

https://dspace.org/

A machine learning approach for assessing sedimentological data's potential for in-situ gold grade prediction in the Witwatersrand basin, South Africa

Mombe, Tariro

2024
https://hdl.handle.net/10539/37438

Downloaded from a DSpace repository. https://dspace.org/

A MACHINE LEARNING APPROACH FOR ASSESSING SEDIMENTOLOGICAL DATA'S POTENTIAL FOR IN-SITU GOLD GRADE PREDICTION IN THE WITWATERSRAND BASIN, SOUTH AFRICA

Tariro Mombe [1612442]

A Dissertation submitted to the

School of Geosciences, University of the Witwatersrand

in fulfilment of the requirements for the degree of

Master of Science in Geology

Supervisor: Prof. Glen Nwaila

Co-supervisor: Dr. Steven Zhang

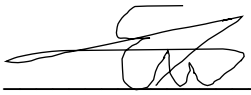
16/03/2023



STATEMENT BY THE AUTHOR

I Tariro Mombe (1612442), am a student registered for GEOL8003A in the year 2023. I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/ or without acknowledging the original source) is wrong.
- I confirm that the work submitted for the assessment for the above course is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  Date: 16/ 03/ 2023

ABSTRACT

Mineral resource estimation is one of the most crucial factors in the mining industry because it extracts information about the quantity and quality of a mineral deposit. Mining operations have traditionally taken a purely geostatistical approach to resource estimation; however, the global technology boom has significantly improved estimation techniques. Machine Learning (ML) has only recently proven relevant in geoscience. Several studies have alluded to it being a revolutionary resource estimation tool with the potential to improve overall prediction accuracies. ML enables the incorporation of geological properties (e.g., lithology, grain size); consequently, resource estimation models produced with geological context are expected to be of higher accuracy. This research aimed to investigate how incorporating sedimentological properties enables in-situ gold grade prediction in the Witwatersrand Basin.

Sibanye-Stillwater provided a gold assay database comprising sample information collected from their Kloof mining operation (South Africa). This dataset was refined and used to produce ML models; geostatistical (kriging) models were later produced for comparison. The experimental component of this dissertation was programmed in Jupyter Notebook using the Python programming language. The Random Forest (RF), AdaBoost, K-Nearest neighbours (KNN) and Elastic Net (EN) were the machine learning algorithms (MLAs) of choice for this study. The original dataset comprised 55 304 datapoints, with sample properties including spatial coordinates, four sedimentological properties (Channel Width, Internal Waste, Conglomerate Percentage and Basal Contact), and the gold grade property. In this study, a sample area was selected; consequently, the final dataset used for predictive modelling comprised 10 682 datapoints (19.31% of original).

The predictive modelling results proved that it is possible to produce high-accuracy mineral resource models when predicting gold grades using MLAs; however, the quality of the data plays a significant role in the extent to which sedimentological properties can be useful in this regard. An extensive quality assurance and quality control (QAQC) workflow had to be implemented to attain good predictions from what was originally a very poor model. Investigations into the spatial distribution of the well-predicting and poor-predicting datapoints revealed a spatial pattern in the ‘bad’ dataset while the ‘good’ dataset had none, suggesting that the poor-predicting samples may have been extracted off-reef. Analysing the differences between the ‘good’ and ‘bad’ datasets thus revealed a prominent data quality issue that may have resulted from errors during sampling. Overall, this study has produced a resource estimation methodology that can be replicated and used on any gold deposit in the Witwatersrand Basin, where similar types of data exist. Therefore, similar deposits in other locations can also use the same methodology.

ACKNOWLEDGEMENTS

- Supervisors

Firstly, I would like to thank my supervisors, Prof. Glen Nwaila and Dr. Steven Zhang. I genuinely appreciate your support and dedication throughout the course of this research. I was never left to fend for myself and consistently received constructive feedback. Thank you for encouraging me to appreciate the small milestones through all the research stages and seeing me through the confusion and obstacles.

