Neutron Porosity	Continuous
LITHO	Lithology	Categorical
STRAT	Stratigraphic information	Categorical

The data presents 3 categorical features and 9 continuous. The categorical variables of lithography have 13 different formations, which 30% of the data correspond to Middle Velkerri Formation. This formation is followed by Kyalla Formation, Upper Velkerri Formation, Moroak Sandstone, and Lower Velkerri Formation with 18%, 18%, 11% and 10%, respectively. As mentioned above, in situations where certain categories are overrepresented, the predictive model could be biased and perform poorly in underrepresented categories. For this reason, we suggest that using MICE with robust machine learning models may solve this problem of imbalanced data.

Table 5. Descriptive Statistics for numerical features in Beetaloo data set.

	count	mean	std	min	0.25	0.5	0.75	max
DEPTH	352143	1414.14	924.82	1.40	613.81	1287.57	2050.69	3920.03
RHOB	227657	2.63	0.17	1.21	2.56	2.62	2.68	4.63
GR	215742	145.88	58.58	3.61	115.01	152.15	181.38	422.92
DT	252249	78.66	13.18	0.22	69.17	77.92	88.41	145.84
RES_10	282225	1.44	0.56	0.00	1.13	1.37	1.67	5.00
SP	200326	-36.58	101.26	-283.75	-124.52	-6.89	41.39	225.62
NPHI	168481	0.15	0.07	0.00	0.11	0.16	0.20	1.00

The descriptive statistics of the numerical variables indicate significant diversity in the data set. For example, well depths range from 1.4 m to 3,920 m with an average depth of approximately

1,414m. The average well depth in the Beetaloo dataset is less than that in the Montney dataset, which could be due to multiple burials and uplifts that the Beetaloo sub-basin has experienced. Moreover, we observe values with wide ranges such as GR, DT, and SP. These may be influenced by various factors related to the geological context. For example, clay-rich formations such as the Velkerri and Kyalla shales tend to have higher GR values.

Figure 22. Missingness Analysis of well-logs in Beetaloo data set.

In the Beetaloo data set, missing values are present in all well-logs in a range of 20% to 52%. The percentages of missing values for RHOB, GR, DT, RES_10, SP, and NPHI are 35%, 39%, 28%, 20%, 43%, and 52%, respectively. We observe a generalized pattern of shortages with a missing not at random mechanism. Therefore, like Montney, data deletion, mean-mode substitution and simple imputation are not recommended alternatives to deal with them. It is important to note that this data set presents well-logs with missing values of around 50%. This situation can affect the performance of the model and consume more time since MICE could struggle predicting the missing values.

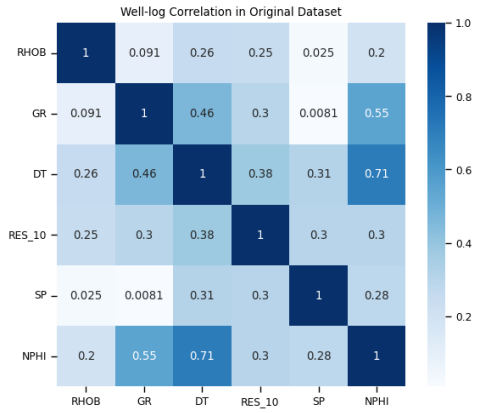


Figure 23. Correlation Matrix of well-logs in Beetaloo data set.

The correlation matrix shows highly correlated well logs as DT and NPHI with 0.71. Similarly, GR and NPHI also present a moderate correlation of 0.55. However, the remaining correlations are moderate to low. It is important to note that MICE considers all features, even if they are poorly correlated, in imputation, but highly correlated well logs may contribute more to missing value predictions.

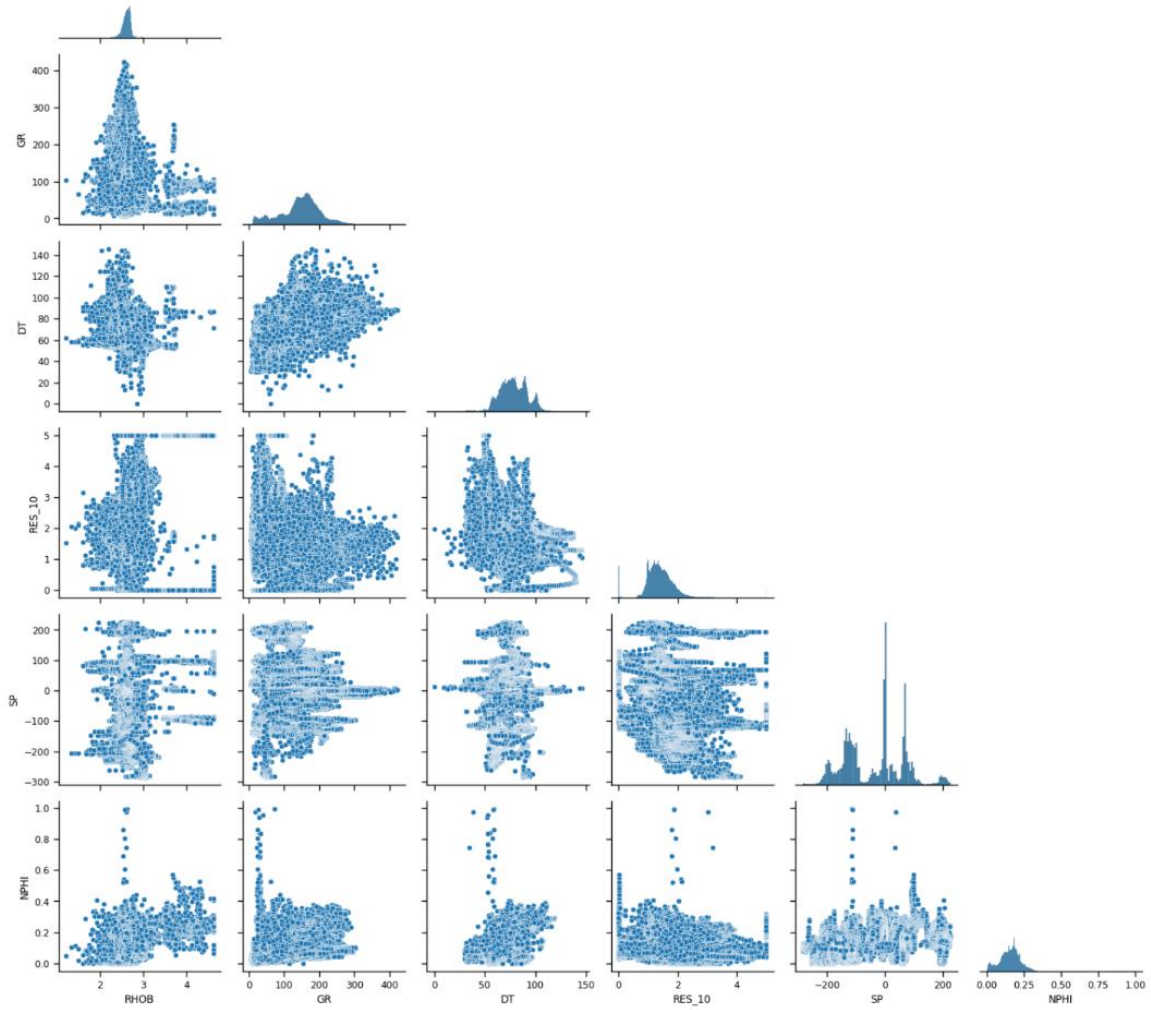


Figure 24. Beetaloo pair plot of well-logs.

The pair plot shows a very complex data set with sparse data and high variability. No clear pattern or trends are identified in the scatterplots. However, MICE can capture hidden relationships in the data set that other techniques can ignore and are visually difficult to interpret.

3.3 FORCE-200

3.3.1 Geological Context

Viking Graben, located in the North Sea, is one of the fields with the largest amount of hydrocarbons in Western Europe. It was formed due to crustal extension throughout the Mesozoic and shares a complex geological history. The geologic history of the Viking Graben is complex with diverse sedimentary environments and stratigraphy (Jackson & Larsen, 2009; Holgate, Jackson, Hampson, & Dreyer, 2013).

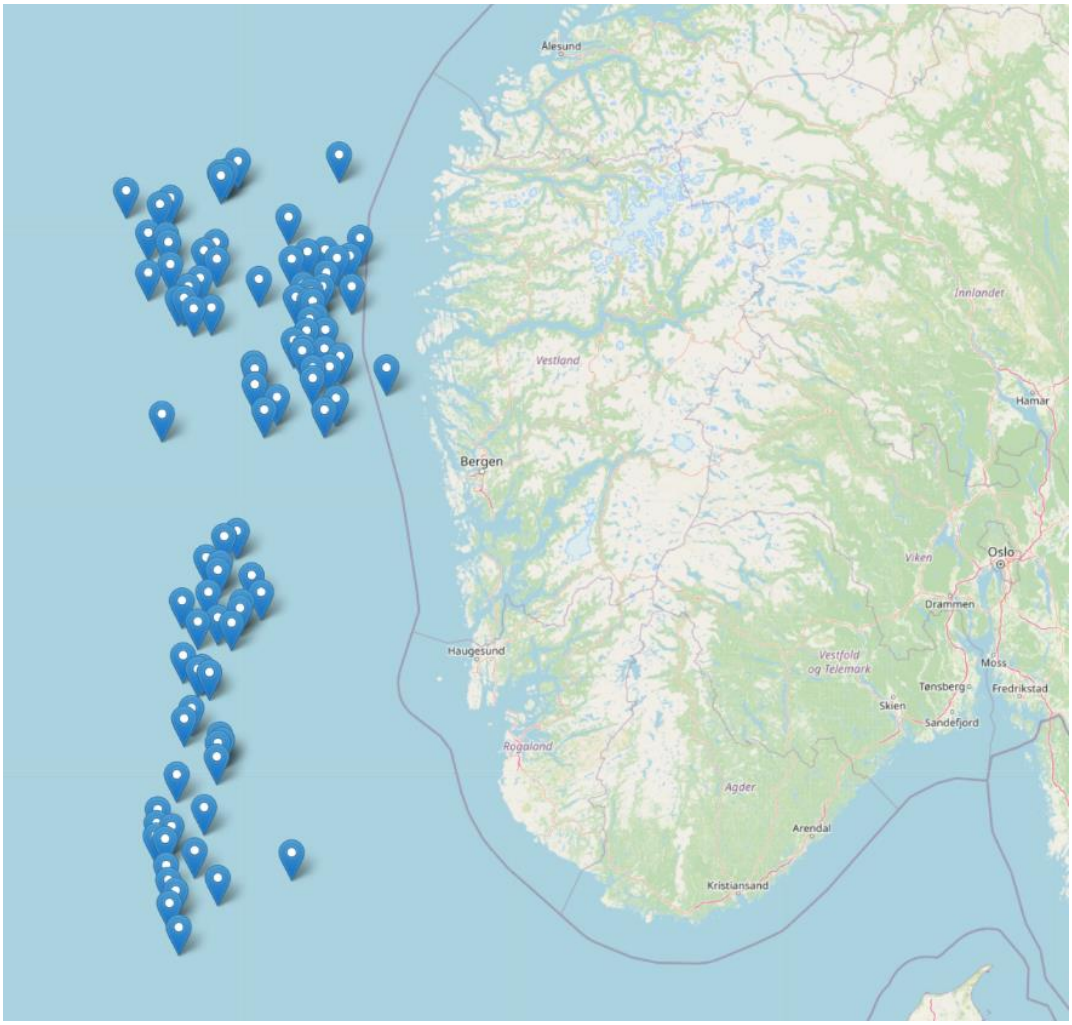


Figure 25. Geographical location of Force-200 Wells.

The South Viking Graben forms a narrow rift basin at the southern end of the Viking Graben. The structural configuration of the South Viking Graben was caused by extension phases during the Permo-Triassic and Late Jurassic (Murillo, Horsfield, & Vieth-Hillebrand, 2019). This Graben presents a network of normal faults in E-W direction. The first step of fault-controlled subsidence occurred approximately in the early to late Permian. During the Permian to Triassic, evaporite-dominated units (Zechstein Group) were deposited in a marine bay. These deposits were followed by shales (Smith Bank Formation) and sandstone-dominated clastic units (Skagerrak Formation) during the Triassic. The Middle and Late Jurassic experienced a fault-controlled subsidence and a eustatic rise in sea level, resulting in the deposition of a deeper succession within the delta plain of the southern Viking Graben (Sleipner Formation) and the passage of shallow marine deposits (Hugin Formation). upwards in shelf deposits (Heather Formation), which in turn are overlain by deep marine deposits (Draupne Formation) (Jackson & Larsen, 2009).

Similarly, the Northern Viking Graben also experienced an initial phase of extension during the Permo-Triassic, followed by multiple extension passes during Middle-Late Jurassic. The basin was characterized by tectonic quiescence and spatially uniform subsidence during the Early Jurassic. However, the Middle-Late-Jurassic rifting event can be divided into several phases of basin-wide rifting and fault-related subsidence. The Northern Viking Graben contains three important Middle-Late Jurassic sandstone formations, namely the Krossfjord, Fensfjord and Sognefjord formations. These formations overlay and interfinger with the Heather Formation, a series of siltstones and mudstones, which is split into three units on the Horda Platform: Bathonian, Callovian, and Oxfordian–Kimmeridgian. These three units either overlap or underlie with the Brent, Krossfjord, Fensfjord, and Sognefjord formations in the stratigraphic column (Holgate, Jackson, Hampson, & Dreyer, 2013).

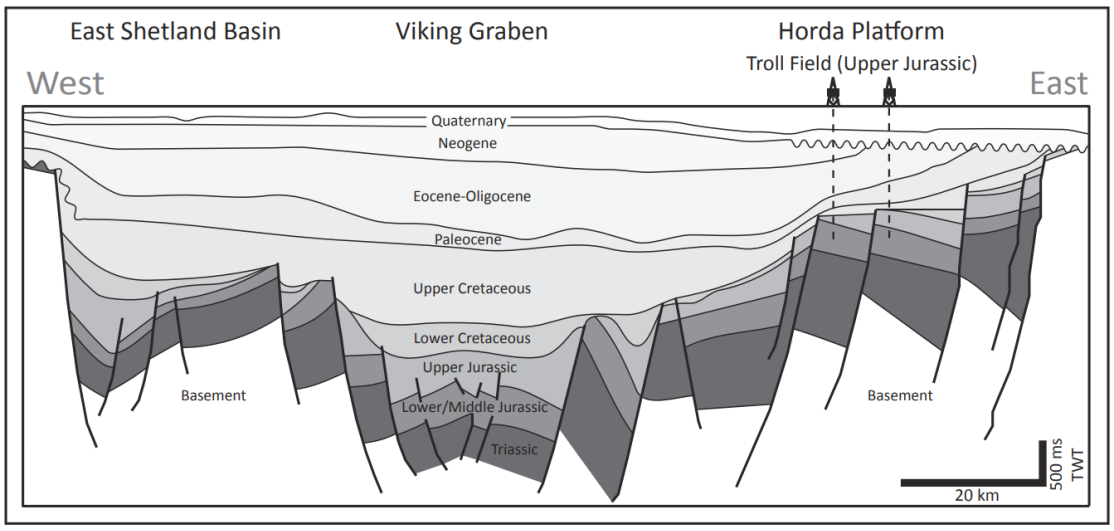


Figure 26. Cross-section from West to East the Viking Graben from (Holgate, Jackson, Hampson, & Dreyer, 2013).

3.3.2 Exploratory Data Analysis (EDA)

I. Selection and Description of the Data Set

The Force-200 data, obtained from the “Machine Learning Lithology Prediction Competition” by FORCE, differs from other datasets in its complex geological context and rigorous pre-processing. The geology of Force-200 is characterized by clayey and sandy sediments and carbonates deposited mainly during the Jurassic and Cretaceous ages. The North Sea presents an intricate rift basin structure, characterized by multiple phases of extension and diverse sedimentary environments throughout different geologic periods. The presence of intricate fault-controlled subsidence, various sandstone formations, and other unique geological features provides a contrasting and challenging environment in which to evaluate the proposed framework.

Additionally, the Force-200 data set has been used in previous research (Hallam, Mukherjee, & Chassagne, 2022), allowing for comparisons and validations with this study. Unlike the Beetaloo and Montney data sets, Force-200 has experienced and substantial pre-processing, extensive pre-processing, following the strict Norwegian Protocol guidelines for reporting well data and the additional cleaning process by experts for the machine learning competition. This intensive preprocessing may differ from the original data acquisition scenario, but it offers a unique opportunity to evaluate the performance of the model in predicting missing values within a different processing context.

The Norwegian protocol pre-processing recommendations mainly cover data cleanup, depth shifting, and interpolation for reporting well data The data is cleaned-up during the creation of the composite log, including removal of sonic cycle skips, normalization of SP, and corrupted data is replaced by other data or by null values. Subsequently, the depth shifting procedure is carried out to ensure precise alignment of data curves, aligning them with an accuracy of 0.2 meters. The gamma-ray log or the first gamma-ray run in hole is used as the reference log for depth shifting. Depth shifting is critical when depth discrepancies between log traces exceed 0.5 meters. Furthermore, the interpolation is employed to address sections with invalid data or no data, but only if the gap between data points exceeds the geologically insignificant distance, typically up to 1 meter (Directorate, 2018).

II. Data Analysis

The Force-200 data sets include separate training and test sets. The train set contains 1,170,511 observations and 98 wells, whereas the test set has 122,397 observations with 10 wells. These two datasets were built and preprocessed by different companies and experts based on the Norway protocol for well data (Directorate, 2018). We keep 16 features of these data sets, removing unnecessary features for this project. Like previous data sets, we conserve well ID, depth, coordinates, well-logs, lithography, and stratigraphy. Additionally, the resistivity logs were also transformed to log base 10.

Table 6. Force features used for the project.

Feature	Description	Data Type
WELL_ID	Identification of the well	Categorical
X	X-coordinate of the well	Continuous
Y	Y-coordinate of the well	Continuous
DEPTH	Depth of the well	Continuous
RHOB	Bulk Density	Continuous
GR	Gamma Ray	Continuous
DTC	Shear Wave Sonic	Continuous
DTS	Compressional Wave Sonic	Continuous
RD10	Log 10 of Deep Resistivity	Continuous
RM10	Log 10 of Medium Resistivity	Continuous
RS10	Log 10 of Shallow Resistivity	Continuous
SP	Spontaneous Potential	Continuous
NPHI	Neutron Porosity	Continuous
PEF	Photo Electric Factor	Continuous
FORMATION	Lithology	Categorical
STRAT	Stratigraphic information	Categorical

As can be seen, the data present 3 categorical characteristics and 13 numerical ones. The categorical variables of lithology have 69 different formations, which 15% correspond to the Utsire formation. This formation is followed by the Kyrre, Lista, Heather and Skade formations with 8%, 6%, 6% and 4%, respectively. Furthermore, the most frequent groups in the data are the Hordaland, Shetland, Viking, Rogaland, Dunlin and Norland groups with 25%, 20%, 11%, 11%, 10% and 10% of the train data. This data set is more balanced compared to Montney and Beetaloo data sets.

Table 7. Descriptive Statistics for numerical features in Force data set.

	count	mean	std	min	25%	50%	75%	max
DEPTH	1170511	2184.09	997.18	136.09	1418.60	2076.60	2864.39	5436.63
RHOB	1009242	2.28	0.25	0.72	2.09	2.32	2.49	3.46
GR	1170511	70.91	34.23	0.11	47.63	68.37	89.04	1076.96
DTC	1089648	113.36	29.99	7.42	87.83	109.59	140.77	320.48
DTS	174613	204.66	71.07	69.16	155.94	188.20	224.65	676.58
RD10	1159496	0.48	0.33	0.01	0.28	0.39	0.55	3.30
RM10	1131518	0.48	0.30	0.00	0.28	0.39	0.57	3.30
RS10	630650	0.50	0.37	0.00	0.27	0.38	0.61	3.34
SP	864247	60.03	76.57	-999.00	32.40	55.39	83.39	526.55
NPHI	765409	0.33	0.13	-0.04	0.24	0.33	0.42	1.00
PEF	671692	6.32	10.96	0.10	3.41	4.31	5.97	383.13

From the descriptive statistics, we can see that the data set covers a wide depth range, from 136 m to 5,436 m, with an average depth of 2,184 m. This means that we have a large variation in depth being analyzed in this data set, allowing us to examine a larger number of geological formations. The wide range of depths in the Force dataset reflects the geologic complexity of the Viking Graben, which experienced crustal extension and compromising diverse rock formations layered across various depths and affected by normal faults. Moreover, we see values with wide ranges such as GR, DTC, DTS, SP, PEF, and resistivity logs. As mentioned above, it is attributed to the complexity of the Viking graben, which resulted in the formation of numerous lithologies in the area. It is also seen that the NPHI has negative values and values equal to 1, which can be caused by tool failures or data acquisition errors. It is important to note that for this data set we did not remove outliers.

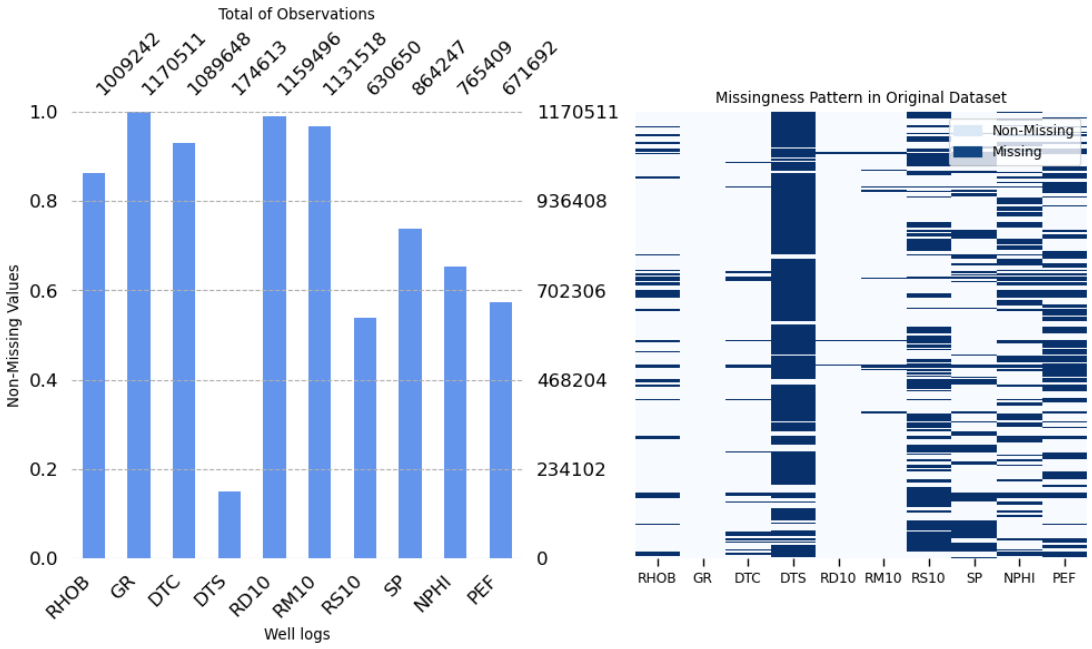


Figure 27. Missingness Analysis of well-logs in Force data set.

