Faculty of Computing and Informatics

Department of Informatics

Leveraging Machine Learning for Enhanced Efficiency in Mineral Processing Through Silica Estimation at Rosh Pinah Zinc Mine

Thesis

Submitted in partial fulfilment of the requirements for the degree of

Master of Data Science

at

Namibia University of Science and Technology

Author: Angula Tuhafeni Kiiga (212083740)

Supervisor: Dr Richard Maliwatu

Date of Submission: 18 April 2025

METADATA

Study Title	Leveraging Machine Learning for Enhanced Efficiency in Mineral Processing Through Silica Estimation at Rosh Pinah Zinc Mine
Student Surname, Name	Kiiga, Angula Tuhafeni
Student Number	21208370
Faculty	Faculty of Computing and Informatics
Department	Department of Informatics
Programme	Master of Data Science (09MADS)
Status	Dissertation
Supervisor/Affiliation	Dr Richard Maliwatu
First Co-Supervisor/Affiliation	N/A
Second Co-Supervisor/Affiliation	N/A
Digital Signature of Student	
Digital Signature of Supervisor	
Digital Signature of Programme Coordinator (or equivalent)	
Digital Signature of FCI HDC Representative	

DECLARATION

I, **Angula Tuhafeni Kiiga-(212083740)**-hereby declare that the work presented in this thesis for the degree of Master of Data Science, titled: "**Leveraging Machine Learning for Enhanced Efficiency in Mineral Processing Through Silica Prediction at Rosh Pinah Zinc Mine"** is my own original work and that I have not previously in its entirety or in part submitted it at any university or other higher education institution for the award of a degree. I further declare that I have fully acknowledged any sources of information I will use for the research in accordance with the Institution's rules.

Signature:  Date: 18 April 2025

ABSTRACT

In mineral processing, real-time silica estimation in ore samples remains a significant challenge which influences operational inefficiencies, increased costs and recovery rate. Traditional methods depend on periodic laboratory sample tests, which, while accurate, introduce delays that exacerbate issues related to delays on changes of the blending ratio to stabilise the silica entering the plant circuit. These conventional methods provide valuable insights into mineral content patterns but fail to timely deliver the information needed for optimal mineral processing efficiency. This study addresses these challenges by introducing machine learning techniques to estimate silica content in the ore feeds.

Four commonly used Machine Learning (ML) models were considered, namely Multiple Linear Regression, Support Vector Regression, Random Forest, and Extreme Gradient Boosting. A dataset consisting of 5967 entries and 10 features collected from mining operations at Rosh Pinah Zinc Mine in Namibia for the past five years (2019-2023) was used to train and test the ML models. The study identifies the key features that significantly influence silica content estimation. Pearson correlation analysis was applied, which shows elements such as Sulphur (-0.54), Zinc (-0.49) and Manganese (-0.49) exhibiting strong negative correlation. Among the four ML models, Random Forest Regression emerged as the highly performing algorithm due to its ability to capture complex and non-linear trends, hence superior performance. The evaluation was obtained using standard machine learning performance metrics, namely Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) scores. The Random Forest model demonstrated 85.4% estimation accuracy, with low error rates of MSE at 0.3821 and MAE at 0.2848.

This study demonstrates the capability of machine learning to offer a timely data-driven approach to enhance ore blending and improve overall mineral processing efficiency. In addition, this enhances decision-making processes in mineral processing, reducing operational costs and increasing recovery rates.

Keywords: mineral processing, comminution, machine learning, ore quality estimation, silica content, AI in mining

ACKNOWLEDGEMENTS

Firstly, I would like to express my profound gratitude to the Almighty God for granting me the strength, wisdom, and courage to complete my master's degree journey. Without His love, grace and protection, this study would not have been possible.

I would like to extend my sincerest thanks to Dr Richard Maliwatu, my supervisor, for his invaluable support, guidance and patience throughout my thesis journey. Despite his tight schedule, he always made time to provide insightful feedback and directed me in the right direction. I am indebted to his commitment and deeply grateful for his mentorship.

To my beloved parents and siblings, words cannot fully express my appreciation for your love and unwavering support. You kept me going during the hard times, and your belief in me encouraged me to achieve this goal. Thank you for lifting my spirits and always being there for me.

Also, not forgetting the Rosh Pinah Zinc plant department. I am so thankful for opening their doors to me and allowing me to access their data which was crucial for this research. Your willingness to share critical information has driven the completion of this thesis. I sincerely appreciate your support in my academic endeavours.

Finally, to my academic friends Ndamona, Julia, Teopolina, Dorian and Diana, who have become more like family, thank you for making this journey so memorable. Those late-night study sessions, endless team calls, and shared struggles have forged bonds that will be cherished forever. Your motivation and shared passion for data science have made this journey not only bearable but also incredibly rewarding. I am truly fortunate to have had you all by my side.

DEDICATION

This thesis is dedicated to everyone who believed in the potential of this research and contributed in their unique ways to its success.

TABLE OF CONTENTS

METADATA ii

DECLARATION iii

ABSTRACT iv

ACKNOWLEDGEMENTS v

DEDICATION vi

TABLE OF CONTENTS vii

LIST OF FIGURES xi

LIST OF TABLES xiii

LIST OF ACRONYMS AND ABBREVIATIONS xiv

LIST OF TERMS xvi

CHAPTER ONE: INTRODUCTION 1

1.1 Background 1

1.2 Problem definition 5

1.3 Research aim and objective 6

1.3.1 Main objective 6

1.3.2 Main research question 6

1.4 Significance of the research 7

1.5 Scope of the study 8

1.6 Organisation of the dissertation 8

CHAPTER TWO: LITERATURE REVIEW 10

2.1 Mineral processing 10

2.1.1 Stages of mineral processing 10

2.1.2 Significance of comminution 12

2.1.3 Efficiency and energy consumption in comminution 12
2.2 Commonly used machine learning methods for ore prediction 13
2.2.1 Multiple Linear Regression 14
2.2.2 Support vector regression 15
2.2.3 Random forest regression 17
2.2.4 Extreme Gradient Boost 18
2.3 ML in the mining industry 20
2.3.1 Digitalisation of the mining industry 20
2.3.2 Machine learning on ore quality estimation 23
2.4 Studies most related to this dissertation 24
2.4.1 Advances and challenges in machine learning for mineral processing 26
CHAPTER THREE: METHODOLOGY 28
3.1 Review: Research aim, objective and questions 28
3.2 Methodological approach 29
3.2.1 Research philosophy 29
3.2.2 Research approach 30
3.2.3 Research strategy 30
3.2.4 Methodological choice 31
3.2.5 Research time horizon 31
3.2.6 Techniques and procedures 32
3.3 Model implementation and performance evaluation 39
3.4 Simulation and comparative analysis 42
3.4.1 Traditional assay approach 42
3.4.2 Machine learning model estimation approach 42
3.5 Ethics considerations 42

3.6 Chapter summary 43
CHAPTER FOUR: RESULTS 45
4.1 Experimental results 45
4.1.1 Overview of data structure 45
4.1.2 Identifying and addressing missing data 45
4.1.3 Correlation insight 50
4.1.4 Evaluating the impact of feature scaling 51
4.1.5 Feature importance analysis 52
4.1.6 Performance evaluation 53
4.1.7 Traditional method vs machine learning approach analysis 60
4.2 Chapter summary 61
CHAPTER FIVE: DISCUSSION OF FINDINGS 62
5.1 Assumption 62
5.2 Identifying and engineering features that impact silica content 63
5.3 Evaluating Machine Learning (ML) Algorithms for Silica Content Estimation 63
5.3.1 Analysis of the models' performance 64
5.3.2 Estimating silica content based on the high-performing model 66
5.4 Accessing how silica content estimation can improve the mineral processing efficiency 67
5.5 Limitations and future work 68
CHAPTER SIX: CONCLUSION 70
6.1 Study's contribution 71
6.2 Concluding remarks 71
REFERENCES 73
APPENDIX A: ETHICAL CLEARANCE CERTIFICATE 79

APPENDICE B: MODEL'S IMPLEMENTATION 80

LIST OF FIGURES

Figure 1.1: Crusher circuit (Source: author). 2
Figure 1.2: Liberation of valuable minerals by size reduction (Source: (Michaux, 2020)) 3
Figure 1.3: Map showing the location of the Rosh Pinah mine. (Source: (Kay, 2017)) 3
Figure 1.4: Microquartzite drill core. (Source: (Jensen et al., 2018)) 4
Figure 2.1: Mineral processing flow diagram. Source: adapted from (Haldar, 2018) 11
Figure 2.2: Multiple Linear Regression. (Source (Thaddeussegura, 2020)) 15
Figure 2.3: Univariate linear SVR (with zero tolerance for error). (Source: (Rasifaghihi, 2023)) 17
Figure 2.4: Random Forest Regression structure. (Source: (Wang et al., 2021)) 18
Figure 2.5: XGBoost structure. (Source: (Wang et al., 2021)) 19
Figure 2.6: Industry revolution transformation 1.0 – 5.0. (Source: (Siddell, 2023)) 21
Figure 2.7: Supervised machine learning for predictive tasks. Source: (Nishadini, 2019) 22
Figure 3.1: Chapter 3 outline map. (Source: author) 28
Figure 3.2: Research Onion framework. (Source: adapted from (Saunders et al., 2019)) 29
Figure 3.3: Research approach. (Source: author) 31
Figure 4.1: Dataset initial concise summary. (Source: author) 45
Figure 4.2: Target variable time series line graph. (Source: author) 47
Figure 4.3: Imputation methods for NaN values. (Source: author) 49
Figure 4.4: Cleaned dataset concise summary. (Source: author) 49
Figure 4.5: Correlation Matrix. (Source: author) 50
Figure 4.6: Features importance. (source: author). 52
Figure 4.7: Evaluation of MLR Model Using `df_all_features` Dataset. (Source: author) 54
Figure 4.8: SVR epsilon and C value grid search. (Source: author) 55

Figure 4.9: Evaluation of SVR Using `df_all_features` Dataset. (Source: author) 56
Figure 4.10: Random Forest tuning. (Source: author) 57
Figure 4.11: Evaluation of RFR Model Using `df_all_features` Dataset. (Source: author) 58
Figure 4.12: Evaluation of xgboost Model Using `df_all_features` Dataset. (Source: author) 59
Figure 4.13: Silica content estimation- traditional method. (source: author) 60
Figure 4.14: Silica content estimation - Machine learning approach. (source: author) 61

LIST OF TABLES

Table 3.1: Summary of the research question and evaluation criteria 44

Table 4.1: Feature scaling evaluation 51

Table 4.2: Feature importance scores. 53

Table 4.3: Multiple Linear Regression evaluation 53

Table 4.4: Kernel’s model confidence 54

Table 4.5: SVR model tuned results 56

Table 4.6: Random Forest model tuned results 57

Table 4.7: Extreme Boost model's evaluation results 59

Table 4.8: Model performance evaluation summary. 61

LIST OF ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
Ca	Calcium
Cu	Copper
df_data	Full Dataset (with all features)
df1_data	Feature Selected Dataset
Fe	Iron
GDP	Gross Domestic Product
HPGR	High Pressure Grinding Rolls
kW	Kilowatt
LASSO	Least Absolute Shrinkage and Selection Operator
MAE	Mean Absolute Error
Mg	Magnesium
ML	Machine Learning
MLR	Multiple Linear Regression
Mn	Manganese
MQZ	Micro quartzite
MSE	Mean Squared Error
Pb	Lead
R^2	R squared (Coefficient of Determination)
RFR	Random Forest Regression
RMSE	Root Mean Squared Error

S	Sulfur
SAG	Semi-Autogenous Grinding
SiO_2	Silica Dioxide
SPF	Sampling Point Features
SVR	Support Vector Regression
XGBoost	Extreme Gradient Boosting
Zn	Zinc

LIST OF TERMS

Ore: Naturally occurring rock that consists of the mineral of interest at an economically feasible concentration.

Ore blending: Is a process of mixing different ores to obtain a constant ore concentration that is suitable for upstream processes.

Waste/gangue: Is the unwanted or impurity associated with ore.

Assay: Is a laboratory process of determining the overall quantity and quality of different minerals present in the ore.

Comminution: Is a process of reducing the size of the rock through crushing, milling and/or grinding.

Lithology: Is the type of rock based on its physical characteristics, induced by its formation.

Mineralogy: Is the chemical composition of rock.

Flotation: Is a mineral processing method of separating ore through its hydrophobic and hydrophilic properties.

CHAPTER ONE: INTRODUCTION

1.1 Background

The mining sector serves as a cornerstone of numerous economies, serving a critical function in the resource supply chain of various industries. To meet the growing demand for resources needed to support global economic growth and ensure a sustainable future, the mining industry must address unprecedented challenges (An et al., 2023). In Namibia, the sector has been and remains the backbone of the economy as reflected by its average annual economic growth, contribution to Gross Domestic Product (GDP), job creation, income generation, and a key source of government fiscal receipts and foreign exchange earnings, among others (Nambinga & Mubita, 2021). According to the Namibia Chamber of Mines (CoM), the total taxes paid amount to N$ 4.401B, with a total 16 147 people directly employed across the country in 2022. The industry's economic significance cannot be overstated, yet it has long been a source of environmental concern. Traditional mining operations have historically struggled to strike an equitable balance among economic viability, depletion of high-grade ore and high cost of operation continue to occur. Among the numerous factors that affect mineral processing efficiency, the silica content in ores stands out as a significant challenge to comminution.

Comminution in mineral processing refers to the process of crushing and grinding ore to liberate valuable minerals, making them easier to extract through subsequent steps like separation and concentration (Kapadia, Comminution in Mineral Processing, 2018). The efficiency of comminution is paramount, as it directly influences the efficiency of subsequent processes during mineral processing. Comminution breaks down the ore to a specific size, enabling the liberation of valuable minerals from the gangue matrix and meeting target size, liberation, and recovery criteria (González, 2020). This process involves a series of crushing and grinding operations designed to break down the ore into smaller, more manageable pieces, as illustrated in Figure 0.1 below. The process begins underground, where ore from the silo are loaded onto a grizzly screen. Oversized materials are broken by a hydraulic hammer, while the undersized materials gravitate to a feeder that feeds the Osbourne jaw crusher. This then reduces the ore size to 150 mm, and the crushed ore is conveyed to the surface. They are again screened, with oversized ore (more than

28mm) crushed by secondary cone crushers, while undersized ore is stockpiled. The stockpile material is further screened, where the oversized ore is crushed by tertiary gyratory crushers before being stockpiled. The undersized material then serves as the feed for the advanced processing plant phases.

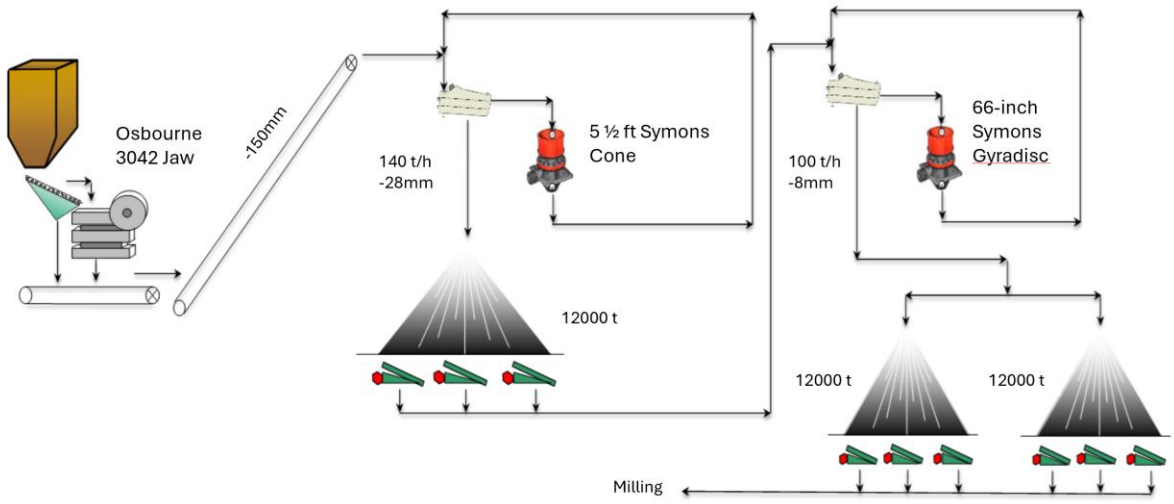

Figure 0.1: Crusher circuit. Source: Author

Among the many factors that affect the efficiency of ore liberation, silica content in ores presents a notable challenge, particularly during comminution. Due to its hardness, silica requires more energy to be broken down compared to softer minerals, which leads to increased wear on comminution equipment and also prolongs the crushing period. Silica particles stand around 6 on the Mohs scale, making it more abrasive, leading to higher energy consumption during the comminution phase. This increased energy requirement not only raises operational costs but also reduces the overall efficiency of the mineral recovery process. The fine silica particles generated during grinding can cause additional challenges in downstream processing stages such as flotation. In flotation, the separation of minerals based on their surface properties can be influenced by the presence of fine silica, which can interfere with the effectiveness of flotation reagents and reduce the recovery rates of target minerals. The liberation of minerals and decrease of particle size, comminution increases the surface area of the solid in question and helps in the pulping of the material (Mkurazhizha, 2018). Figure 0.2 illustrates the concept of ore liberation, which involves the breaking down of ore in smaller particles. This increases the exposed surface area of the valuable mineral as it is freed from the surrounding gangue, making it easier for the targeted

mineral to be accessible for upstream processing phases like flotation. The performance of a mineral processing plant is greatly influenced by the nature of the ore being processed and its liberation characteristics (Anzoom et al., 2024). This will determine the comminution period and also the throughput size, for optimal liberation.

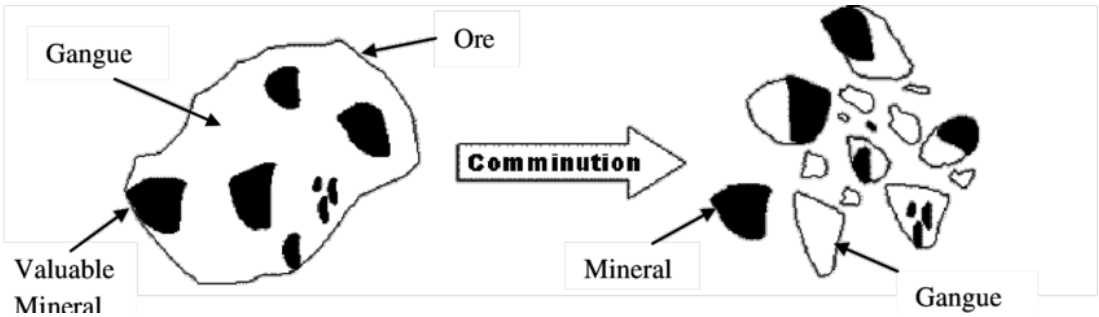

Figure 0.2: Liberation of valuable minerals by size reduction. Source: Michaux (2020)

Rosh Pinah Zinc Mine in Namibia is among the mines driving the silicified ore management plan through ore blending techniques. Established in 1969, the mine has maintained continuous operations for over five decades, currently yielding an impressive daily output of around 2,000 tons of zinc and lead concentrate. It is among the limited number of underground mines still in active production within Namibia, situated on the outskirts of the town of Rosh Pinah in the ǁKaras Region in the southern part of the country, as shown in Figure 0.3 below.

Figure 0.3: Map showing the location of the Rosh Pinah mine. Source: Kay (2017)

The geotechnical domains at Rosh Pinah zinc mine are mainly dominated by Microquartzite, especially the high-grade ore. Figure 0.4 illustrates a drill core containing Microquartzite, which

is always described as a highly silicified, glassy, black, or dark grey rock composed predominantly of crystalline silica having a particle size < 20 microns and does not scratch with steel (Jensen et al., 2018). This is the main source of silica in the mine's ore, among different geotechnical domains found at the mine.

Figure 0.4: Microquartzite drill core. Source: Jensen et al. (2018))

Silica dioxide (SiO2), also known as silica, is an abundant mineral on Earth, which is an associate of sand, quartz and most rock types. Additionally, it mostly plays the role of gangue mineral in many mining operations, except for sand mining, which is the main product of interest. Its presence alongside the valuable mineral poses huge challenges during the liberation process in the following ways:

- *Processing difficulty*: Since silica does not hold economic value, it needs to be separated from the desired mineral during processing. This can be an energy draining process, which complicates all downstream processes, depending on the specific minerals involved and the liberation methods used.
- *Lower ore grade*: The occurrence of silica dilutes the concentration of the valuable mineral in the ore. This means that more ore needs to be mined to increase the grades, which will require additional processing to obtain the same amount of the desired mineral. This affects the overall efficiency and profitability of the mine.
- *Technical challenges*: Depending on the properties of the silica and the desired mineral, it can hinder specific processing techniques. For instance, silica can interfere with flotation, a common separation method during mineral processing, and/or require finer grinding, which increases energy consumption.

Mining operations are significantly impacted by the silica content present in ores, driving the industry to develop new advanced ore blending and extraction techniques through innovation. However, complex geological formations pose a challenge to the accurate quantification of ore associations, thereby hindering precise blending. Ore blending is a widely used method which involves mixing ores from different parts of a mine to ensure a steady and predictable feed to the mineral processing plant while accounting for variability in grade and mineral composition (Zhang et al., 2023). To address this issue, this study proposes the implementation of a machine-learning model to estimate silica content, thereby facilitating improved ore blending controls at Rosh Pinah Zinc Mine.

1.2 Problem definition

In the mining industry, it is standard practice to utilise periodic laboratory tests to evaluate ore quality and optimise mineral extraction processes. However, accurately estimating silica content within ore samples presents a significant challenge. Traditional laboratory analysis, conducted every six hours, results in delays, high processing costs, and increased chances of human error. These issues are further compounded by fluctuations in ore feed and shifting output demands, necessitating frequent adjustments in processing techniques. The presence of silica in ore can cause processing difficulties, affecting crushing, grinding, and froth flotation processes, and reducing overall efficiency. The financial losses incurred from valuable ore ending up in tailings underscore the need for a solution.

Laboratory analysis is an essential process for determining mineral composition, which provides insightful information for metallurgical studies, optimising mineral processing processes and aiding in the overall long-term resource management and mine planning (Joel & Maliwatu, 2024). The implementation of a near-real-time estimation model for silica estimation and improved ore blending estimations could have a significant impact on mineral processing efficiency. This approach will enhance ore quality, regardless of the mine's complex ore composition, encompassing impurities, and fluctuating ore grades (Jensen et al., 2018). Instant and accurate silica estimations would enable real-time decision-making, leading to cost reduction, waste minimisation, and increased productivity through optimised ore blending. This proactive approach has the potential to address the operational challenges posed by silica content and significantly enhance the mine's overall performance and profitability.

1.3 Research aim and objective

The aim of this research is to enhance the effectiveness of mineral processing processes by advancing silica content estimation techniques using machine learning algorithms. This is governed by the following objectives.