- Sibanye-Stillwater

I would like to thank Phillip Staley (Project Geologist) from Sibanye-Stillwater for providing the sample data used in this research. This study would not have been possible without you.

- Funders

I would also like to thank the NRF (National Research Fund) and CIMERA (Centre of Excellence for Integrated Mineral and Energy Resource Analysis) for funding my master's studies; without their help, I would not have been able to further my tertiary studies.

- Friends and Family

Finally, I would like to express my sincere gratitude to my family and friends for their overwhelming support throughout my tertiary studies. To my parents, thank you for being my pillar of strength and always showing me unconditional love. To my friends, Gabby and Khutsiso, thank you for providing a shoulder to lean on during those difficult times; I could not have stayed sane without you.

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES xi

ACRONYMS AND ABBREVIATIONS i

1. INTRODUCTION 1

1.1. Motivation 1

1.2. Aims and Anticipated Outcomes 2

1.3. Objectives 2

1.4. Data context 2

1.5. Research scope 4

1.6. Thesis outline 4

2. LITERATURE REVIEW 6

2.1. Geological setting 6

2.1.1. Goldfields 7

2.1.2. Kloof operation 10

2.2. Resource Estimation 11

2.2.1. Inverse Distance Weighting (IDW) 12

2.2.2. Kriging 13

2.3. Machine Learning: Introduction 14

2.3.1. Terminology 16

2.4. Machine learning modelling: Application 17

2.4.1. Interpretability 18

2.4.2. Predictive modelling 19

2.4.3. Project management and model development 21

2.4.4. Data understanding 23

2.4.5. Model fitting 24

2.4.6. Model audit 25

3. METHODOLOGY 28

3.1. Dataset 28

3.2. Data pre-processing 29

3.3. Experiment 1 30
3.3.1. Random Forest Algorithm 30
3.3.2. AdaBoost Algorithm 32
3.3.3. K-Nearest Neighbours Algorithm 33
3.3.4. Elastic Net Algorithm 34
3.3.5. Cross-Validation 36
3.4. Experiment 2 36
3.4.1. K-Means clustering 36
3.5. Post-prediction investigations 38
3.5.1. Residuals Analysis 38
3.5.2. Obtaining predictions from the entire 'good' dataset 38
3.6. Assessing the efficiency of sedimentological features in gold grade prediction 39
3.6.1. Feature Selection 39
3.6.2. Feature Importance 39
3.7. Kriging Block Model 39
3.7.1. Ordinary Kriging 39
3.7.2. Validation 41
4. RESULTS 43
4.1. Explanatory Data Analysis 43
4.1.1. Feature Analysis 46
4.1.2. Target Analysis (Au g/t) 48
4.1.3. Correlation Matrix 49
4.2. Full Subset 50
4.2.1. Experiment 1 50
4.2.2. Experiment 2 51
4.2.3. Feature Importance 54
4.3. Post-prediction Investigations 55
4.3.1. 'Good' Dataset 57

4.3.2. ‘Bad’ Dataset 62
4.3.3. RE Investigations 67
4.4. Assessing the importance of sedimentological features in gold grade prediction 72
4.4.1. Feature Selection 72
4.4.2. Feature Importance 73
4.5. Kriging block model 75
4.5.1. Model Validation 77
5. DISCUSSION 80
5.1. Evaluation of the results of experiments 1 and 2 80
5.2. Importance of sedimentological data for gold grade prediction 80
5.3. Importance and reliability of sedimentological data in resource estimation 82
5.4. The effect of data quality on gold prediction 82
5.5. Comparison of geostatistical and ML techniques 83
5.6. Implications of this study for the exploration of Witwatersrand-type gold deposits 85
5.7. Limitations and recommendations 86
6. CONCLUSIONS 87
7. REFERENCE LIST 89
8. APPENDICES 99
APPENDIX A: SILHOUETTE SCORE PLOTS 99
APPENDIX B: TRAINING SCORE PLOTS (EXPERIMENT 2) 101