From the missing plots, we identified a generalized missing pattern with a missing not at random mechanism, highlighting the need for a robust method for dealing with missing data such as MICE. We observe that DTS is almost missing with 85% of the data missing. The RS10, NPHI and PEF also have a high percentage of missing data, 46%, 35%, 43% of the data are missing data respectively. As mentioned, a high percentage of missing data can lead to increased uncertainty in the imputed values and biased results, especially if the data is missing not at random.

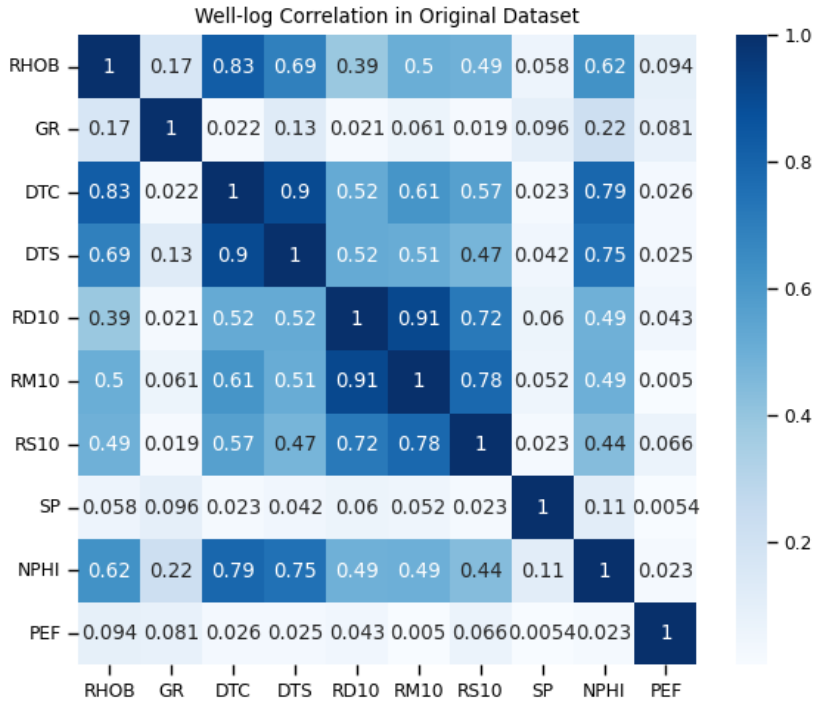


Figure 28. Correlation Matrix of well-logs in Force data set.

The correlation matrix indicates that there are strong correlations between several of the well-log variables. For example, there is a strong correlation between DTC and DTS of 0.85, which is expected as both measure sound speed through rocks but in different ways. Similarly, RD10 and RM10 are highly correlated with 0.9, as both are measures of resistivity at different depths. These well-logs with high correlation can be leveraged to impute missing values more precisely.

However, it is important to mention that highly correlated features can cause multicollinearity, resulting in biased results and overfitting when using linear regression models. In this case, multicollinearity may occur when two or more well-logs are highly correlated, and one well-log can be estimated linearly from other log with high performance. For instance, in the Force data set, the well logs of DTC and DTS are likely to present multicollinearity. RD10 and RM10 indicate high correlation; as a result, they can also cause multicollinearity.

Therefore, high correlations should be interpreted cautiously in MICE, as they can be both advantageous and problematic. High correlations can be advantageous in MICE because they can provide valuable information in the imputation. However, they can also be problematic due to the potential for multicollinearity, which can affect the performance and reliability of the imputed values.

On the other hand, it is observed that GR, SP and PEF have low correlations with other variables. For instance, the highest correlation in SP is 0.11, showing weak linear relationship with other logs. However, MICE considers all features in the imputation, even if they are poorly correlated.

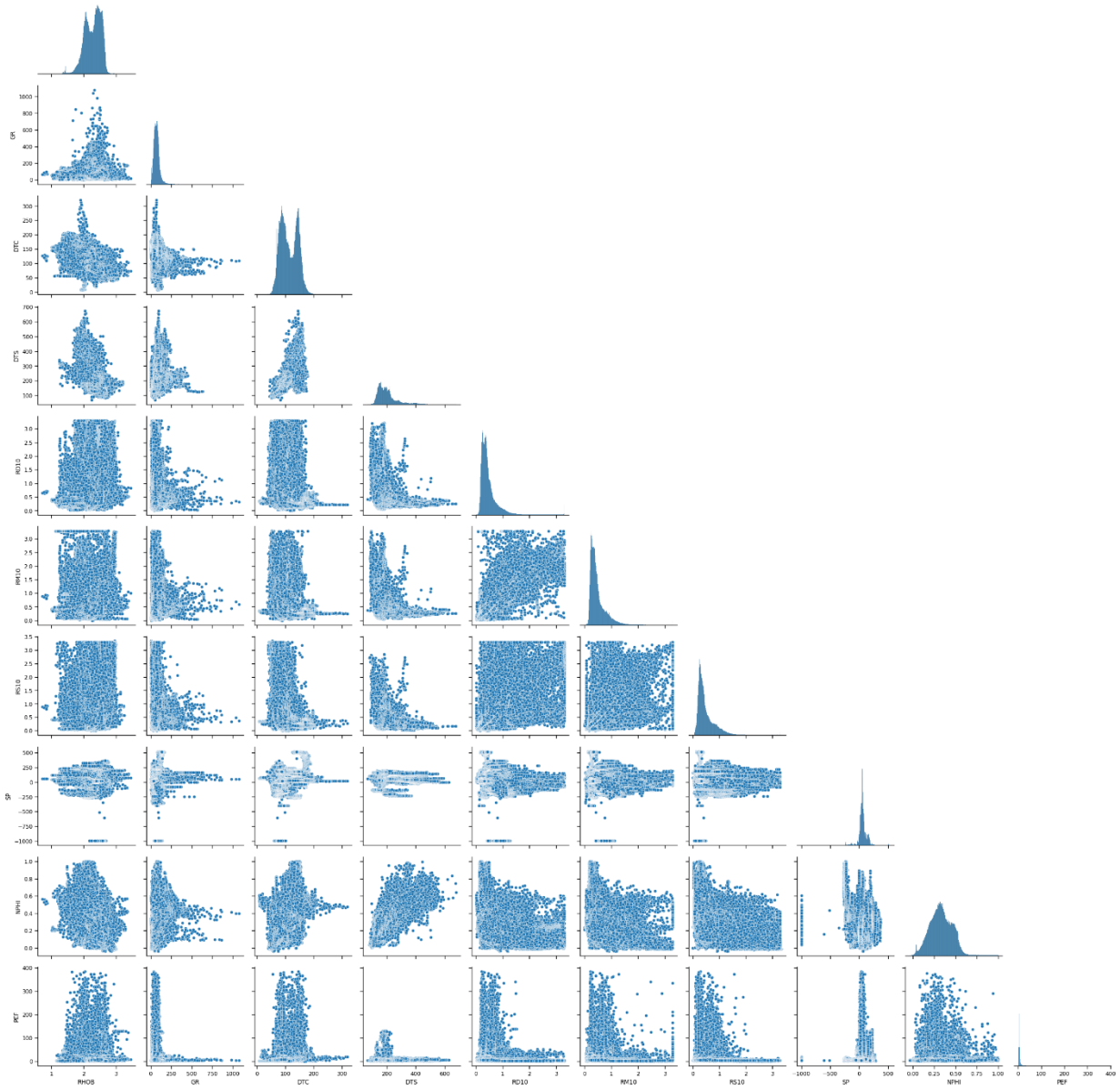


Figure 29. Force pair plot of well-logs.

The pair plot visualizes the complex relationships and distributions for the Force data set. The data is also noisy and scattered with high variability in the data. However, we can identify the relationship in the scatter plot for DTC and DTS, showing a strong relationship. This relationship and the high correlation observed could be explained because both measures are sonic logs. There may be other relationships and trends in the data that we are not able to see, but MICE can capture to perform the predictions.

4 | METHODOLOGY

This project aims to propose a framework for evaluating the performance of MICE in predicting well-log data with missing values, using three separate data sets from different sedimentary basins. The methodology is developed within a Python programming environment, implementing various machine learning models. This section is divided into three parts, which will explain the selected models, the proposed framework, and the computational environment used for modeling.

4.1 SELECTED MODELS

4.1.1 MICE

The multivariate imputation by chained equations (MICE) is used in this methodology for its ability to keep the relationships between variables makes it suitable for the complex structure of the well-log data. This method allows the integration with machine learning algorithms enabling a unified, flexible, and robust approach that can estimate all well-logs with missing values simultaneously. Additionally, previous studies have shown promising results using MICE with well-log data (Hallam, Mukherjee, & Chassagne, 2022).

4.1.2 Machine Learning Algorithms

The following regressor models are specifically chosen for their proven effectiveness in predicting missing values and compatibility with MICE:

- **K-Nearest Neighbors (KNR)**
  - Selection Reason: Offers reasonable performance without extensive tuning.
  - Attributes: Excellent reference method.
- **Bayesian Ridge (BR)**
  - Selection Reason: Recommended for handling missing data in MICE (Buuren & Groothuis-Oudshoorn, 2011).
  - Attributes: Incorporates prior information about parameters and constructs good prior distributions.

- **Random Forest (RF)**
  - Selection Reason: Addresses overfitting and sensitivity to outliers.
  - Attributes: Enhances performance and generalization.
- **XGBoost (XGB)**
  - Selection Reason: Known for high predictive performance.
  - Attributes: Can handle missing values without additional imputation strategies.

4.2 WORKFLOW

The workflow of this research consists of five main steps: data pre-processing, data splitting, model training, testing imputation, and final evaluation. These steps are structured to facilitate the evaluation of MICE performance with various machine learning models to predict well-logs with missing values. An overview of the entire workflow is shown in the Figure 30, summarizing the key processes. The subsequent sections provide detailed information about each step of the proposed approach.

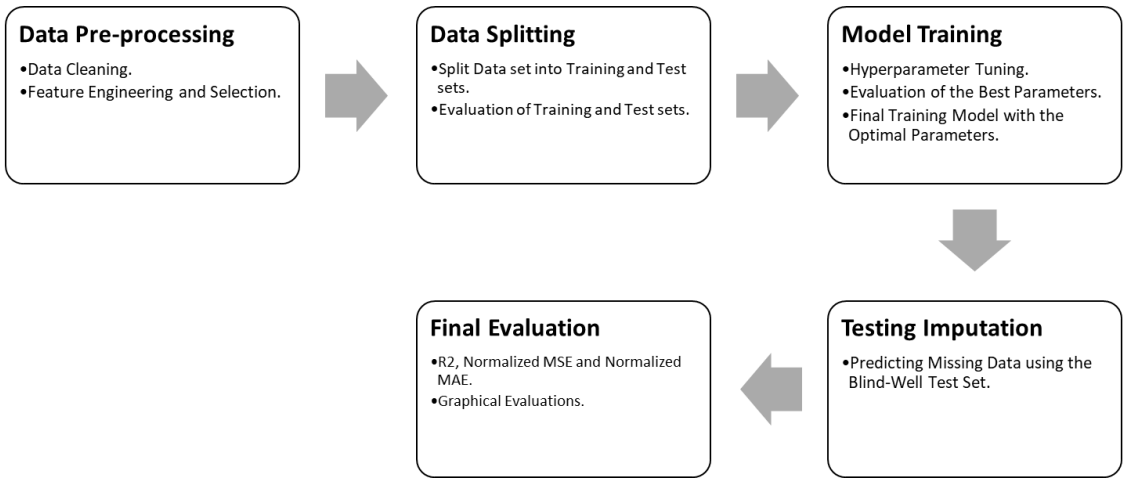


Figure 30. Workflow Diagram Representing the Processes of the Methodology.

4.2.1 Data Pre-processing

Data pre-processing is the initial step in the proposed framework. This step consists of cleaning and preparing the data sets for model training. This process includes renaming columns, dropping unnecessary columns, replacing outliers with missing values, and converting well-logs. Specifically, resistivity logs are transformed using base 10 logarithm, a common practice to reduce data variability and skewness. This transformation can also improve the performance of machine learning algorithms. Moreover, categorical variables, such as wells and sequential stratigraphy,

are converted into numerical forms suitable for machine learning models using label encoding. This encoding technique assigns a unique integer label to each category. It is important to note that handling missing values is intentionally omitted in this step since the framework aims to predict all missing values simultaneously.

4.2.2 Data Splitting

After pre-processing, the data set is divided into a training set (~80%) and a “blind-well” test set (~20%). This split is done by selecting certain wells from the original data set to create a “blind-well” test set. These wells are used as unseen data that will assess the generalization of the model.

Before proceeding with model development, the training and test sets are evaluated to ensure that the data is representative. This evaluation includes the following analyses:

- **Location:** Well location plotting ensures that the spatial distribution of training and test wells covers all relevant regions of the area.
- **Missingness Pattern:** Heatmaps are used to compare the missingness patterns in both the training set and the original data set, ensuring similarity. In addition, test well-logs are analyzed to ensure sufficient data to evaluate the models.
- **Distributions:** Density plots help verify that statistical properties of the well-logs in the training set are preserved as in the original data set.
- **Correlation:** Correlation heatmaps provide insights into the relationships between different well-logs in both the original and training sets, ensuring representativeness.

By evaluating the training and test sets, we ensure that the training set is adequate to build a reliable and generalizable model, which has unseen data to evaluate its performance. The details of these analyses can be found in the appendix DData Analysis.

4.2.3 Model Training

The model training involves teaching a model to recognize patterns in training data; therefore, the model can make predictions when presented with new data. This process is implemented using MICE with the four selected machine learning models. The model training consists of three steps: hyperparameter tuning, evaluation of the best parameters, and the final training of the model.

I. Hyperparameter Tuning

This step allows the identification of the best combination of parameters for a given model, using a 5-fold cross-validation procedure as shown below:

1. **K-Fold Splitting:** The data set is divided into 5 parts, 4 are used for training and 1 for validation. This ensures that the model is evaluated on unseen data at each fold.

2. **Simulating Missing Data:** In each fold, a copy of the training data is made, and specific points are set as missing (NaN) based on validation values. Therefore, it is replicated how missing values occur in real-world scenarios where some well-logs might be entirely missing.
3. **Normalization:** The training data with simulated missing values is scaled within the interval [0, 1]. The original training data is also scaled using the same transformation to avoid leakage. This ensures that the evaluation metrics are on comparable scale since the well-logs have different units.
4. **Imputation:** The missing data is predicted using the MICE method with different machine learning models. The imputation is performed only for the scaled training data with simulated missing values.
5. **Evaluation:** The performance of each model is evaluated with the scaled data by comparing the original values with the imputed values. It uses the normalized mean squared error (NMSE) as the metric. The combination of parameters that yields the lowest NMSE is considered the best.

Note: By scaling the training data with simulated missing values, it is ensured that the training data does not have information from the validation data. This process is done for each fold, avoiding leakage between training and validation data.

II. Evaluation of the Best Parameters

This step consists of evaluating the performance of the optimal parameters obtained from the hyperparameter tuning using the custom cross-validation procedure described above. The evaluation considers three metrics: R-Squared (R2), Normalized Mean Squared Error (NMSE), and Normalized Mean Absolute Error (NMAE). Time is also recorded to assess the efficiency of each model.

III. Final Training Model

Based on previous results, the final model is trained using the entire training set with the best combination of parameters. This model is trained on the scaled training data using normalization.

4.2.4 Testing Imputation

This step is essential to ensure that imputation models can effectively predict missing values using the blind-well test set. The step performs the following three processes:

I. Scaling

The blind-well test set, selected during the initial data splitting, is first scaled using the same normalization scaler fitted with the training data, ensuring the well-logs are on a similar scale. The scaled values are stored separately for later comparison.

II. Custom Cross-Validation to Impute

Like the model training, the custom cross-validation is implemented to simulate and estimate missing values for all well-log combinations as follows:

1. **KFold Splitting:** The KFold method splits the well-log combinations into 5 parts, creating multiple folds for cross-validation.
2. **Simulating Missing Data:** In each fold, a copy of the unscaled blind-well test data is created, and specific well-logs are set as missing values (NaN).
3. **Scaling:** The copied blind-well test data, with simulated missing values, is scaled using the transformation applied to the training data.
4. **Imputation:** The trained imputation model fills the missing values within the scaled blind-well test data, and the results are stored.
5. **Inverse Transformation:** The imputed values are transformed back to their original scale for subsequent evaluation.

III. Results of the Blind-Well Test Set

The results are assembled into a detailed data set containing the original values, the original scaled values, the scaled imputed values, and the imputed values for each well-log. This consolidated information is used for the final evaluation of the performance of the models, which will be carried out in the next step.

4.2.5 Final Evaluation

The model performance is evaluated based on three metrics: R-squared (R2), Normalized Mean Squared Error (NMSE), and Normalized Mean Absolute Error (NMAE). It is important to note that the evaluation metrics are computed with scaled data; therefore, it is normalized MSE and normalized MAE. The main metric used for comparison is the NMSE, which can effectively compare the performance across all the imputed features. These metrics provided insights into the precision and error associated with the imputations from each model.

Moreover, graphical visualizations are performed to identify trends or intervals where the model performs better or worse, which the evaluation metrics cannot fully capture. These visualizations include plots of well-logs comparing original values versus imputed values, and scatter plots of true values versus predicted ones. In addition, the computational time is analyzed from each train model.

By comparing the various imputation models on different data sets, the results can be analyzed to assess the performance of MICE, answering our research question and providing a clear perspective on how MICE performs predicting missing values in well-logs.

4.3 COMPUTATIONAL ENVIRONMENT

The models were trained and evaluated on a system with the following specifications:

- **Processor:** 11th Gen Intel(R) Core(TM) i5-11400H @ 2.70GHz 2.69 GHz
- **Memory:** 15.84 GB DDR4 RAM
- **Graphics Card:** NVIDIA GeForce RTX 3050
- **Storage:** 500 GB SSD
- **Operating System:** Windows 11 Home, 64-bit
- **Software Environment:** Python 3.9.16, Scikit-learn 1.2.1

5 | RESULTS AND DISCUSSION

5.1 RESULTS

5.1.1 Metric Evaluation of MICE

For the evaluation of the imputed well-logs, we mainly use NMSE and R2 metrics. This evaluation is carried out with different number of iterations, values of 1, 10 and 20. In this study, the machine learning models are performed with the default hyperparameters, and MICE with the same random state.

I. Force-200

The Force-200 data set presents favorable results in predicting the missing values using the MICE approach, see Figure 31. These results are consistent with a previous study by (Hallam, Mukherjee, & Chassagne, 2022), which also considered MICE as imputation method evaluating the performance only for DTC, DTS and RHO.

The MICE results from 1 to 20 iterations indicate that the models perform well using this approach since most of the R2 values are positive. However, the SP log consistently yields negative R2 scores across all the models and number of iterations. Furthermore, GR presents negative R2 scores, indicating poor performance in the KNR and BR models with 1 iteration. Across most of the well-logs, XGB and RF consistently outperform KNR and BR.

Comparing the performance of MICE with 1 to 20 iterations, we observe that the difference in R2 scores and NMSE values are relatively minimal. For KNR, BR, RF, and XGB, NMSE and R2 values for most of the well-logs tend to remain stable regardless of the number of iterations. Although increasing the number of iterations does not yield significant enhancements, the computational resources and time required for running multiple iterations become notably high. This inefficiency makes higher iterations less practical without substantial performance improvements.

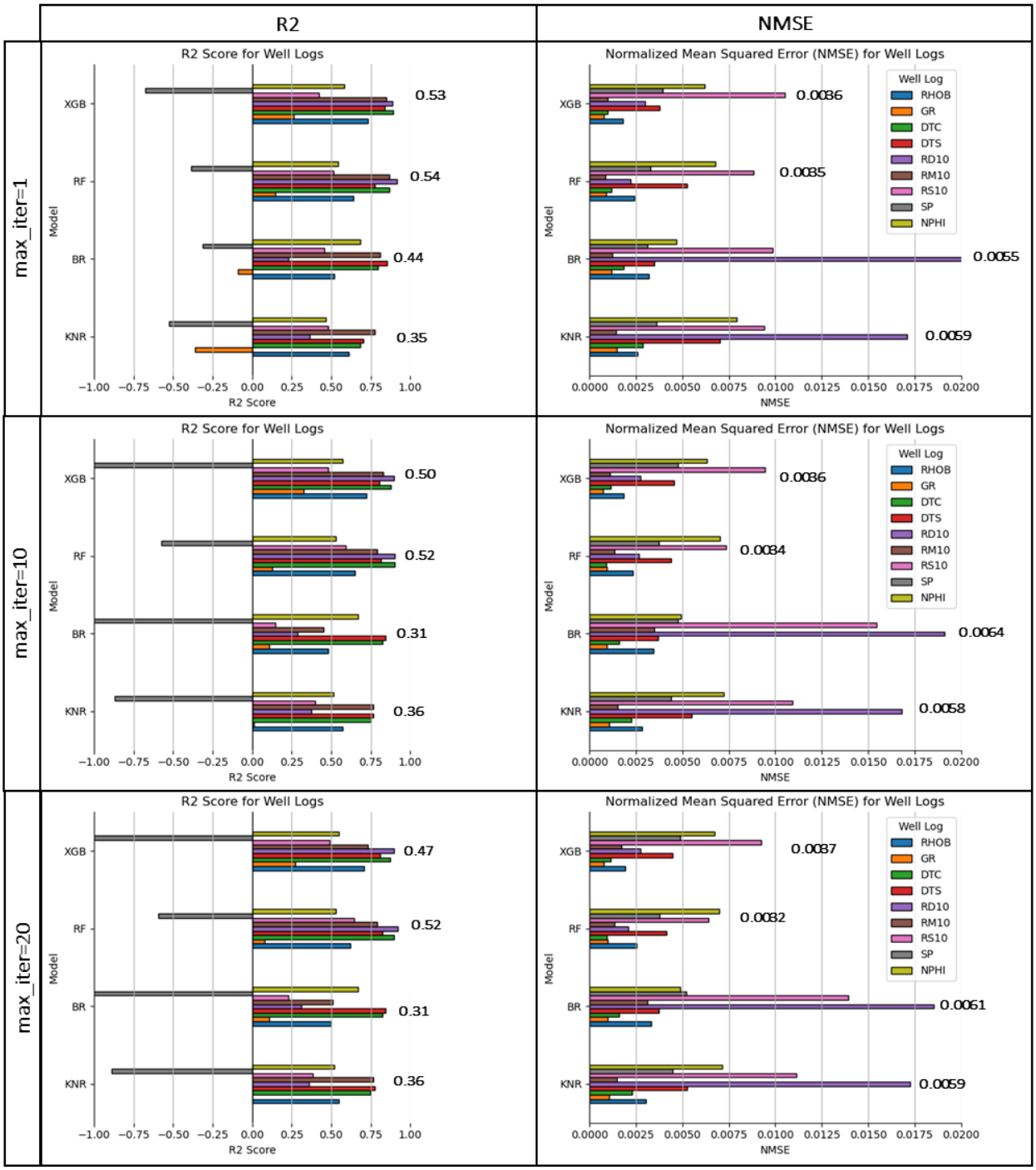


Figure 31. Force-200 Results of MICE with 1 to 20 Iterations using R2 and NMSE metrics.