1.3.1 Main objective

To apply a Machine Learning (ML) algorithm to estimate the silica content in the ore feed before entering the advanced stages in the processing plant. The main objective of this research is broken down into the following sub-objectives:

1.3.1(a) Sub-objectives

I. To identify and engineer relevant features that impact silica content, based on the mineral composition.

II. To evaluate various machine learning algorithms, namely Multiple Linear Regression, Random Forest Regression, Support Vector Regression and XGBoost, to determine the effective model for silica content estimation.

III. To assess how silica content estimation in the ore feed can enhance the efficiency of advanced mineral processing stages.

1.3.2 Main research question

How can Machine Learning (ML) algorithms effectively estimate the silica content in the ore feed? The main research question is broken down into the following sub-questions:

1.3.2(a) Research questions

I. Which mineral composition features most significantly influence the estimation of silica content in ore feed?

Based on the Pearson correlation analysis, the most impactful features are minerals such as Sulphur (-0.54), Zinc (-0.49) and Manganese (-0.49), which show a strong negative correlation with Silica. Including these elements as predictors could improve the model's accuracy, as their levels might inversely affect SiO2 concentrations.

II. ***Which machine learning algorithm, among the commonly used machine learning algorithms, namely Multiple Linear Regression, Random Forest Regression, Support Vector Regression, and XGBoost, offers the highest accuracy in estimating silica content in the ore feed?***

The Random Forest Regression consistently outperformed the other three algorithms, which indicates its ability to handle complex, non-linear relationships in the data effectively. This indicated performance superiority, achieving lower RMSE, MAE and higher R^2 score. The Random Forest Regression model achieved an estimation accuracy of 85% with the highest R2 score compared to the other models. Its error rates are also the lowest, with RMSE at 0.3821 and MAE at 0.2848.

III. ***How can silica content estimation in ore feed enhance mineral processing efficiency?***

Accurate estimation of silica content plays a significant role in optimising mineral processing efficiency due to its direct impact on ore blending, crushing, grinding, flotation and other major plant performance. This ensures a stable feed entering the plant and spikes minimisation, which reduces the wearing of equipment, and increases grinding time and energy consumption. Additionally, this provides continuous insights into silica patterns which outperform the traditional six-hour laboratory assays. This machine learning approach allows proactive actions through the timely responses to silica levels fluctuations, ore blending optimisation, adjustments on milling rates and grinding media to address ore hardness challenges. Overall, accurate silica estimation promotes greater profitability by improving efficiency and minimising operational instability.

1.4 Significance of the research

Comminution is the starting point and serves as an important stage in ore beneficiation. Given the complexities and uniqueness of ore, it is crucial to maintain consistent silica content in the ore feed entering the plant circuit. Variability in silica content can lead to major challenges, including the need for frequent adjustments across all stages of the processing plant, a drop in Kilowatt (kW), and excess use of reagents, which in turn impact operational efficiency and cost-effectiveness.

To address this, it is essential to implement a data-driven approach that ensures accurate silica content measurement in the feed. Historically, ore lab assays, which are the tests of the ore to determine its content, have been conducted at intervals of six (6) hours, but this interval has recently been extended to twelve (12) hours, thus increasing the potential instability of the silica content and associated operational inefficiencies.

This research project assesses the implementation of a machine learning (ML) model to estimate silica content in the feed, aiming to maintain a fixed feed ore composition through an informed ore blend adjustment. This study seeks to enhance the accuracy and timeliness of silica content estimations, ultimately reducing the need spikes and drops of grades or kW and improving overall plant performance.

1.5 Scope of the study

The scope of this study assesses the impact of machine learning (ML) on the mine's operation. This focuses on the silica content estimation in the ore feed, which is geographically centred on the ore found at Rosh Pinah Zinc Mine. This not only provides a significant insight for the mine but extends to the mining industry in general in terms of technological enhancement and long-term sustainability. Additionally, this study aligns with the increasing trends in the sector, where the ML techniques are increasingly used to enhance different operational efficiencies, improve sustainability and cost reduction. The results of this study are aimed at informing not only the Rosh Pinah Zinc mine's operational strategies but also motivating the wider discourse on the enhancement of various mining operations globally.

1.6 Organisation of the dissertation

The rest of the dissertation is organised as follows:

CHAPTER TWO: **LITERATURE REVIEW**

> The literature review chapter explores the fundamental description of mineral processing, focusing on the comminution phase, and the application of ML in the mining industry. Additionally, it also discusses the related studies pertinent to this research

CHAPTER THREE: **METHODOLOGY**

This chapter discusses the methodological approach employed on this study and implementations of the models to achieve the objectives on this study.

CHAPTER FOUR: **RESULTS**

The study's results chapter delineates the applied methodological framework outcomes, reviewing the selected model's performance.

CHAPTER FIVE: **DISCUSSION**

The results of this study are further discussed in this chapter, emphasising their alignment with the research objectives and questions. Additionally, it also focuses on the limitations and recommendations for future research.

CHAPTER SIX: **CONCLUSION**

This chapter closes off the research with an overall summary and contribution of this study to the mining industry.

CHAPTER TWO: LITERATURE REVIEW

This chapter provides an overview of the significant concepts essential to understanding this study, structured into three main sections. The first section introduces the fundamental principles of mineral processing, which outlines the stages involved in liberating valuable minerals from ores and discusses the importance of comminution in optimising recovery and the reduction of costs. The second section discusses the application of Machine Learning (ML) in the mining industry, focusing on how these technologies have transformed mineral processing through improved predictive accuracy and operational efficiency. The last section reviews the related studies that focused on the problems addressed in this study (silica content estimation) and the method used to solve them. This chapter sets the foundation for understanding the study's context and the technological advancements that underpin the research.

2.1 Mineral processing

Mineral processing is also known as ore dressing, which involves the process of liberating economically valuable minerals from their ores. As provided in the introduction chapter, comminution is a significant stage during mineral processing, since well-sized ore particles ensure a better separation and extraction of ore from the waste material, which allows a maximum recovery rate of valuable minerals. Additionally, this process enables the extraction of minerals from the rock that have a good economic value in the market (Kapadia, 2018). This section explores mineral processing stages, recent advancements and emerging trends.

2.1.1 Stages of mineral processing

As illustrated in Figure 0.1, the mineral processing flow includes several important stages to extract minerals of interest from the ore and ensure that they are ready for subsequent processes of refinement. These stages are as follows:

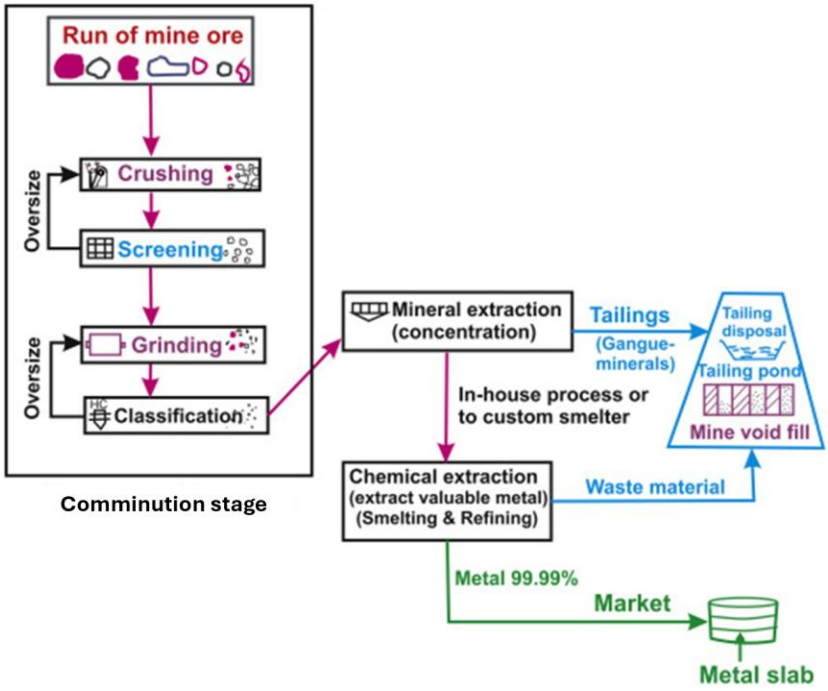

Figure 0.1: Mineral processing flow diagram. Source: Adapted from Haldar (2018)

Crushing: It is the initial stage of mineral processing, which involves large rocks being broken down into smaller, more manageable pieces. This stage is crucial for the efficiency of subsequent grinding and the overall separation processes (Haldar, 2018).

Grinding: This process involves a further size reduction, following the undersize from crushing. It grinds the ore into finer particles, which helps to expose a larger surface area of the valuable mineral from the surrounding gauge material to enhance the efficiency of this stage. This is accomplished through various types of mills, which employs an amalgamation of impact, attrition, and shear mechanisms to reduce ore particles to micron sizes, adopting various grinding media such as ore pebbles in AG (autogenous) mills, a mix of steel balls and pebbles in SAG (semi-autogenous) mills, steel balls in ball mills, steel rods in rod mills, or ceramic media in stirred mills (Whitworth et al., 2022).

Separation: This process differs depending on the type of ore being processed, which includes gravity separation, magnetic separation, and/or flotation, using the ore characteristics to liberate valuable minerals from gangue. These methods exploit ores' differences in physical or chemical properties to achieve effective separation. These

properties can be colour (optical sorting), density (gravity separation), magnetic or electric (magnetic and electrostatic separation), and physicochemical (flotation separation), which is mostly used in processing base metal ores (Gruner & Lorig, 2024).

Concentration: This is the last stage of mineral processing, which aims to increase the percentage of valuable minerals in the ore by removing excess unwanted material. This stage mostly includes thickening, filtering, and drying to prepare the material for further processing or sale (Gruner & Lorig, 2024)

2.1.2 Significance of comminution

Comminution is a crucial phase in mineral processing, as discussed in the introduction under the background subsection. It is the most energy-consuming stage across the mining stages. This process of crushing and grinding ore is the most energy-intensive step in mining, consuming about 53% of the mine's energy and accounting for at least 10% of production costs (MINING.COM, 2019). Therefore, the efficiency of the comminution stage directly impacts the overall performance of the mine, influencing recovery rates and overall operational costs.

- **Particle Liberation:** Comminution increases the exposure of valuable minerals which was locked up in the RoM feed to the plant, which enhances the efficiency of the separation processes (Wikedzi & Leißner, 2021).
- **Impact on Downstream Processes:** The size and uniformity of particles produced during comminution affect the efficiency of downstream processes like flotation, where finer particles can lead to better separation outcomes (Guven, 2022).

2.1.3 Efficiency and energy consumption in comminution

The effectiveness of comminution processes remains an important stage in mineral processing, which has a major impact not only on the mineral extraction efficiency but also on the overall operational costs and environmental sustainability.

- **Energy-intensive nature**

  The comminution process, being one of the largest energy consumers in the mining and mineral processing industries, is particularly costly due to its low energy

efficiency, as it requires substantial mechanical input to generate new surface area (Chimwani, 2021). The hardness of the ore, particularly in cases of imbalanced silica content due to a late adjustment of ore blending ratio, directly influences the energy required for comminution due to a prolonged grinding/milling time.

- **Economic and environmental implications:**

  Efficient comminution reduces operational costs and contributes to more sustainability within the mining industry. The mining energy value chain is significantly impacted by comminution, with energy-efficient system designs and technologies offering substantial reductions in overall energy consumption (Klein et al., 2018).

- **Technological advancements**

  Recent advancements in comminution technology, including the use of high-pressure grinding rolls (HPGR) as an energy-efficient alternative to SAG mills and vertical stirred mills as a substitute of the traditional ball mills, have significantly reduced the total energy consumption (Vijfeijken, 2023). These innovations are essential for reducing the environmental impact of mining operations, as a reduction in consumption translates to lower greenhouse gas emissions, hence a smaller carbon footprint. Additionally, the introduction of Artificial Intelligence and Machine Learning has contributed to the improvement of various processes within the mining industry.

2.2 Commonly used machine learning methods for ore prediction

Machine learning is widely used in concentration estimation to deepen the understanding of large laboratory data. ML makes use of algorithms which allow computers to learn patterns from the provided data. This can be extended to estimation making, whereby its performance is improved as the model is trained and tuned. ML is mainly categorised into supervised learning, unsupervised learning, and reinforcement learning, each is applied for different applications based on the type of data provided. ML techniques mostly used include Support Vector Regression (SVR), K-Nearest Neighbours (KNN), Feedforward Artificial Neural Networks (ANN), Random Forest

Regression (RFR), Gradient Boosting Regression (GBR) and Linear Regression (Huynh et al., 2022). Here are some of those techniques:

2.2.1 Multiple Linear Regression

Multiple linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables (Bozkus, 2022). It is a fundamental statistical technique that models the relationship between two or more independent variables and a dependent variable by fitting a linear equation to observed data. In mineral processing, this algorithm is commonly used to estimate the concentration of valuable minerals based on multiple input variables. Furthermore, this is mostly used due to simplicity and interpretability, making it a popular choice for initial model development, though it may not always capture the complexity of mineral processing data.

Multiple Linear Regression is one of the two concepts of linear regression that is used to build a linear equation for analysing the correlation between a dependent variable (also known as the target or outcome variable) and two or more independent variables (predictors or explanatory variables) (Sahu, 2023). In silica estimation, it can estimate mineral associates based on a single factor, such as the ore feed grades. Moreover, while the Multiple Linear Regression model performs well in linearly separable phenomena with faster computational capability, it struggles in non-linearly separable cases (Osei, 2019).

The estimations are calculated using the multiple linear regression algorithm, which is defined by equation (0.1).

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots\cdots\cdots + \beta_p x_p + \epsilon \tag{0.1}$$

This equation is used to fit a correlation between a numerical dependent variable denoted by 'Y' and a set of independent variables referred to as predictors '$x_1, x_2 \ldots\ldots x_p$'. The coefficient parameters denoted by '$\beta_0 \ldots\ldots\ldots \beta_1$' is estimated by the least squares method, which finds the *p*-dimensional plane that best cuts the plane closest as shown in Figure 0.2 below. During regression

modelling, the algorithm not only estimates the coefficients but also selects which independent variable should be included in the model.

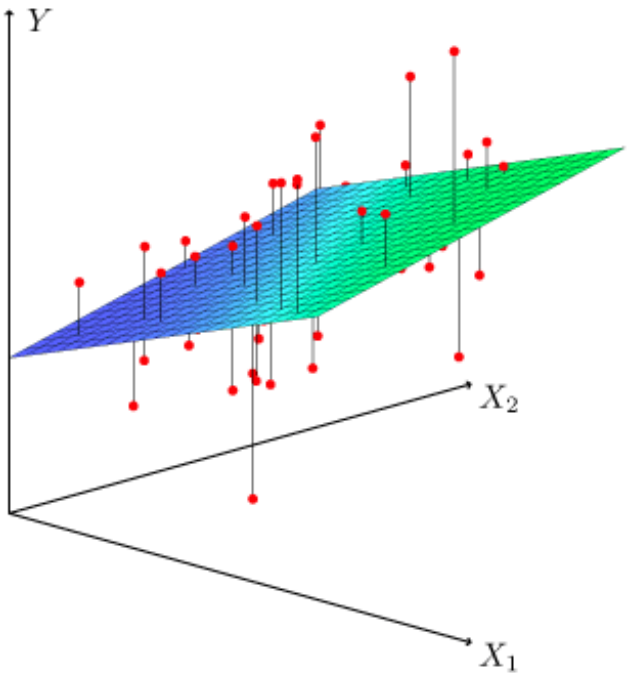

Figure 0.2: Multiple Linear Regression. Source: Thaddeussegura (2020)

2.2.2 Support vector regression

Support vector regression is a machine learning technique used to examine the relationship between predictor variables and a continuous dependent variable by learning the importance of each variable without relying on traditional model assumptions (Zhang & O'Donnell, 2020). It is a variant of Support Vector Machines (SVM) that is designed to predict continuous numeric values or for regression tasks (Verma, 2023). This regression supports linear and non-linear datasets, where it maps the input features into a higher-dimensional space and aligns a hyperplane that best represents the relationship between the input and output variables. In mineral processing, SVR can handle non-linearities in the data, which is most possible due to the uniqueness of ore, thus making it useful for estimating ore compositions in terms of relationship complexity between the input variables and the desired output.

SVMs are mostly used for classification, whereby their primary focus is obtaining the hyperplane separating the two classes fed into the system. However, SVR makes use of the error tolerance of the model to find the best fit of the data within that margin. In SVR, the goal is to obtain the best fit that accurately predicts the target variable while reducing complexity to avoid overfitting (Rasifaghihi, 2023). As shown in Figure 0.3, the function $f(x) = wx + b$ aims for a plane as flat as possible while maintaining a high deviation of the error tolerance denoted by ε in comparison to the actual values for all training data. To reduce the risks of overfitting, the function is kept as flat as possible, which indicates less sensitivity to changes made to the input data.

They are applied through a loop whereby each kernel is applied one by one in each iteration. Each kernel type is initialised by the SVR model and stored for each iteration. The model is then trained using the features and target train sets. The model's performance was then evaluated using the coefficient of determination of the estimation generated using different kernels. This allows the selection of the kernel which yields the best fit to the training data.

Hyperparameters: The SVR model's performance and flexibility are governed by several key hyperparameters:

- **Kernel Function:** The Radial Basis Function (RBF) kernel is chosen for its ability to handle non-linear relationships by transforming the feature space into a higher-dimensional space.
- **Regularisation Parameter (C):** The regularisation parameter controls the trade-off between achieving a low error on the training data and maintaining a smooth decision boundary. A specific value of C will be tuned during model optimisation.
- **Kernel Coefficient (Gamma):** This parameter defines the influence of a single training example. The value of gamma will be selected to optimise the model's performance.

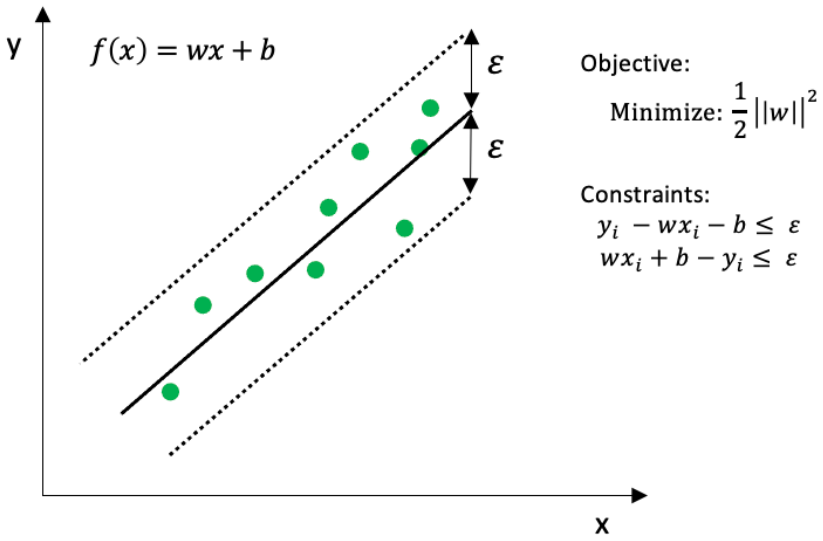

Figure 0.3: Univariate linear SVR (with zero tolerance for error). Source: Rasifaghihi (2023)

Loss Function: The epsilon-insensitive loss function is employed to create a margin of tolerance where no penalty is given to errors within a specified epsilon range. This loss function emphasises errors outside the epsilon margin, allowing the model to focus on significant deviations from the estimated values.

2.2.3 Random forest regression

Random forest regression is a supervised learning algorithm that employs ensemble learning, a technique combining estimations from multiple machine learning algorithms to achieve more accurate regression estimations (Chaya, 2020). It generates multiple decision trees and then combines them to produce a more accurate and stable estimation. It enables insight to be obtained into how important each variable can be in the context of making an estimation about the ore quality, given the substantial size of such datasets and the numerous numbers of input variables. This capability of handling non-linear relationships and complex interactions between variables makes RF particularly useful for application in mineral processing.

This method trains multiple classifying decision trees on various subsamples, benefiting from an averaging mechanism to enhance predictive accuracy and mitigate overfitting, with training samples randomly selected with replacement and maintaining the size of each new training set

equivalent to the original dataset (Dogan et al., 2019). As shown in Figure 0.4, each sub-tree model in a random forest performs random sampling with replacement from the training data, averages results from all sub-models, runs in parallel without any dependency, and differs in tree construction by using distinct subsets of the data. The process of averaging out multiple machine learning algorithms' estimations improves the estimation accuracy.

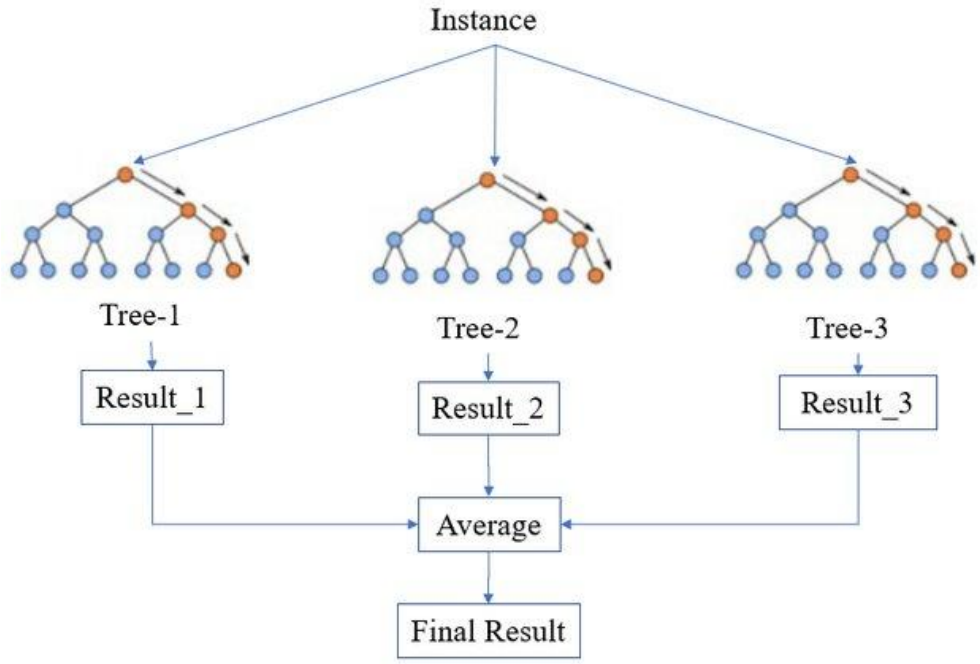

Figure 0.4: Random Forest Regression structure. Source: Wang et al. (2021)

2.2.4 Extreme Gradient Boost

Gradient Boosting Regressor is an ensemble model that is an iterative collection of sequentially arranged tree models, so that the next model learns from the error of the former model (Otchere et al., 2022). This method is commonly used in capturing non-linear relationships and interactions between complex variables. In mineral processing, Gradient Boosting Regression can be used to estimate ore quality with high accuracy, especially when the data contains complex patterns and relationships, which makes it extremely hard to be captured by simpler models.