LIST OF FIGURES

Figure 1.1: Locality map of the study area, included is the location of the Witwatersrand Basin (A); its goldfields (B); its cross-section (C); and its stratigraphic column with the position of the thin Kloof Reef outlined (adapted from Frimmel and Nwaila, 2020). 3
Figure 1.2: Research Scope 4
Figure 1.3: Thesis outline. 5
Figure 2.1: Locality map illustrating the geological setting of the Witwatersrand Basin (Frimmel and Nwaila, 2020). 7
Figure 2.2: Generalised stratigraphic column of the Central Rand Group (adapted from Frimmel et al., 2009). 10
Figure 2.3: Generalised relationship between mineral resources and reserves (*a*dapted from Dumakor-Dupey and Arya, 2021). 11
Figure 2.4: Generalised illustration of the IDW technique, where Zi is the known point and the unknown points are in green. $Z1$would be assigned the highest weight (adapted from Coleman, 2021). 12
Figure 2.5: Simplified illustration of the semi-variogram (adapted from Rossi and Deutsch, 2014 and Coleman, 2021). 14
Figure 2.6: Simplified diagram illustrating the process followed in ML (adapted from Cate et al., 2017 and Samson, 2019). 15
Figure 2.7: Common Supervised Learning algorithms (adapted from Samson, 2019 and Jung and Choi, 2021). 16
Figure 2.8: Common Unsupervised Learning algorithms (adapted from Samson, 2019 and Jung and Choi, 2021). 16
Figure 2.9: A simplified overview of the different predictive modelling processes in typical statistical modelling compared to machine learning. In typical modelling (A), much weight is placed on domain understanding, while in machine learning (B), much focus is placed on the predictive modelling process (adapted from Inglis et al., 2021). 21
Figure 2.10: Flow chart illustrating an iterative variant of the CRISP-DM tool-agnostic predictive modelling procedure. The process is initiated by data preparation and followed by data understanding, model assembly and model audit. These three phases can be iterated. Then finally, the model delivery step concludes the process (adapted from Wickham and Grolemund, 2017 and Hall et al., 2019).. 23
Figure 2.11: Graphical explanation of the CoD metric (adapted from Alencar et al., 2017 and Botchkarev, 2018). 27
Figure 3.1. Simplified sketch of the physical relationship between the Internal Waste (cm) and Conglomerate Percentage. 29
Figure 3.2: Illustration of a decision tree showing the Root, Decision and Leaf Nodes (adapted from Luckert and Schaefer-Kehnert, 2016). 31

Figure 3.3: Simplified sketch illustrating the KNN algorithm's process. The training data's values are represented by red, green and blue, while black represents a value from the testing set. The coordinates are arbitrary feature dimensions. In subfigure A, the distance between the training and testing datapoints is computed. In subfigure B, we assume that k = 3, for example; in this case, 5,1 and 6 will be selected as our point's nearest neighbours based on their proximity (adapted from Kramer, 2013). 34
Figure 3.4: K-Fold Cross-Validation (adapted from Kapageridis, 2005). 36
Figure 3.5: Graphical representation of the K-Means algorithm's clustering process. A K of 2 is assigned, and the clustered datapoints are represented in blue and orange. The median line (dashed) is generated by computing the median between the two centroids (adapted from Tan et al., 2021). 37
Figure 3.6: Samples weighed by variogram in OK estimation (adapted from Coleman, 2021). 40
Figure 3.7: SLOR and bias (adapted from Kentwell, 2022). 42
Figure 4.1: Spatial map illustrating the distribution of the original/raw dataset. Datapoints are colour-coded by their gold grade. The red box marks the subset area that this project will focus on. 44
Figure 4.2: Spatial map illustrating the distribution of the selected subset block. Datapoints are colour-coded by their gold grade. 45
Figure 4.3: Spatial maps illustrating the distribution of the datapoints, colour-coded by individual features. The black circle on the conglomerate percentage map marks an area where some sort of cluster has been observed. 46
Figure 4.4: Boxplot and histogram illustrating the distribution of Channel Width data. 47
Figure 4.5: Boxplot and histogram illustrating the distribution of the Internal Waste data. 47
Figure 4.6: Boxplot and histogram illustrating the distribution of the Percentage Conglomerate data. 48
Figure 4.7: Boxplot and histogram illustrating the distribution of the gold grade data. 48
Figure 4.8: Feature correlation heatmap. 49
Figure 4.9: Graphical illustrations of the full subset's first round of testing for each of the four algorithms. 50
Figure 4.10: Spatial map visualising the full subset's clustered datapoints when K = 3. 52
Figure 4.11: Graphical illustrations of the full subset's second round of testing, visualising each algorithm's CoD scores. 53
Figure 4.12: Spatial map visualising the full subset's clustered datapoints when spatial features are incorporated. 54
Figure 4.13: Graphical illustrations of the RF and AdaBoost algorithms' feature importance statistics for the full subset. 55
Figure 4.14: Graphical illustrations of the RF and AdaBoost algorithms' feature importance statistics for the full subset (including spatial features). 55