II. Montney

The results of MICE in the Montney data set show that XGB has the lowest NMSE and the best R2 scores compared to the other models. However, none of the models predict with good performance the SP logs, which present largely negative R2 values regardless of the number of iterations. Additionally, the number of iterations in MICE imputation does not lead to any significant

change in the performance of R2 and MSE in the four models tested. The results were marginally better with 1 iteration compared to 10 iterations, suggesting that increasing the number of iterations may not necessarily lead to improvements in predictive performance.

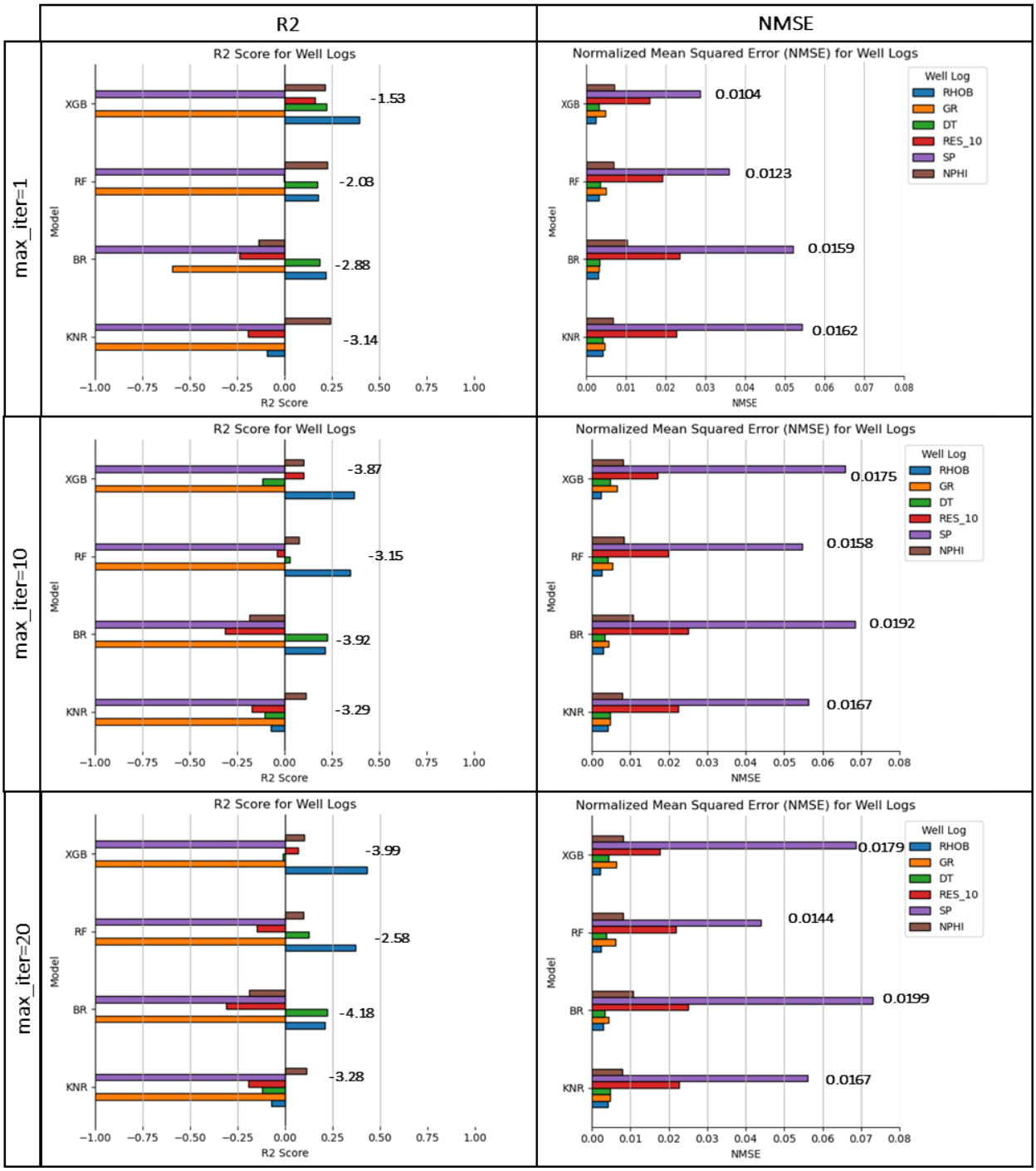


Figure 32. Montney Results of MICE with 1 to 20 Iterations using R2 and NMSE metrics.

III. Beetaloo

The MICE approach with different iterations parameters, from 1 to 20, present variations in the performance of the predictions model used in KNR, BR, RF and XGB. Similarly, to Force-200 and Montney, the results also indicate that increasing the number of iterations does not lead to significant improvements in the estimations and performance. Additionally, none of the models performed well consistently for all the well-logs such as GR, DT, REST_10, SP, and NPHI. For instance, SP log presented the worst R2 scores in all models tested regardless of the number of iterations. The XGB model has the best performance among the other algorithms.

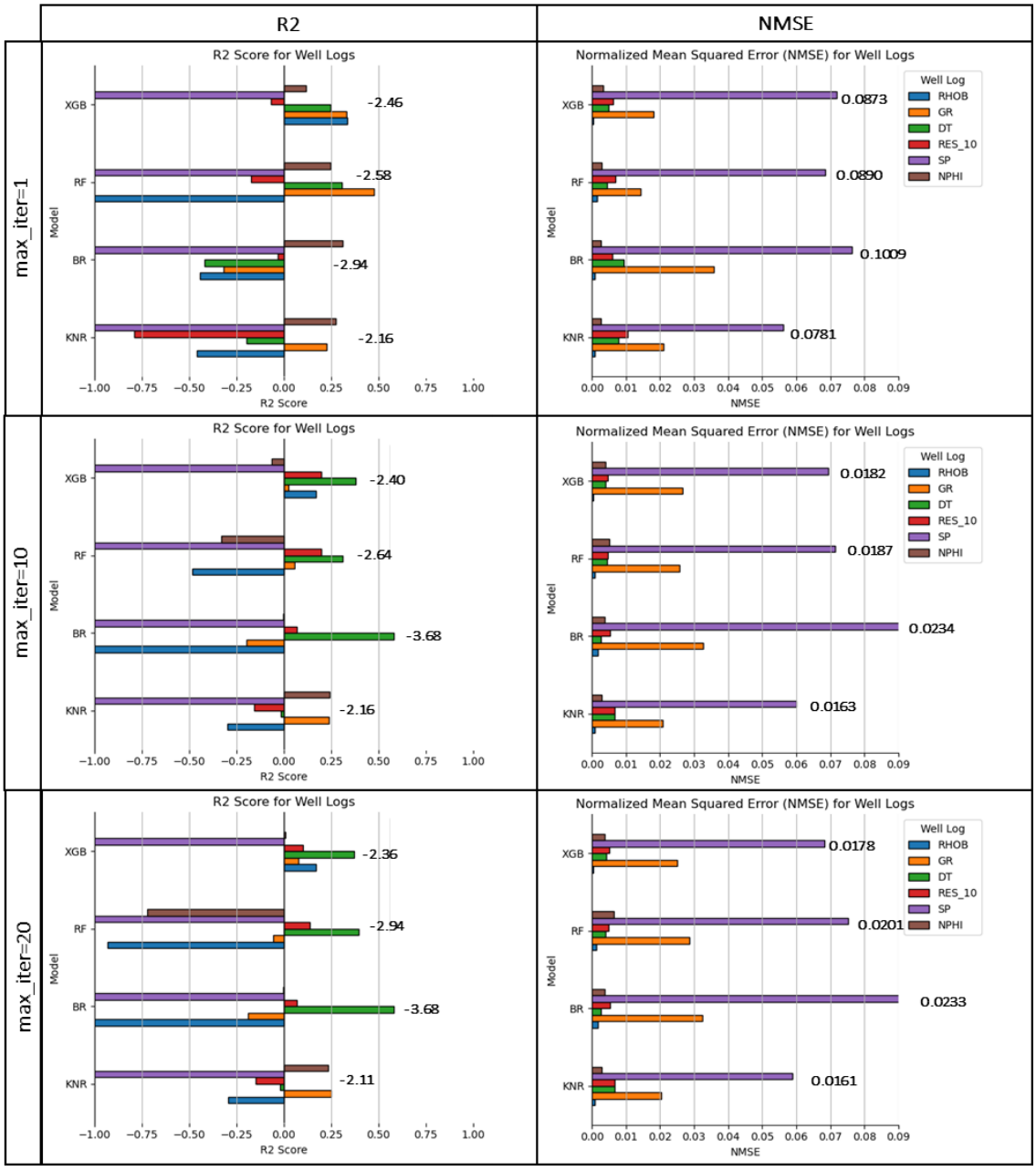


Figure 33. Beetaloo Results of MICE with 1 to 20 Iterations using R2 and NMSE metrics.

5.1.2 Blind-Wells Performance

The MICE performance on individual blind wells is evaluated using the average of NMSE and R2. This evaluation considers the models for each data set using 10 iterations. Additionally, the number of observations and the fraction of missing values are taking into account in this analysis.

I. Force-200

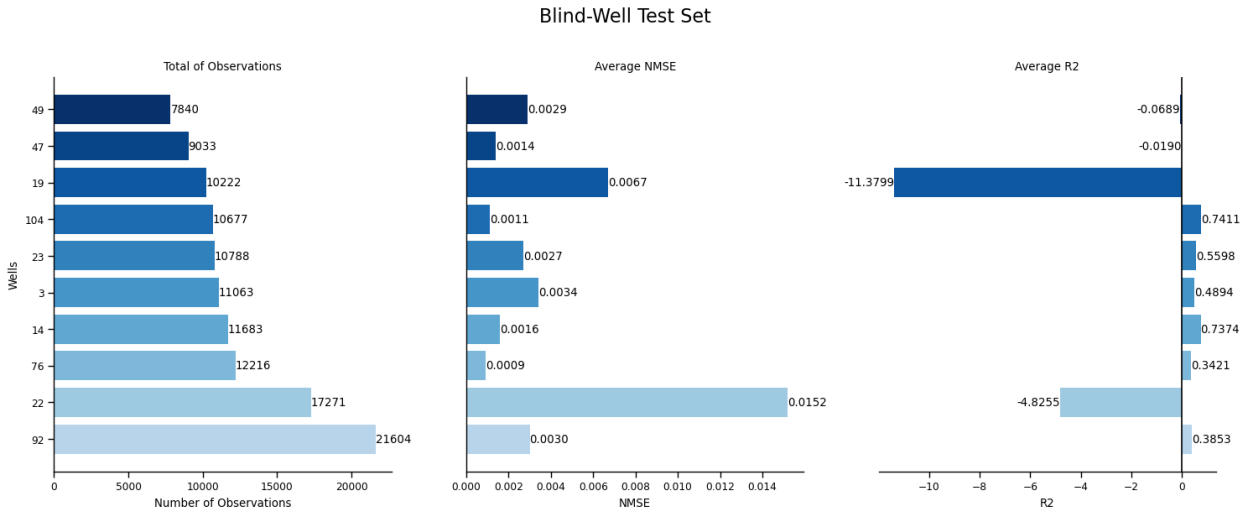


Figure 34. Force-200 Blind Wells Performance using MICE.

Performance across different wells exhibits considerable variability. While certain wells show promising predictions with positive R2 values, others present extremely negative R2 values. Well performance ranges from a low NMSE, such as the well 76 with a value of 0.0009, to a high NMSE, such as the well 22 with a value of 0.0152. The well 14 shows the highest positive R2 value of 0.7374. In contrast, the well 19 exhibits a substantially negative R2 value of -11.3799, suggesting possible deficiencies in the generalization of the specific model to that well.

Furthermore, the well 92 has the highest number of observations and performs relatively well with an R2 of 0.3853. However, the wells 22 and 19 have significantly negative R2 values although they have more than 10,000 and 17,000 observations respectively, indicating poor performance.

Additionally, the well 19 have complete log data and the well 22 presents missing values in some logs such RS10 and SP, 64% and 47% respectively. Notably, these wells exhibit the worst R2 scores. On the other hand, wells 104, 23, 3, 14, 76, which have RS10 and SP entirely missing, present positive values in R2, indicating a better performance. For further details refer to the appendix DData Analysis and the notebooks [GitHub].

II. Montney

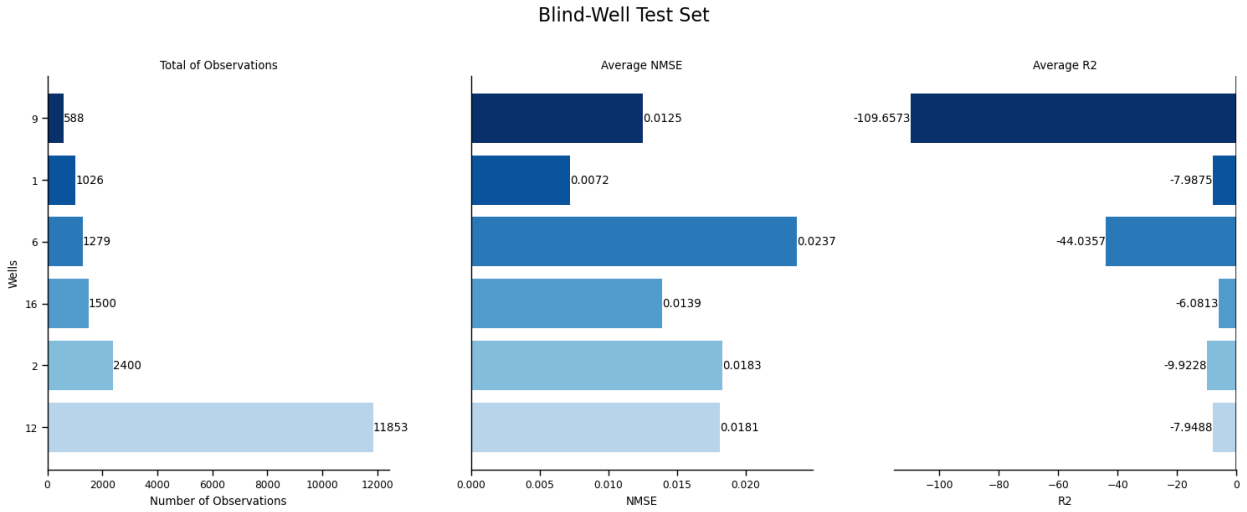


Figure 35. Montney Blind Wells Performance using MICE.

The Montney data set presents negative R2 values in all the well-logs, indicating the poor performance in the predictions. It is important to note that these wells are almost complete, only well 2 has 35% of the NPHI log missing. Particularly, well 6 has the worst performance with an R2 of -44.0357 despite having more than 1,000 observations.

III. Beetaloo

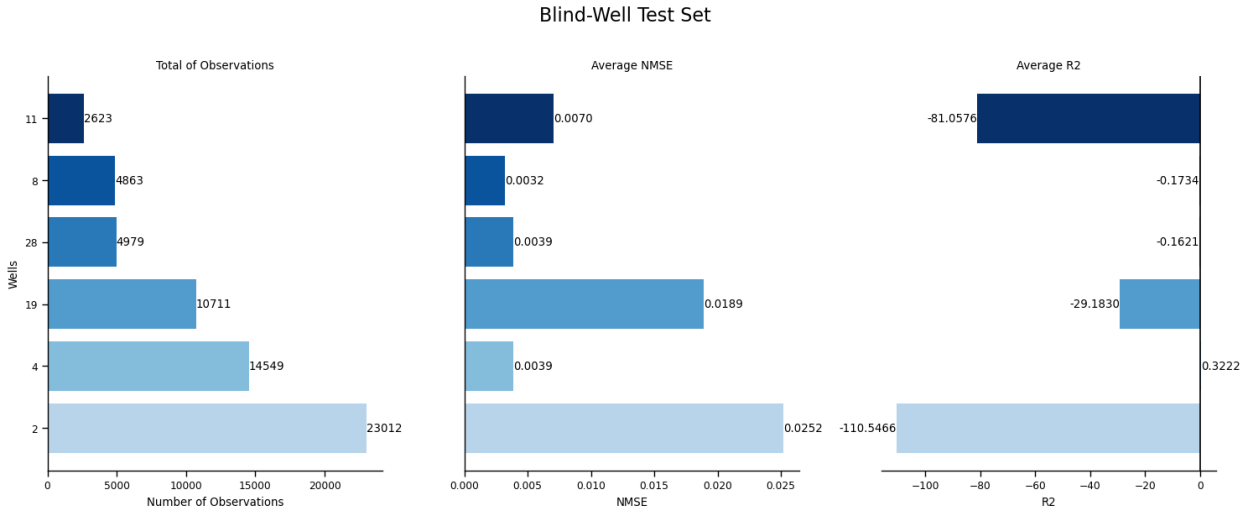


Figure 36. Beetaloo Blind Wells Performance using MICE.

Among the wells examined in this data set, well 2 stands out for its remarkably poor performance despite having complete log data and the highest number of observations. The average R2 of this well is -110.5466. On the other hand, well 4 presents better performance with an average R2 close to 0.32. Although this well has no SP log and shows about 15% missing data in other logs, the well 4 has the best performance of this data set.

Based on the blind wells analysis, the number of observations and log integrity in a well do not appear to have a consistent relationship with the MICE performance along different datasets.

5.1.3 Lithostratigraphy Performance

Like the previous analysis of blind wells, the MICE performance of the models is evaluated with respect to the lithostratigraphic units for each data set. This evaluation also considers the total number of observations, fraction of missing values and the average of NMSE and R2.

I. Force-200

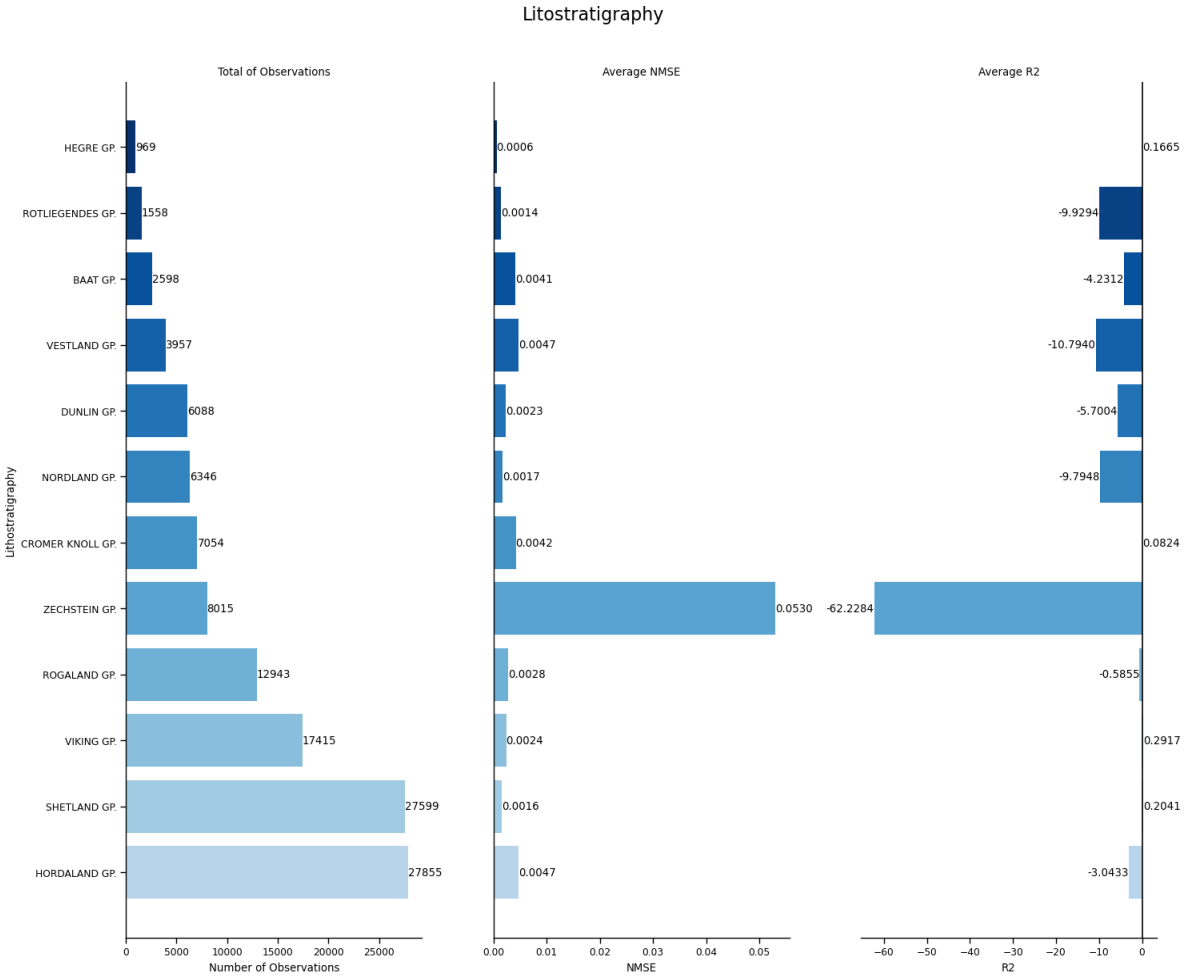


Figure 37. Force-200 Lithostratigraphy Performance using MICE.