Unlike traditional boosting algorithms, XGBoost enhances model performance by incorporating advanced features such as regularisation, sparsity awareness, and parallel processing, thus making

it highly effective for large-scale data and complex patterns. XGBoost operates by sequentially adding new models to correct the errors made by the previous ones. At each iteration, a new tree is trained to minimise a loss function, and the overall model is updated by adding the estimations of the new tree to the existing ensemble. This iterative process continues until the model reaches a specified number of iterations or the performance ceases to improve significantly.

The algorithm minimises the loss function by adjusting the parameters of each tree, allowing the model to fit the training data more accurately while controlling overfitting through regularisation techniques like L1 (Lasso) and L2 (Ridge) penalties. To ensure that the model generalises well to unseen data, XGBoost incorporates regularisation and cross-validation. Regularisation penalises the complexity of the trees, preventing them from fitting too closely to the training data, which could lead to overfitting. Figure 0.5 shows the structure of XGBoost, whereby Tree-1 feeds Tree-2, which then feeds Tree-3. This is to reduce the residual.

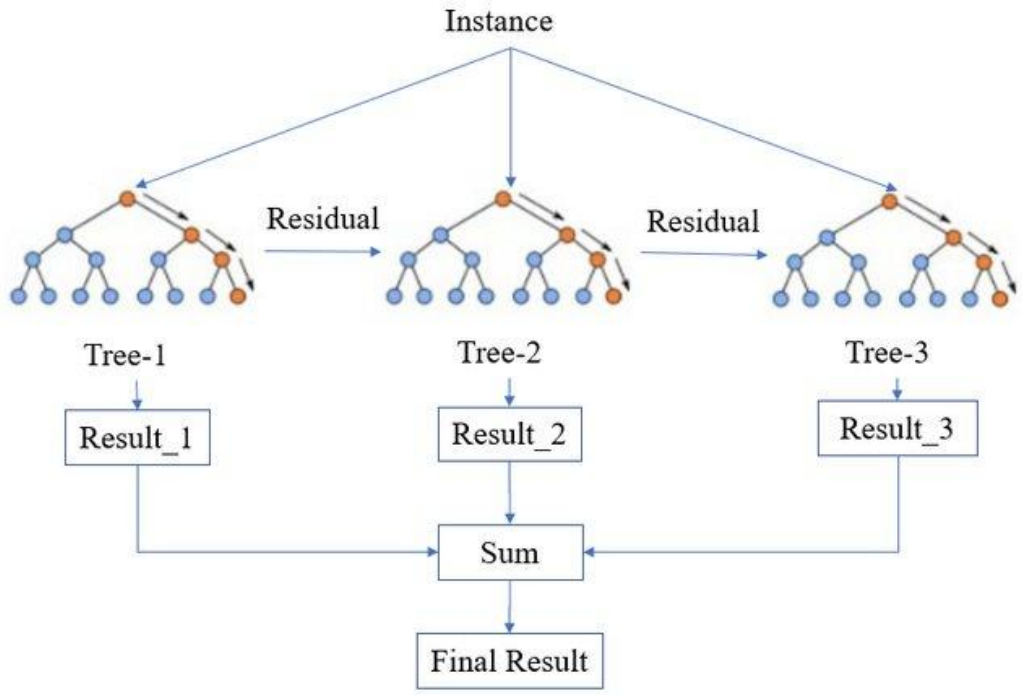

Figure 0.5: XGBoost structure. (Source: Wang et al. (2021)

The performance and flexibility of the XGBoost model are governed by several key hyperparameters:

- **Learning Rate:** Controls the step size at each iteration while moving towards a minimum of the loss function. A smaller learning rate makes the model more robust to overfitting, but it requires more iterations to converge.
- **Number of Trees (n_estimators):** Specifies the number of trees to be added in the model. More trees can improve performance, but might lead to overfitting if not properly regularised.
- **Maximum Depth (max_depth):** Determines the depth of each tree. Deeper trees can capture more complex patterns but may also lead to overfitting if not constrained.
- **Subsample:** Represents the fraction of samples used to train each tree. By using a subset of the data, this parameter introduces randomness and helps in preventing overfitting.
- **Column Subsample (colsample_bytree):** Specifies the fraction of features used for building each tree, thus introducing additional randomness and reducing correlation among trees.

Loss function: XGBoost employs a regularised objective function that combines the traditional loss function which is the mean squared error (MSE) with regularization terms to penalise model complexity. This approach balances the model's ability to fit the training data while maintaining generalisability to new data.

2.3 ML in the mining industry

2.3.1 Digitalisation of the mining industry

Mining, a long-standing activity with a strong traditional profile, has experienced the greatest changes in its history in the past 100 years, particularly in efficiency and safety (Rascón, 2024). It has witnessed a significant transformation, with the integration of advanced technologies such as AI and ML into various aspects of mining operations (Mapp, 2023). The mining industry has seen an increase in technological advancements associated with automation, drones, Internet of Things (IoT), sensors, AI, and ML in recent years. This has transformed the industry through four (4) major industrial revolutions as indicated in Figure 0.6.

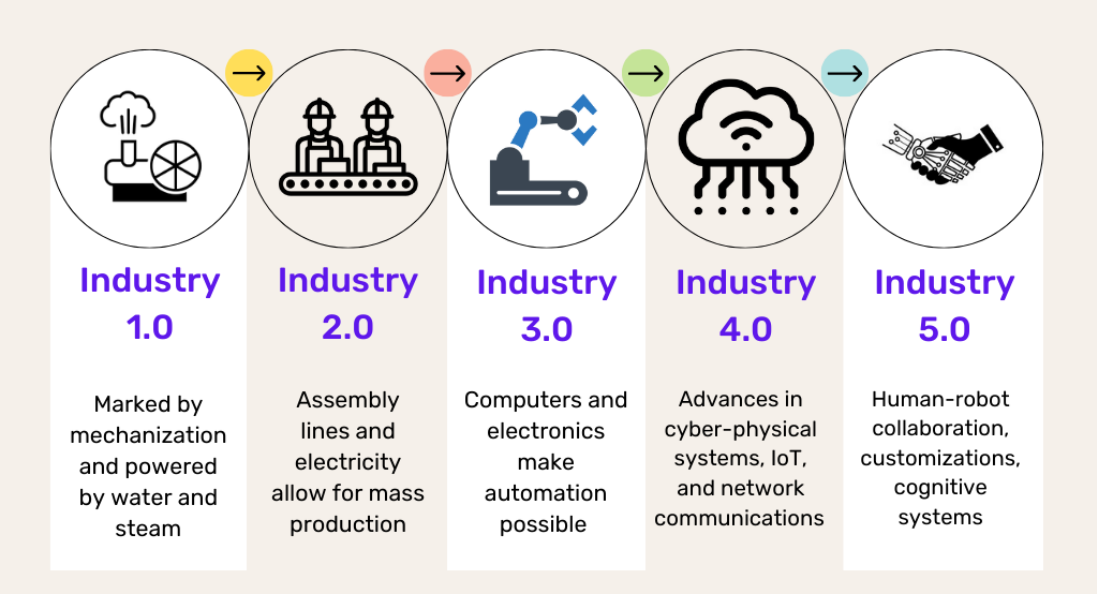

Figure 0.6: Industry revolution transformation 1.0 – 5.0. Source: Siddell (2023)

The first revolution started with the mechanisation of hydraulic drilling and the use of the steam engine. The second innovation was the transition from the steam engine to the use of electricity, which was a pool of mass production. This was followed by the third revolution, that is, the introduction of computers and automation (mining robots) within the industry. Mining 4.0 brought a cutting edge to the industry, which enhances the transformation of the industry through the integration of AI and ML into mining operations (Mapp, 2023). Currently, we are witnessing the 5IR referred to as Mining 5.0. Industry 5.0 has advanced from Industry 4.0, which brought the integration of human capability with cutting-edge technologies like robotics. According to Demir et al. (2019), Industry 5.0 emphasises not only the efficiency and automation but focuses on the balance between human creativity and machine precision, with the main vision including the human-robot collaboration and leaning towards bioeconomy.

According to a study conducted by Skenderas and Politi (2023), "Artificial Intelligence (AI) contributes to the mining industry's (i) exploration; the process may become more efficient by identifying and analysing potential sites and processes; and (ii) predictive site maintenance, where AI can run numerous different scenarios highlighting potential issues and ensuring the safety and stability on the site" (p. 3). It represents a shift from traditional practices to innovative, technology-driven solutions. Mining has evolved through distinct eras, from mechanisation to automation, and now it stands on the cusp of Mining 5.0, with a focus on enhanced AI, robotics, and green

investments (Zhironkin & Ezdina, 2023). The application of AI has led to more efficient resource extraction, real-time data access, predictive analysis, improved safety measures, and automation, paving the way for a promising future of mining innovation (Garbini, 2023).

These industrial revolutions have led to the integration of advanced technology, such as the application of machine learning to the mining sector, from estimating machinery maintenance to real-time monitoring of systems. ML is commonly referred to as a subset of AI and encompasses techniques that discern patterns through the process of learning from organised historical data (Johnston et al., 2019). Kanade (2022) also defines Machine learning as "a discipline of artificial intelligence (AI) that provides machines with the ability to automatically learn from data and past experiences while identifying patterns to make estimations with minimal human intervention". ML uses a trained dataset to develop a model through analysing the patterns within the dataset. Once trained, the ML model analyses patterns within a newly introduced data and estimates the correspondence as illustrated in Figure 0.7. ML algorithms use computation methods to learn directly from data instead of relying on any predetermined equation that may serve as a model (Kanade, 2022).

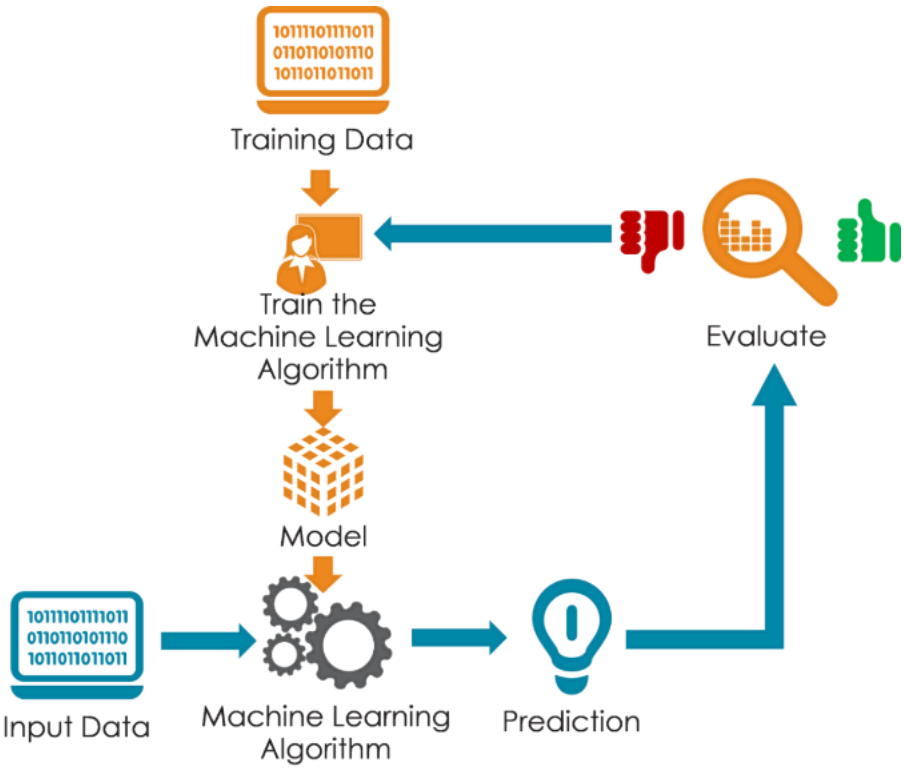

Figure 0.7: Supervised machine learning for predictive tasks. Source: Nishadini (2019)

The mining sector heavily relies on data from various sources, such as exploration activities, drill results, and laboratory assay results, to identify areas with promising mineral deposits (Sahota, 2023). All this data is available for models learning, and promotes improvements in the mining industry through predictive, descriptive, prescriptive, generative, hybrid and anomaly detection algorithms.

2.3.2 Machine learning on ore quality estimation

Mineral processing is a critical phase in mining operations, where ore is extracted from the host rock and then subjected to a range of physical and chemical procedures to liberate valuable minerals from the associated gangue materials. Traditionally, mineral processing has relied on empirical models and operator experience to optimise operations. However, the integration of ML techniques offers the potential to address the complexity of ore deposits by providing data-driven insights for process optimisation (Ali, 2022). Researchers have explored the development of predictive models that can accurately estimate mineral composition within ore samples. Notably, studies by Osei (2023) and Montanares et al. (2021) have demonstrated the feasibility of ML models to enhance the estimation and control of silica concentrate levels in mineral processing. By utilising predictive modelling, these models consider multiple input variables to estimate silica concentration, enabling proactive adjustments in the processing workflow to effectively control and prevent high silica levels from entering the advanced upstream processes in the processing plant.

In recent years, the predictive modelling of ore quality, particularly silica content, has gained significant attention. Studies by Osei (2019) and Montanares et al. (2021) have shown the feasibility of using ML to enhance the estimation and control of silica concentrate levels in mineral processing. ML includes a variety of algorithms, each with unique qualities and applications in ore associate estimation. Predictive models fall under the supervised learning, whereby it is trained on labelled data to generate an estimation on test data. Depending on the designed output, different models are used, whether continuous or categorical:

- **Linear Regression:** Predicts continuous outcomes based on input features.
- **Logistic Regression:** Predicts binary outcomes (for example, yes/no or male/female).
- **Decision Trees:** Generate decisions by partitioning data into branches based on feature values.

- **Support Vector Machines:** Find the hyperplane that separates different classes.

Among the widely used techniques, multiple linear regression, random forest, and Artificial Neural Networks (ANN) are explored on their capability to predict ore content and specifically forecast silica concentrate levels in iron ore froth flotation processing plants (Montanares et al., 2021). Regression techniques are widely used in building predictive models to establish relationships between variables and making estimations based on observed data, most suitable for continuous data (Dawood, 2023).

2.4 Studies most related to this dissertation

The introduction of machine learning in the mining industry has become increasingly pivotal, significantly impacting the predictive accuracy and operational efficiency of various processes, including mineral processing activities, for instance, silica content estimation. There are several case studies demonstrating the transformative impact of machine learning in these areas, illustrating its potential to revolutionize the section.

One prominent study by Osei (2019) investigated the feasibility of deploying machine learning algorithms to predict silica content in ore feeds. The study sought to reduce the dead time caused by the conventional laboratory testing to assay the silica content entering the plant circuit. It leveraged extensive historical data from mineral processing activities to train various predictive models that were later deployed in real-time settings and evaluated their estimation performance. The researcher sought key variables that influence silica concentration by applying two feature selection methods: Random Forest and backwards elimination. Silica estimation was achieved using a combination of machine learning and deep learning, through Multiple Linear Regression, Random Forest, and Artificial Neural Network models. Notably, the study by Joel and Maliwatu (2024) also used those ML techniques to predict metal concentrations in copper deposits, which can substitute the traditional laboratory analysis, which can be time-consuming, costly and has a negative influence on the mining process due to increased turnaround times. This allowed for better control over mineral processing stages, leading to improved consistency and efficiency. However, the study also highlighted several challenges, such as considering additional parameters, especially of the froth's physical characteristics. Additionally, the need for continuous model updates was emphasised, particularly as ore characteristics can vary over time, necessitating adaptive models to maintain predictive accuracy.

Harsha et al. (2021) made a valuable contribution by showcasing the application of machine learning techniques, particularly Multiple Linear Regression, Random Forest, Decision Trees, and LSTM, for predicting silica impurity in the mining environment. Their research demonstrated the potential of these models to improve operational efficiency by enabling real-time estimation of silica content, which traditionally requires time-consuming laboratory analysis. The successful implementation of these predictive models in a web application underscores the role of machine learning in enhancing decision-making processes in the mining industry.

Srinija et al. (2022) also contributed to the field by successfully assessing machine learning techniques (Random Forest model) to predict the estimated percentage of silica in iron ore during the advanced froth flotation process. Their study verified the application of machine learning to enhance operational efficiency in mineral processing, achieving a high estimation accuracy of 91% within a short time frame. The application proved to be valuable for real-time analysis, which is critical in the fast-paced environment of mineral processing plants. Moreover, Pural (2023) developed a data-driven soft sensor to also predict silicate content in iron ore flotation concentrate, making use of machine learning models such as Random Forest, Ridge Regression, and Multi-Layer Perceptron. Despite these improvements, the study acknowledged that further validation with a broader set of parameters, such as froth physical characteristics, is necessary to fully realise the model's predictive capabilities and ensure its generalisability across different processing plants (Srinija et al., 2022).

Chen et al. (2023) explored the impact of mineralogical composition on the uniaxial compressive strength of sedimentary rocks: sandstones, carbonates, and shales. The study established that mineral composition significantly influences rock strength, with quartz (silica) content being the best predictor for sandstone strength and dolomite for carbonate strength. Statistical analyses and regression methods supported these findings that mineral content influences the estimation of rock strength in different lithologies. These results provide an understanding of how mineralogical differences contribute to rock strength, emphasising the importance of lithology-specific models for accurate estimations (Chen et al., 2023). Lithology is defined as the description of the rock in terms of its physical characteristics.

These case studies collectively highlight the impact machine learning has on enhancing operational efficiency and improving product quality in mining operations. The documented successes provide

compelling evidence of technology's potential to revolutionise mineral processing stages. While the benefits of predictive control systems are well established, with numerous studies reporting positive outcomes, integrating these systems into existing mining procedures, particularly in older mining operations, presents a notable challenge. Old systems often lack compatibility with modern machine learning technologies, necessitating upgrades or the development of mixed solutions that can bridge the gap between outdated and new systems. To address these gaps, the present research aimed to focus on the optimisation of models through assessing feature selection and architectural model techniques.

2.4.1 Advances and challenges in machine learning for mineral processing

In recent years, the application of machine learning techniques in mineral processing has gained significant traction. Key techniques include predictive modelling, feature selection, and optimisation of operational efficiency (Osei, 2019; Harsha et al., 2021). Several models, such as Multiple Linear Regression, Random Forest, and Artificial Neural Networks, have been employed to predict and control ore quality, particularly silica levels in mineral processing (Montanares et al., 2021). The application of these methods has shown effectiveness in improving the accuracy of estimations and the overall efficiency of mining operations. However, the review reveals that challenges persist in feature selection, handling complex model architectures, and the need for continuous model updates (Srinija et al., 2022; Pural, 2023).

Even though significant progress has been achieved, there are still challenges to address, such as the limited amount of research on the thorough impact of machine learning across various stages of the mining industry. Most studies focus on specific tasks, such as silica prediction and ore blending, without fully exploring the broader potential of these technologies to revolutionise the entire mineral processing workflow (Liu et al., 2021). Moreover, the review also identifies several challenges and research gaps, including the strengths:

2.4.1(a) Summary of findings

The reviewed studies have shown the effectiveness of ML models such as Random Forest, Artificial Neural Networks, and Gradient Boosting in ore quality estimation, specifically in the estimation of silica content. These models have demonstrated adequate accuracy and the ability to process complex datasets, thus offering improvements over traditional empirical models.

Additionally, the integration of AI and ML into the mining industry has led to notable improvements, including real-time estimations and improved decision-making. Case studies such as those by Osei (2019) and Srinija et al. (2022) have proven the capability of these technologies to improve operational efficiency and control impurity levels.

2.4.1(b) Limitations and gaps in prior studies

Most of the studies reviewed have indicated that feature selection for predictive modelling has been a significant challenge. Ore compositions' complexity, together with the various influencing factors across the mineral processing, has made it hard to consistently select features that enhance model performance. This challenge is compounded by the fluctuation of ore characteristics across different mining domains, leading to issues in generalising findings. Moreover, the literature review highlighted difficulties in the selection of fixed model architectures and hyperparameters, particularly for complex models such as regression algorithms. To improve these models' performance, extensive tuning and specialised expertise are required, making them resource-intensive and time-consuming. As a result, the applicability of these models will be limited by factors such as computational resources and the unique conditions of the mining operations, which can vary significantly from different ore domains.

Several studies have been exploring, with few efforts into fine-tuning ML algorithms for optimal estimations. Therefore, there is a need for more comprehensive research focusing on optimising existing models for specific mineral processing applications. The literature also lacks systematic approaches to model architecture selection. More in-depth studies are needed to introduce best practices for selecting model architectures for different mineral processing scenarios. Furthermore, there is a notable gap in integrating ML technologies into mining operations. The compatibility of modern ML systems remains an issue, especially with legacy infrastructure, and a hush environment remains a significant challenge, which requires hybrid solutions.

CHAPTER THREE: METHODOLOGY

This chapter outlines the methodological framework employed in this study, detailing the approach employed to attain the research objectives. The methodology is designed to align closely to enhance mineral processing by advancing silica content estimation techniques through the application of a machine learning algorithm. The chapter begins with a brief review of the research aim, objectives, and questions, thereby ensuring a clear connection between the research focus and the methods chosen. This is followed by the "Research Onion" framework by Saunders et al. (2007), which governs the structure and systematic process applied during this research. This approach involves peeling away layers of the research process, from philosophical underpinnings to specific techniques and procedures, thereby ensuring that each layer is carefully considered and aligned with the overall research objectives. The chapter map is shown in Figure 0.1.