Figure 4.15: Spatial distribution of the ‘good’ vs ‘bad’ datasets, represented in orange and blue respectively. 57
Figure 4.16: Graphical illustrations of the good subset’s first round of testing for each of the four algorithms. 58
Figure 4.17: Spatial map visualising the ‘good’ dataset’s clustered datapoints when K = 5. 59
Figure 4.18: Graphical illustrations of the ‘good’ dataset’s second round of testing, visualising each algorithm's CoD scores. 60
Figure 4.19: Spatial map visualising the ‘good’ dataset’s clustered datapoints when spatial features are incorporated. 61
Figure 4.20: Graphical illustration of the ‘good’ dataset’s final randomised prediction model. 62
Figure 4.21: Graphical illustrations of the bad dataset’s first round of testing for each of the four algorithms. 64
Figure 4.22: Spatial map visualising the ‘bad’ dataset’s clustered datapoints when K = 5. 65
Figure 4.23: Graphical illustrations of the ‘bad’ dataset’s second round of testing, visualising each algorithm's CoD scores. 66
Figure 4.24: Spatial map visualising the ‘bad’ dataset’s clustered datapoints when spatial features are incorporated. 67
Figure 4.25: Spatial Map of the full subset colour coded by the RE. 68
Figure 4.26: Scatter plot investigating the relationship between the RE and the gold grade (Au g/t). 69
Figure 4.27: Scatter plot investigating the relationship between the RE and the Conglomerate Percentage. 70
Figure 4.28: Scatter plot investigating the relationship between the RE and the Internal Waste. 71
Figure 4.29: Scatter plot investigating the relationship between the RE and the Channel Width. 71
Figure 4.30: Graphical illustrations of the RF algorithm’s feature importance statistics for the ‘good’ and ‘bad’ datasets. 73
Figure 4.31: Graphical illustrations of the AdaBoost algorithm’s feature importance statistics for the ‘good’ and ‘bad’ datasets. 74
Figure 4.32: Graphical illustration of the RF’s feature importance statistics for the ‘good’ and ‘bad’ datasets when spatial features are incorporated. 75
Figure 4.33: Graphical illustration of the AdaBoost’s feature importance statistics for the ‘good’ and ‘bad’ datasets when spatial features are incorporated. 75
Figure 4.34: Kriging block semi-variogram model, including parameters. 76
Figure 4.35: Block model estimates for the full (a) and ‘good’ (b) datasets. 77
Figure 4.36: KEFF results for the full (a) and ‘good’ (b) datasets. 78
Figure 4.37: SLOR results for the full (a) and ‘good’ (b) datasets. 79