In the Force-200 data set, the Shetland and Viking groups present the best R2 score, with R2 values of 0.2041 and 0.2917 respectively. These groups have the lowest percentage of missing values in the DTS log compared with other groups. The Shetland group has 73% of the data missing in the DTS log, while the Viking group has 71%. It is important to note that Cromer Knoll group has positive R2 values, and this group also presents one of the three least percentage of missing values in the DTS log with 71% of the data missing.

On the other hand, the Hordaland group has the highest number of observations with negative R2 scores, indicating poor performance. This group has the highest missing values in log such DTS, NPHI and RS10 with values of 96%, 57& and 50% respectively.

The Zechstein group also has the worst performance than the other groups. This group presents the highest NMSE of 0.0530 and most negative R2 of -62.2284, suggesting poor model predictions for this lithostratigraphy. This group has DTS, SP and RM10 missing with 94%, 83% and 79% respectively.

II. Montney

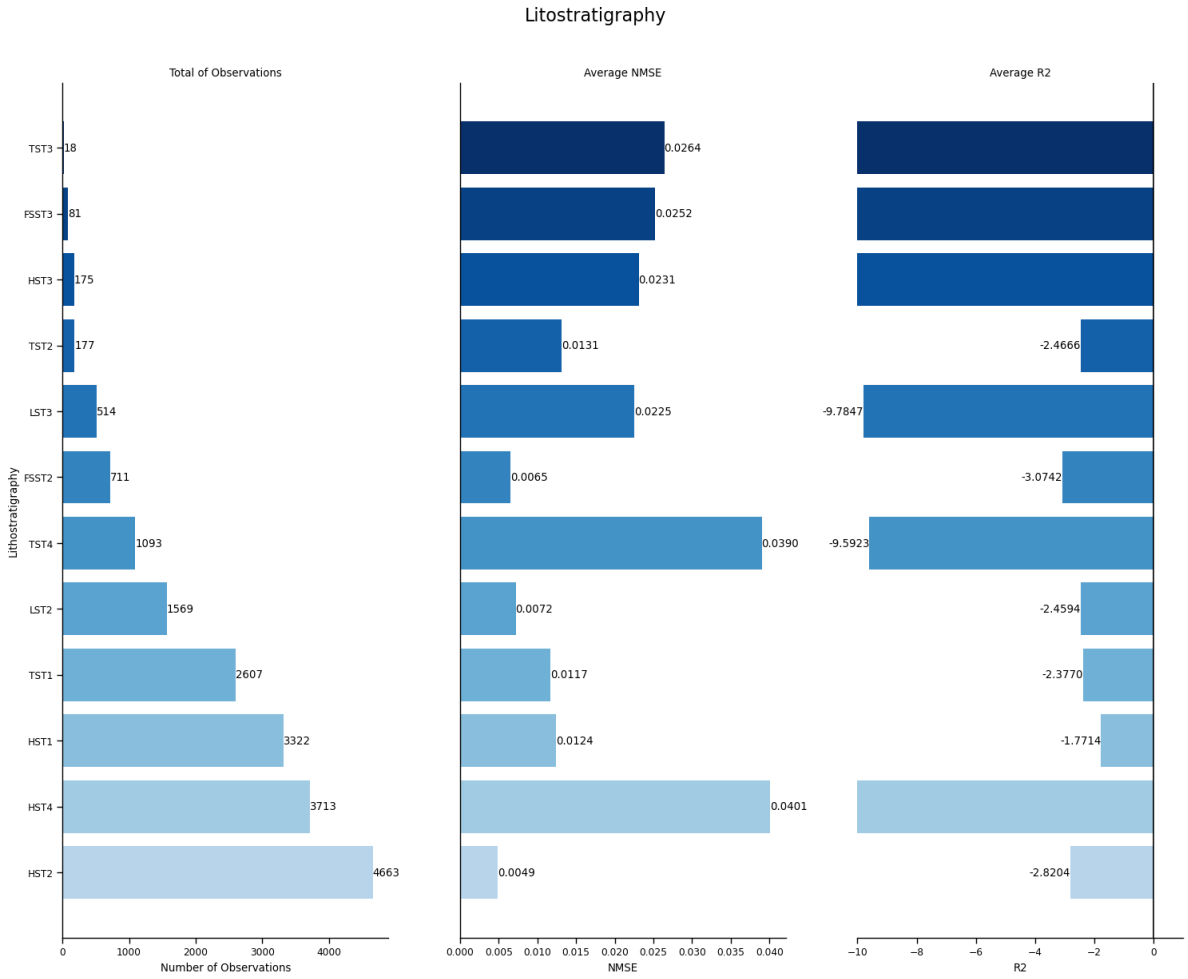


Figure 38. Montney Lithostratigraphy Performance using MICE.

In this data set, all the stratigraphic units performed poorly regardless of the number of observations. The HST2 unit stands out with the lowest NMSE; however, it presents a negative R2 score. Despite the higher number of observations and maintaining relatively strong data integrity, HST2 performs poorly in predicting missing values.

Moreover, HST4 records the highest NMSE value among the stratigraphy units, indicating larger discrepancies between the predicted values and original values. HST4 is the second stratigraphy with the highest number of observations and presents a high percentage of missing values in logs. For instance, GR, DT, NPHI, and SP logs are missing in approximately 49%, 43%, 32% and 32% respectively.

III. Beetaloo

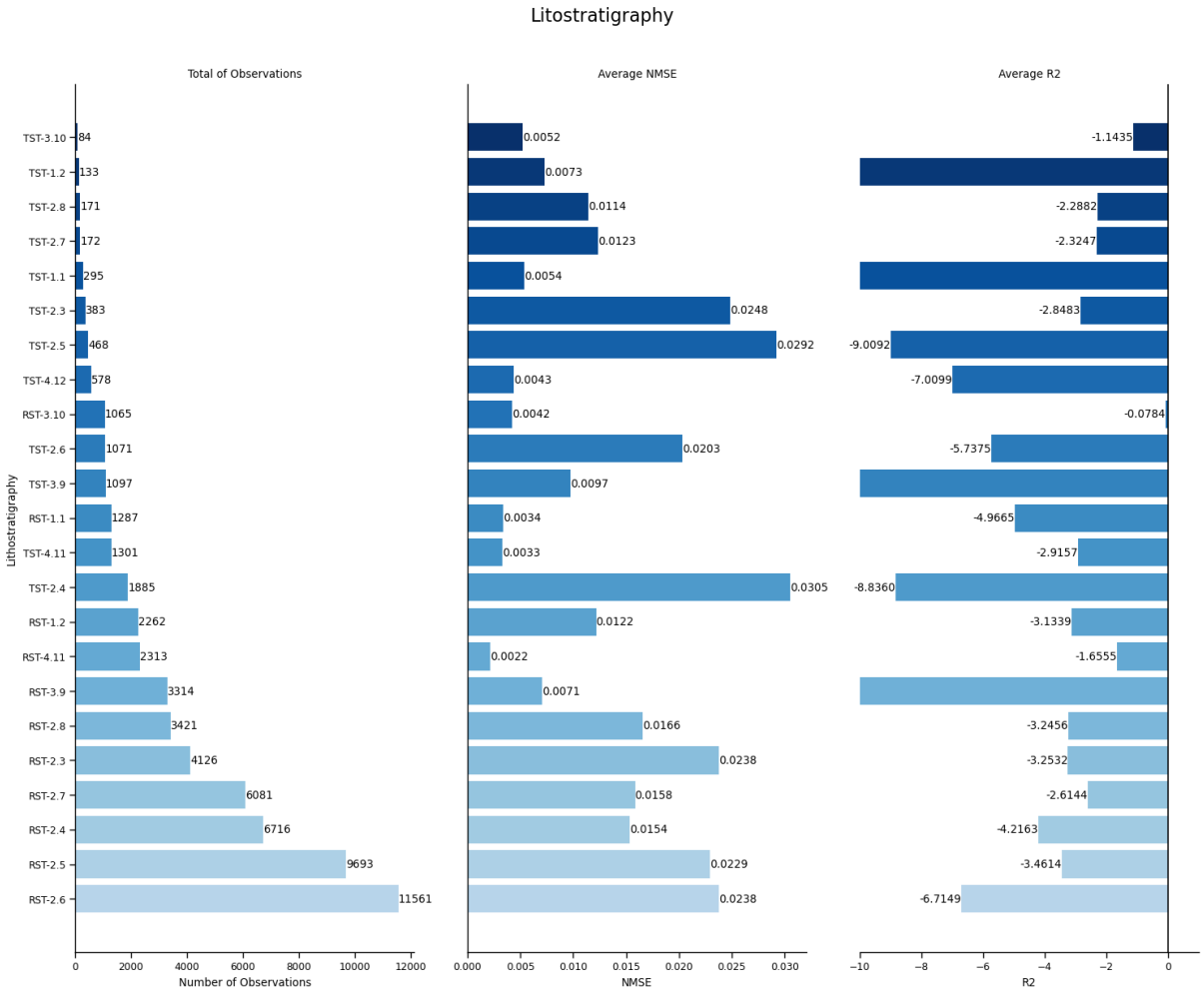


Figure 39. Beetaloo Lithostratigraphy Performance using MICE.

The Beetaloo data set also presents negative average R2 values across all the stratigraphy units, indicating poor performance of the predictions. Among these units, the RST-3.10 stratigraphy stands out with the least negative R2 score of -0.0784, while the TST-3.9 unit has the largest negative R2 score, which is -1,578.3262.

It is worth noting that a large number of observations for a given stratigraphic unit does not necessarily correlate with better performance such as RST-2.6 and RST-2.5. Additionally, the Beetaloo data set, based on the exploratory data analysis, has high percentages of missing

values across all well-logs. For example, RHOB, GR, SP, and NPHI log have more than 50% of the data missing. Therefore, in the stratigraphy performance analysis, data integrity can return negative values of R2, leading to poor model performance.

5.1.4 Graphical Evaluation of MICE

The graphical evaluation uses well-logs and scatter plots to visualize the original values versus the imputed values. This analysis can identify patterns that the evaluation metrics cannot capture.

I. Force-200

Based on the blind well analysis, the well 19 presents the worst negative average R2 score even though this is the only well that contains all the logs complete. In contrast, the wells 104, 23, 3, 14, 76, which lack RS10 and SP logs, present positive values in R2, indicating a better performance. For this reason, the wells 19 and 14 have been chosen for the graphical evaluation of the imputation results.

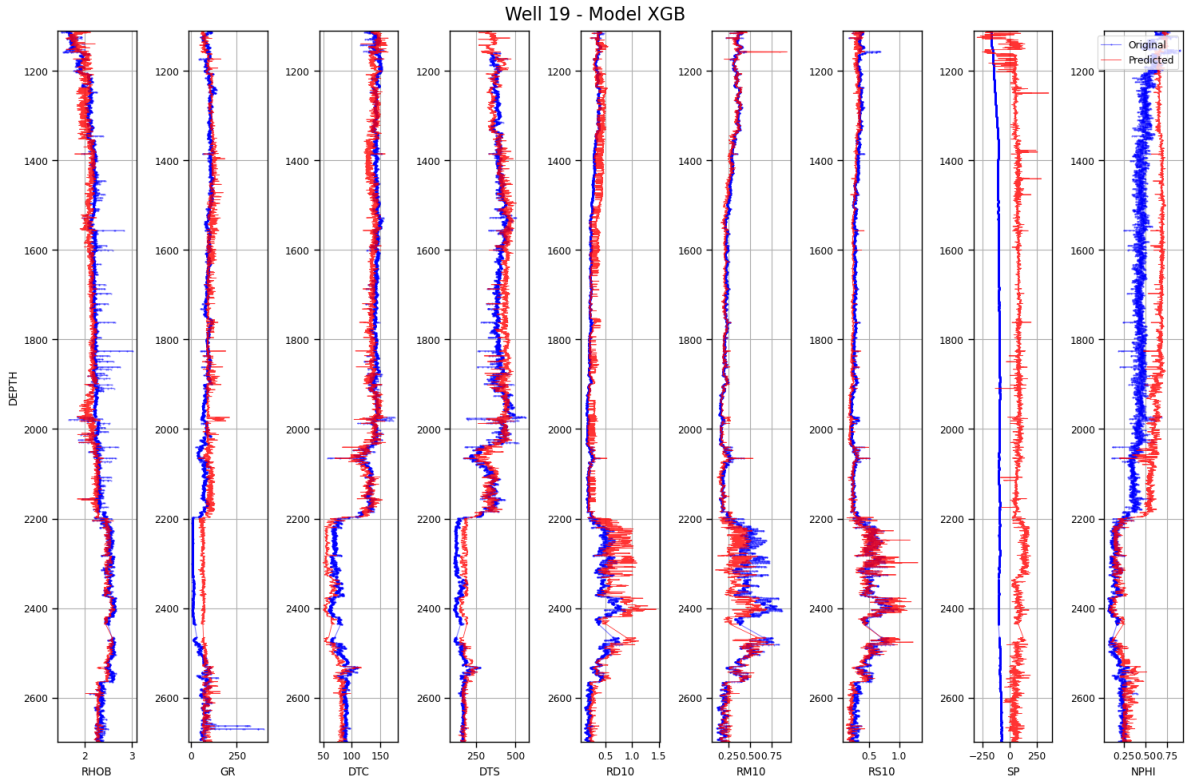


Figure 40. Well-Log Plot of the Well 19 with Original Values and Predicted Values using XGB in the Force-200 Data set, the blue lines are original logs, and the red lines are the predicted logs.

It can be seen in the Figure 40 that the original records such as DTC, DTS, RD10, RM10 and RS10 are very similar to each other. This underlines the strong correlation that exists between these logs, suggesting that the presence of one log can provide information about others logs.

Regarding the imputed logs, it can be observed that the predicted logs follow the trend of the original logs. However, there are shifts and peaks with outliers in some intervals, indicating the limitation of MICE in predicting missing well-log data. For instance, the estimated log of SP is shifted and spikey compared to the original log, which is smooth and flat.

On the other hand, the Figure 41 shows the scatter plots of the well 19, which can be seen that DTC, DTS, RD10, RM10, RS10 present a more linear relationship compared to the other logs. Although RHOB and NPHI logs have linear relationships, they are noisier and more dispersed than the sonic and resistivity logs. For instance, the noisiest areas correspond to the Hordaland group, which presents negative average R2 scores. It is important to mention that this group has the highest number of observations and the highest missing values in logs such as RHOB, DTS and NPHI where the data is dispersed in the scatter plots.

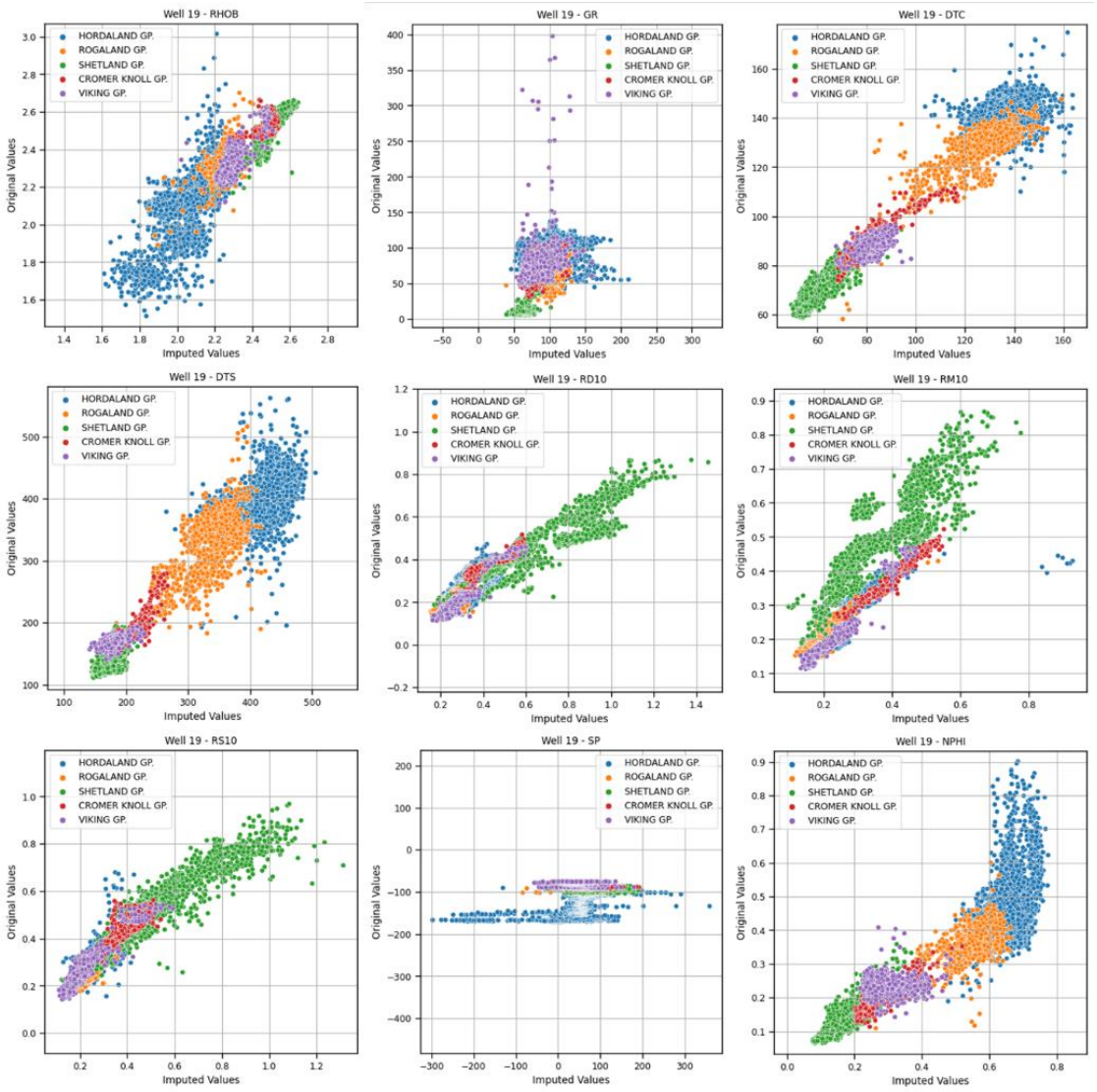


Figure 41. Scatter Plots - Original Values versus Imputed Values of the Well 19, Force-200 data set.

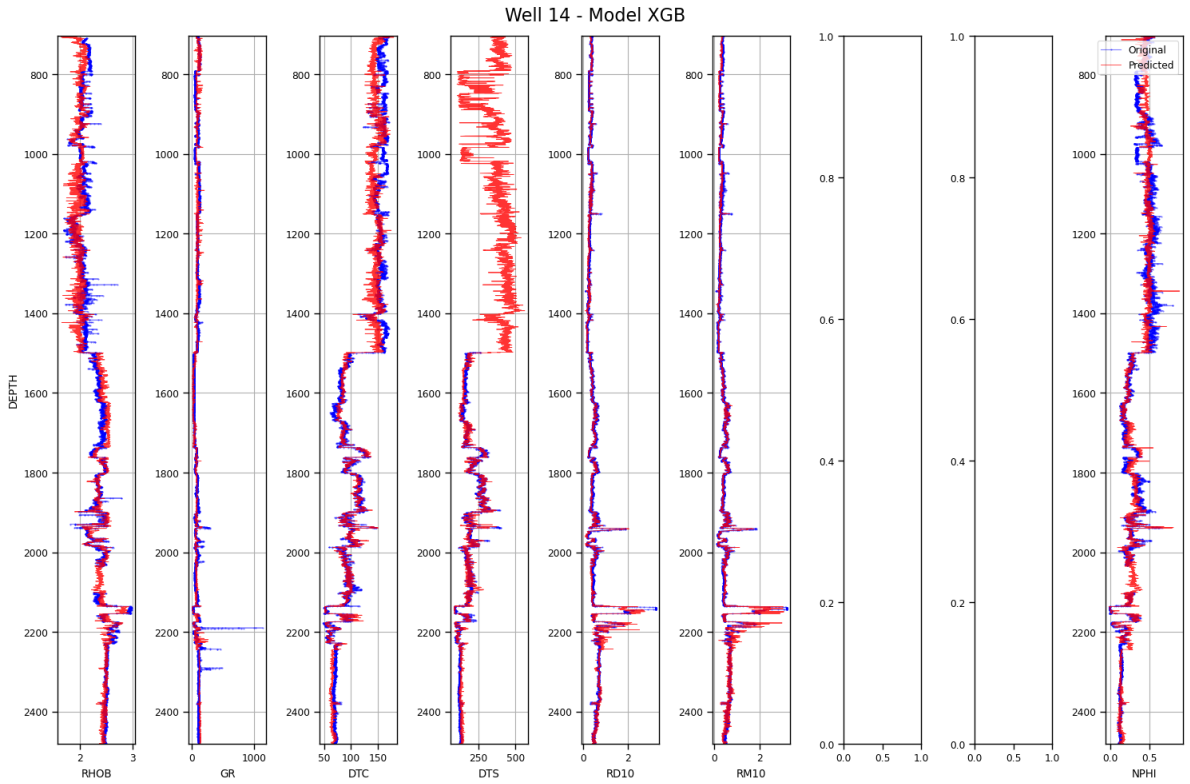


Figure 42. Well-Log Plot of the Well 14 with Original Values and Predicted Values using XGB in the Force-200 Data set, the blue lines are original logs, and the red lines are the predicted logs.

The well-logs from well 14 reflect similarities to the findings from well 19. These logs show that the original data from DTC, DTS, RD10 and RM10 are nearly identical. However, the well 14 lacks RS10 and SP logs, which could explain the good performance in predicting the missing values. Since MICE struggles in predicting SP, the absence of this log could improve imputation performance. Additionally, the predictions are generally consistent in this well, but there are peaks and outliers at various depth intervals in different imputed logs.

Moreover, the Figure 43 shows scatter plots for the well 14, which validate the strong linear relationships for logs such as DTC, DTS, RD10, and RM10, like the well 19. GR and NPHI maintain their linearity, but they are more dispersed than the sonic and resistivity logs. It can be observed that MICE also struggles in predicting GR logs where there is no linear correlation between the original values versus the predicted values.