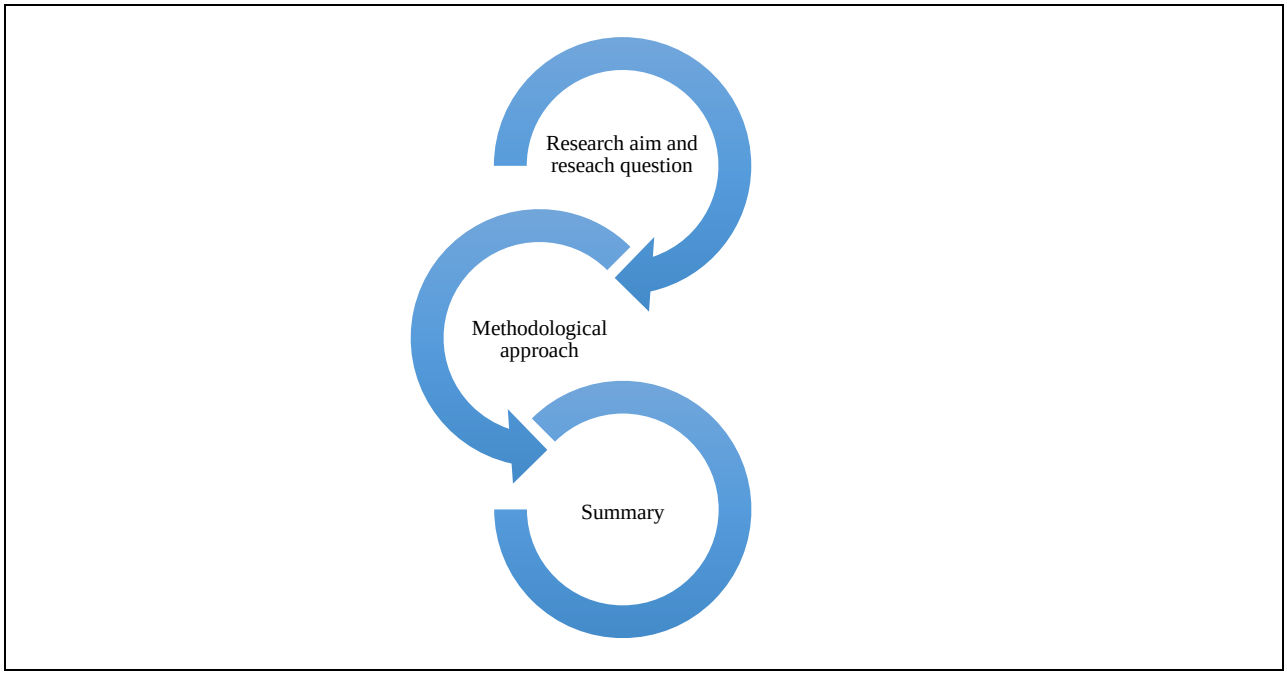

Figure 0.1: Chapter 3 outline map. Source: author

3.1 Review: Research aim, objective and questions

The research aimed to enhance mineral processing by advancing silica content estimation through machine learning algorithms. The main objective was to apply ML algorithms to estimate silica content in ore feed before processing. Sub-objectives include identifying relevant features,

evaluating various ML algorithms (Multiple Linear Regression, Random Forest Regression, Support Vector Regression, XGBoost) for estimation accuracy, estimating silica content with the best algorithm, and exploring ML's feasibility for ore blending control. Key research questions addressed the influence of geological features on silica estimation, the accuracy of different ML algorithms, the impact of the predictive model on processing efficiency, and the potential benefits and challenges of ML in ore blending control. These questions form the groundwork of the research, guiding the methodology and ensuring that the selected methods are well aligned with the research objectives. The literature review was structured to explore these questions, especially on the necessary theoretical knowledge and context for the methodological choices made in this study.

3.2 Methodological approach

The research's methodological approach was guided by the six sections (onion research framework) of decisions: Research philosophy, Research approach, Research design which constitutes the methodological choice and research strategy, time horizon, and tactic, which covers the techniques and procedures used for data collection and the analysis aspects (Saunders et al., 2019). The six layers are shown in Figure 0.2 below, whereby each layer gives insights into the methodological decisions driving the research, thus ensuring details and coherence in the research process.

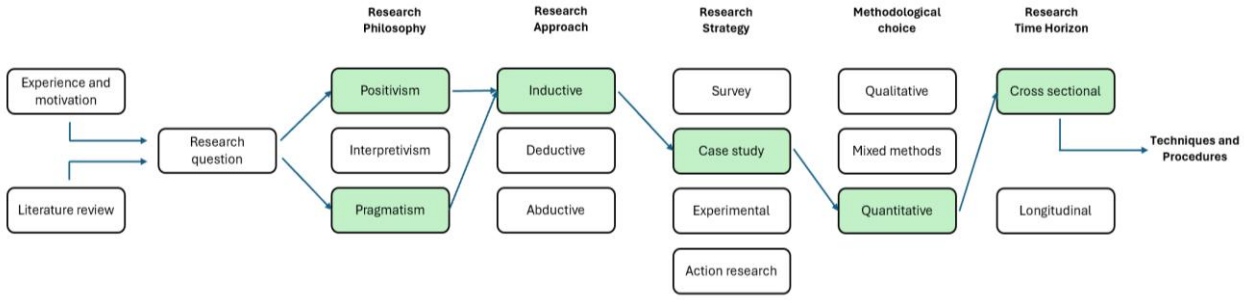

Figure 0.2: Research onion framework. Source: adapted from Saunders et al. (2019)

3.2.1 Research philosophy

Recognising the philosophical basis of any research is crucial as it profoundly influences the development of the research methods (Jansen, 2023). Pragmatism and positivism were the selected

research philosophies for this study. Pragmatism takes a more practical and flexible approach, emphasising the usefulness and applicability of research findings, rather than rigid philosophical positions. This approach can acknowledge the complex nature of the real-time silica content estimation and ore blending optimisation in a real-world context, allowing for the integration of both quantitative and qualitative methods. Positivism is crucial for the development of the predictive model to be used for silica estimation, as it underpins objectivity, empirical observation and the quantitative data, which uncovers patterns and regularities from the historical assay data of the ore feed composition.

3.2.2 Research approach

Research approach dictates the decision used for both data collection and analysis of any study. The main approaches are inductive, deductive, and abductive. An inductive bottom-up approach was adopted because the conclusions of an effective predictive model of silica content in the ore feed were based on the statistical analysis drawn from historical laboratory assay data collected and then analysed. The knowledge and insights were used to build from the data itself, which was applied to develop an effective model for the silica content estimation.

3.2.3 Research strategy

The research strategy defines how the researcher aims to carry out the study, about how the research question will be answered (Saunders et al., 2007). Research questions play a pivotal role in developing the strategy applied in any research. It should always be answerable in terms of getting evidence or data for justification. To answer the "*how leveraging ML technique can improve efficiency in mineral processing, focusing on using a high-performing predictive model to estimate the silica content?"* a mixed case study and archival approach was used. The case study approach allows for an in-depth examination of unique and complex situations. Case study research serves to deeply engage with phenomena and their surrounding contexts, thereby offering the advantage of providing rich and detailed insights into specific cases (Araujo, 2020). The specific case study revolved around Rosh Pinah Zinc Mine, with a primary emphasis on the analysis of the laboratory assay of ore concentrate and feed. The historical laboratory assay data were used to assist in gathering sufficient data for analyses, which enabled the pattern identification and development of an efficient predictive model, hence, the archival approach.

3.2.4 Methodological choice

This study adapted a quantitative approach, which complemented the study's research philosophy. Quantitative study involves numerical data, where statistical analysis is applied to produce empirical and objective evidence expressed in numerical formats. Generally, this approach is utilised to test hypotheses, identify patterns, and make estimations (McLeod, 2023). This study followed a quantitative approach, as it primarily relied on numerical data and statistical analysis of the ore feed composition. The research approach is illustrated in Figure 0.3 below.

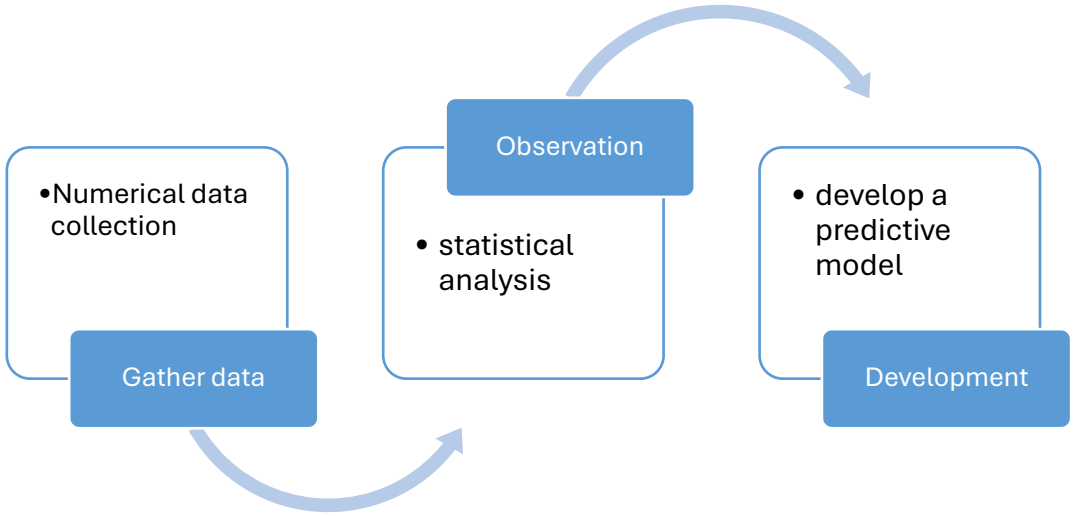

Figure 0.3: Research approach. Source: Author

3.2.5 Research time horizon

Time horizon refers to the duration within which the study is intended to be completed (Saunders et al., 2019). In this research, a cross-sectional time horizon was adopted for the collection of assay data on ore feed composition. A cross-sectional study is a 'snapshot' study. It refers to a phenomenon or a cross-section of the population which is studied for one time (Setia, 2016). This approach enabled the study of data collected at a specific point in time to identify associations between silica and other ore components. Focusing on data from a single moment, assessed the present state and interrelationships between these variables, which unlocked insights into the immediate effects of ML solutions, providing a snapshot of the current situation without the need for an extended study duration. This approach was suitable given the duration of this research, compared to the longitudinal approach.

3.2.6 Techniques and procedures

Techniques and procedures are the last yet crucial stage of the elusive onion research framework by Saunders et al. (2007). It is the overall inner circle of the research onion strategy, constitutes of 'tactics' which refers to features about the finer details of data collection and analysis (Tengli, 2020). This includes the procedures followed during the literature review, data collection, processing and analysis. The techniques and procedures used guide the researcher to achieve the objectives of the study.

3.2.6(a) Data sources and collection process

According to Taherdoost (2021), data collection is a process of gathering data aiming to gain insights regarding the research topic of interest ethically. It is further described as "a main stage in research which can overshadow the quality of achieving results by decreasing the possible errors which may occur during a research project" (Taherdoost, 2021, p. 11). In addition to a well-methodological framework, it is crucial to obtain high-quality data to enable researchers to obtain accurate results. This process involved the type of collection method, data type as per methodological choice and the source of data. Ethical consideration is required before commencing with this stage. It involved a detailed ethical clearance application, which included the research proposal and the informed consent form. An ethical clearance certificate was issued after a review from the NUST Faculty Research Ethics Committee (F-REC), which was used for data collection.

The historical data related to ore composition, processing parameters, and laboratory analyses were gathered from the plant department's database at Rosh Pinah Zinc Mine. This extensive database includes all ore sample tests at different stages of the ore processing. For this study, our focus is on the initial stage, where the ore feed is sampled for laboratory testing before entering the upstream ore upgrading process. These samples serve as the input dataset used to develop the silica levels in the ore feed. Human factors play a critical role in this process. The collection, preparation, and analysis of data involves collaboration between various departments and personnel, each contributing their expertise to ensure data accuracy and relevance. This source was selected due to its high relevance, accuracy, and availability as described below.

- Relevance: The selected sources provide directly relevant information for estimating silica content, as they include laboratory measurements of mineral concentrations.
- Accuracy: These sources are reliable and have been verified through historical usage and departmental credibility.
- Availability: The data from these sources is accessible and has been collected over a significant period, ensuring a comprehensive dataset.

The plant department staff is responsible for meticulously recording ore compositions and processing parameters. Laboratory technicians conduct detailed assays to measure the silica content, following strict protocols to maintain consistency and reliability. The dataset covers the period from 01 October 2019 to 31 October 2023. The next section describes the tools used and the structure of the gathered dataset, which served as the foundational bedrock for training and validating the machine learning (ML) model, ensuring that it encompassed the breadth of information needed to make accurate estimations regarding silica content within ore samples.

3.2.6(b) Dataset structure

The input dataset consists of a time series laboratory assay dataset of ore composition measurements captured over a period of five (5) years. This dataset encompasses various mineralisation characteristics of the ore feed, which are assayed at an interval of six (6) hours across the four (4) shifts in a day. It consists of 13 columns representing 5967 entries, which include the ‘Day’, ‘Shift’ and ‘Stamp’ measures, which will not be part of the input features. The remaining 10 features are our input dataset to estimate the target value (SiO2). The dataset column descriptions are as follows:

- **Day:** The date the measurements were taken.
- **Shift:** The shift during which the measurements were taken.
- **Stamp:** This indicates the exact date and time during which the measurements were taken, with shifts.

Features:

- **Zn (SPF):** The concentration of Zinc in the sample, measured in percentage.
- **Pb (SPF):** The concentration of Lead in the ore feed, measured in percentage.
- **Cu (SPF):** The concentration of Copper in the ore feed, measured in percentage.

- **Fe (SPF):** The concentration of Iron in the ore feed, measured in percentage.
- **Ag (SPF):** The concentration of Silver in the ore feed, measured in percentage.
- **Mg (SPF):** The concentration of Magnesium in the ore feed, measured in mg/L.
- **Mn (SPF):** The concentration of Manganese in the ore feed, measured in percentage.
- **Ca (SPF):** The concentration of Calcium in the ore feed, measured in percentage.
- **S (SPF):** The concentration of Sulfur in the ore feed, measured in percentage.

Target:

- **SiO2 (SPF):** The concentration of Silica (SiO2) in the ore feed, measured in percentage.

3.2.6(c) Preprocessing and Data Exploration

Data exploration and pre-processing serve as an initial phase in the data analysis process. This involves a detailed investigation of the dataset to understand its attributes and unveil preliminary patterns, relationships, or anomalies. This phase is significant for summarising the dataset's primary features, making use of various visualisation techniques, before further statistical analysis. The step of converting the raw data into a usable format for machine learning purposes is illustrated below. For this study, a Google Collab cloud-based Python notebook was used to write and execute Python code to develop a predictive ML model for silica content estimation.

All preprocessing and exploration computation were done using an HP EliteBook running on an Intel(R) Core (TM) i7-8550U CPU 1.80GHz 1.99 GHz, with a 16.0 GB (15.8 GB usable) RAM. This process was achieved using the steps explained below. This ensured the underpinning of the dataset landscape and preparation for the predictive models.

- **Loading the data**

  The collected data is imported into a data manipulation and analysis environment using the pandas library "pd.read_excel('Lab_assay.xlsx')" function to read the data from an Excel stored file saved as 'Lab_assay' into a pandas DataFrame. This served as a structured transformation where the data was stored in rows and columns, which enabled the data for various data processing activities

  During the data importing process, any entries that were not captured are automatically treated as missing values by pandas. These missing values, by default, are denoted either

Not a Number (NaN) for numerical data or Not a Time (NaT) for datetime data, which acts as a special placeholder. This ensured an accurate reflection of the Dataframe by indicating the presence of incomplete data. The ability to handle missing values is one of the key strengths of the pandas library, allowing for robust and flexible data pre-processing.

- **Understanding the dataset**

  After the dataset is imported, understanding the concise summary through the ".info()" function is applied. This will help guide all the preprocessing work on the dataset before fitting it to the models in terms of converting the features into their respective datatypes suitable for statistical analyses.

- **Handling missing data**

  Handling missing data played a crucial role in the analysis quality. There are various methods of handling missing data, which helps to retain the sample population and avoid the central tendency imputation as the composition only follows a trend within a short interval.

  The 'Day' column was forward-filled with the most recent non-null value, while the 'Stamp' column was updated using a combination of the shift dictionary mapping and the forward-filled 'Day' column. This study will visually access the four (4) methods: fill NaN with 0 (".fillna(0)"), fill NaN with mean value (".fillna(mean)"), Last Observation Carried Forward (LOCF) fill (".ffill()") and interpolation fill ".interpolate(method='linear')". These options are visualised through plotting a line plot using the matplotlib "plt.plot()" visualisation library. A suitable method will be applied to the entire dataset.

- **Ordering the dataset**

  The chronological order and time interval were analysed through sorting the dataset and taking a detailed view of the dataset. The equidistant timestamps was observed through extraction of the hours from the Stamp column then counted the occurrence of the 6 hour interval using the ".isin([0,6,12,18])" function and matched on the total entries. This is crucial for time series analysis.

3.2.6(d) Feature selection

Feature selection is crucial in machine learning and predictive modelling as it helps identify the most relevant variables for estimating the target outcome. Effective feature selection improves model accuracy, reduces overfitting, and enhances model interpretability by removing irrelevant or redundant features. In this study, selecting the right features is essential to accurately predict silica content (SiO2) in ore feed. This process involves a rigorous exploration of the data to identify and engineer relevant features that exert influence on silica content. This is the final stage of data preparation before fitting them to any model. This study encompasses linearity, feature selection, scaling and splitting the dataset. The method used for feature selection is a correlation matrix, a feature importance matrix and domain knowledge.

I. Correlation analysis

Correlation analysis was conducted to assess the linearity between each feature and the target variable (SiO2). This process is achieved through a correlation matrix, which assesses the linear relationship between each feature and the target variable (SiO2). A correlation matrix is created to evaluate the strength and direction of the relationships. Pearson correlation was employed in this study to assess the linear relationship between variables. This was done through the ".corr(method='pearson')" which is a function from the "scipy.stats" library. Pearson correlation is calculated using the equation (0.1) below.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (0.1)$$

Where:

x_i and y_i are the individual sample points

$\bar{x}$ and $\bar{y}$ are the means of x and y

The Pearson correlation coefficient r ranges from +1 to -1, which is a scale of either a strong positive, strong negative or weak relationship, where.

- +1 indicates a perfect positive linear relationship.

- 0 indicates no linear relationship.
- -1 indicates a perfect negative linear relationship.

The highly positive and negative were assessed to guide the selection of features. This helps identify which features have a strong correlation with SiO2, guiding their inclusion or exclusion in the model.

II. Feature importance metrics

Feature importance metrics weigh how much each feature contributes to the predictive performance of the model. Random Forest model (RF) was applied for this study to reduce the dataset's feature dimension. This method provided a score indicating the contribution of each feature to building an efficient and manageable regression model. Features with higher importance scores are considered more relevant. This is achieved using the '*.feature_importances_'* attribute, which computes the importance score of each feature fitted to the model. Features were evaluated based on these scores, with a cutoff of 0.05 applied. Features with importance scores above this threshold were retained for modelling.

Additionally, features were selected based on the mineralogy and lithology of the ore composition. The "Stamp" column, representing timestamp data, was converted to an index to facilitate time-based analysis, allowing the removal of the "Day" and "Shift" columns, which were deemed less relevant for this study.

3.2.6(e) Feature Scaling

Feature scaling is a vital stage of data pre-processing as it converts the numerical features in a dataset to the same range. As if left unaddressed, the differences can lead to the model assigning more importance to the features with bigger scales. This results in biased outcomes, skewness and suboptimal model performance. Therefore, feature scaling helps balance the features' influence by scaling them to the same range, which enhances models' performance. The dataset comes with different scales (for example, the SiO_2 (SPF) is measured on percentage, while Ag (SPF) uses mg/L). Therefore, the dataset is scaled to ensure that each feature contributes equally to the analysis. This study considers two (2) types of features scaling techniques. Each technique was individually applied to the models, and their impact was evaluated to help determine the best fit got this study.

Standardisation (Z-score): This method aims to scale the dataset to have a mean of 0 and a standard deviation of 1. Shown below is the formula (0.2) applied to determine the Z-score.

$$Z_{score} = \frac{X - \mu}{\sigma} \tag{0.2}$$

Where:

X : Original value

μ : Mean of the feature

σ : Standard deviation of the feature

Normalisation (Min-Max Scaling): This method aims to scale the dataset to fall within a range of 0 and 1. Shown below is equation (0.3) applied for normalisation.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{0.3}$$

Where:

X : Original value

X_{min} : Minimum value of the feature

X_{max} : Maximum value of the feature

3.2.6(f) Data partitioning

Split train and test data was the final stage before the data was fit to the models. The clean data was divided into train and test datasets. This study's dataset is partitioned into training and testing using the traditional 80:20 split ratio using the '*train_test_split'* function from the scikit-learn library. The two datasets were used to fit and evaluate the developed machine learning. These partitioned datasets were critical for both the training of the machine learning models and their

subsequent evaluation. This enables a robust assessment of the model's performance on unseen data.

3.3 Model implementation and performance evaluation

Machine learning (ML) enables certain computational tasks to be performed with no explicit programming instructions. Therefore, ML identifies pattern patterns through the process of learning from organised historical data (Johnston et al., 2019). It uses statistical algorithms that allow self-learning through identifying patterns and making estimations. This study adopts supervised machine learning algorithms to predict the silica content in the ore feed, then, through an evaluation matrix, the suitable algorithm will be adopted for the implementation of the predictive model for the Rosh Pinah Zinc mine. Among various supervised ML regression algorithms, four (4) regression algorithms were employed on the laboratory assay historical data. Regression analysis may be defined as a type of predictive modelling technique which investigates the relationship between a target and predator (Pirge, 2020).

The nature of mineral resources (ore) lies in their uniqueness, due to their geological formation, composition, occurrence and distribution. This requires a sophisticated ML technique to determine the complexity of relationships, interpretability and robustness to noise that comes with ore compositions. This study adopts four regression techniques, namely *Multiple Linear Regression, Support Vector Regression, Random Forest Regression and Gradient Boost Regression through the XGBoost*. These models' capability was used to select the desired models, as they offer diverse complementary strengths, from simplicity (Multiple Linear Regression), adaptability to non-linear datasets (Support Vector Regression), flexibility to noise (Random Forest Regression) and high performance (XGBoost). Additionally, these four models are commonly used, and they effectively evaluate how best different machine learning techniques handle this regression task while ensuring a robust model selection for silica estimation.

The model's implementation was carried out using Python via Google Collab, leveraging common libraries such as sckit-learn, TensorFlow and XGBoot. Python's inbuilt machine learning extensive ecosystem provided most of the necessary tools required for data preprocessing, model training and performance evaluation. MLR, RFR and SVR were implemented through scikit-learn, which offered a simple and efficient interface for training and evaluating these algorithms. The XGBoot library was applied for the gradient boosting framework, which is known for its high-performance

boost. Additionally, Pandas and NumPy were utilised for data manipulation and preprocessing tasks. This includes feature scaling, missing data handling and data portioning. Seaborn and Matplotlib were essential for generating data visualisations for easier interpretations. The models were initialised, trained datasets were fitted and evaluated using the following regression model evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R^2) to assess the estimation accuracy and model robustness.

- **Root Mean Square Error** calculates the average error by computing the residual between predicted and observed values. The error measurement is of the same units as the original data, since it square roots the mean square error. RMSE values were evaluated to determine the performance of the models. Equation (0.4) was used to calculate those values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (0.4)$$

  Where:

  RMSE = Root Mean Square Error

  i = variable i

  n = number of observations

  y_i = actual value observation

  $\hat{y}_i$ = estimated value

- **Mean Absolute Error** calculates the error by computing the difference of the average absolute difference between the actual and estimated observation. MAE takes the absolute residual values, which gives all errors equal weight. MAE values were evaluated to determine the performance of the models. Equation (0.5) was used to calculate the MAE values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

(0.5)

Where:

MAE = Mean Absolute Error

n = number of observations

i = variable i

y_i = actual value observation

$\hat{y}_i$ = estimated value

- **R-Square** calculates the coefficient of determination, which indicates how well the model's estimations match the actual data. Equation (0.6) was used for these calculations. This was computed to determine the variance proportion of the dependent variable that is predictable from the independent variable.