Figure 5.1: Figure illustrating the sample distribution between the full, ‘bad’ and ‘good’ datasets. The RF prediction score for each set is included. 83
Figure 5.2: Factors affecting the choice of algorithm (adapted from Dumakor-Dupey and Arya, 2021). 85
Figure A1: Graph visualising the full dataset’s silhouette scores. 100
Figure A2: Graph visualising the ’bad’ dataset’s silhouette scores. 100
Figure A3: Graph visualising the ‘good’ dataset’s silhouette scores. 101
Figure B1: Graphical illustrations of the full dataset’s second round of training, visualising each algorithm's CoD scores. 102
Figure B2: Graphical illustrations of the ‘bad’ dataset’s second round of training, visualising each algorithm's CoD scores. 103
Figure B3: Graphical illustrations of the ‘good’ dataset’s second round of training, visualising each algorithm's CoD scores. 104

LIST OF TABLES

Table 3.1: RF model hyperparameters..... 31
Table 4.1: Raw dataset illustrating the identification column, spatial coordinates, the predictive properties, and the target property. 43
Table 4.2: Python extract evaluating the features' data types and the occurrence of non-null values. 45
Table 4.3: Descriptive statistics for the full subset. 49
Table 4.4: First round of test scores for the full subset. 50
Table 4.5: Full subset's final hyperparameters. 51
Table 4.6: Second round of test scores for the full dataset. 52
Table 4.7: Second round of test scores for the full dataset when spatial features are incorporated. 54
Table 4.8: CoD test scores according to Relative Error (RE) intervals. 56
Table 4.9: Descriptive statistics for the 'good' subset. 57
Table 4.10: First round of test scores for the 'good' dataset. 58
Table 4.11: 'Good' dataset's final hyperparameters 59
Table 4.12: Second round of test scores for the 'good' dataset. 60
Table 4.13: Second round of test scores for the 'good' dataset when spatial features are incorporated. 61
Table 4.14: Final parameters for the randomised prediction of the 'good' dataset. 62
Table 4.15: Descriptive statistics for the 'bad' dataset. 63
Table 4.16: First round of test scores for the 'bad' dataset. 63
Table 4.17: 'Bad' dataset's final hyperparameters. 64
Table 4.18: Second round of test scores for the 'bad' dataset. 65
Table 4.19: Second round of test scores for the 'bad' dataset when spatial features are incorporated. 67
Table 4.20: 'Good' dataset's feature selection testing scores. 72
Table 4.21: 'Bad' dataset's feature selection testing scores. 72
Table 4.22: Summary statistics of kriging estimates for the full and 'good' datasets 76
Table 4.23: Average SLOR and KEFF results for the full and 'good' datasets 77

ACRONYMS AND ABBREVIATIONS

Au	Gold
CRISP-DM	Cross-Industry Standard Process for Data Mining
DT	Decision Tree
EN	Elastic Net
IDW	Inverse Distance Weighting
KNN	K-Nearest Neighbours
Max	Maximum
Min	Minimum
ML	Machine learning
MLA	Machine learning algorithm
RE	Relative Error
RF	Random Forest
Std	Standard deviation

1. INTRODUCTION

Resource estimation is the imperative foundation upon which all mine plans/designs are built. Traditionally, resource estimation in the mining industry has been described as a geostatistical problem where the amount of resource in an area is predicted using a set of point-wise samples from specific locations within a particular deposit (Mpanza, 2015; Granek, 2016; Kaplan and Topal, 2020). Most previous studies (Rahimi et al., 2018; Marwanza et al., 2018) have taken a purely geostatistical approach to resource estimation, with most methods usually entailing some form of kriging (Mpanza, 2015; Zhang et al., 2021a). These geostatistical methods typically assume data retrieved from small samples to be representative of large ore deposits – under the assumption of spatial autocorrelation of regionalised variables (Young, 2008; Samson, 2019; Kaplan and Topal, 2020). The suitability of certain geostatistical techniques for in-situ metal grade estimation is, however, contextual and application specific (Lindsay et al., 2021). Furthermore, a trade-off exists between sample size, cost, and computation time (Dutta et al., 2010; Zhang et al., 2021a). These estimation processes also require considerable manual data manipulation, which increases the possibility of unforeseeable errors and analytical partiality (Samson, 2019; Zhang et al., 2021a; Kaplan and Topal, 2020).