In the scattered plots of the well 14, the areas with the greatest dispersion and noise correspond to groups such as Hordaland and Zechstein. Both groups have low performance in previous analyses. This can be seen particularly in the scatter plots of RHOB, RD10, RM10 logs in the following figure.

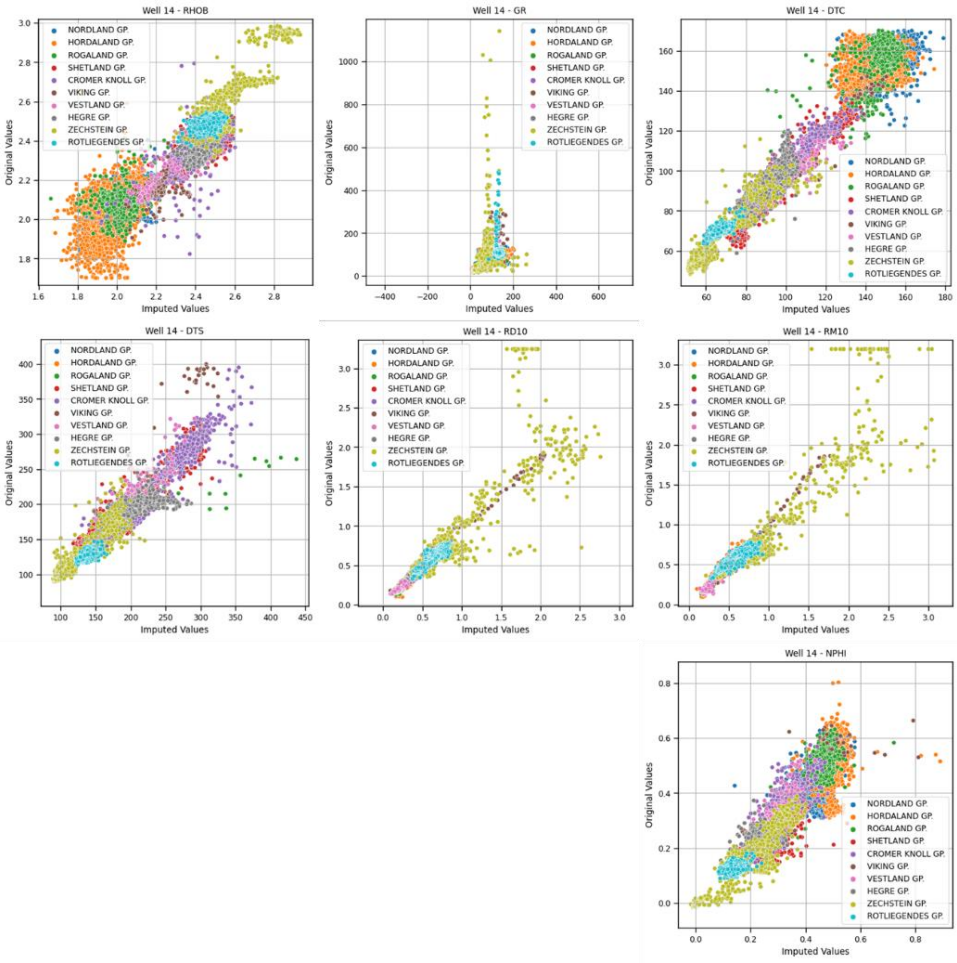


Figure 43. Scatter Plots - Original Values versus Imputed Values of the Well 14, Force-200 data set.

II. Montney and Beetaloo

The wells of the Montney and Beetaloo data sets perform poorly in predicting missing values. Specifically, the Montney data set reveals that all the wells have negative R2 values, and the well 16 has the best performance among others. On the other hand, in the Beetaloo data set, only the well 4 have positive R2 score, while the remaining wells have negative R2 values. Based on these findings, the Montney well 16 and the Beetaloo well 4 are selected for graphical analysis.

As can be seen in the Figure 44, the Montney well 16 is presented. These well-logs show the limitations of MICE in predicting missing logs on this data set. Although the estimated values follow the general trend of the original values, numerous peaks and outliers are observed in most of the intervals. Additionally, certain changes in the original values are not estimated by the algorithm; specifically, the marked circles in the DT log underline the inability of MICE to precisely estimate these changes. It seems that the algorithm tends to use the mean value of the log to predict these changes, leading to imprecision and unreliable results.

On the other hand, the imputed SP log is very spikey and in certain intervals is shifted compared to the original log, which is smooth and almost flat.

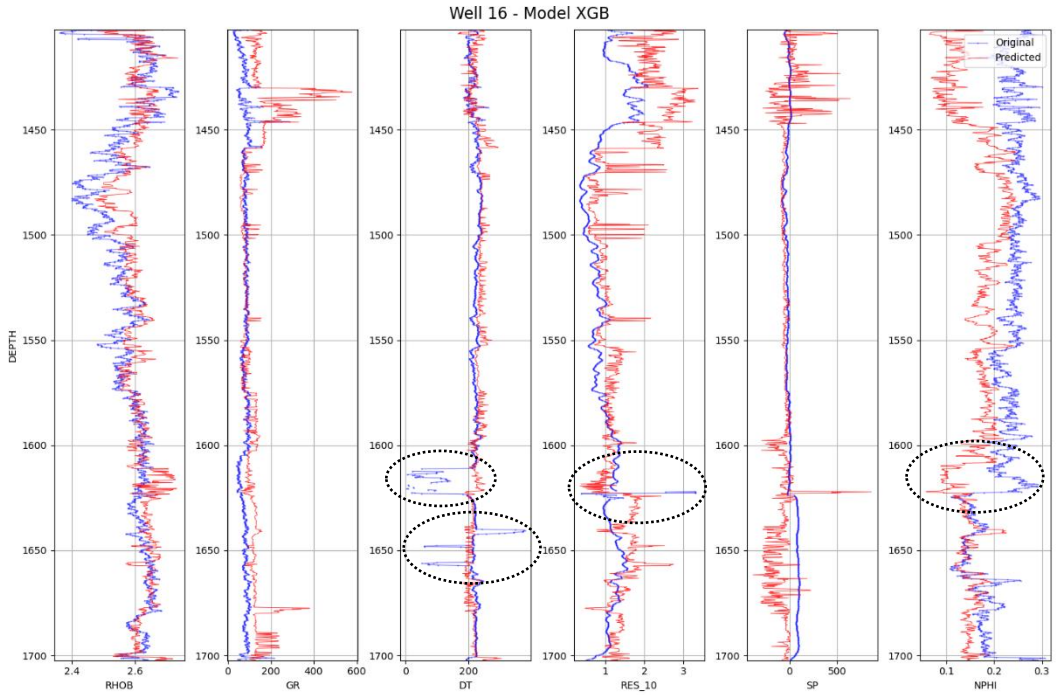


Figure 44. Well-Log Plot of the Well 16 with Original Values and Predicted Values using XGB in the Montney Data set, the blue lines are original logs, and the red lines are the predicted logs. The circles remark on the poor performance of MICE in predicting missing values.

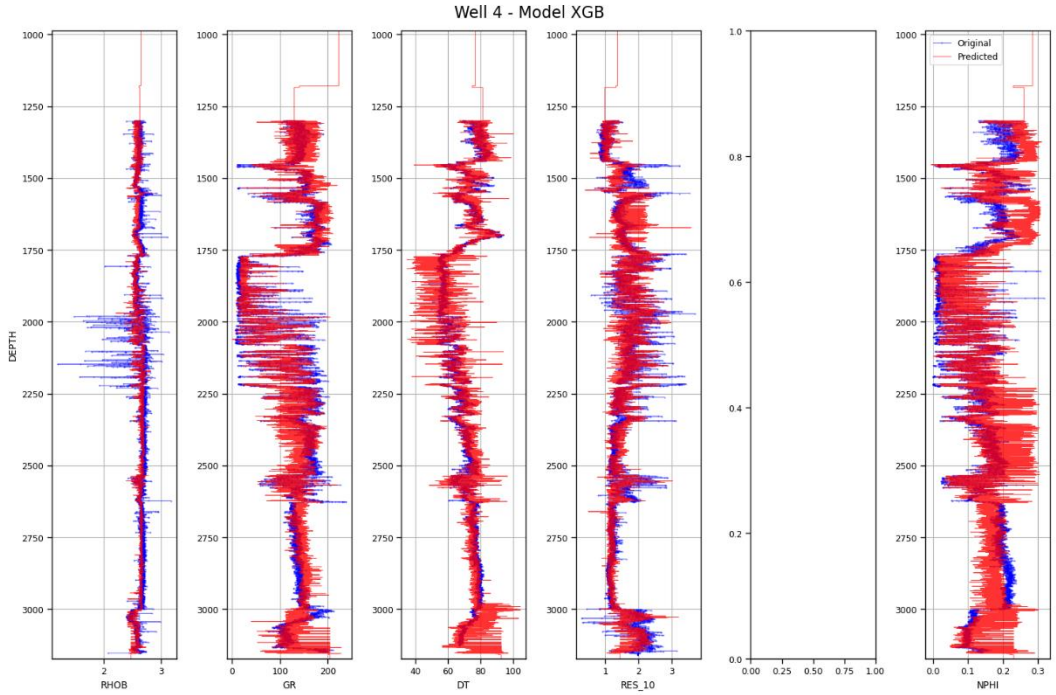


Figure 45. Well-Log Plot of the Well 4 with Original Values and Predicted Values using XGB in the Beetaloo Data set, the blue lines are original logs, and the red lines are the predicted logs.

The Figure 45 illustrates the logs of the Beetaloo well 4. Like Montney, it can be seen that MICE does not predict precisely the missing logs. Although the estimated logs follow the trend of the original logs, there are many outliers and spikes in all the logs at various depths, indicating the poor performance of the algorithm.

It is important to highlight that the absence of the SP log in this well could contribute to the positive R2 score it has obtained this well 4. This observation can be extended to other wells within the Beetaloo data set that also lack the SP log, such as wells 8 and 28. Even though these wells show negative R2 values, they are very close to a score of 0.

5.1.5 Computational Efficiency

We recorded the time during the cross-validation phase to comprehensively measure the computational efficiency of each model and data set.

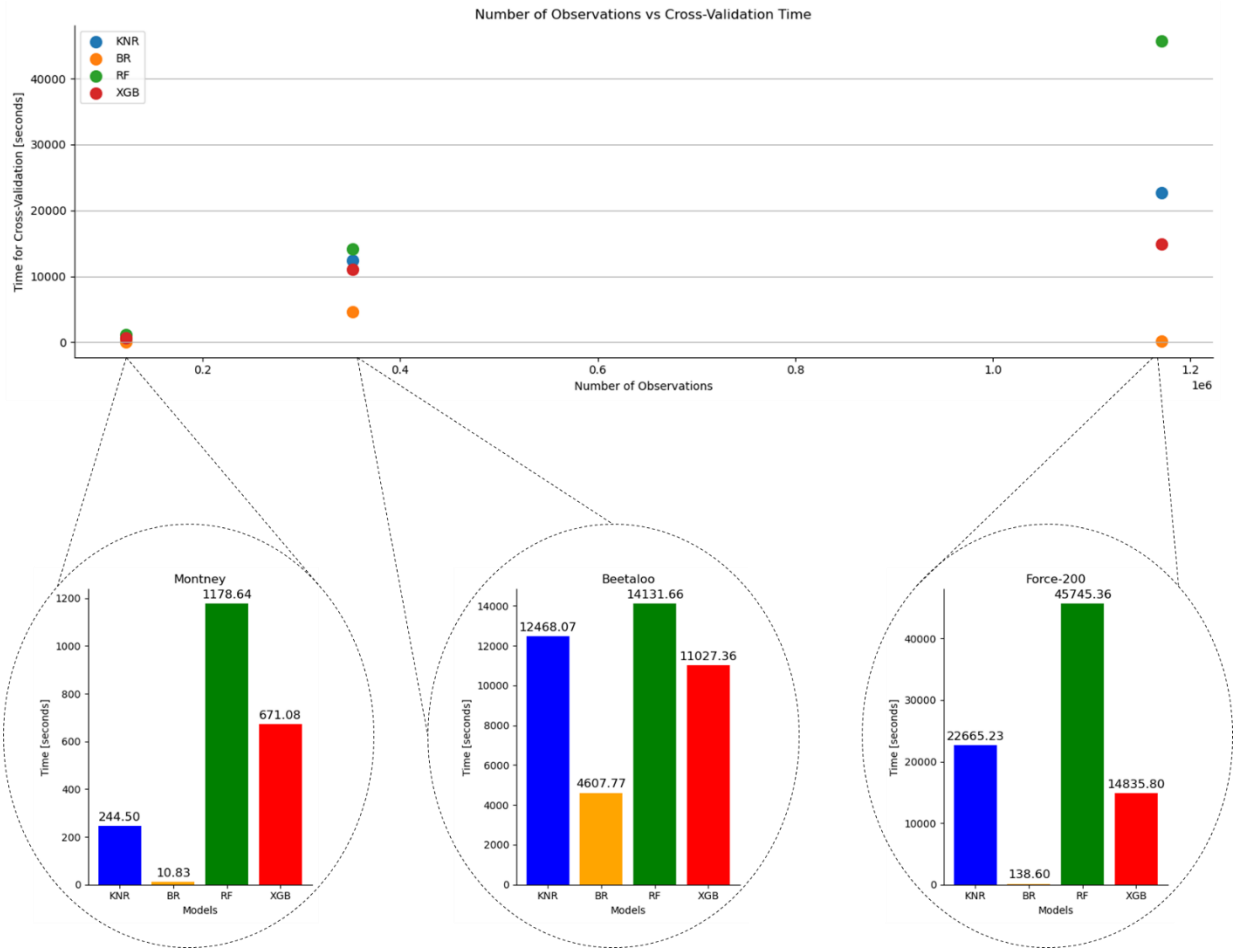


Figure 46. Computational Efficiency of the Models, including Zoom Plots for each Data set.

The Force-200 data set is the largest with over a million observations, whereas Montney is the smallest data set with approximately a hundred observations. As expected, larger data sets tend

to require more computational time than the smaller ones, given the increased data size and potential model complexity.

The BR algorithm presents the most time-efficient across all data sets. For instance, BR took only 138.60 seconds to complete the cross-validation in the Force-200 data set despite it is the largest. However, BR required 4,607.77 seconds in the Beetaloo data set, which has only 300,000 observations approximately.

On the other hand, the XGB model has variable computational time based on the data set, but it is more time efficient than KNR and RF on larger data sets. For example, the time required for KNR increases considerably as the size of the data set grows. While it only took about 4 minutes for the Montney dataset, it took over 3 hours for Beetaloo and over 6 hours for Force-200. Additionally, RF is the method that requires the longest time on all data sets. This could be due to the nature of the algorithm, which involves bootstrapping and building multiple decision trees.

5.1.6 Hyperparameter Tuning

The hyperparameter tuning process is an essential step in machine learning models to improve their performance. However, it is important to take into account the trade-off between the improvement in performance and the computational time consumed in the tuning process. This section discusses the results obtained from the sensitivity analysis and hyperparameter tuning performed.

I. Sensitivity Analysis

Since hyperparameter tuning can be computationally expensive, particularly for ensembles models such as RF and XGB. Therefore, we performed a sensitivity analysis of these algorithms to limit the number of parameters and the range of values for tuning. The sensitivity analysis was applied in the Montney and Beetaloo data sets first. During this process, we evaluate the performance of each well-log individually using different parameters, varying one parameter at a time.

Additionally, a sensitivity analysis for RF is performed for the maximum depth of the tree using the Force-200 data set. The primary reason was that setting maximum depth of the tree as default, which has no limit of depth in RF, was leading to computational issues. It was necessary to set a value for this parameter to perform the other estimations. Another reason was the high impact of the maximum depth of the tree on the performance of the models. We considered that it was important to evaluate this parameter for all data sets. It is essential to note that it was necessary to take a sample from the Force data set for this analysis due to the long time required to carry out this analysis. The sample was taken considering the same strategy for the data split in this project, which is based on wells and makes sure if the data is representative. For further information about this refer to the notebook of the Force-200 [GitHub].

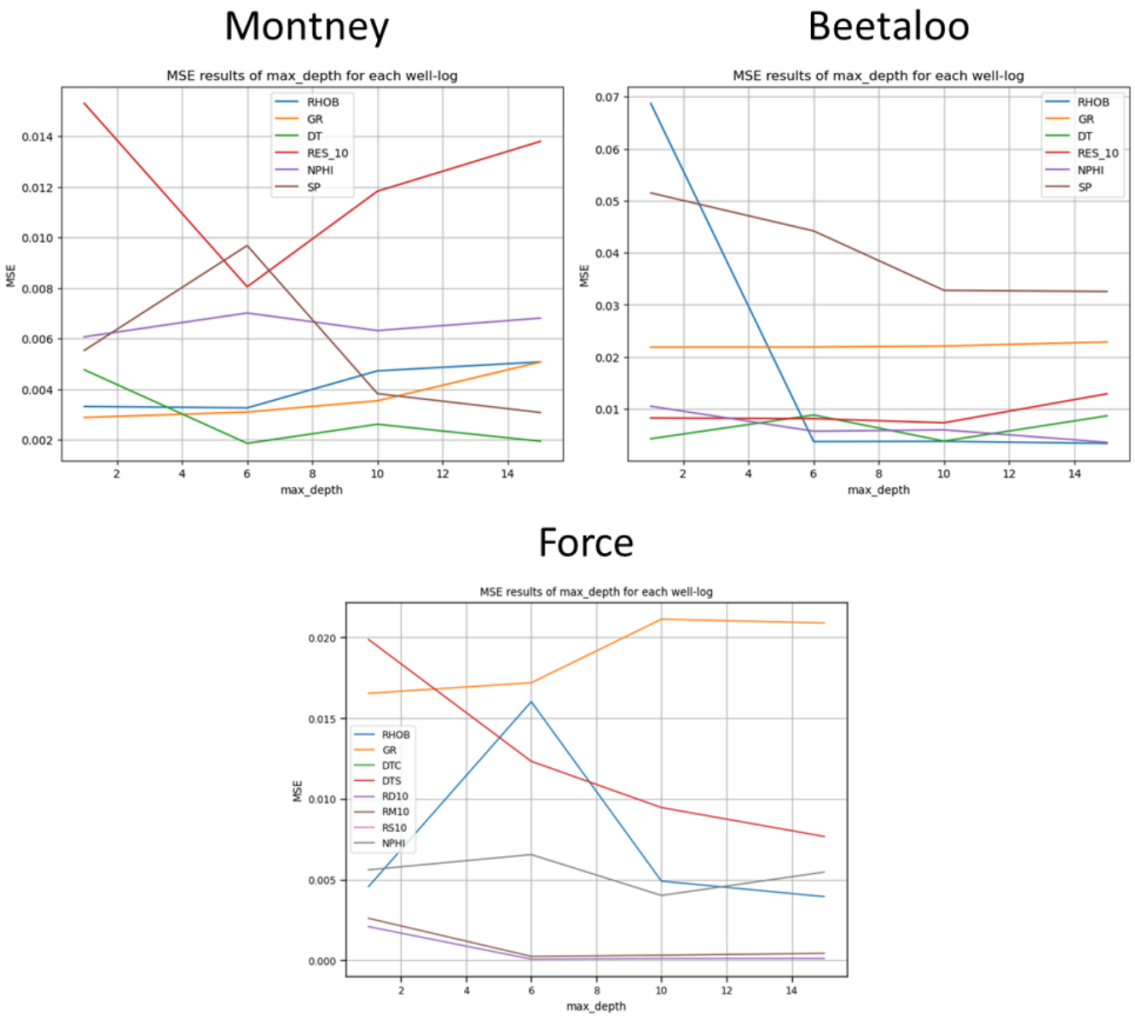


Figure 47. Sensitivity Analysis of the Force-200, Montney and Beetaloo data sets, using max_depth for RF model.

As can be seen in the Figure 47, the results demonstrate that each well-log has unique performances, complicating the task of finding the optimal combination of parameters that simultaneously improves the performance of all the features. For instance, there is not a common maximum depth of the tree that improves the performance simultaneously for all the well-logs in the Force-200 data set.

Random Forest model for each dataset shows that depth around 10 appears to provide the best performance in terms of MSE. A depth of 1, which means a very shallow trees, the models performed poorly, which is considerable as the model is likely underfitting the data, and not capturing the underlying complexity of the data sets. In contrast, a very deep tree (maximum depth of the tree = None) can lead to overfitting, but this setting cannot be tested for this study.

II. Tuning

The hyperparameter tuning was carried out on four machine learning models: KNR, BR, RF, and XGB for two different locations, Montney and Beetaloo. The Table 8 shows the results of the hyperparameter tuning for XGB in the Montney data set.

Table 8. Results of Hyperparameter Tuning for XGB in the Montney data set.

Parameters	NMSE
('max_depth=6', 'reg_alpha=0', 'reg_lambda=2')	0.0045
('max_depth=10', 'reg_alpha=0', 'reg_lambda=1')	0.0051
('max_depth=7', 'reg_alpha=0', 'reg_lambda=1')	0.0053
('max_depth=10', 'reg_alpha=0.5', 'reg_lambda=2')	0.0055
('max_depth=6', 'reg_alpha=0.5', 'reg_lambda=0')	0.0055
('max_depth=6', 'reg_alpha=0', 'reg_lambda=0')	0.0055
('max_depth=10', 'reg_alpha=0.5', 'reg_lambda=1')	0.0057
('max_depth=7', 'reg_alpha=0.5', 'reg_lambda=2')	0.0057
('max_depth=10', 'reg_alpha=0.5', 'reg_lambda=0')	0.0059
('max_depth=7', 'reg_alpha=0.5', 'reg_lambda=0')	0.0059
('max_depth=7', 'reg_alpha=0.5', 'reg_lambda=1')	0.0060
('max_depth=6', 'reg_alpha=0.5', 'reg_lambda=1')	0.0060
('max_depth=6', 'reg_alpha=0', 'reg_lambda=1')	0.0061
('max_depth=6', 'reg_alpha=0.5', 'reg_lambda=2')	0.0061
('max_depth=7', 'reg_alpha=0', 'reg_lambda=2')	0.0063
('max_depth=10', 'reg_alpha=0', 'reg_lambda=0')	0.0064
('max_depth=10', 'reg_alpha=0', 'reg_lambda=2')	0.0071
('max_depth=7', 'reg_alpha=0', 'reg_lambda=0')	0.0072
Average	0.0059

The XGB model tuning shows the normalized MSE recorded as 0.0045 using that using parameters max depth=6, regularization term alpha=0, regularization term lambda=2. This model shows slight sensitivity towards its hyperparameters as we can observe some variation in the performance. It is important to note that the best values of max depth and alpha were the default values. However, the parameter lambda, which the default value is 1, shows the lowest scores, but with different values of max depth. In this analysis, it also observed that values between 6 and 10 shows better performance than other values.