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(0.6)

Where:

R^2 = R-Square

n = number of observations

i = variable i

y_i = actual value observation

$\hat{y}_i$ = estimated value

$\bar{y}$ = mean of the actual value

$SS_{residual}$ = residual sum of squares

SS_{total} = total sum of squares

To further interpret the model's performance, visualisations were used to plot the actual vs estimated values which are display in the results chapter with their corresponding code snippets provided in the appendix.

3.4 Simulation and comparative analysis

A simulation approach was applied to compare the silica content quantification between the traditional assay method and machine learning with a high-performance power in this study. The simulation was applied to evaluate the practicality of an adopting estimation model over the conventional time-based techniques.

3.4.1 Traditional assay approach

In the traditional method, a chemical laboratory assay test was used to measure the silica content at an interval of six hours. The accuracy of this method comes at a cost of time delay due to the required time for laboratory sample preparation and analysis. A day's data was used as the benchmark for comparing with the machine learning model's information generated. A graph was generated to visualize the silica content trend, highlighting possible information deduced from it

3.4.2 Machine learning model estimation approach

The high-performing machine learning model was used to estimate silica content based on the grade control data. The model's estimation capability allowed an immediate understanding of the silica content variation through the provision of a continuous estimation at a finer resolution. The graph generated shows the finer trends of silica content, providing dynamic fluctuations within the feed.

3.5 Ethics considerations

This study's activities carried out posed minimal risk, due to the study's primary involvement of sensitive datasets for analytical tasks and model training. Secondary data collection was applied, as the laboratory assay data gathered are already captured by the laboratory personnel and they are being used for daily grade tracing by the grade control team, which makes it a review of archived records (published source), as per (Taherdoost, 2021) describes what secondary data is. Safeguarding sensitive operational data was guided by the IT department mines procedures. This involves secure data storage and using a personal laptop or work's desktop computer only to

protect data privacy, upholding ethical standards. Strong credentials and the use of an authenticator were applied to prevent unauthorised access to the mines' data. This study is governed by FREC/15/24 ethical clearance certificate provided in the appendix.

3.6 Chapter summary

Methodology plays a vital role in guiding the researchers to achieve the study's aims and answer the questions set for that study. This was designed to guide each step to directly contribute to achieving the study's overall aim of estimating silica content in the ore feed by answering the research questions. Table 0.1 succinctly outlines the activities involved and the criteria considered to answer the research questions.

Research question	Research activity	Evaluation criteria
I. Which mineral composition features most significantly influence the estimation of silica content in ore feed?	Conducted Pearson correlation analysis for the identification of the most impactful features affecting silica content estimation.	Analysed the correlation coefficients to determine correlations between SiO2 and other minerals. The key features were those with either a strong negative or positive relationship.
II. Which machine learning algorithms, among the commonly used machine learning algorithms, namely Multiple Linear Regression, Random Forest Regression, Support Vector Regression, and XGBoost, offer the highest accuracy in estimating silica content in the ore feed?	Used the assay data to train and evaluate four ML models to estimate silica content.	Applied the evaluation matrix: RMSE, MAE, and R^2. Compared scores across the models to assess estimation accuracy.

III. How can silica content estimation in ore feed enhance mineral processing efficiency?	Used the best-performing model to estimate silica content and analyse its impact on mineral processing.	Demonstrated the accurate silica content estimation and discussed potential optimisation of mineral processing processes.

Table 0.1: Summary of the research question and evaluation criteria

CHAPTER FOUR: RESULTS

This chapter presents the experimental scenarios' results, focusing on the outcomes of the main task of the study. This includes evaluation comparison from different features, finding the optimal parameters and the different models' outcomes and highlights estimates obtained using the best model.

Experimental results

4.1.1 Overview of data structure

After the data was uploaded, it was observed that all the variables were stored in the correct datatype as shown in Figure 0.1 below. "Day" and "Stamp" are stored in a datetime format, "Shift" in an object format, and models input features as a numeric format. The data is already stored in an equivalent equidistant timestamp of six (6) hours and in chronological order. Therefore, additional data preparation (resampling and decomposition) was not required during the preparation stage.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5968 entries, 0 to 5967
Data columns (total 13 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Day         2587 non-null   datetime64[ns]
 1   Shift       5968 non-null   object
 2   Stamp       5956 non-null   datetime64[ns]
 3   Zn (SPF)    5935 non-null   float64
 4   Pb (SPF)    5935 non-null   float64
 5   Cu (SPF)    5935 non-null   float64
 6   Fe (SPF)    5935 non-null   float64
 7   Ag (SPF)    5935 non-null   float64
 8   Mg (SPF)    5935 non-null   float64
 9   Mn (SPF)    5935 non-null   float64
 10  SiO2 (SPF)  3524 non-null   float64
 11  Ca (SPF)    5934 non-null   float64
 12  S(SPF)      3520 non-null   float64
dtypes: datetime64[ns](2), float64(10), object(1)
memory usage: 606.2+ KB
```

Figure 0.1: Dataset initial concise summary. Source: Author

4.1.2 Identifying and addressing missing data

The concise summary above indicates a discrepancy between the total entries of the overall dataset in comparison to the feature's entries. The 'Day' column was forward-filled with the most recent non-null value, while the 'Stamp' column was then updated using a combination of the shift

dictionary mapping and the forward-filled 'Day' column. Plotting the time series reveals a significant gap in the edge data for both the target and Sulfur columns, indicating missing data. This occurred between 2021-05-20 and 2019-10-31 and is indicated with a red block as illustrated in Figure 0.2. The dataset's start date was adjusted by dropping entries from that period.

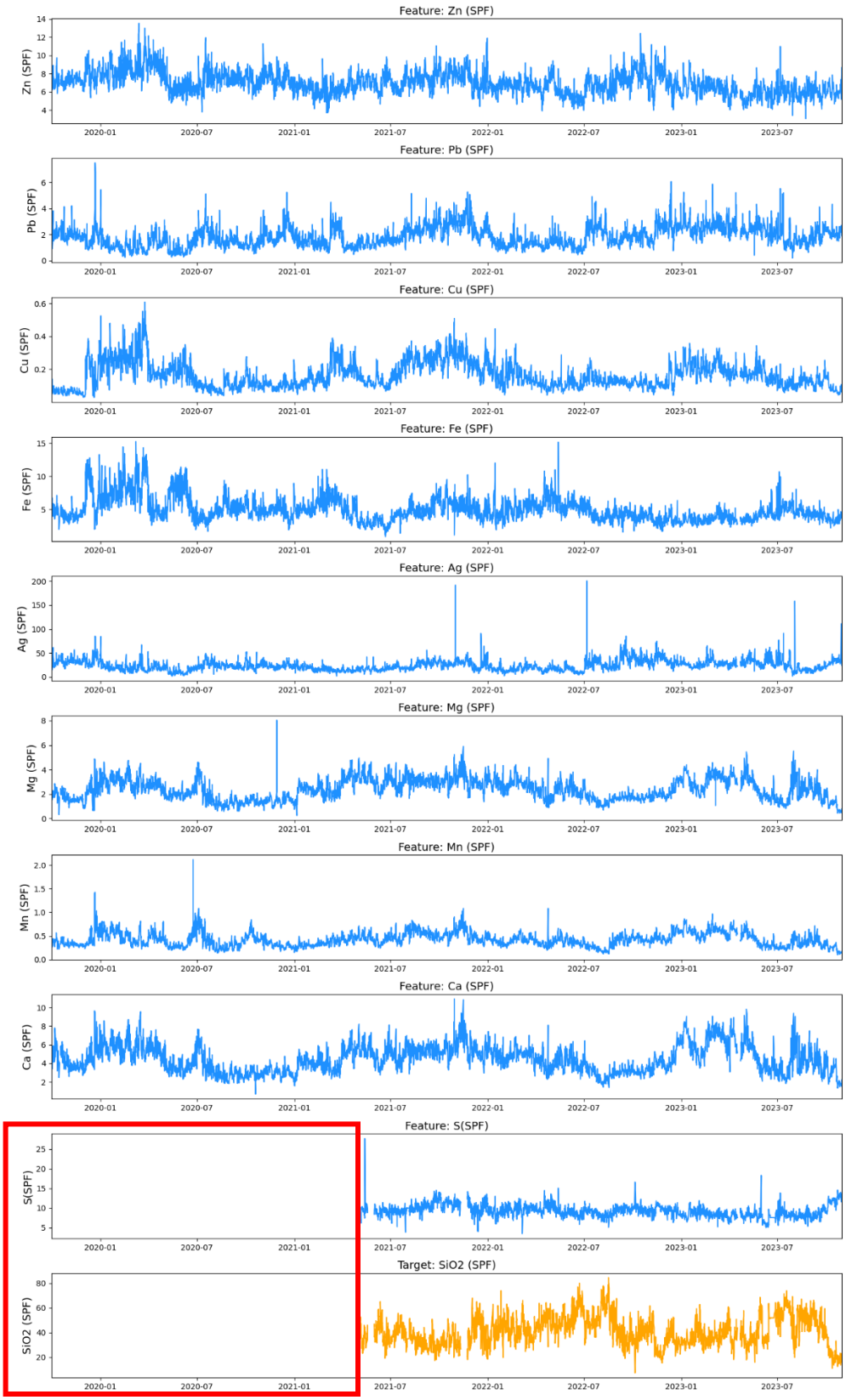

Figure 0.2: Target variable time series line graph. Source: Author

Furthermore, there were minor missing values in the dataset. Four imputation methods were applied and visualised, with a focus on the period from 2021-05-01 to 2021-10-01, for enhanced clarity. Each method has a different impact on how the data is interpreted.

- **Method 1** – fill NaN with Zero

  This option is not applicable as there should be a quantity of SiO2 due to the nature of the ore body. Furthermore, filling with 0 can introduce bias by artificially lowering the value, especially in series where a value of 0 is not plausible
- **Method 2** – fill NaN with Mean Value

  This option is also not a suitable idea as ore composition follows a certain trend based on the domain being mined during that period. Filling with the means reduces variability and can hide trend effects.
- **Method 3** – fill Nan with Last Value

  This method make use of a forward fill ".ffill()" function, forward fill uses a limited neighboring value making it not suitable for this dataset. Forward filling preserves the most recent data point but can introduce artificial trends.
- **Method 4** – fill NaN with Linearly Interpolated Value

  This method uses the interpolation ".interpolate(method='linear')" function. This turns out to be the best option as it gives a smoother transition yet retains a better representation of trends. Additionally, interpolation offers a balance by estimating the missing values based on existing data, maintaining the series' continuity.

All those NaN values were handled using a linear interpolation, which was chosen from the 4 imputations options due to a visual analysis observation in Figure 0.3.

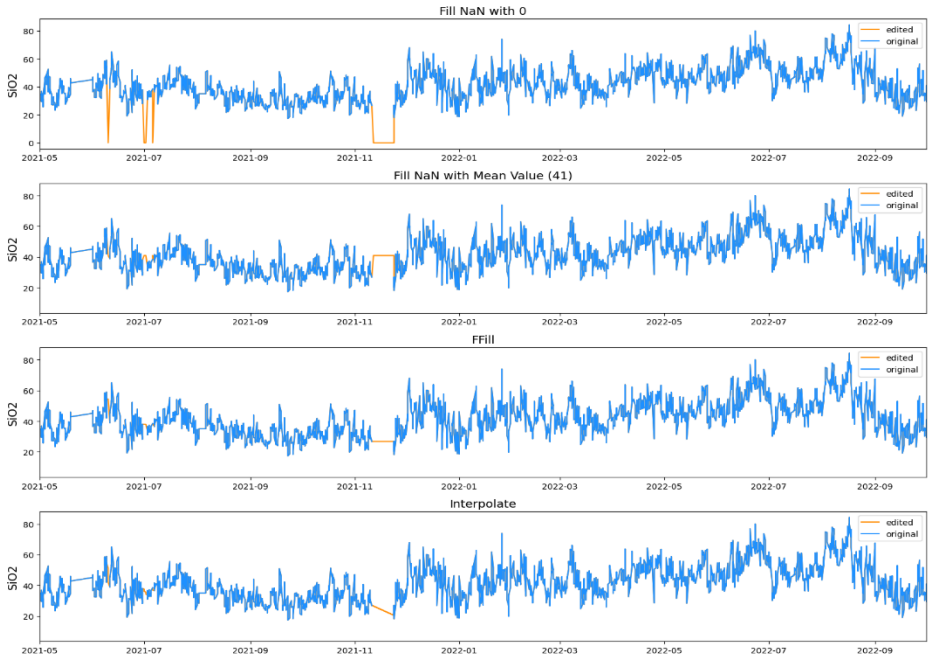

Figure 0.3: Imputation methods for NaN values. Source: Author

Additional prepositions were done, dropping the 'Day' and 'Shift', then converting the 'Stamp' column to the DataFrame index. After cleaning the dataset, the entries declined from 5967 to 3608 which are the input values that were used for models' development. The new concise summary is displayed in Figure 0.4.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 3608 entries, 2021-05-01 00:00:00 to 2023-10-31 18:00:00
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Zn (SPF)    3608 non-null   float64
 1   Pb (SPF)    3608 non-null   float64
 2   Cu (SPF)    3608 non-null   float64
 3   Fe (SPF)    3608 non-null   float64
 4   Ag (SPF)    3608 non-null   float64
 5   Mg (SPF)    3608 non-null   float64
 6   Mn (SPF)    3608 non-null   float64
 7   SiO2 (SPF)  3608 non-null   float64
 8   Ca (SPF)    3608 non-null   float64
 9   S(SPF)      3608 non-null   float64
dtypes: float64(10)
memory usage: 310.1 KB
```

Figure 0.4: Cleaned dataset concise summary. Source: Author

4.1.3 Correlation insight

The Pearson correlation coefficient r calculates the relationship between the ore associates. Figure 0.5 shows the correlation matrix among the 10 variables, namely SiO2 (SPF), Zn (SPF), Pb (SPF), Cu (SPF), Fe (SPF), Ag (SPF), Mg (SPF), Mn (SPF), Ca (SPF), and S (SPF).

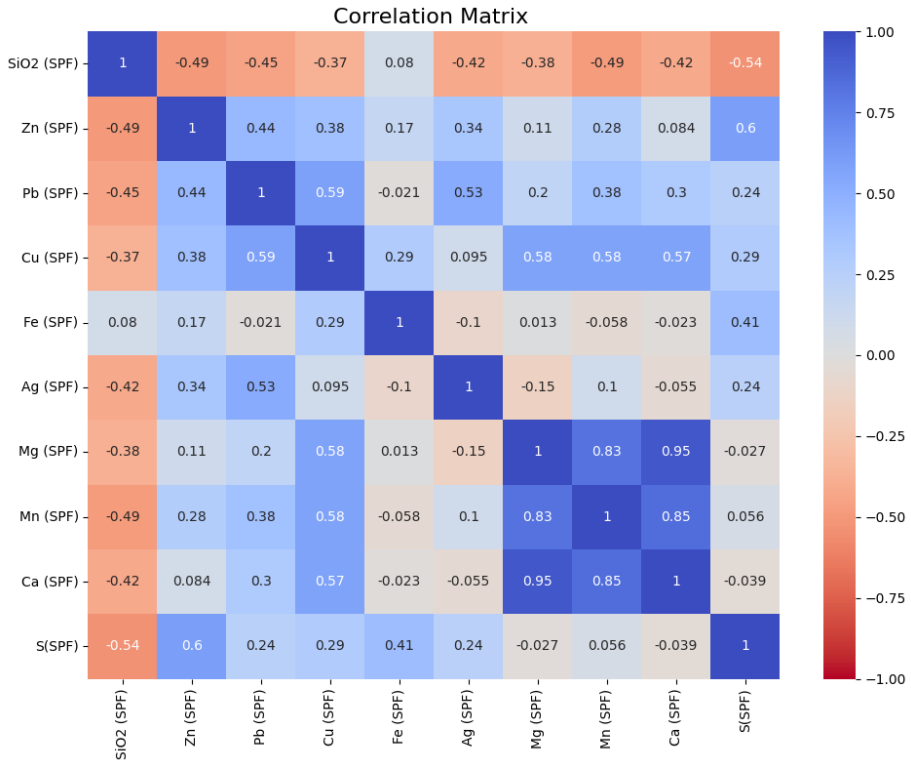

Figure 0.5: Correlation Matrix. Source: Author

4.2.3(a) Negative correlations

Sulphur (-0.54) shows the strongest negative correlation with SiO2 among other minerals. This suggests that as Sulpur levels increase, Silica tends to decrease. Therefore, Sulphur could be an important predictor in a model, especially in environments where Sulphur and Silica are inversely related due to geochemical processes. Zinc (-0.49) and Manganese (-0.49) show a moderately strong negative correlation with SiO2. Including these elements as predictors could improve the model's accuracy, as their levels might inversely affect SiO2 concentrations. Lead (-0.45) and Silver (-0.42), they also show a significant negative correlation with SiO2. Their presence could help tune the estimation model, particularly in contexts where there are spikes.

4.2.3(b) Positive correlations

Iron (0.08) has a weak positive correlation which translates to a little direct influence on SiO2, so it might not be a strong predictor on its own, but its associate can boost its involvement.

4.1.4 Evaluating the impact of feature scaling

Given the observation from the model's performance using the two feature scaling methods, the Z-score scaling method shows a high performance in comparison to MinMaxScaler across all models. This is shown in Table 0.1.

Model	Feature scaling	RMSE	MAE	R Score
MLR	Z-score	0.5544	0.4094	0.6934
	MinMaxScaler	6.4818	4.7861	0.6934
SVR	Z-score	0.3932	0.2884	0.8458
	MinMaxScaler	6.5211	4.8273	0.6867
RFR	Z-score	0.3821	0.2848	0.8544
	MinMaxScaler	4.4712	3.3335	0.8541
XGboost	Z-score	0.3851	0.2901	0.8521
	MinMaxScaler	4.4984	3.3949	0.8523

Table 0.1: Feature scaling evaluation

- For **MLR**, Z-score scaling provides significantly better RMSE and MAE compared to MinMaxScaler.
- For **SVR**, Z-score scaling yields substantially better RMSE, MAE, and R^2 scores.
- For **RFR** and **XGboost** while the R^2 scores are similar for both scaling methods, the RMSE and MAE are noticeably better with Z-score scaling.

The performance metrics (RMSE, MAE, and R^2) steadily exhibited better results with **Z-score** standardisation, leading to its selection for the analysis that follows.

4.1.5 Feature importance analysis

The feature selection is governed by the features' importance score. Figure 0.6 shows the importance score on how much each feature contributes to the predictive power of the model. The importance score is based on the Gini importance (mean decrease in impurity), which evaluates the importance of each feature through the impurity reduction across all decision trees in the applied Random Forest model.

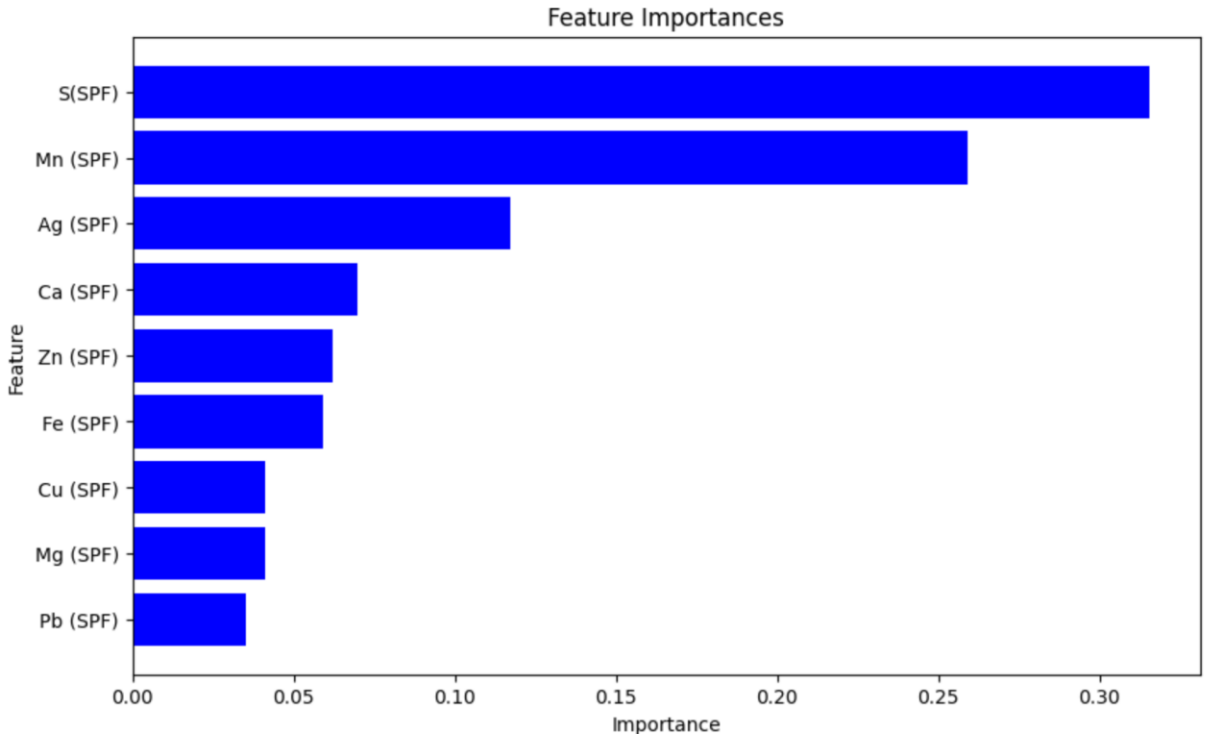

Figure 0.6: Feature importance. Source: Author

S(SPF) has the highest importance value (0.317612), which indicates that it is the most influential feature in estimating the target variable. The second most significant is Mn (SPF) with a significant importance value (0.252144). Ag (SPF) (0.117287) and Ca (SPF) (0.073351) are moderately important. Features like Pb (SPF) (0.034511), Mg (SPF) (0.039495), and Cu (SPF) (0.043331) have the lowest importance values, suggesting they contribute less to the model's estimations. The complete importance score for all features is shown in Table below.

Feature	Importance
S(SPF)	0.318391
Mn (SPF)	0.25487
Ag (SPF)	0.114162
Ca (SPF)	0.072408
Fe (SPF)	0.061615
Zn (SPF)	0.060039
Cu (SPF)	0.041501
Mg (SPF)	0.041177
Pb (SPF)	0.035838

Table 4.2: Feature importance scores

4.1.6 Performance evaluation

This section covers the evaluation of model estimation based on the three metrics, namely, RMSE, MAE, and R^2 of the different models used. The predictive models were trained using the pre-processed dataset: df_all_features: consider all the features, and df_important_features: consider features with importance value above 0.05. Furthermore, the output from different hyperparameters is also presented.

4.1.6(a) Results of Multiple Linear Regression

The results of the Multiple Linear Regression model across the 2 datasets are shown in Table 0.2. This demonstrates the impact of feature selection on the model's performance. It provides a straightforward approach to understanding the relationships between features and the target variable.