The process of hyperparameter tunning was repeated for the Beetaloo data set with similar observations. All the models presented minor or insignificant variations in NMSE with different parameters; therefore, the improvement obtained from hyperparameter tuning in MICE in these two scenarios do not justify the time and computational resources spent. For this reason, it is not carried out hyperparameter tuning for the Force-200 data set. For further details refer to the appendix AHyperparameter Tuning and notebooks [GitHub].

5.2 DISCUSSION

The focus of the project was to evaluate the performance of MICE in predicting missing well-log data in various sedimentary basins. To begin with, we evaluate various number of iterations in each data set. We use 1, 10 and 20 iterations to analyze the impact on the predictions. According to Buuren & Groothuis-Oudshoorn (2011), it is recommended to use 10 to 20 iterations to obtain reliable results.

Nevertheless, the number of iterations in MICE does not lead to any significant change in the performance of R2 and NMSE in all the models tested. This challenges the idea that increasing the number of iterations would lead to an improvement in the imputation performance, at least with the data sets used in this study. Consequently, it is inefficient to increase the number of iterations since the computational resources and the time needed to execute the multiple iterations significantly increase.

Comparing the performance between data sets, the Force-200 data set reveals favorable results in predicting missing values using MICE in well-logs. These results are consistent with a previous study by Hallam, Mukherjee, & Chassagne (2022), which also considered MICE as imputation method evaluating the performance only for DTC, DTS and RHO. However, Hallam, Mukherjee, & Chassagne did not extend their evaluation to other logs, included in our study. Moreover, the authors only use benchmark data sets from the same region which have undergone extensive preprocessing steps. For this reason, we proposed to work with various data sets with different preprocessing steps, locations, and sedimentary basins.

Furthermore, we propose a methodology that uses cross-validation to replicate how missing values occur in real-world scenarios where some well-logs may be missing entirely. This method allows us to simultaneously evaluate all well-logs and provides a proposal for the limitation of some studies that introduce random missing values for the tests, which do not represent real-world situations. Therefore, by working with different sedimentary basin data sets and implementing a more realistic approach to assess missing values, we can evaluate the performance of MICE and its generalization in a more robust way.

Using the proposed approach, we observed negative values of R2 particularly in SP and GR logs. This indicates poor performance of MICE in predicting missing values in these well logs in all data sets, including the Force-200 data set. This can be explained by the weak correlation of SP and GR with the other well-logs, and the complex geological conditions of each data set; as a result, the algorithm cannot capture relationships between the logs and has difficulty precisely imputing their missing values.

Comparing the performance of the 3 data sets, the Force-200 presents better results in R2 and NMSE. Montney and Beetaloo exhibits average negative values for R2 in all their models and

high values for the NMSE. This performance difference between data sets may be explained by Montney and Beetaloo are not extensive preconditioned as Force-200. Data preprocessing could explain the high correlations between well-logs in the Force-200 data set, which are not observed in Montney and Beetaloo. The Force-200 data set has undergone extensive pre-processing, following the strict guidelines of the Norwegian Protocol for Reporting Well Data and additional cleaning process by experts for machine learning competition. The Norwegian protocol preprocessing recommendations mainly cover data cleanup, depth shifting, and interpolation for reporting well data (Directorate, 2018). As a result, this procedure could ultimately infer the performance of the imputation of missing values in Force-200.

Based on the MICE performance analysis, XGB often outperforms other algorithms such as KNR, BR, RF, in terms of imputing missing values using MICE with different number of iterations. This can be seen on the evaluation metrics NMSE and R2 applied to different well-logs and data sets. For instance, in the Beetaloo and Montney data sets, XGB predicts with positive R2 values most of the well-logs such as RHOB, DT, and NPHI. The consistent performance of the XGB model suggests that gradient boosting techniques could be further explored and optimized for these types of data sets.

In addition, we evaluate the performance of individual blind wells for each data set using NMSE and R2 metrics, which shows that data quality may have influenced better predictions in the models. For instance, we observe that certain wells with fewer observations perform better than larger wells. It is also observed that the missing values are relevant in some wells and data sets, which could affect the performance of the model. For example, the Beetaloo data set has the worst imputation performance compared to Force-200 and Montney. This could explain because Beetaloo presents the highest percentage of missing values in all the well-logs, ranging from 20% to 52% of the data missing. This implies that the higher the percentage of missing data, the more difficult it is to impute the missing values, as the relationships between the variables may not be precisely represented.

On the other hand, the evaluation of the lithostratigraphic units reveals a pattern similar to the of the analysis of blind wells. Performance discrepancies between these units are likely due to a combination of factors, including data quantity and quality, geological complexities, and extent of data coverage. A higher volume of observations within a specific stratigraphic unit does not necessarily correlate with better model performance. This underscores the importance of both data quality and geologic context for accurate predictions in lithostratigraphic units.

It is important to note that any of these models reached convergence, including models with 20 iterations. Since MICE is an iterative process and requires convergence to produce reliable results, this lack of convergence can also explain the poor performance of the models. When there is an issue with the convergence, this could be due to insufficient number of iterations, high percentage of missing values or small sample sizes.

The evaluation of MICE is enhanced by analyzing both well-log and scatter plots since these graphical insights reveal patterns that might not be evident through mere evaluation metrics. These plots show the limitations of MICE in predicting missing well-log data precisely, particularly for logs with complex relationships. The presence or absence of specific well-logs in certain wells significantly influences the imputation performance, impacting both evaluation metrics and graphical representations. For instance, the absence of SP seems to lead to an improvement in the imputation performance in some cases, as observed in higher R2 values. The reason for this improvement may be related to the fact that SP presents the most negative values compared to other records; therefore, when there is no SP, the average performance of that well or stratigraphy improves. However, it is important to consider further analysis and test different combinations of well-logs to evaluate the performance.

On the other hand, the computational efficiency section provides valuable information on the practicality of these models. As data size grows, computational demands increase. However, certain models, such as BR, showed surprising efficiency on the longest data sets, making it a potential candidate for future large-scale applications. The performance of XGB also stands out for its balance between efficiency and performance, especially in contrast to models like the KNR and RF, which required more time.

The process of hyperparameter tuning reveals that well-logs have different performances with one set of parameters. This variability in the performance indicates that each well-log has unique characteristics that require a custom model to obtain optimal predictions, showing a great limitation of MICE to optimize performance across all features simultaneously. The best hyperparameters for each well-log may differ from each other, creating a challenge to find a single optimal set for the overall model. For example, we observe that improving the performance of one well-log sometimes negatively affects others logs. Therefore, there is a need for a more flexible approach that can better adapt to the unique characteristics of each well-log. Potential future strategies could involve implementing feature-specific models or exploring advanced machine learning methods, such as deep learning and neural networks, capable of learning complex feature interactions.

This study underscores the complexity of imputing missing data from well-logs in heterogeneous sedimentary basins. Although machine learning algorithms with MICE offer automated solutions, the findings emphasize the need for more research to minimize user input. Challenges arise from the limited and diverse nature of well-logs, the wide ranges of values, and the complex relationships within the data. In particular, the extensive preprocessing, as seen on the Force-200 dataset, can yield better results. Since the understanding the subsurface is complex, future strategies could include custom models for different well-logs or advanced techniques such as deep learning and neural networks to better capture complex feature interactions.

Limitations:

This study is limited by the use of data sets with limited missing measurements, which may not fully represent scenarios with entirely missing well-logs. The exploration of hyperparameters was restricted due to practical limitations, potentially missing optimal combinations. Hyperparameter tuning was carried out for specific data sets, excluding the Force-200 data set due to its size and resource demands. The reliance on a single benchmark data set with strong correlations raises questions about the applicability of MICE to more diverse data sets with varying pre-processing levels. The absence of uncertainty analysis in the models, influenced by consistent random states, and the computational complexity of such analysis present further limitations. Moreover, the methodology employed for cross-validation, aiming to replicate how missing values manifest in real-world scenarios, only represents situations where some well-logs may be missing entirely and does not account for gaps or intervals where data might be absent. This may not wholly reflect the diversity of missing data patterns seen in practice. Moreover, the methodology used for cross-validation, with the goal of replicating how missing values manifest in real-world scenarios, only represents situations where some well-logs may be completely missing and does not account for gaps or intervals where data may be missing. This may not fully reflect the diversity of missing data patterns seen in practice. While the study comprehensively evaluated all well-logs in each dataset, it did not assess different combinations of well-logs for model performance. Lastly, there was no direct comparison between MICE and conventional machine learning strategies that independently predict single well-logs.

6 | CONCLUSIONS

6.1 CONCLUSIONS

In conclusion, this study aimed to evaluate the performance of Multivariate Imputation by Chained Equations (MICE) in predicting missing well-log data in various sedimentary basins. The research was conducted using three different data sets from distinct geological contexts, with minimal user input and preprocessing. The results showed that MICE, when combined with machine learning algorithms such as XGBoost (XGB), Random Forest (RF), K-Nearest Neighbors (KNR), and Bayesian Ridge (BR), can provide valuable insights into the prediction of missing well-log data. The XGB algorithm frequently outperformed other techniques, especially in imputing missing values using MICE across varying iterations. However, the performance of MICE varied across different data sets and well-logs, highlighting the complexity of imputing missing data in heterogeneous sedimentary basins.

The study revealed that the number of iterations in MICE did not significantly impact the performance of the models, challenging the idea that increasing the number of iterations would lead to improved imputation performance. The results also indicated that data quality, preprocessing, and geological complexities play a crucial role in the performance of MICE. The Force-200 data set, which underwent extensive preprocessing, demonstrated better imputation performance compared to the Montney and Beetaloo data sets.

The evaluation of MICE performance was complemented with graphical visualizations, which showed the limitations of MICE in predicting missing well-log data precisely, particularly for logs with complex relationships. The presence or absence of specific well-logs in certain wells significantly influenced the imputation performance, impacting both numerical metrics and graphical representations.

The study also highlighted the challenges of finding a single optimal set of hyperparameters for the overall model, as each well-log has unique characteristics that require a custom model to obtain optimal predictions. This suggests a need for more flexible approaches that can better adapt to the unique characteristics of each well-log, such as implementing feature-specific models or exploring advanced machine learning methods like deep learning and neural networks.

6.2 RECOMMENDATIONS AND FUTURE RESEARCH

Despite the limitations of this study, it provides insights into the applicability of MICE for predicting missing well-log data in different geological contexts. The findings emphasize the need for more research to minimize user input and develop more robust and flexible approaches to imputing missing data in well logs. This new approach should facilitate and guide the user in preprocessing methods and the selection of appropriate imputation methods, machine learning algorithms and hyperparameters. This could involve developing decision frameworks that analyze well-log characteristics and suggest the most appropriate approach based on available data.

Furthermore, more advanced imputation techniques can be explored beyond traditional machine learning algorithms. Techniques such as deep learning have shown promising results in handling complex data relationships. For instance, investigate the viability of training neural networks to impute missing well-log data that includes geological relationships and well-logs interactions for more accurate predictions in heterogeneous sedimentary basins.

Another area of future research could focus on understanding the reasons behind the poor performance of certain wells and lithostratigraphy units. This may involve investigating the quality of the data, identifying missing features relevant to those wells or stratigraphic units, or exploring alternative modeling techniques. For wells and stratigraphic units with strongly negative R2 values, reassessing model applicability or considering collecting more data or additional features could help improve predictions.

In addition, it would be valuable to examine the impact of preprocessing on different datasets, similar to the comprehensive preprocessing performed on the Force-200 dataset, which follows Norwegian protocol guidelines. Investigate whether preprocessing techniques can improve MICE performance on data sets with lower log correlations, such as Montney and Beetaloo. This can provide information about the generalization of preprocessing strategies.

To further evaluate the performance of MICE, it is recommended to explore various combinations of well-logs, including variations in the number of well-logs used within the models. Additionally, it would be important to contrast MICE with conventional machine learning methodologies that predict individual well-logs, allowing the evaluation of effectives and performance of both approaches in different sedimentary basins.

Finally, it is necessary to investigate further a cross-validation approach that effectively replicates real-world scenarios. These scenarios involve the presence of missing intervals within certain well-logs, and even extend to cases where entire well logs are missing. This approach would allow for a more realistic assessment and provide a comprehensive evaluation of MICE performance and its ability to address complex challenges in the context of subsurface data.

7 | REFERENCES

Bangert, P. (2021). *Machine Learning and Data Science in the Oil and Gas Industry.* Gulf Professional Publishing. doi:https://doi.org/10.1016/B978-0-12-820714-7.00001-7

Belyadi, H., & Haghighat, A. (2021). *Machine Learning Guide for Oil and Gas Using Python.* Elsevier. doi:https://doi.org/10.1016/C2019-0-03617-5

Berglund, P. A., & Heeringa, S. (2014). *Multiple imputation of missing data using SAS.* SAS Institute.

Buuren, S. v., & Groothuis-Oudshoorn, K. (2011). mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3). Retrieved from http://www.jstatsoft.org/

Crombez, V., Kunzmann, M., Faiz, M., Piane, C. D., Munday, S., & Forbes, A. (2022). Stratigraphic architecture of the world's oldest shale gas play: The 1400-1200 Ma Velkerri and Kyalla formations in the Beetaloo Sub-Basin. *EarthArXiv*. doi:https://doi.org/10.31223/X5SS6R

Crombez, V., Rohais, S., Baudin, F., & Euzen, T. (2016). Facies, well-log patterns, geometries and sequence stratigraphy of a wave-dominated margin: insight from the Montney Formation (Alberta, British Columbia, Canada). *BULLETIN OF CANADIAN PETROLEUM GEOLOGY*. doi:https://doi.org/10.2113/gscpgbull.64.4.516

Darling, T. (2005). *Well Logging and Formation Evaluation.* Gulf Professional Publishing. doi:https://doi.org/10.1016/B978-0-7506-7883-4.X5000-1

Directorate, N. P. (2018). *Guidelines for reporting well data to authorities after completion.* Norwegia Petroleum Directorate. Retrieved from https://www.npd.no/globalassets/1-npd/regelverk/veiledninger/guidelines-for-reporting-welldata-to-authorities-after-completion-blue-book.pdf

Dixneuf, P., Errico, F., & Glaus, M. (2021). A computational study on imputation methods for missing environmental data. *arXiv*. doi:https://doi.org/10.48550/arXiv.2108.09500

Ducros, M., Sassi, W., Vially, R., Euzen, T., & Crombez, V. (2017). 2-D Basin Modeling of the Western Canada Sedimentary Basin across the Montney-Doig System: Implications for

Hydrocarbon Migration Pathways and Unconventional Resources Potential. *The American Association of Petroleum Geologists*. doi:https://doi.org/10.1306/13602027M1143703

Erler, N. S. (2019). *Bayesian Imputation of Missing Covariates.* Erasmus University Rotterdam.

Evenick, J. C. (2018). *Introduction to Well Logs and Subsurface Maps (2nd Edition).* PennWell. Retrieved from https://app.knovel.com/hotlink/toc/id:kpIWLSME08/introduction-well-logs/introduction-well-logs

Faiz, M., Crombez, V., Piane, C. D., Lupton, N., Camilleri, M., & Langhi, L. (2021). Petroleum systems model for source-rock-reservoir evaluation in the Beetaloo Sub-basin. *Australasian Exploration Geoscience Conference (AEGC)*.

Feng, R., Grana, D., & Balling, N. (2021). Imputation of missing well log data by random forest and its uncertainty analysis. *Computers & Geosciences, 152*. doi:https://doi.org/10.1016/j.cageo.2021.104763

Gallatin, K., & Albon, C. (2023). *Machine Learning with Python Cookbook, 2nd Edition.* O'Reilly Media, Inc.

Galli, S. (2022). *Python feature engineering cookbook.* Packt Publishing.

Hallam, A., Mukherjee, D., & Chassagne, R. (2022). Multivariate imputation via chained equations for elastic well log imputation and prediction. *Applied Computing and Geosciences, 14*. doi:https://doi.org/10.1016/j.acags.2022.100083.

Holgate, N. E., Jackson, C. A.-L., Hampson, G. J., & Dreyer, T. (2013). Sedimentology and sequence stratigraphy of the Middle–Upper Jurassic Krossfjord and Fensfjord formations, Troll Field, northern North Sea. *Petroleum Geoscience*. doi:https://doi.org/10.1144/petgeo2012-039

Huyen, C. (2022). *Designing Machine Learning Systems.* O'Reilly Media, Inc.

Jackson, C.-L., & Larsen, E. (2009). Temporal and spatial development of a gravity-driven normal fault array: Middle–Upper Jurassic, South Viking Graben, northern North Sea. *Journal of Structural Geology*. doi:https://doi.org/10.1016/j.jsg.2009.01.007

Jafari, R. (2022). *Hands-On Data Preprocessing in Python.* Packt Publishing.

Jeyaraman, B. P., Olsen, L. R., & Wambugu, M. (2019). *Practical Machine Learning with R.* Packt Publishing.

Leke, C. A., & Marwala, T. (2019). *Deep Learning and Missing Data in Engineering Systems.* Springer Cham. doi:https://doi.org/10.1007/978-3-030-01180-2

Little, R. J., & Rubin, D. B. (2020). *Statistical analysis with missing data.* John Wiley & Sons, Inc.

Liu, H. (2015). *Principles and Applications of Well Logging.* Springer Berlin, Heidelberg. doi:https://doi.org/10.1007/978-3-662-54977-3

Lopes, R. L., & Jorge, A. M. (2018). Assessment of predictive learning methods for the completion of gaps in well log data. *Journal of Petroleum Science and Engineering*. doi:https://doi.org/10.1016/j.petrol.2017.11.019

Michimae, H., & Emura, T. (2022). Bayesian ridge estimators based on copula-based joint prior distributions for regression coefficients. *Computational Statistics*. doi:https://doi.org/10.1007/s00180-022-01213-8

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python : a guide for data scientists.* O'Reilly Media.

Murillo, W. A., Horsfield, B., & Vieth-Hillebrand, A. (2019). Unraveling petroleum mixtures from the South Viking Graben, North Sea: A study based on δ13C of individual hydrocarbons and molecular data. *Organic Geochemistry*. doi:https://doi.org/10.1016/j.orggeochem.2019.103900

Piane, C. D., MacRae, C., Rickard, W., Crombez, V., Faiz, M., & David, N. D. (2021). Diagenetic controls on the reservoir quality of organic-rich shales of the Mesoproterozoic Velkerri Formation (Beetaloo Basin). *Australasian Exploration Geoscience Conference (AEGC)*.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. doi:https://doi.org/10.1093/bioinformatics/btr597

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. doi:https://doi.org/10.1002/sim.4067

APPENDIX A

DATA ANALYSIS

The following figures illustrate key aspects of the data splitting strategy to ensure that the data is representative and well-suited for model evaluation. For further information refer to the notebooks.

A.1 MONTNEY

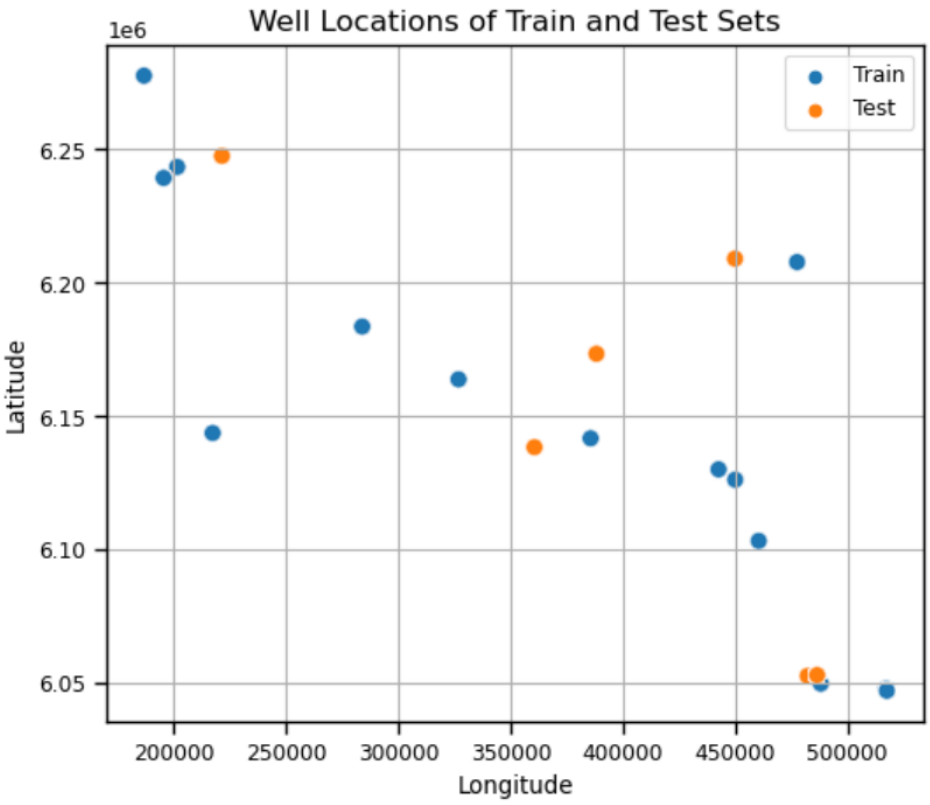


Figure 48. Geo graphical distribution of the wells in the training and test sets for Montney.