Dataset	RMSE	MAE	R2 Score
df_all_features	0.5544	0.4094	0.6934
df_important_features	0.5578	0.4111	0.6897

Table 0.2: Multiple Linear Regression evaluation

Overall, the model performance across the two datasets shows a slight variation in RMSE, MAE and R^2 Score. It shows a marginally lower RMSE and MAE, with a higher R^2 Score when fitted

with df_data dataset compared to df_all_features dataset. On that basis, df_all_features model indicates a marginally high estimation accuracy due to the lower average error, whereby a greater portion of its variance in the target variable is well accounted for at 69%. Figure 0.7 shows the comparison of the silica's (SiO2) actual test values indicated in the blue line and the estimated values indicated in the orange line. This graph indicates that the test and estimation line fairly following each other fairly, demonstrating that a moderate model performs moderately due to the frequent diversion in the fluctuations and peaks across the two lines. This indicates how the model captures both short-term and long-term trends.

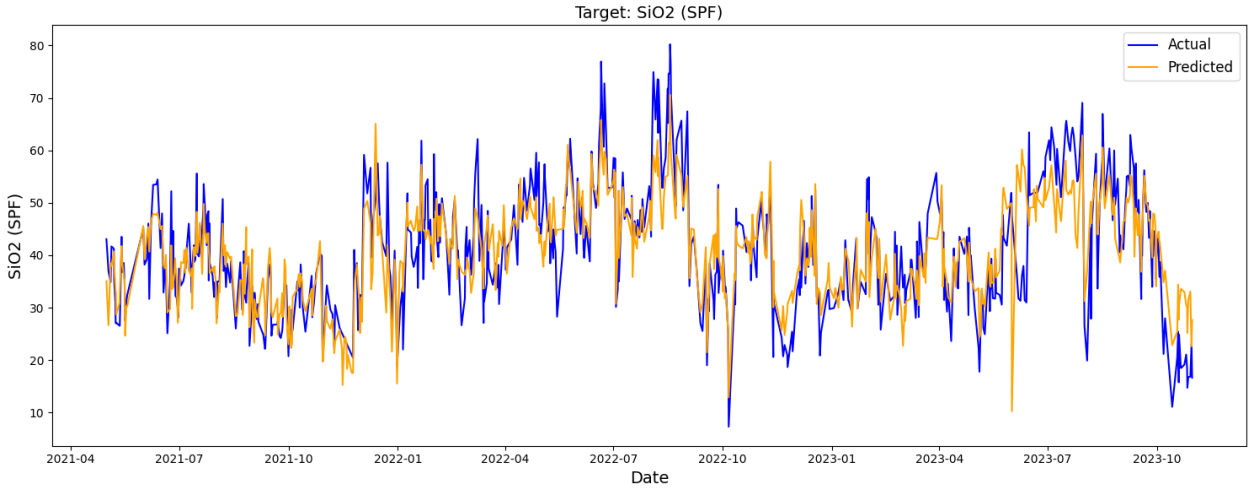

Figure 0.7: Evaluation of MLR Model Using `df_all_features` Dataset. Source: Author

4.1.6(b) Results of support vector regression

This section looks at the results obtained from the SVR model when the 2 datasets were applied. In machine learning kernel is a function that's applied in algorithms that transforms data into a higher dimensional space, ensuring simplicity to model the relationships between features. Choosing the kernel that best fit each dataset is governed by how well each kernel explains the variance, which is illustrated in Table 0.3.

Dataset	linear	Poly	rbf	sigmoid
df_all_features	0.6285	0.6030	0.8537	-2028.3104
df_important_features	0.6308	0.4391	0.8244	-1299.8922

Table 0.3: Kernel's model confidence

The rbf kernel performed marginally well compared to the other kernels across both datasets. This shows a strong capture of complex patterns effectively. Feature selection only improved the linear kernel, while negatively impacting the poly, rbf and sigmoid kernels. The sigmoid kernel performed quite low in either scenario. The model's tuning grid search mean test score for the specified epsilon and C value is shown in Figure 0.8.

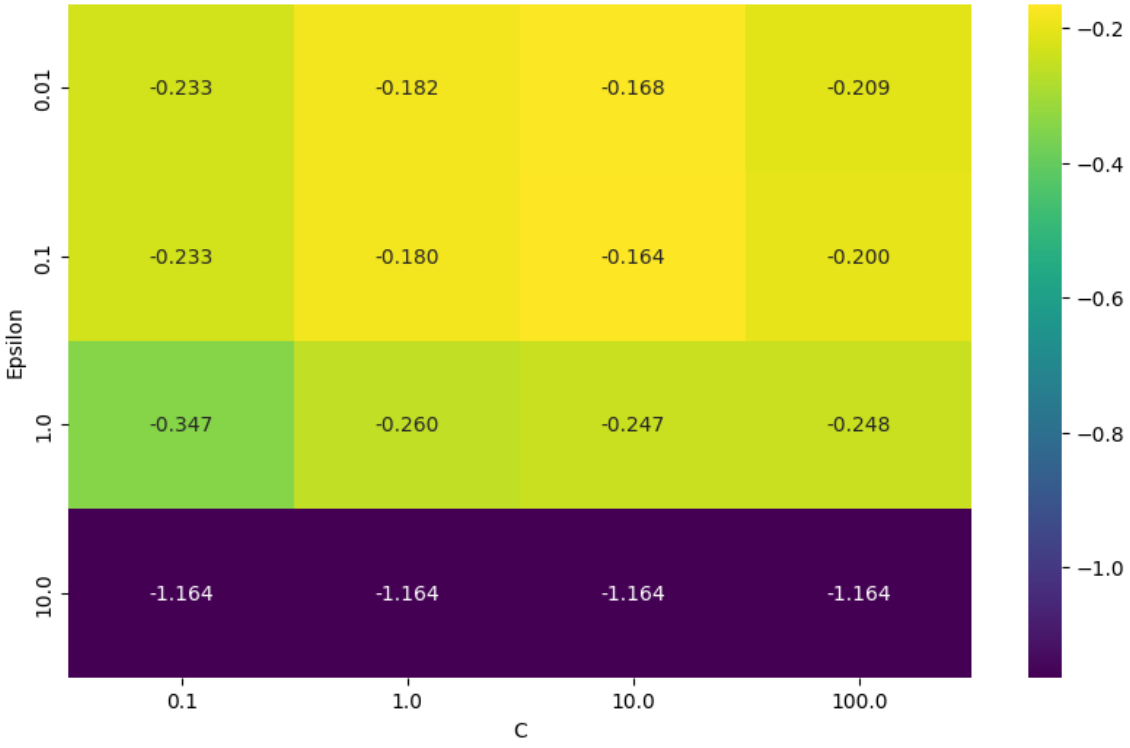

Figure 0.8: SVR epsilon and C value grid search. Source: Author

GridSearchCV was performed with a combination of 16 parameters, which were evaluated across all possible combinations with a 5-fold cross-validation, resulting in a total of 80 fits. The least negative combination of epsilon and C determines their optimal values. In Figure 0.8, the least negative score is -0.164, which links to the optimal regularisation and epsilon to be 10 and 0.1, respectively. This yields a strong performance of the rbf SVR model tuning and performance evaluation for all the datasets. Feature selection slightly reduces the model's performance, compared to the inclusion of all features, which has a lower error rate and slightly higher R^2 score, as shown in Table 0.4 below. The RMSE increased from 0.3932 to 0.4284 on the feature-selected dataset, while the R^2 dropped.

Dataset	RMSE	MAE	R2 Score
df_all_features	0.3932	0.2884	0.8458
df_important_features	0.4284	0.3206	0.8169

Table 0.4: SVR model tuned results

Finally, the SVR tuned model was applied to the 2 datasets, and the performance results are displayed in the SVR model tuned results above. Figure 0.9 shows the comparison of the silica's (SiO2) actual test values indicated in the blue line and the estimated values indicated in the orange line. This graph indicates that the test and estimation lines closely follow each other, demonstrating that a good model performs due to the similar fluctuations and peaks across the two lines. This indicates the ability of the model to capture both short-term and long-term trends. However, there are few instances where there are huge deviations between the estimated and actual values.

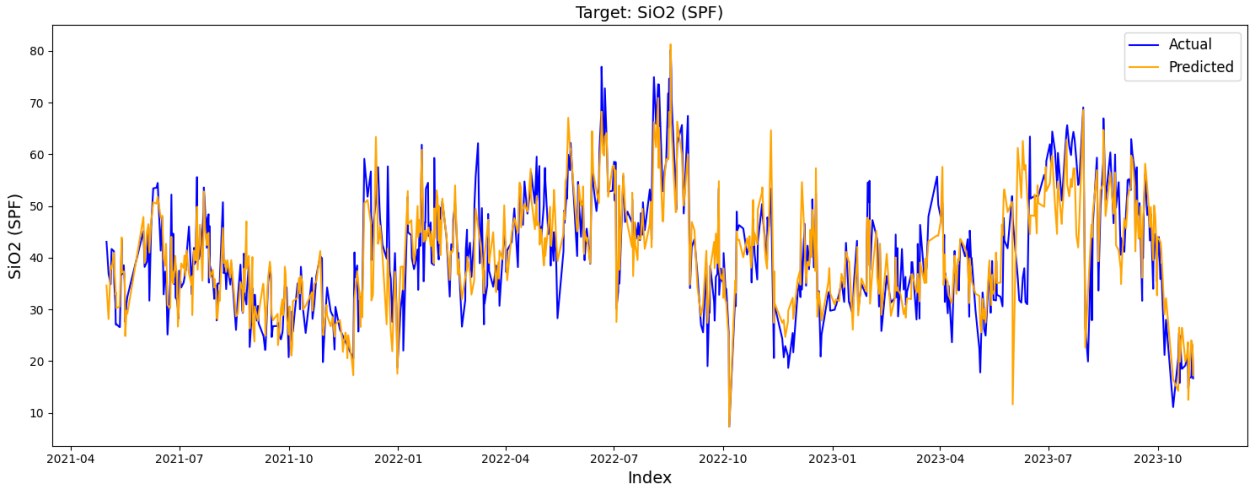

Figure 0.9: Evaluation of SVR Using `df_all_features` Dataset. Source: Author

SVR is a powerful regression technique which is well-suited to datasets with non-linear relationships. This model, particularly with the RBF kernel, showed strong performance,

4.1.6(c) Results of Random Forest Regressor

Figure 0.10 shows the different Cross-Validation RMSE (CV RMSE) values when the df1_data is applied. Describing the number of trees (n_estimators) in the forest, it is evident that 250 is the

optimal number of trees for this model. This is obtained from the smallest value of CV RMSE, to determine the optimal model, which is 0.4374 CV RMSE.

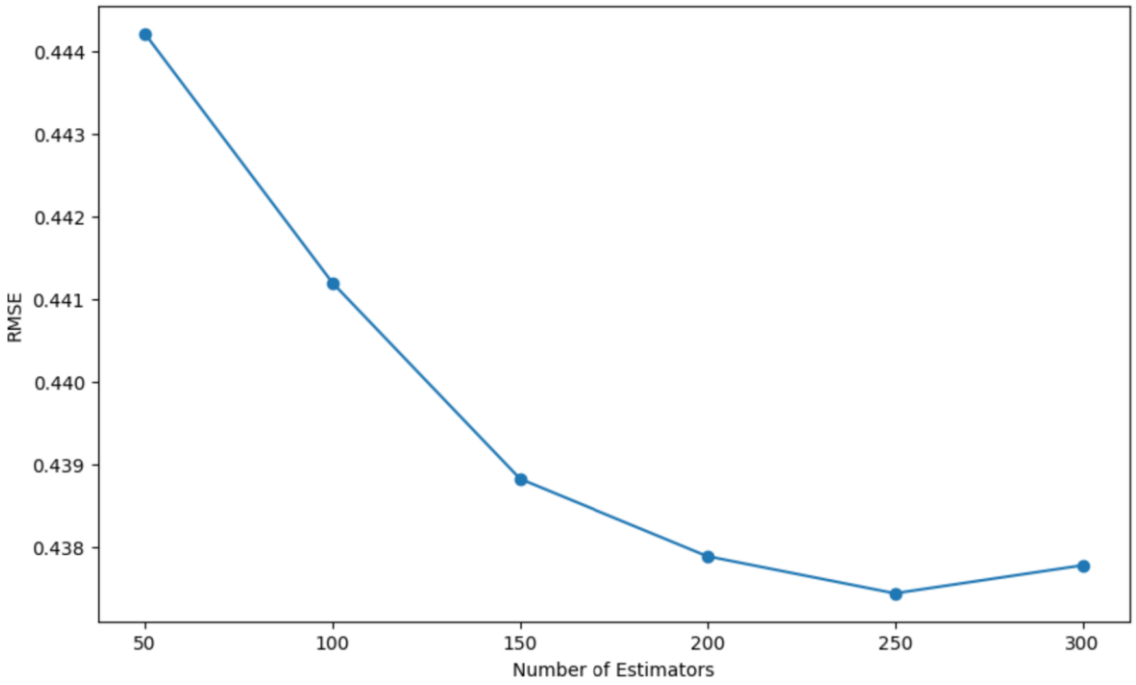

Figure 0.10: Random Forest tuning. Source: Author

To analyse the random forest performance, we look at the evaluation metrics on the test set. This is shown in Table 0.5, where the feature selected dataset still shows a slightly high error rate, and a drop in its model performance. The 'df_all_features' dataset model training was used to analyse the estimated vs actual silica content, and the average error rate of 0.3821 RMSE.

Dataset	RMSE	MAE	R2 Score
df_all_features	0.3821	0.2848	0.8544
df_important_features	0.4049	0.3027	0.8365

Table 0.5: Random Forest model tuned results

Figure 0.11 shows the comparison of the silica's (SiO_2) actual test values indicated in the blue line and the estimated values indicated in the orange line. This graph indicates the test and estimation lines closely following each other, demonstrating that a good model performs due to the similar fluctuations and peaks across the two lines. This indicates the ability of the model to capture both

short-term and long-term trends. However, there are few instances where there are deviations between the estimated and actual values.

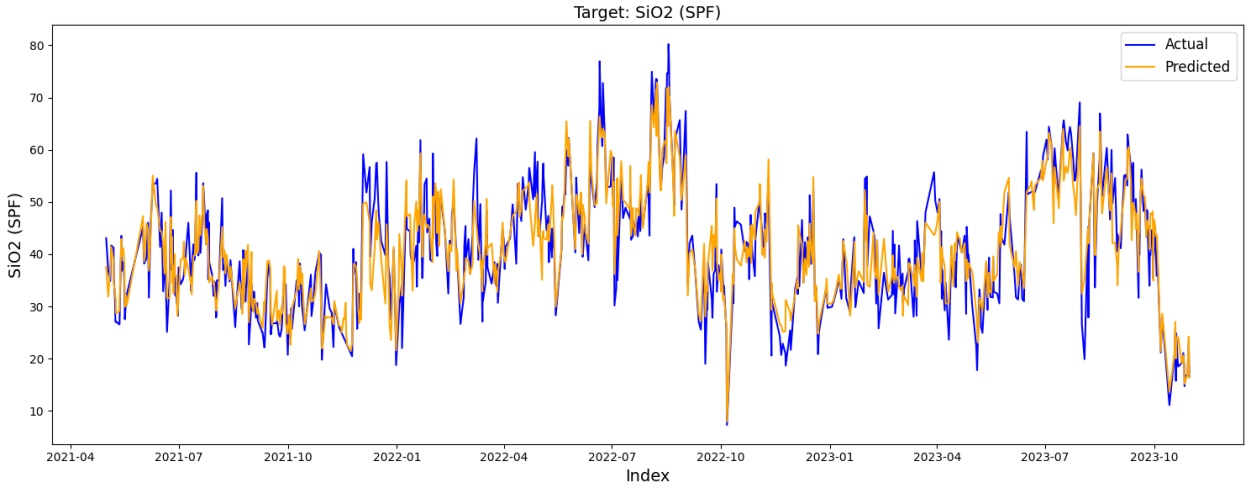

Figure 0.11: Evaluation of RFR Model Using `df_all_features` Dataset. Source: Author

Random Forest model displays a strong performance, coupled with its ability to handle non-linear relationships and provide feature importance scores, makes it an excellent choice for silica estimation.

4.1.6(d) Results of Extreme gradient boosting (XGboost) This section covers the xgboost model results when applied to df_data and df1_data, which covers the influence of feature selection, is observed. It also includes the optimal parameters selected through the GridSearchCV, and the estimations vs test plot.

GridSearchCV was performed with a combination of 243 parameters which was evaluated across all possible combinations with a 5-fold cross validation, resulting to a total of 1215 fits. The 2 dataset we applied giving the same optimal parameters for learning_rate, max_depth, n_estimators and subsample as follows.

- **learning_rate:** 0.05
- **max_depth:** 7
- **n_estimators:** 150

- **subsample:** 0.9

However, the colsample_bytree for df_all_features is 0.9 while for df_important_features is 1. The CV MSE indicates that the df_all_features has a lower error rate of -0.1896 compared to the -0.1659 for df_important_features. The model's evaluation based on df_all_features and df_important_features dataset optimal parameters was used to evaluate the model's performance. The results are shown in Table 0.6 below.

Dataset	RMSE	MAE	R2 Score
df_all_features	0.3851	0.2901	0.8521
df_important_features	0.4061	0.3081	0.8355

Table 0.6: Extreme Boost model's evaluation results

Figure 0.12 shows the xgboost models' evaluation, focusing on the target test dataset and the model's estimation. The model used is trained on the df_data as it yields better performance compared to df1_data. Additionally, this graph indicates the test and estimation line closely following each other, demonstrating that a good model performs due to the similar fluctuations and peaks across the two lines. This indicates the ability of the model to capture both short-term and long-term trends. However, there are instances where there are deviations between the estimated and actual trends.

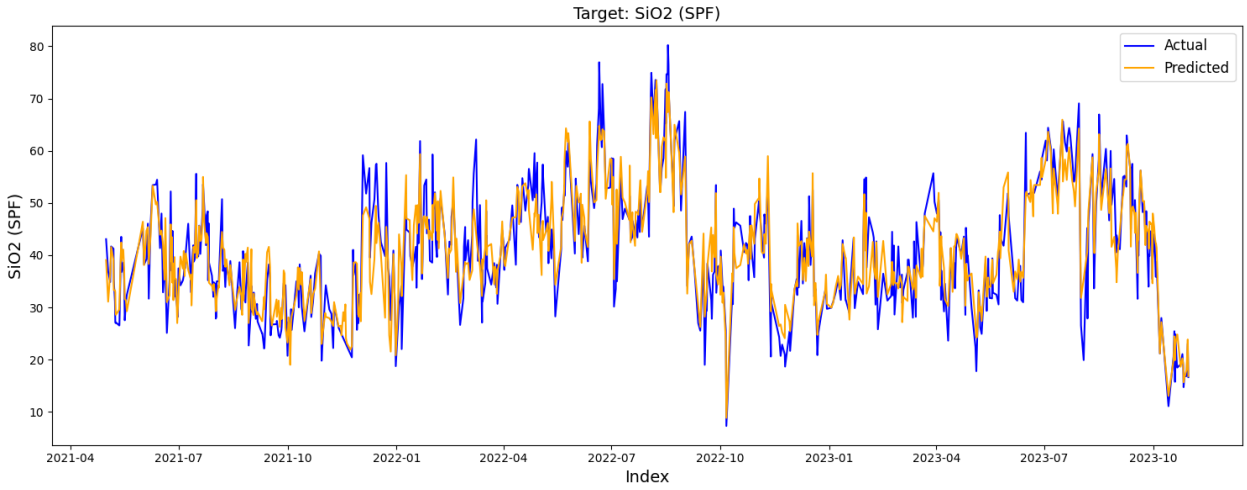

Figure 0.12: Evaluation of xgboost Model Using `df_all_features` Dataset. Source: Author)

4.1.7 Traditional method vs machine learning approach analysis

This section covers the comparison of the insights obtained from the analysis traditional approach and the application of machine learning techniques to estimate the silica content.

4.1.7(a) Tradition Method Results

The traditional method results in looking at the silica content measurements measured at a six-hour interval. They are presented in a time series plot, as illustrated in Figure 0.13.

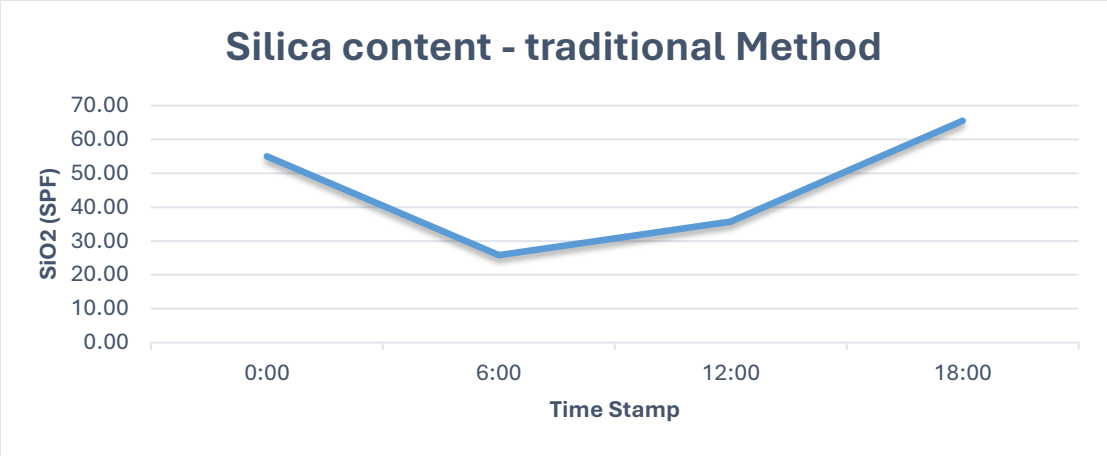

Figure 0.13: Silica content estimation- traditional method. Source: Author

This shows that silica content in the feed dropped from 55.05% from mid-night to 25.81% at 06:00, which then started picking up from 35.75% to 65.51% at midday and 18:00 respectively. This is an indication of an increasing silica content in the processing circuit, which negatively influences the extraction of valuable minerals from the ore and operational costs.

4.1.7(b) Machine Learning Model Results

Random Forest Regressor outperformed all the models used in this study; hence, it was applied to estimate the silica content. The estimation generated by RFR shows a detailed insight into the silica content, producing a finer granularity in the data with an 85.44% R^2 score. Estimations between the traditional method are plotted, giving a deeper insight into the silica content trends as illustrated in Figure 0.14. During the first and second phase (00:00 to 06:00 and 06:00 to 12:00) majority of the silica content was above 40%, while the third phase (12:00 to 18:00) demonstrated a low silica content level.

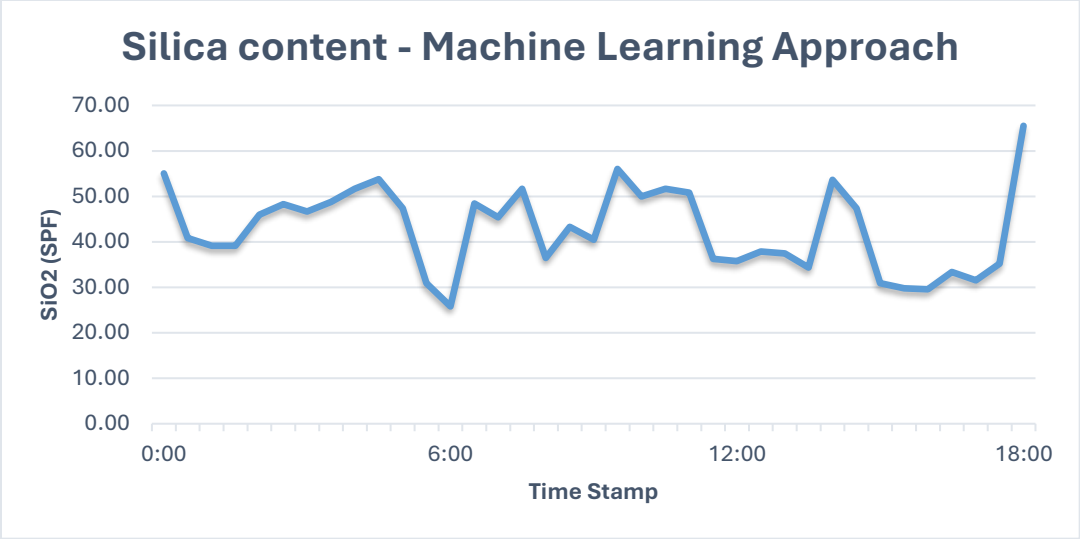

Figure 0.14: Silica content estimation - Machine learning approach. Source: Author

Chapter summary

This chapter presented the experimental results of the study. The data preprocessing steps ensured that the dataset was suitable for model training. The models were evaluated using various metrics, and it was found that the full dataset generally provided better predictive performance than the feature-selected dataset. Among the models, Random Forest showed the most promise for estimating silica content, demonstrating its ability to handle complex, non-linear relationships in the data as indicated in Table 0.7 below.

Dataset	RMSE	MAE	R^2 Score
MLR	0.5544	0.4094	0.6934
SVR	0.3932	0.2884	0.8458
RFR	0.3821	0.2848	0.8544
XGBoost	0.3851	0.2901	0.8521

Table 0.7: Model performance evaluation summary.

CHAPTER FIVE: DISCUSSION OF FINDINGS

This study sets out to apply a regression Machine Learning (ML) based approach for estimating silica content in the ore feed. A readily available and accurate estimation of silica content in the feed is crucial as it directly impacts the overall efficiency of mineral processing operations, particularly during the comminution phase. Accurate silica estimation is crucial in mineral processing, as it offers transformative opportunities in this area through the enhancement of traditional methods with more precise and timely estimation of mineral compositions, particularly for advanced mineral processing stages like flotation (Szmigiel et al., 2024). The estimation models contribute to data-driven decisions, which will optimise the overall mineral process, reducing operational costs and enhancing recovery rates at Rosh Pinah Zinc Mine. To achieve this goal, the following objectives were defined:

I. To identify and engineer relevant features that impact silica content, based on the mineral composition.
II. To evaluate various machine learning algorithms namely Multiple Linear Regression, Random Forest Regression, Support Vector Regression and XGBoost, to determine the effective model for silica content estimation.
III. To evaluate how silica content estimation in the ore feed can enhance the efficiency of advanced mineral processing stages.

5.1 Assumption

The dataset of 5967 entries obtained from Rosh Pinah Zinc Mine archives for a period of five years (2019-2023) is assumed to be accurate and representative of the overall ore feed variability. It is further assumed that the correlation between the mineral compositions and silica is consistent, due to the uniqueness of the ore. This ensures stability and consistency in the estimation of silica content from dynamic mineral composition contents. The acquisition of the silica content estimation can be integrated into the existing mining operations for timely silica estimations, which will ensure an immediate actionable insight for ore blending controls. Ultimately, the study assumes that the selected model can adapt to and account for specific mineralogical characteristics of the ore for reliable and accurate estimations.

5.2 Identifying and engineering features that impact silica content

The first objective focuses on identifying and engineering relevant features that impact silica content, which are mineral composition and the ore's geological characteristics. Through a systematic review and subsequent data analysis on the assay recordings, this study assessed the influence of mineral composition and geological attributes on the silica content in the feed. The ore domain, which includes lithology and mineral composition, indicates the direct correlation between the silica levels and the associated minerals. Feature engineering helps enable a transformed dataset (variable), which best represents the hidden pattern that affects silica content in the ore feed.

By examining these features, this study set the foundation for selecting a high-accuracy predictive model. Feature selection is critical for the enhancement of models' performance, and related studies like Osei (2019) state that the robustness and efficiency of a model can be boosted by dimensionality reduction, where the identification and extraction of the most significant features is applied. Considering the results, although the selection of relevant features did not yield the anticipated outcomes, it should not be overlooked, as feature selection generally improves model performance and contributes to developing robust models. This ensures a full focus on highly related features to the target variable for the predictive tasks. The model performance includes an evaluation of feature selected features and the application of all features. This insight focused on the influence of variable selection and the performance of the models. This involved the evaluation of neglected minerals that have a low correlation rate, which could influence the estimation accuracy. Similarly, studies such as Li et al. (2024) highlighted the influence of feature selection in non-linear geological datasets. Additionally, the inclusion of both less and high-essential features during model training ensured a high-level evaluation on the feature selection in mineral processing applications, as demonstrated by Harsha et al.'s (2021) study. Feature importance is shown in Chapter 4, Figure 0.6.

5.3 Evaluating Machine Learning (ML) Algorithms for Silica Content Estimation

The second objective involved assessing the selected ML algorithms: Multiple Linear Regression, Random Forest Regression, Support Vector Regression, and XGBoost, to identify the high-performing model for estimating silica content in the ore feed. Through experimentation and model

tuning, each of the 4 models was trained and tested on both the full dataset and a feature-selected dataset based on their level of importance. The evaluation metrics included RMSE, MAE, and R^2 score, revealed that Random Forest Regression consistently outperformed the other models, which indicates its ability to handle complex, non-linear relationships in the data. Additionally, this indicates that the estimation of silica content involves correlation between features that is so complicated for simpler linear models, such as Multiple Linear Regression. The inherent variability was handled very well by Random Forest Regression, indicated by its superior performance in terms of low error rates and model's accuracy compared to the other models that were evaluated.

This observation has a huge implication for the mining industry as it handles unique and complex ore compositions. Ensembled models like Random Forest Regression ensure accurate estimation of silica content, which helps with prompt reliable decision making. This also encourages further experiments on the application of machine learning techniques to refine predictive accuracy.

5.3.1 Analysis of the models' performance

Multiple Linear Regression (MLR): MLR performed the lowest compared to other models, acquiring high RMSE and MAE values with low accuracy (R^2 scores). The non-linearity of the assay data has an impact on this model's performance due to its inherent shortcomings in capturing non-linear trends between the dependent and independent variables. This correlation matrix, shown in chapter 4 (Figure 0.5), shows that the features exhibit weak to moderate correlation with the target variable (SiO_2 (SPF)), which ranges between 0.08 to -0.54. Features with moderate correlations like Zn (SPF), Mn (SPF) and S (SPF) indicate some predictive capacity, however, they are more likely to be suppressed by non-linear interaction among the features, which influence MLR performance capability. Additionally, the high inter-feature correlations among features like Mn-Ca: 0.85, Mg-Ca: 0.95, Mg-Mn: 0.83, might have confused the model due to the challenge of effectively disentangling each feature's contribution toward the silica estimation. Though the correlation matrix analysis the linear relationship, it omits the non-linear dependencies. This excludes the feature's non-linear impact on the estimation, which MLR fails to capture. Therefore, the combination of low linear correlation, inter-feature among estimators and the model's inability to capture non-linear trends contributes to MLR's low performance. This finding is like a study by Li et al. (2024), which highlighted similar challenges of linear regression models when applied to

non-linear geological datasets. Similarly, Auret and Aldrich (2012) also indicated that even though linear regression is routinely used in the mineral processing phases, its performance is negatively affected by the linearity of the variables.

Support Vector Regression (SVR): SVR models demonstrated a moderate performance, achieving fair RMSE and MAE values, with an R^2 score better than MLR's model. However, this outcome differs from those of Li et al. (2024), which indicates that SVR outperforms ensemble methods in the hydrocardon geological dataset. Though SVR can handle non-linear trends effectively with the aid of kernel functions, its performance was likely reduced by the nonlinearity of the dataset. As of MLR, there is an indication of some estimation's potential, but the low correlation can still influence the performance of the model due to the model's inability to effectively distinguish each feature's contribution. Even though SVR has a theoretical advantage of capturing complex trends, non-linear relationships, its moderate performance is based on the dataset feature interactions and optimal tuning of hyperparameters like C, epsilon and gamma.

Random Forest Regressor (RFR): The RFR model demonstrated a robust performance, acquiring low RMSE and MAE values together with a high R^2 score, which indicates and justifies its estimation power. The model's capability to capture non-linear relationships and multicollinearity among features boosts its success in silica content estimation. This gives it an advantage over other models, as RFR can handle complex dependencies and still thrive through weak to moderate linear correlation (as they are shown in Chapter 4, Figure 0.5). Moreover, a fundamental aspect of RFR's strength lies in its ensemble nature through averaging estimations from multiple decision trees. This reduces overfitting and improves the model's ability to generalise better to unseen data by capturing diverse patterns. Additionally, the ensemble approach spreads out the learning process across multiple trees, where it focuses on each feature's subset and untangles its impact on silica estimation. Therefore, RFR's ability to adapt to non-linearity, ensemble capability and noise resilience underscores the reason why it performed better than MLR, SVR and XGboost. This observation aligns with studies such as Auret and Aldrich (2012) and Srinija et al. (2022), which highlighted that the robustness of RFR models in capturing the importance of each variable, even with datasets that are subjected to high noise. However, these

results differ from those of Joel and Maliwatu (2024), who found K-Nearest Neighbour (KNN) outperforming RFR and SVR.

Extreme gradient boosting (XGboost): XGboost also performed exceptionally well, achieving low RMSE and MAE values alongside a high R^2 score. This makes it one of the top-performing models, slightly behind RFR in terms of the estimation evaluations. XGboost acquired marginally higher RMSE (0.3851 vs 0.3821), MAE (0.2901 vs 0.2848) and slightly lower R^2 score (0.8521 vs 0.8544) as shown in Chapter 4, Table 0.7. The gradient boosting framework empowers the model's estimation ability through incremental construction of multiple trees to effectively correct residual errors and nonlinear trends. XGboost has demonstrated a strong ability to handle data structures' complexity, non-linearity and multicollinearity with remarkable efficiency. However, the ensemble nature of RFR provides a slight edge on this study's silica content estimation. Despite this, XGboost remains one of the strongest-performing models, which offers robust estimation capabilities.

5.3.2 Estimating silica content based on the high-performing model

Estimating silica content based on the high performing model, was achieved by deploying the Random Forest Regression model. Apart from it delivering the lowest error rates, it also provides the highest R^2 score among the tested models. Additionally, this shows the robustness of the model in term of accuracy and the ability of precise generalizing to unseen data. Practically, this mine can rely on these estimations due to the model's reliability and accuracy for silica content estimation in the ore feed without the limited traditional sampling and analysing intervals.

Deploying this model in a practical setting of Rosh Pinah Zinc Mine will lead to greater achievement in terms of mineral processing efficiency. This model's estimation accuracy rates offer a tangible benefit in terms of ore blending control to maintain a consistent ore feed to the advanced processing plant stages. As it will allow a data-driven adjustment to the blending ratio and processing parameters based on the present silica content levels. These underscore the significance of implementing Machine Learning to enhance efficiency in mineral processing.

The findings of this dissertation indicate the ability of a machine learning model to estimate the silica content in the ore feed, with an accuracy of 85% (Random Forest Regression model). This

has posed a significant implication in the mining industry. Additionally, the adaptation of predictive machine learning models has demonstrated the ability to enhance mineral processing operations, reduce operational costs and improve recovery rates. By enhancing the estimation precision, mining operations can attain steady and efficient processing, which eventually leads to increased profitability and sustainability. A study by Chen et al. (2023) demonstrated that the high silica content is directly proposition to comminution energy consumption. This strengthens the potential benefit of a timely silica estimation on the overall operational benefits, especially mineral processing phase. The evolving mining industry requires the integration of advanced techniques, such as the application of AI and ML will be critical in maintaining effectiveness and addressing challenges associated with traditional mining practices. Furthermore, given the imminent challenges caused by the exhaustion of good-quality deposits, deepening of ore, and the approaching thermodynamic limits, it is imperative to improve the efficiency of mineral operations to mitigate rising production costs and ensure a sustainable supply of essential raw materials (Vidal et al., 2021).

5.4 Accessing how silica content estimation can improve the mineral processing efficiency

Determining the silica content before the feed enters the processing stages plays a crucial role in the overall mineral processing efficiency, as it has a direct impact on the ore blending controls, comminution, flotation and metals recovery. An informed estimation of the silica content in the ore feed will guide the plant operators to make data-driven adjustments to the ore blending ratios, which will ensure a stable feed entering the plant. A consistent feed to the plant ensures a stable processing circuit, which minimises spikes that lead to equipment wear and tear, energy consumption and excessive grinding time.

The integration of the Random Forest Regressor model provides a finer and continuous insight into the silica content behaviour, enhancing more informed and data-driven decision making. The model's high estimation ability (85.44% R^2 score) offers an advantage over the traditional method (six-hour interval assays) in terms of the silica content trends. As shown in Chapter 4, Figure 0.13, the traditional method misses the critical variations within the laboratory tests as in the first phase (00:00 to 06:00), there is a decline in silica levels, but the finer details are not displayed. In contrast, the application of RFR provided a finer insight, which can effectively guide the responses to treat

the silica that has entered the circuit and proactively manage silica about to enter the circuit. This enables a dynamic adjustment of the ore blending ratios, ensuring a consistent ore feed. Additionally, RFR ensure the identification of silica content spikes, which is crucial in mitigating potential issues in advanced upstream processes.

High silica levels are known for increasing ore hardness, which can be mitigated in various ways:

- **Blending ratio adjustment:** Ore blending ratio can be adjusted to low the silica concentration by increasing the low silica ore in the blend. This will help lower silica entering the circuit.
- **Milled stockpile adjustment:** Same with blending ratio adjustment, the milled stockpile can be optimised to lower silica content that enters the milling circuit, minimising liner wear.
- **Milling rate modification:** The milling rates can be reduced due to the increased hardness caused by the presence of high silica ore.
- **Grinding media optimization:** Introduction of additional grinding media enhance ore size's reduction, maintaining the milling rates despite an increase in energy consumption.

Moreover, operational stability ensures optimal silica levels due to its impact on the metal's concentration and recovery. High silica content leads to spikes in the crushing circuit due to its hardness, which affects the recovery rates, wearing of liners and increases energy consumption. Accurate silica estimation allows the grade control to timely adjust the ore blending ratio to drop the silica level, which not only enhances the comminution process but also reduces operational costs. This translates into increased profitability through enhanced mineral processing efficiencies and reduces the mineral processing operational costs.

5.5 Limitations and future work

Despite the models' impressive performance, this study faced several drawbacks. The main limitation of this dissertation is the model's performance evaluation based on data from Rosh Pinah Zinc Mine. Despite providing valuable insights, the models' evaluation limits the generalisability to other mines. Additionally, missing data was handled using linear interpolation, which may introduce errors if the missed data is non-random. Furthermore, this study exclusively focuses on

silica content estimation, leaving out other impurities and by-products that also impact mineral processing efficiency.

Even with these limitations, this study demonstrated the capability of implementing machine learning within the mineral processing operations due to its promising insights on silica content estimation. Future research could explore evaluating datasets from diverse mining contexts, including complex feature selection techniques and expand the estimation scope to inclusion of by-product minerals and impurities that influence the overall mineral processing efficiency.

CHAPTER SIX: CONCLUSION

This dissertation demonstrates the empirical possibility of developing a machine learning algorithm using a 5-year dataset from the laboratory ore feed assay at Rosh Pinah Zinc Mine, which was assessed in this study. The study aimed to apply a data-driven model which estimates the percentage of silica content in the ore feed before entering the advanced stages of ore beneficiation. In addition, the study sought to identify the significant variables (minerals) that influence silica content in the ore feed, thereby optimising the mineral processing operation.

All 10 features (minerals) in the dataset were examined to assess their significance on the predictive task. This was accomplished through feature selection techniques and correlation analysis, ensuring the manageability of the model. Based on feature importance outcome, a refined dataset was created, focusing on the most relevant variables guided by the p-value set at 0.005. Sulfur, Silver and Manganese were observed to be the most significant variables influencing silica content, while Copper, Lead and Magnesium were realised to be less significant. Consequently, 4 machine learning algorithms were applied to the redefined and original dataset for training and selection of the optimal model for silica estimation. The models were:

- Multiple Linear Regression
- Support Vector Regression
- Random Forest Regression
- Extreme Gradient Boosting

The models were evaluated using the three performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-squared (R^2) score. The outcomes from the performance metrics indicate that the Random Forest model consistently outperformed other models regarding the error metrics at 0.3821 RMSE, 0.2848 MAE and accuracy rate at 85.4%. Strong estimation capabilities were also shown from Support Vector Regression and Extreme Gradient Boosting models, with a low performance from Multiple Linear Regression. Random Forest demonstrated a higher accuracy and lower error rates, making it the optimal model for estimating silica content in the ore feed. However, the adaptation of the Random Forest model may still need to be tuned for different mining operations based on the uniqueness of ore compositions.

6.1 Study's contribution

This study makes several contributions to the mineral processing field and mining industry:

I. **Feature importance insight:** This study analysis has identified key variables (sulfur, silver and manganese) that have a high influence on silica content. This provides actionable insights on future ore blending and processing adjustments.

II. **Silica content estimation:** The study successfully applied machine learning on estimating the silica content in the ore feed, which has been seen limited to tradition laboratory assay tests and data acquired during exploration. The superior performance by Random Forest Regressor model demonstrated the ability to estimate silica content in the ore feed, with an R^2 score of 0.8544.

III. **Application to real time decision making:** This research demonstrated the impact of accurately estimating silica content in the ore feed, which shows potential integration of machine learning into operational workflow on ore blending for a consistent feed and improving mineral processing efficiency.

IV. **Advancing methodology in mineral processing:** This study showcases how advance machine learning models can be integrated to the mineral processing field, which paves the way for more applications of machine learning models on different mining industry domains.

6.2 Concluding remarks

This study has demonstrated the estimation capability of the Random Forest Regressor model on a complex, non-linear dataset. It has also shown the transformative potential of machine learning within the mining industry by leveraging machine learning techniques and domain-specific knowledge. This bridges the gap between traditional data analytics and practical usage in mineral processing processes. Additionally, the model's ability to accurately estimate demonstrates the value of implementing data-driven approaches to improve efficiencies, maintain constant ore feed composition and improve mines' sustainability.

Considering this study's findings, it can serve as a blueprint for comparable applications in other mining processes. With ongoing advancements in computational power and the availability of data, the ability to integrate machine learning into the mining industry holds immense promise.

This dissertation not only focused on the model's estimation capabilities but also highlighted opportunities to continue exploring innovations to address modern mining challenges. The path forward is clear: embrace data, enhance machine learning, harness technology, and drive progress to create a more efficient and sustainable future.

REFERENCES

Ali, O. (2022, May 16). *Using artificial intelligence for mineral processing and exploration*. https://www.azomining.com/Article.aspx?ArticleID=1678

An, Z., Zhao, Y., & Zhang, Y. (2023). Mineral exploration and the green transition: Opportunities and challenges for the mining industry. *Resource Policy, 86*(104263). https://doi.org/https://doi.org/10.1016/j.resourpol.2023.104263

Anzoom, S. J., Bournival, G., & Ata, S. (2024). Coarse particle flotation: A review. *Minerals Engineering*, *206*(108499). https://doi.org/https://doi.org/10.1016/j.mineng.2023.108499

Araujo, L. (2020). Case study research: Opening up research opportunities. *RAUSP Management Journal*, 100-111.

Auret, L., & Aldrich, C. (2012). Interpretation of nonlinear relationships between process variables by use of random forests. *Mineral Engineering, 35*, 27-42. https://doi.org/10.1016

Bozkus, E. (2022, December 6). *Multiple Linear Regression (MLR) in Python.* https://eminebozkus.medium.com/multiple-linear-regression-mlr-in-python-cf22b37ea5c5

Chaya. (2020). *Random forest regression.* Level Up Coding.

Chen, Z.-L., Shi, H.-Z., Xiong, C., He, W.-H., Wang, H.-Z., Wang, B., . . . Ge, K.-Q. (2023). Effects of mineralogical composition on uniaxial compressive strengths of sedimentary rocks. *Petroleum Science*, 3062-3073.

Chimwani, N. (2021). A review of the milestones reached by the attainable region optimisation technique in particle size reduction. *Minerals, 11*, 1280. https://doi.org/10.3390/min11111280

Dawood, M. (2023, May 27). *Building predictive models with regression techniques*. https://www.linkedin.com/pulse/building-predictive-models-regression-techniques-muhammad-dawood/

Demir, K. A., Doven, G., & Sezen, B. (2019). Industry 5.0 and human-robot co-working. *Procedia Computer Science*, 688-695.

Dogan, A., Birant, D., & Kut, A. (2019). Multi-target regression for quality prediction in a mining Process. *2019 4th International Conference on Computer Science and Engineering (UBMK)* (pp. 639-644). Samsun: IEEE.

Garbini, S. (2023, June 29). *Mining 4.0: How geoai innovation drives digital transformation*. https://picterra.ch/blog/mining-4-0-how-geoai-innovation-drives-digital-transformation/

González, J. J. (2020). *The impact of classification efficiency on comminution performance and flotation recovery.* Julius Kruttschnitt Mineral Research Centre.

Gruner, H., & Lorig, C. H. (2024, June 15). *Mineral processing*. https://www.britannica.com/technology/mineral-processing

Guven, O. (2022). The effect of shape and roughness on flotation and aggregation of quartz particles. *Physicochemical Problems of Mineral Processing*. https://doi.org/10.37190/ppmp/154021

Haldar, S. K. (2018). *Mineral processing*. https://www.sciencedirect.com/topics/earth-and-planetary-sciences/mineral-processing

Harsha, S. S., Prasad, V., & Raghuveer, K. (2021). *Prediction of silica impurity using deep learning techniques for the mining environment.* REVISTA GEINTEC.

Huynh, T.-M. T., Ni, C.-F., Su, Y.-S., Nguyen, V.-C.-N., Lee, I.-H., Lin, C.-P., & Nguyen, H.-H. (2022). Predicting heavy metal concentrations in shallow aquifer systems based on low-cost physiochemical parameters using machine learning techniques. *International Journal of Environmental Research and Public Health, 19(19)*, 12180. https://doi.org/10.3390/ijerph191912180

Jansen, D. (2023, June). *Research philosophy & paradigms.* https://gradcoach.com/research-philosophy/

Jensen, T., Blakley, I. T., Jacquemin, T., & Patel, A. A. (2018). *Technical report on the Rosh Pinah Mine, Namibia.* Roscoe Postle Associates Inc.