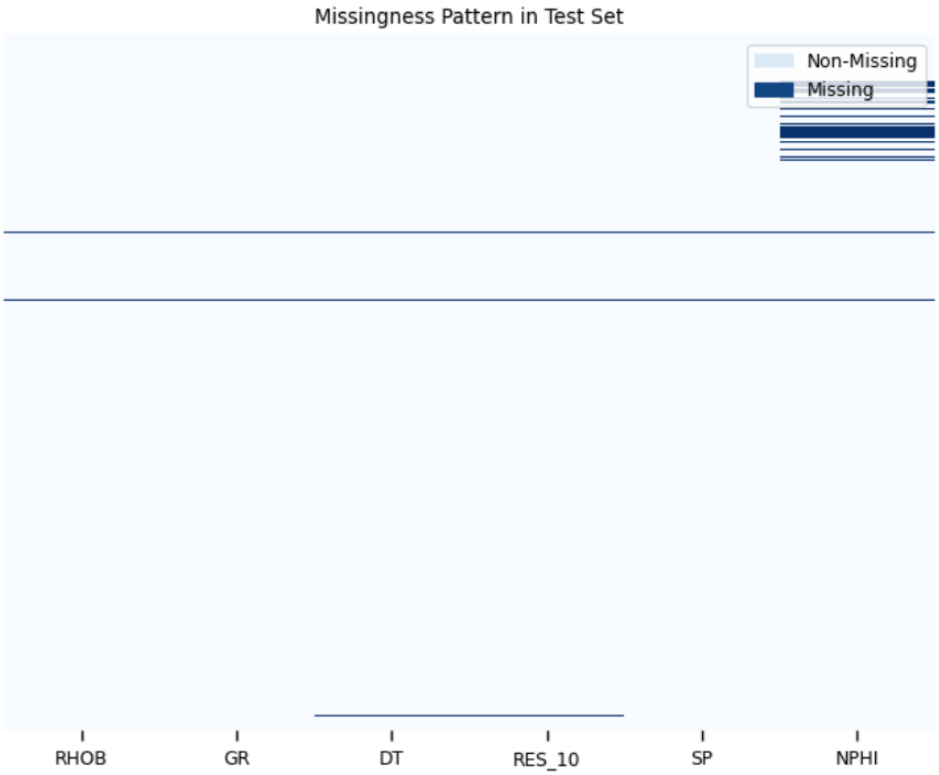


Figure 49. Pattern of missing values in the test set for Montney.

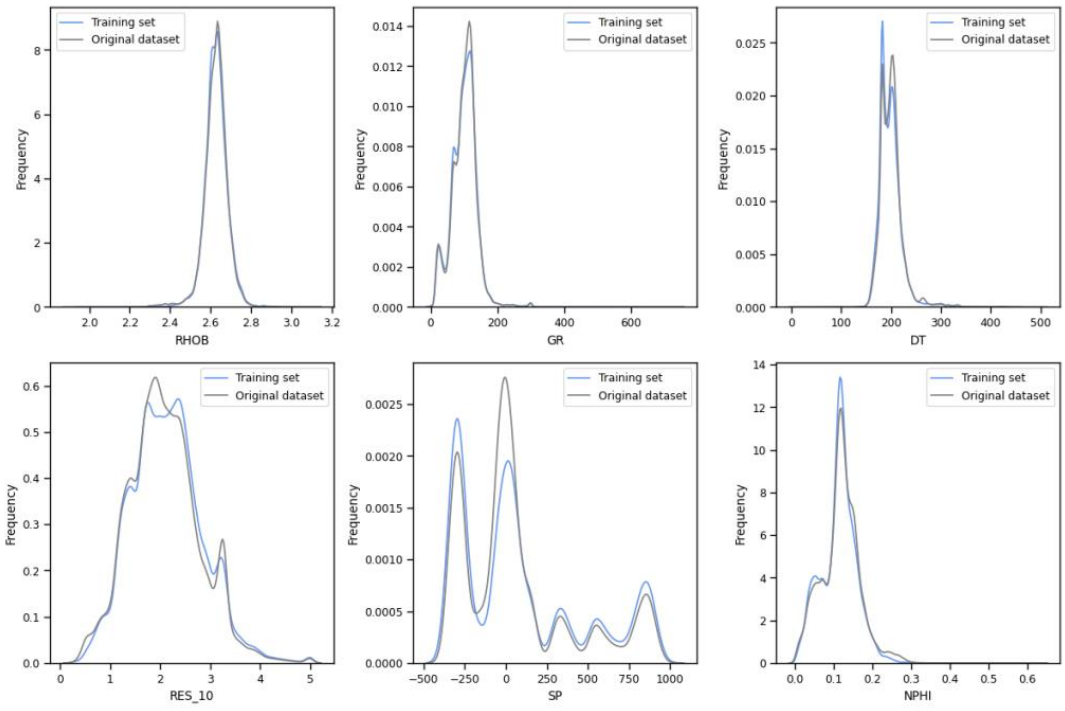


Figure 50. Distribution of well-logs features in the original dataset and training set for Montney.

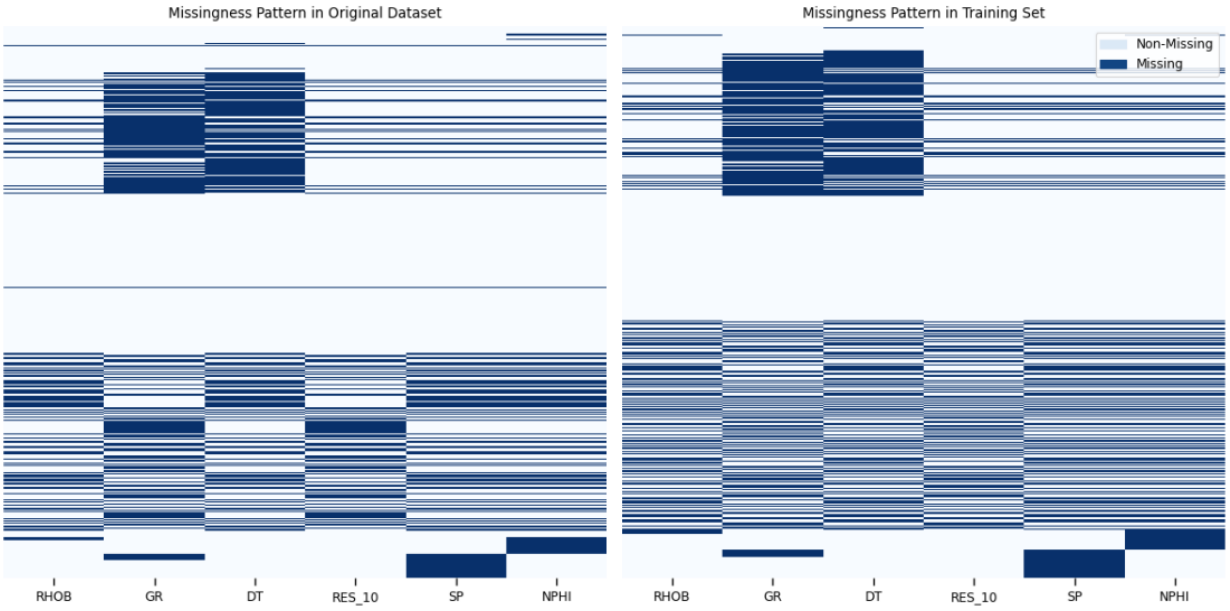


Figure 51. Missing data patterns between the original data set and training set for Montney.

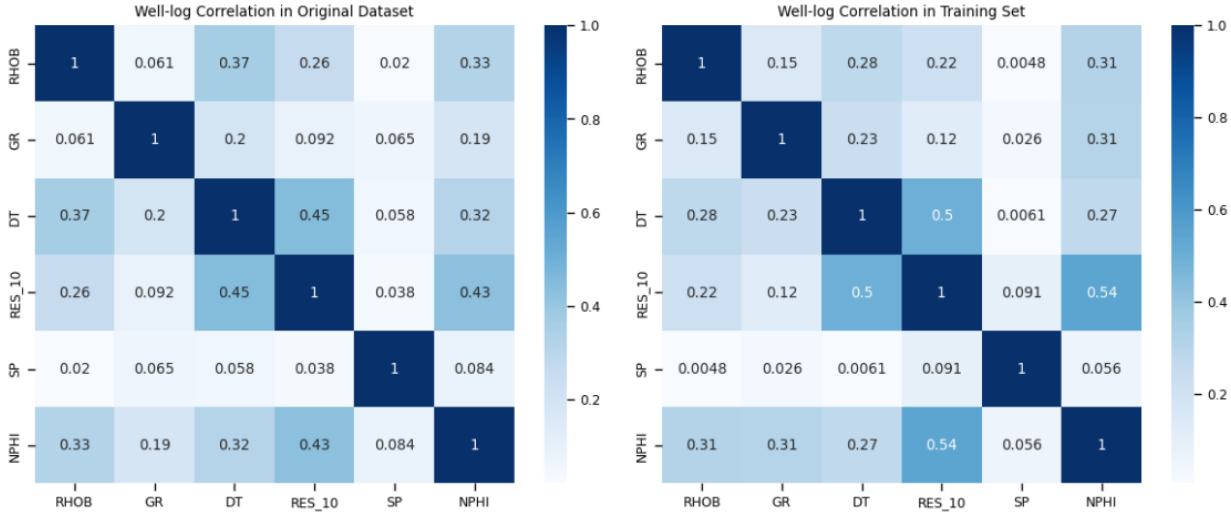


Figure 52. Correlation matrices for the original data set and the training set for Montney.

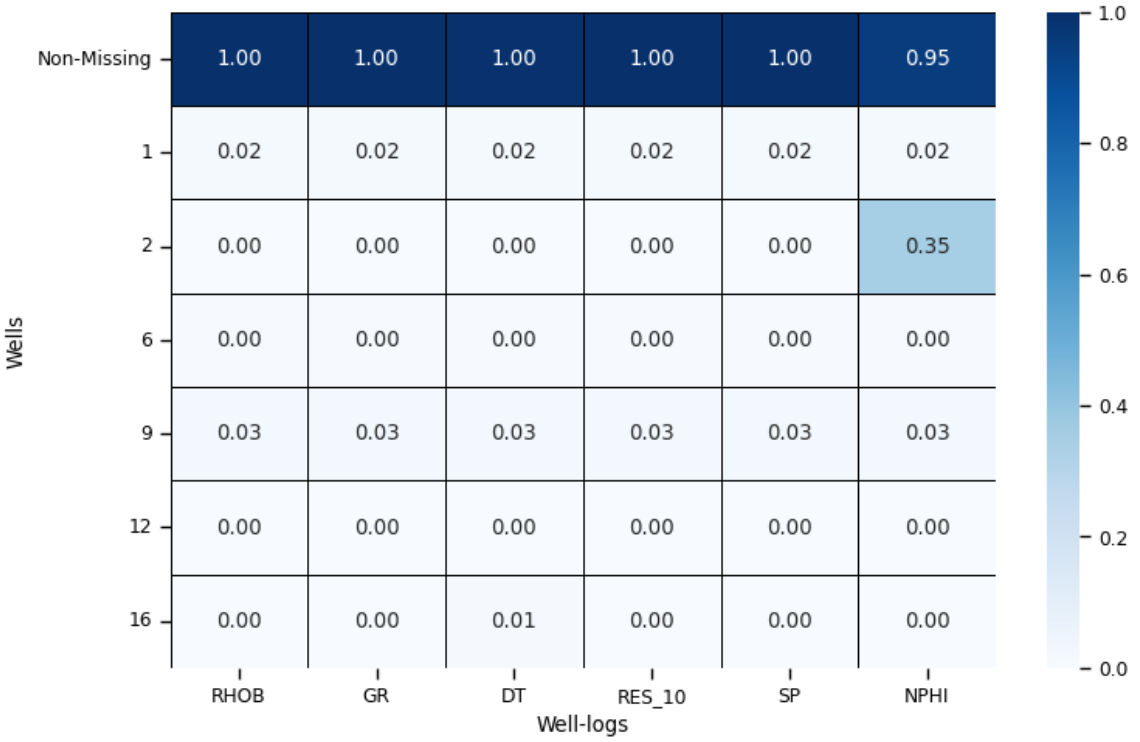


Figure 53. Fraction of missing data in wells from the Montney data set.

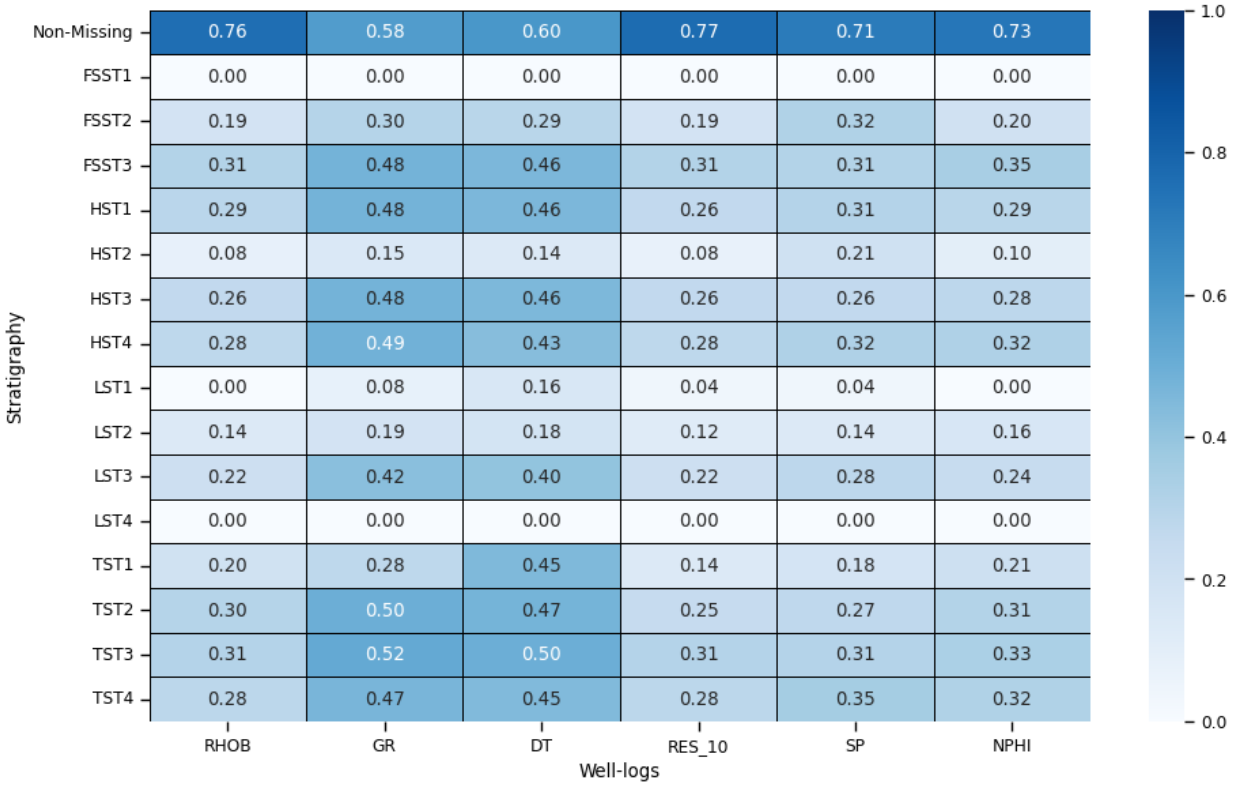


Figure 54. Fraction of missing data in stratigraphy units from the Montney data set.

A.2 BEETALOO

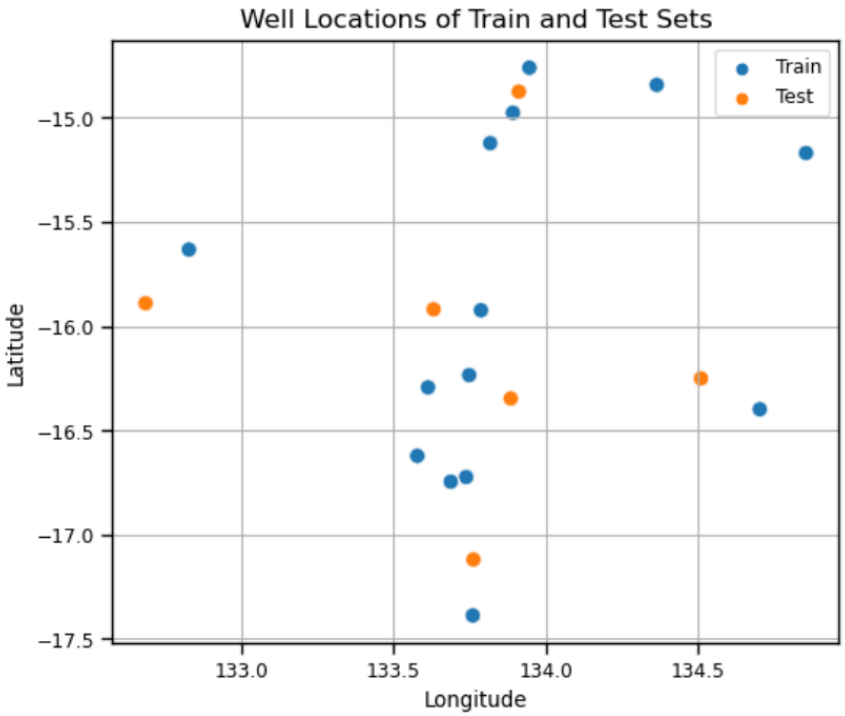


Figure 55. Geographical distribution of the wells in the training and test sets for Beetaloo.

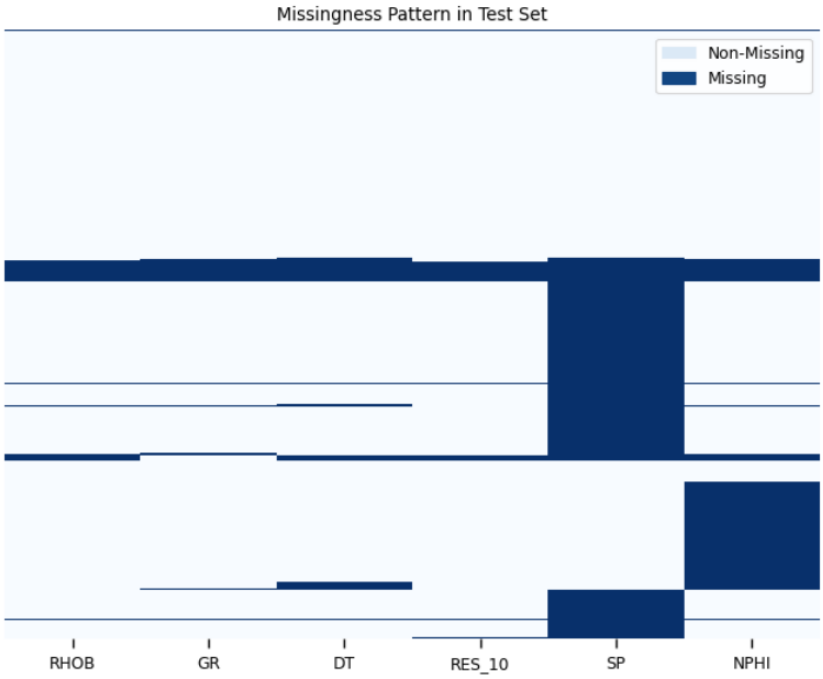


Figure 56. Pattern of missing values in the test set for Beetaloo.

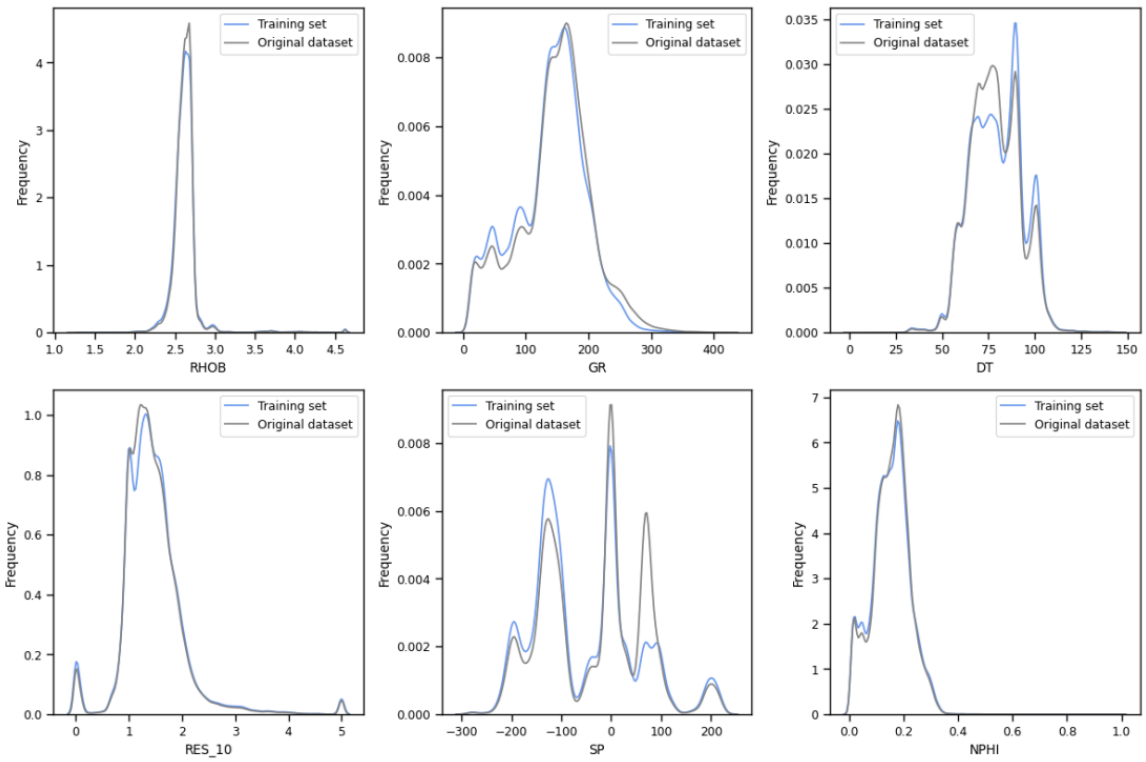


Figure 57. Distribution of well-logs features in the original dataset and training set for Beetaloo.

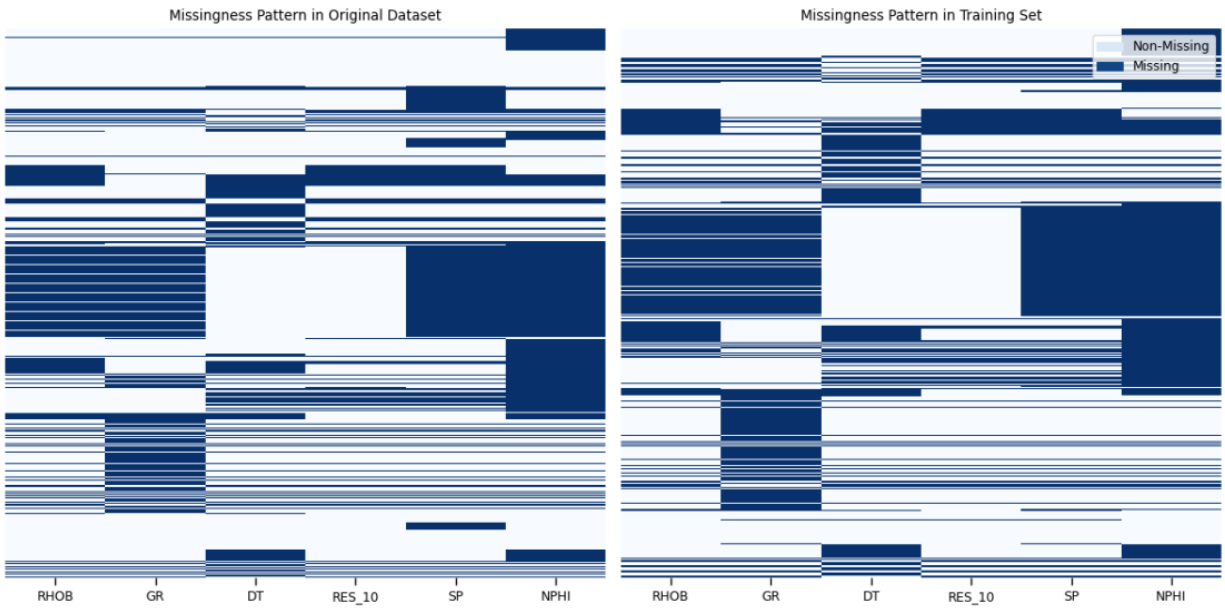


Figure 58. Missing data patterns between the original data set and training set for Beetaloo.