Joel, L., & Maliwatu, R. (2024). Exploring machine learning on geochemistry data. *5th African International Conference on Industrial Engineering and Operations Management.* Johannesburg: IEOM Society International. https://doi.org/10.46254/AF05.20240204

Johnston, R., Ahuja, S., Levesque, M., Lindley, C., McDonald, T., O'Brien, M., . . . Yusufali, F. (2019). *Foundations of AI: A framework for AI in mining.* Global Mining Guidelines Group (GMG).

Kanade, V. (2022, August 30). *What is a support vector machine?* Spiceworks.

Kapadia, S. (2018). *Comminution in mineral processing.* https://doi.org/10.13140/RG.2.2.34991.89760

Kapadia, S. (2018, November). *Comminution in mineral processing*. https://doi.org/10.13140/RG.2.2.34991.89760

Kay, J. (2017, August 2). *Namibia: Miners in wildcat strike*. https://libcom.org/article/namibia-miners-wildcat-strike

Klein, B., Wang, C., & Stefan, N. (2018). Energy-efficient comminution: Best practices and future research needs. *Green Energy and Technology*, 197-211.

Li, Y., Li, X., Guo, M., Chen, C., Ni, P., & Huang, Z. (2024, October). Regression analysis and its application to oil and gas exploration: A case study of hydrocarbon loss recovery and porosity prediction, China. *Energy Geoscience, 5*(4). https://doi.org/100333

Liu, B., Zhang, D., & Gao, X. (2021). A method of ore blending based on the quality of beneficiation and its application in a concentrator. *Applied Science*.

Mapp, M.-R. G. (2023, February 24). *A mining we will go*. https://irishtechnews.ie/a-mining-we-will-go/

Mcleod, S. (2023, December 18). *Qualitative vs quantitative research methods & data analysis*. https://www.simplypsychology.org/qualitative-quantitative.html

Michaux, S. (2020). How to set up and develop a geometallurgical program FINAL v6. *ResearchGate*.

Miller, A., Panneerselvam, J., & Liu, L. (2022). A review of regression and classification techniques for analysis of common and rare variants and gene-environmental factors. *Neurocomputing*, 466-485.

MINING.COM. (2019, June 21). *Grinding down energy consumption in comminution*. https://www.mining.com/sponsored-content/grinding-down-energy-consumption-in-comminution/

Mkurazhizha, T. H. (2018). *The effects of ore blending on comminution behaviour and product quality in a grinding circuit.* Svappavaara.

Montanares, M., Guajardo, S., Aguilera, I., & Risso, N. (2021). Assessing machine learning-based approaches for Silica concentration estimation in Iron Froth flotation. *2021 IEEE International Conference* (pp. 1-6). Valparaíso: IEEE. https://doi.org/10.1109/ICAACCA51523.2021.9465297

Nambinga, V., & Mubita, L. (2021). *The impact of mining sector to the Namibia economy.* National Planning Commission of Namibia.

Nishadini, M. (2019, April 19). *Introduction to machine learning*. https://medium.com/@malsha_nishadini/introduction-to-machine-learning-19539f54d145

Osei, E. K. (2019). *Machine learning-based quality prediction in the froth flotation process of mining.* Dalarna University.

Otchere, D. A., Ganat, T. O., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022, January). Application of the gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering, 208.*

Pirge, G. (2020, December 26). *Comparison of regression analysis algorithms*. https://gursev-pirge.medium.com/comparison-of-regression-analysis-algorithms-db710b6d7528

Pural, Y. E. (2023). *Developing a data-driven soft sensor to predict silicate impurity in iron ore flotation concentrat*e. Physicochemical Problems of Mineral Processing.

Ranstam, J., & Cook, J. A. (2018). LASSO regression. *British Journal of Surgery*, 1348.

Rascón, Á. C. (2024). *Mining 4.0, an evolving industr*y. Gerencia de Riesgos y Seguros.

Rasifaghihi, N. (2023, April 21). *From theory to practice: Implementing support vector regression for predictions in Python*. https://medium.com/@niousha.rf/support-vector-regressor-theory-and-coding-exercise-in-python-ca6a7dfda927

Sahota, N. (2023, 05 15). *AI in mining: Smart excavation with algorithms*. https://www.neilsahota.com/ai-in-mining-smart-excavation-with-algorithms/

Sahu, A. (2023). *Mastering multiple linear regression: Unlocking the power of predictive modelling.* Analytics Vidhya.

Saunders, M. N., Lewis, P., & Thornhill, A. (2019). Understanding Research philosophies and approaches. *Research Methods for Business Students*, 122.

Saunders, M., Lewis, P., & Thornhill, A. (2007). *Research methods for business students.* Financial Times/Prentice Hall.

Setia, M. S. (2016). Methodology series module 3: Cross-sectional studies. *Indiana Journal of Dermatology, 61(3)*, 261-264. https://doi.org/https://doi.org/10.4103/0019-5154.182410

Siddell, K. (2023, October 31). *Computer vision will be essential to the industry 5.0 Era*. https://alwaysai.co/blog/industry5.0

Srinija, K., Nikita, A., & Pushpa, B. (2022). Estimate the percentage of silica in iron ore using the machine. *International Journal & Magazine of Engineering, Technology, Management and Research*, 42-46.

Szmigiel, A., Apel, D. B., Skrzypkowski, K., Wojtecki, L., & Pu, Y. (2024). Advancements in machine learning for optimal performance in flotation processes: A review. *Minerals*, *14*(4), 331. https://doi.org/https://doi.org/10.3390/min14040331

Taherdoost, H. (2021). Data collection methods and tools for research: A step-by-step guide to choosing data collection techniques for academic and business research projects. *International Journal of Academic Research in Management (IJARM)*, 10-38.

Tengli, M. B. (2020, August 27). *Research onion: A systematic approach to designing research methodology*. https://www.aesanetwork.org/research-onion-a-systematic-approach-to-designing-research-methodology/

Thaddeussegura. (2020, September 3). *Multiple linear regression in 200 words*. https://medium.com/@thaddeussegura/multiple-linear-regression-in-200-words-data-8bdbcef34436

Verma, N. (2023, November 2). *An introduction to support vector regression (SVR) in machine learning*. https://medium.com/@nandiniverma78988/an-introduction-to-support-vector-regression-svr-in-machine-learning-681d541a829a

Vidal, O., Boulzec, H. L., Andrieu, B., & Verzier, F. (2021, December 21). Modelling the demand and access of mineral resources in a changing world. *Sustainability*, 11. https://doi.org/https://doi.org/10.3390/su14010011

Vijfeijken, M. v. (2023). *Flowsheet of the future: High-pressure grinding rolls, vertical stirred mill, coarse.* Swiss Tower Mills Minerals.

Viswa. (2023, July 31). *Unveiling decision tree regression: Exploring its principles, implementation*. https://medium.com/@vk.viswa/unveiling-decision-tree-regression-exploring-its-principles-implementation-beb882d756c6

Wang, W., Chakraborty, G., & Chakraborty, B. (2021). Predicting the risk of chronic kidney disease (CKD) using a machine learning algorithm. *Applied science*, 202.

Whitworth, A. J., Forbes, E., Verster, I., Jokovic, V., Awatey, B., & Fox, A. P. (2022). Review on advances in mineral processing technologies suitable for critical metal recovery from

mining and processing wastes. *Cleaner Engineering and Technology, 7*. https://doi.org/10.1016/j.clet.2022.100451

Wikedzi, A., & Leißner, T. (2021). Mineral Liberation: A Case Study for Buzwagi Gold Mine. *Tanzania Journal of Science, 47*. Retrieved from https://dx.doi.org/10.4314/tjs.v47i3.2

Zhang, F., & O'Donnell, L. (2020). Machine Learning. In F. Zhang & L. O'Donnell, *Methods and applications to brain disorders* (pp. 123-140). Academic Press.

Zhang, Z. X., Sanchidrian, J. A., Ouchterlony, F., & Luukkanen, S. (2023). Reduction of fragment size from mining to mineral processing: A review. *Rock Mechanics and Rock Engineering, 56*, 747-778. https://doi.org/https://doi.org/10.1007/s00603-022-03068-3

Zhironkin, S., & Ezdina, N. (2023). Review of Transition from mining 4.0 to mining 5.0 innovative technologies. *Applied Sciences*, 4917.

APPENDIX A: ETHICAL CLEARANCE CERTIFICATE

NAMIBIA UNIVERSITY
OF SCIENCE AND TECHNOLOGY

FACULTY RESEARCH ETHICS COMMITTEE (F-REC)

DECISION/FEEDBACK ON THE RESEARCH PROPOSAL

Dear Angula Tuhafeni Kiiga (212083740)

RESEARCH TOPIC: LEVERAGING MACHINE LEARNING TO ENHANCE EFFICIENCY IN MINERAL PROCESSING AND BLENDING CONTROLS AT ROSH PINAH ZINC MINE.

Supervisor (if applicable): **Dr R. Maliwatu**

Qualification registered for (if applicable): Master of Data Science

(Reference number of applications: **FACULTY RESEARCH ETHICS COMMITTEE REGISTRATION NUMBER: FREC - 15/24)**

Re: Ethical screening application No: **FREC - 15/24**

The Faculty of **Computing and Informatics** Ethics Screening Committee of the Namibia University of Science and Technology reviewed your application for the above-mentioned research. The research as set out in the application has been:

Approved	X

(Indicate with an X, and N/A if not applicable and proceed)

We would like to point out that you, as a researcher, are obliged to maintain the ethical integrity of your research, adhere to the ethical guidelines of NUST, and remain within the scope of your research proposal and supporting evidence as submitted to the F-REC. Should any aspect of your research change from the information as presented to the F-REC, which could affect the possibility of harm to any research subject, you are under the obligation to report it immediately to your supervisor or F-REC as applicable in writing. Should there be any uncertainty in this regard, you must consult with the F-REC.

We wish you success with your research and trust that it will make a positive contribution to the questfor knowledge at NUST.

Any ethical issues that need to be highlighted?	Why are these issues important?	What must/could be done to minimize the ethical risk?
No	N/A	N/A

Recommendation: The application is approved.

Sincerely,

Prof. Suama L Hamunyela
Chairperson: Faculty Ethics Screening
CommitteeTel: +264-61-207-2922
CC: Co-supervisor: None

NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY
P/Bag 13388 Windhoek NAMIBIA
2024 -04- 08
OFFICE OF THE DEAN
FACULTY OF COMPUTING & INFORMATICS

APPENDICE B: MODEL'S IMPLEMENTATION

Multiple Linear Regression (MLR) Model

All the required libraries were installed and imported into the script. The important libraries include '*linear_model'* from scikit-learn where the '*LinearRegression'* class is imported from. This regressor was used to create the linear regression model, which predicts the target variable (SiO2) based on the multiple input features (Zn, Pb, Cu, Fe, Ag, Mg, Mn, Ca and S). Furthermore, '*sklearn.metrics'* module was imported which contains the functions used to evaluate the performance of this regressor model. The three specific functions imported from the 'metrics' submodule are '*import mean_squared_error, mean_absolute_error, r2_score'*.

Defining the Multiple Linear Regressor (mlr) Model

Initializing and fitting the model was achieved through creating an instance '*LinearRegression()*' model which is fitted to the training dataset of the target and features. The '*.fit*' method trains this model using the train dataset. The predictions are calculated using the multiple linear regression algorithm which is define by equation (0.1) below.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots\cdots\cdots + \beta_p x_p + \epsilon \tag{0.1}$$

Assumption made for linear regression equation are also applied on this model:

- Linear relationship: Dependent and independent variable follows a linear relationship
- Multivariate normality: Normally distribution across all variables
- Anti-correlation: Little to no correlation among independent variables
- Homoscedasticity: Equal distribution of residuals across the regression line or plane.

Training the model and making Predictions

Once the model is trained, it is then used to make the predictions by fitting it to the features test datasets. and compares the prediction generated against the actual target values. The '*predict'* method is used to generate the model's predicted values based on the features from the test set.

```
[46] #Multiple Linear Regressor
```

```
[47] #WITH ALL FEATURES

     model_mlr = LinearRegression()
     model_mlr.fit(X_train, y_train)

     # Make predictions on the test set
     y_pred_mlr = model_mlr.predict(X_test)

     # Evaluate the model
     mse_mlr = mean_squared_error(y_test, y_pred_mlr)
     rmse_mlr = np.sqrt(mse_mlr)
     mae_mlr = mean_absolute_error(y_test, y_pred_mlr)
     r2_mlr = r2_score(y_test, y_pred_mlr)
     percentage_accuracy = r2_mlr * 100


     #print
     print(f'RMSE: {rmse_mlr:.4f}')
     print(f'MAE: {mae_mlr:.4f}')
     print(f'R² Score: {r2_mlr:.4f}')
     print(f'Percentage Accuracy based on R²: {percentage_accuracy:.2f}%')
```

Support Vector Regression (SVR)

All the required libraries were installed and imported into the script. Important libraries include '*svm'* from scikit-learn where the support vector regression class is imported from. Hyperparameter tunning tools were installed from the '*sklearn.model_selection*', from which the '*gridSearchCV'* class was imported.

Defining the SVR Model

The Support Vector Regression (SVR) model with a rbf kernel was defined. This kernel specifies the transformation of data making it possible to find a linear relationship in a higher-dimensional space.

Setting Up the Parameter Grid

A parameter grid is defined, where a specified range of hyperparameters are defined and tested during the optimization process. The '*svm.SVR*' class implements SVR, whereby the optimal hyperparameters of kernel function, regularization parameter (C) and coefficient (ε) were determined. The four common kernel functions used include linear, poly (polynomial), radial basic function (rbf) and sigmoid. The parameters are as follows:

C: The commonly used regularization parameter with values is defined, this value determines the level of regularization where a high value implies less regularization. The values are [0.1, 1, 10, 100].

Epsilon: The commonly used epsilon-tube width is defined. This implies the width on which no penalty is awarded to the errors. The values are [0.01, 0.1, 1, 10].

Initializing GridSearchCV:

GridSearchCV is applied to determine the best combination of the defined hyperparameters by performing a comprehensive search across the set grid. The search involves:

- ✓ The SVR model which was defined above.
- ✓ The search over the defined parameter grid.
- ✓ Cross-validation of 5 folds to ensure that the model generalizes well to new data.
- ✓ A negative mean squared error as the scoring metric is used, this is to maintain a consistent use of optimization algorithms that expect a higher score to represent a better model.

Training the model and retrieving the best parameters

- The grid search process is applied to the training dataset. This involves fitting the SVR model with each combination of parameters and evaluating its performance using cross-validation.
- After the grid search is completed, the best parameters that resulted in the highest performance are retrieved. These parameters are considered optimal for the given SVR model and dataset.

Making Predictions

The SVR model is then redefined using the optimal parameters obtained during the grid search. This optimized model is used to make predictions on the test data.

[54]
```python
# Define the SVR model for df_data
svr = SVR(kernel='rbf')

# Define the parameter grid
param_grid = {
    'C': [0.1, 1, 10, 100],
    'epsilon': [0.01, 0.1, 1, 10]
}

# Initialize the GridSearchCV object
grid_search = GridSearchCV(estimator=svr, param_grid=param_grid, cv=5, scoring='neg_mean_squared_error', verbose=2, n_jobs=-1)

# Fit the grid search to the data
grid_search.fit(X_train, y_train)

# Get the best parameters
best_params = grid_search.best_params_
print(f'Best parameters: {best_params}')

# Use the best estimator to make predictions
best_svr = grid_search.best_estimator_
y_pred_svr = best_svr.predict(X_test)

# Evaluate the model
mse_svr = mean_squared_error(y_test, y_pred_svr)
rmse_svr = np.sqrt(mse_svr)
mae_svr = mean_absolute_error(y_test, y_pred_svr)
r2_svr = r2_score(y_test, y_pred_svr)
percentage_accuracy = r2_svr * 100

#print
print(f'RMSE: {rmse_svr:.4f}')
print(f'MAE: {mae_svr:.4f}')
print(f'R² Score: {r2_svr:.4f}')
print(f'Percentage Accuracy based on R²: {percentage_accuracy:.2f}%')
```

Random Forest Regressor (RFR)

All the required libraries were installed and imported into the script. Important libraries include 'RandomForestRegressor' from scikit-learn. Hyperparameter tunning tools were installed from the 'sklearn.model_selection', from which the 'gridSearchCV' class was imported.

Defining the rfr Model

The rfr model was created using the '*RandomForestRegressor*' module which comes from an '*sklearn.ensemble*' library. This model's prediction comes from various decision trees which are built then ensemble the predictions. A random state of 42 is set to ensure the results are reproducible.

Setting Up the Parameter Grid

The parameter grid where the hyperparameter will be tuned on. The possible number of trees '*n_estimator*' are specified. The '*n_estimator*' parameter for this study includes values [50, 100, 150, 200, 250, 300].

Initializing GridSearchCV:

GridSearchCV is applied in determining the best combination of the specified hyperparameters by performing a comprehensive search across the various number of trees. The search involves:

- ✓ The RandomForestRegressor '*estimator*' model which is initialized and tuned.
- ✓ The search over the defined parameter grid.
- ✓ Cross-validation of 5 folds to ensure that the model generalizes well to new data.
- ✓ A negative mean squared error as the scoring metric is used, this is to maintain a consistent use of optimization algorithms that expect a higher score to represent a better model.

Retrieving the best parameters and training the model

- The grid search process is applied to the training dataset. This involves fitting the rfr model with each combination of parameters and evaluating its performance using cross-validation.
- After the grid search is completed, the different outputs are assessed and the best parameters that resulted in the highest performance is retrieved.

Making Predictions

The rfr model is then redefined using the optimal parameters obtained during the grid search. This optimized model is used to make predictions on the test data.

```
# Parameter grid for GridSearchCV
param_grid = {
        'n_estimators': [50, 100, 150, 200, 250,300],
    'max_depth': [None, 10, 20, 30, 40, 50],
}

# Initialize Random Forest regressor
rfr = RandomForestRegressor(random_state=42)

# Perform GridSearchCV
grid_search_rf = GridSearchCV(estimator=rfr, param_grid=param_grid, cv=5, scoring='neg_mean_squared_error')

# Fit the grid search to the data
grid_search_rf.fit(X_train_top, y_train)

# Get the best parameters
best_params = grid_search_rf.best_params_
print(f'Best n_estimators: {best_params["n_estimators"]}')
print("Best CV RMSE:", np.sqrt(-grid_search_rf.best_score_))

# Use the best estimator to make predictions
best_rf = grid_search_rf.best_estimator_
y_pred_rf = best_rf.predict(X_test_top)

# Evaluate the model
mse_rf = mean_squared_error(y_test, y_pred_rf)
rmse_rf = np.sqrt(mse_rf)
mae_rf = mean_absolute_error(y_test, y_pred_rf)
r2_rf = r2_score(y_test, y_pred_rf)
percentage_accuracy = r2_rf * 100

# Print evaluation metrics
print(f'RMSE: {rmse_rf:.4f}')
print(f'MAE: {mae_rf:.4f}')
print(f'R² Score: {r2_rf:.4f}')
print(f'Percentage Accuracy based on R²: {percentage_accuracy:.2f}%')
```

XGBoost (Extreme Gradient Boosting)

All the required libraries were installed and imported into the script. Important libraries include 'XGBRegressor' from xgboost class. Hyperparameter tuning tools were installed from the 'sklearn.model_selection' module, from which the GridSearchCV was imported.

Defining the XGBoost Regressor (xgb) Model

The xgb model is created using '.*XGBRegressor*', which comes from the xgboost library. This model builds gradient boosted decision trees and ensembles their predictions. A random state is set to 42, which ensures the results are reproducible.

Setting Up the Parameter Grid

The parameter grid specifies the hyperparameters to be tuned, the possible values are defined, as illustrated in below. This study's parameter grid includes:

- ✓ The adjustment on how much the model learns per boost step. Learning_rate values set at [0.05, 0.1, 0.15].
- ✓ The specified maximum depth per tree. The value of max_depth set at [3, 5, 7]
- ✓ The number of tress or boosting stages used. The value of n_estimators set at [50, 100, 150]
- ✓ The portion of samples used per tree during training. The value of subsample set at [0.8, 0.9, 1.0]
- ✓ The portion of features used per tree during training. colsample_bytree: [0.8, 0.9, 1.0]

```
[54] param_grid = {
        'learning_rate': [0.05, 0.1, 0.15],
        'max_depth': [3, 5, 7],
        'n_estimators': [50, 100, 150],
        'subsample': [0.8, 0.9, 1.0],
        'colsample_bytree': [0.8, 0.9, 1.0],
    }
```

Initializing GridSearchCV

GridSearchCV is applied to help determine the ultimate combination of the specified hyperparameters by performing an extensive search, the code snippet is shown below. The search involves:

- ✓ Tuning the XGBRegressor model, which is defined above.
- ✓ Searching over the specified parameter grid.
- ✓ Performing 5-fold cross-validation to ensure the model generalizes well to new data.
- ✓ Using negative mean squared error as the scoring metric to maintain consistency with optimization algorithms that expect a higher score to represent a better model.

```
[60] #all feature

     # Initialize XGBoost regressor
     xgb_model = xgb.XGBRegressor(objective='reg:squarederror', random_state=42)

     # Initialize GridSearchCV
     grid_search1 = GridSearchCV(estimator=xgb_model, param_grid=param_grid,
                                 cv=5, scoring='neg_mean_squared_error', verbose=1)

     # Perform Grid Search CV on training data
     grid_search1.fit(X_train, y_train)

     # Print the best parameters and best score
     # Changed from grid_search to grid_search1 to reflect the fitted object
     print(f'Best parameters: {grid_search1.best_params_}')
     print(f'Best CV MSE: {grid_search1.best_score_}')
```

Retrieving the Best Parameters and Training the Model

- The grid search process is applied to the training dataset. This involves fitting the XGB model per combination of the specified parameters and evaluating thier performance using cross-validation.
- After the grid search is completed, the best parameters that resulted in the highest performance are retrieved.

Making Predictions

The xgb model is then redefined using the optimal parameters obtained from the grid search. This optimized model is used to make predictions on the test data, the code snippet is shown below

```
[62] #All features
     # Predict using the best model
     best_modelXB = grid_search1.best_estimator_
     y_pred_gbr = best_modelXB.predict(X_test)

     # Evaluate the model
     mse_gbr = mean_squared_error(y_test, y_pred_gbr)
     rmse_gbr = np.sqrt(mse_gbr)
     mae_gbr = mean_absolute_error(y_test, y_pred_gbr)
     r2_gbr = r2_score(y_test, y_pred_gbr)
     percentage_accuracy = r2_gbr * 100


     #print
     print(f'RMSE: {rmse_gbr:.4f}')
     print(f'MAE: {mae_gbr:.4f}')
     print(f'R² Score: {r2_gbr:.4f}')
     print(f'Percentage Accuracy based on R²: {percentage_accuracy:.2f}%')
```