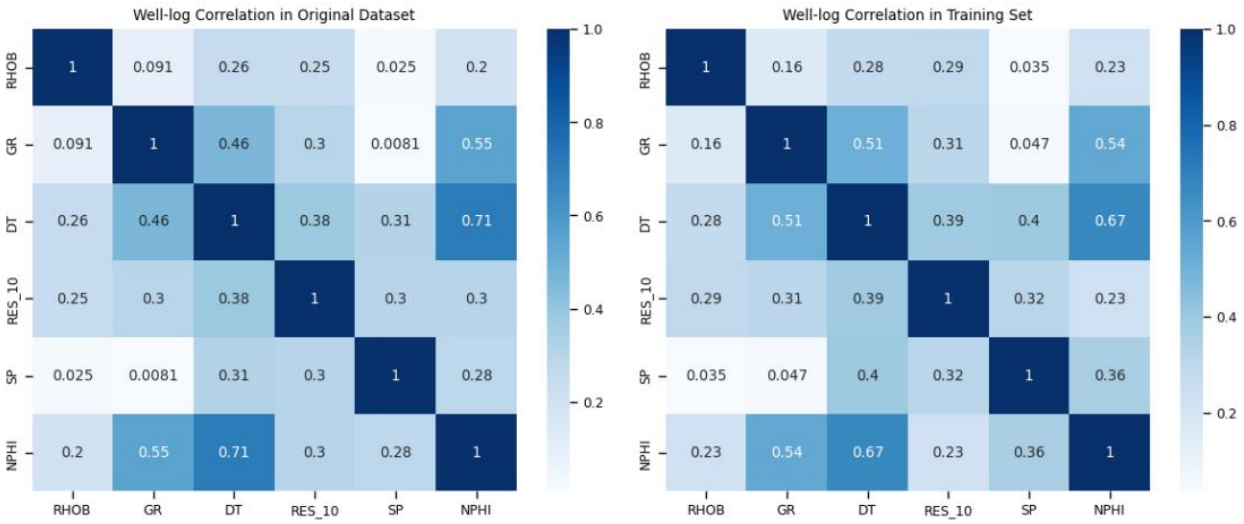


Figure 59. Correlation matrices for the original data set and the training set for Beetaloo.

Figure 60. Fraction of missing data in wells from the Beetaloo data set.

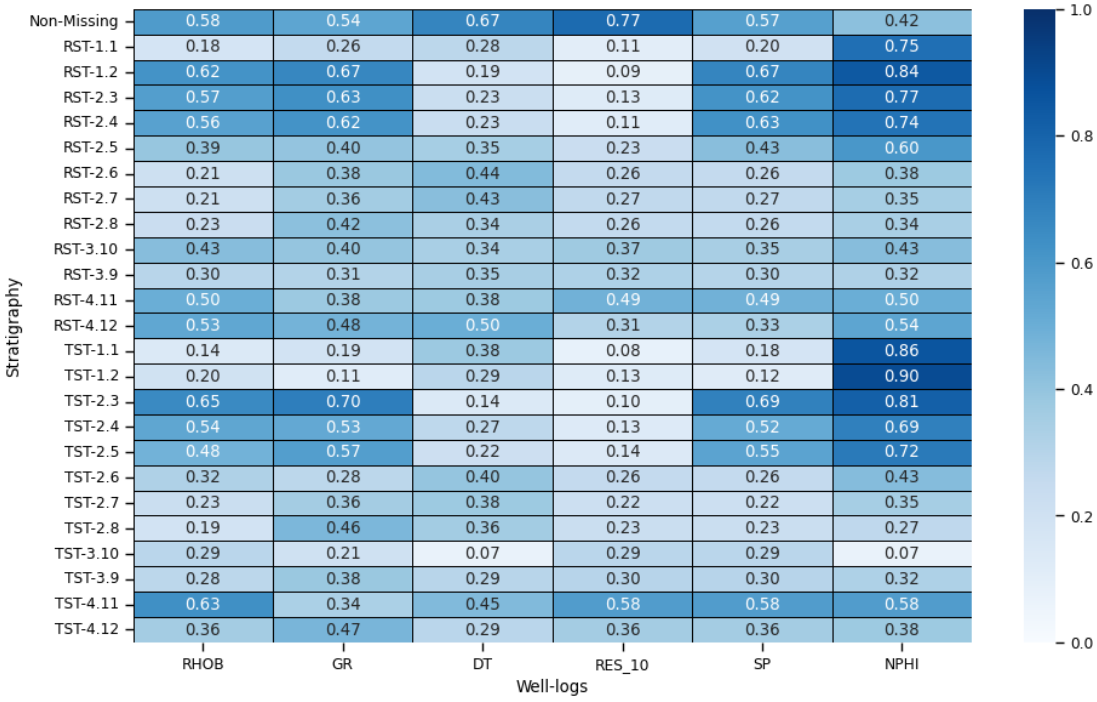


Figure 61. Fraction of missing data in stratigraphy units from the Beetaloo data set.

A.3 FORCE-200

Figure 62. Fraction of missing data in wells from the Force-200 data set.

Figure 63. Fraction of missing data in stratigraphy units from the Force-200 data set.

APPENDIX B

HYPERPARAMETER TUNING

A.1 SENSITIVITY ANALYSIS

Table 9. Hyperparameters and Values used for the sensitivity analysis.

Algorithm	Hyperparameter	Tunned Values	Default Value
Random Forest	max_depth	{1, 6, 10, 15}	None
	n_estimators	{1, 10, 50, 100, 200}	100
	min_samples_split	{2, 6, 10, 15}	2
	min_samples_leaf	{1, 5, 10, 15}	1
XGBoost	max_depth	{1, 3, 6, 7, 10, 15}	6
	n_estimators	{1, 10, 50, 100, 200}	100
	min_child_weight	{0, 0.5, 1, 2}	1
	learning_rate	{0, 0.05, 0.1, 0.3, 0.5}	0.3
	gamma	{0, 0.5, 1}	0
	reg_alpha	{0, 0.5, 1}	0
	reg_lambda	{0, 1, 2}	1

A.1.1 Montney

I. Random Forest

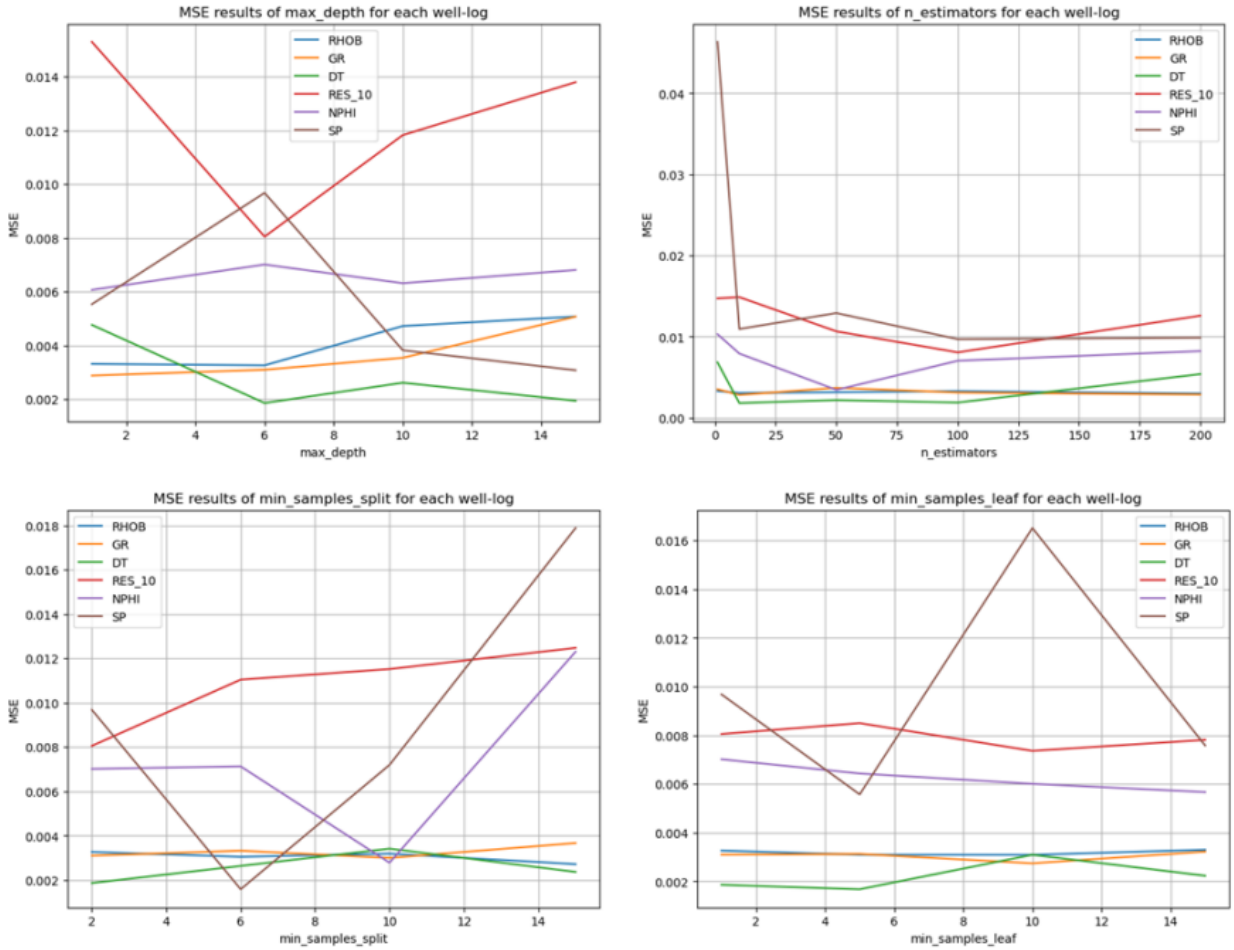


Figure 64. Sensitivity Analysis RF for Montney.

I. XGBoost

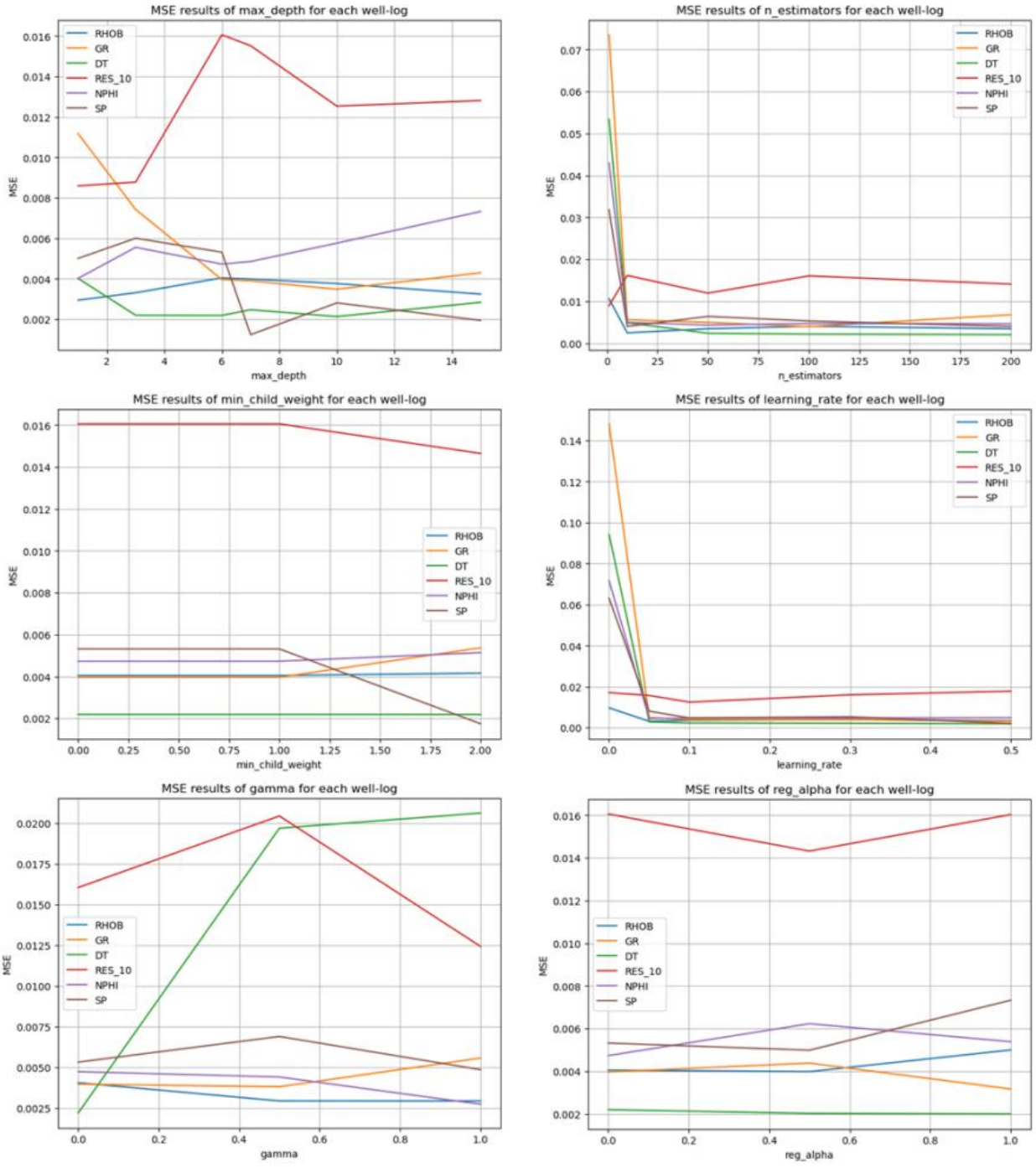


Figure 65. Sensitivity Analysis XGB for Montney.

A.1.2 Beetaloo

I. Random Forest

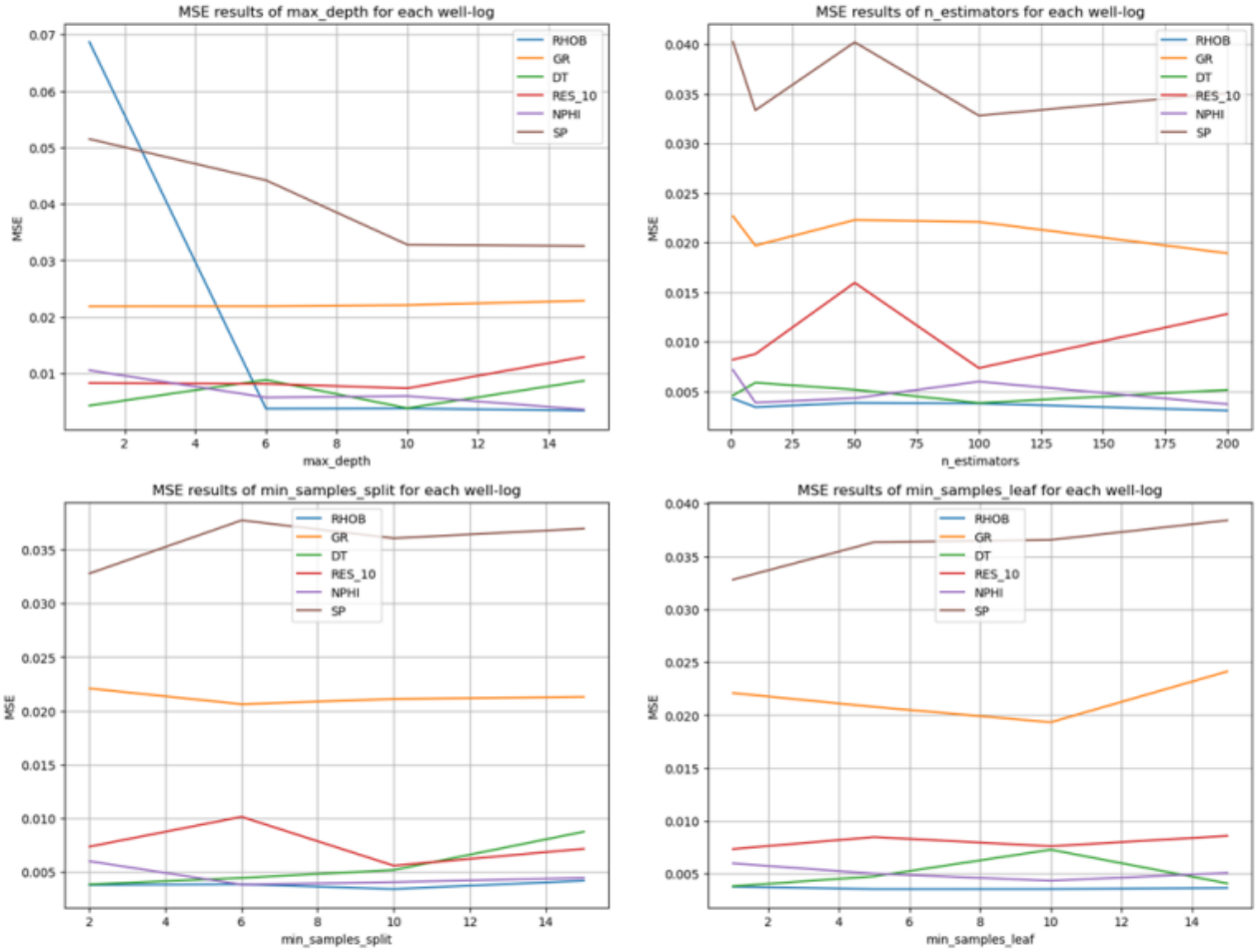


Figure 66. Sensitivity Analysis RF of Beetaloo.

II. XGBoost

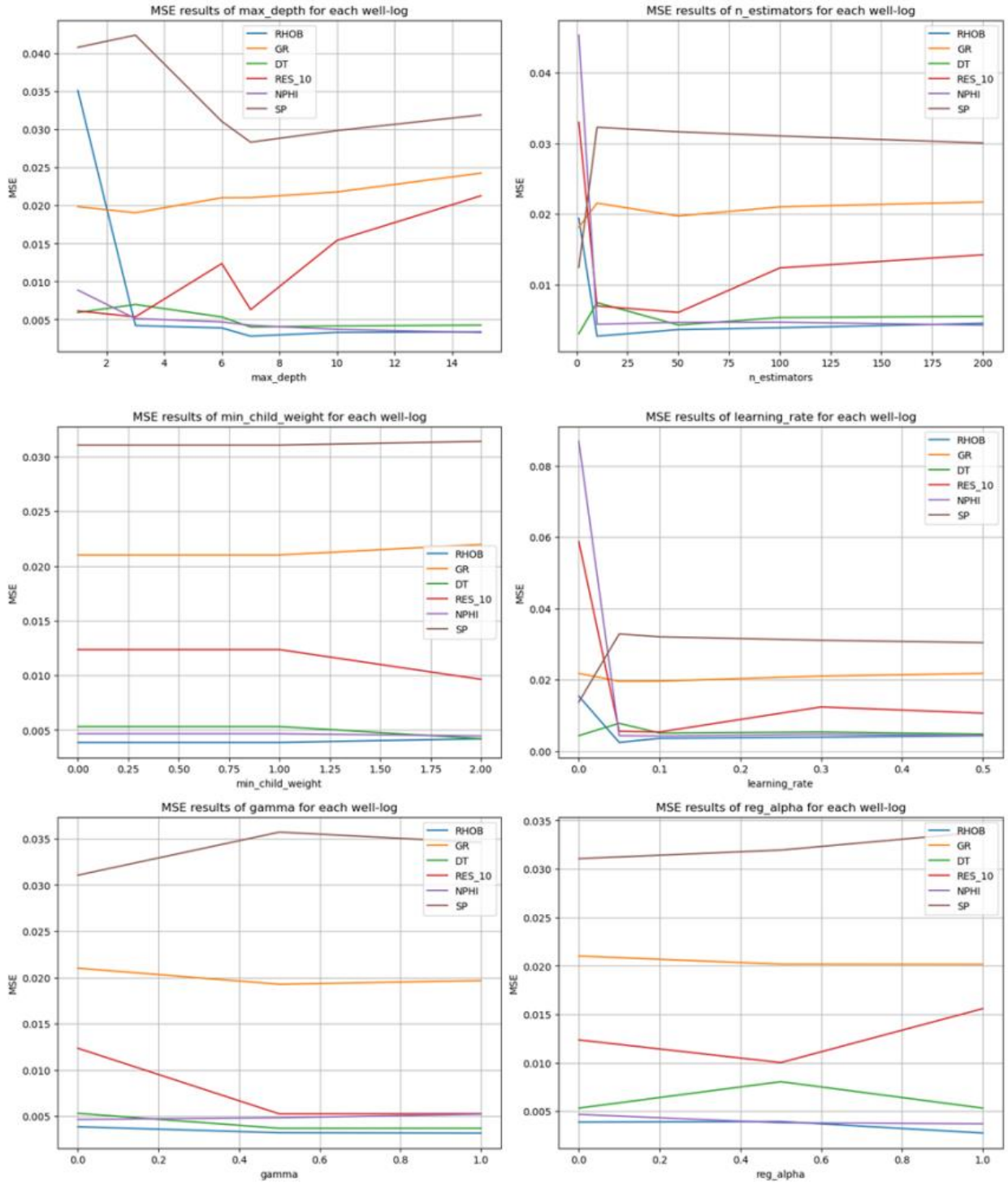


Figure 67. Sensitivity Results of XGB for Beetaloo.