The last few decades have seen a steady introduction of machine learning (ML) across a variety of industries such as manufacturing, healthcare, and finance; mining operations are one of the latest to join in (Jung and Choi, 2021). The integration of machine learning algorithms (MLAs) with geospatial and geostatistical techniques serves to provide a more advanced and potentially more reliable approach to resource estimation. Machine learning utilises computing power along with MLAs to automatically distinguish and leverage patterns in data (Samson, 2019; Zhang et al., 2021a; Kaplan and Topal, 2020). This approach can be faster and more accurate because less manual data processing and interpretation is required (thus reducing human error, among other sources of error); furthermore, it allows for larger and higher-dimensional datasets (Lindsay et al., 2021). Instead of the programmatic approach: where the user programs a computer to automate certain calculations by formulating models and algorithms based on the specific task; the ML approach gives the task and data to a machine through the use of MLAs, while the user understands, selects, implements, and validates a number of appropriate MLAs (Samson, 2019; Zhang et al., 2021a; Kaplan and Topal, 2020). The machine then creates models and establishes model-specific parameters using the assigned algorithm (Dutta et al., 2010; Kaplan and Topal, 2020). This approach is also highly interdisciplinary since the fundamental techniques are data-driven and not discipline- or hypothesis-driven.

1.1. Motivation

The Witwatersrand Basin has previously been a subject of gold estimation studies (e.g., Rahimi et al., 2018); however, very few have been conducted using integrated geological data and machine learning algorithms (e.g., Zhang et al., 2021a). This approach serves to identify some of the errors in

the mineral resource estimation process at a data level, as it will account for attributes that have previously been neglected (McMillan et al., 2019). Errors in resource estimation have led to lower gold (Au) recovery and thus revenue losses. This study directly addresses this problem in part because it looks at a more general and flexible estimation approach that will return more accurate and efficient results. Ore reserve estimation has a considerable impact on financial decision-making in mining projects; as a result, reliable estimates of mineral resource grades, along with suitable measures of uncertainty, play an essential role in preventing revenue loss.

1.2. Aims and Anticipated Outcomes

Aim

This research aims to study the properties of sedimentological data and their importance as predictive features for in-situ gold grade prediction using MLAs in the Witwatersrand Basin. A set of MLAs will be implemented to rank the importance of sedimentological variables, which will thereafter be used for gold prediction. With this integration, vital rock characteristics can be incorporated when estimating the in-situ concentrations of gold.

Anticipated Outcomes

i. Through working knowledge, we know that the spatial distribution of gold orebodies and their grade is controlled by geological factors. Consequently, I anticipate that there is likely to be a strong causal relationship between sedimentological parameters and in-situ gold grades.

ii. I anticipate that in-situ gold grade predictions made using MLAs may potentially be more accurate than those made by standard geostatistical methods. In addition, combining MLAs with geostatistical techniques may provide the most optimal results.

1.3. Objectives

i. To understand relationships between qualitative rock descriptions (e.g., sedimentological descriptions) and quantitative in-situ gold grades using a legacy database.
ii. To determine the most suitable MLA(s) for the Witwatersrand Basin's in-situ gold grade predictions.
iii. To compare gold grade prediction results from ML with estimation results from geostatistical techniques.

1.4. Data context

The data used in this project originates from Sibanye-Stillwater's Kloof gold operation. The mine is located in South Africa's Gauteng province, about 50 km west of Johannesburg, near the towns of Randfontein and Westonia (Frimmel and Nwaila, 2020). Geologically, it is situated in the western line of the Witwatersrand Basin (Fig. 1.1). The samples used in this study were extracted from the Kloof operation's Kloof Reef.

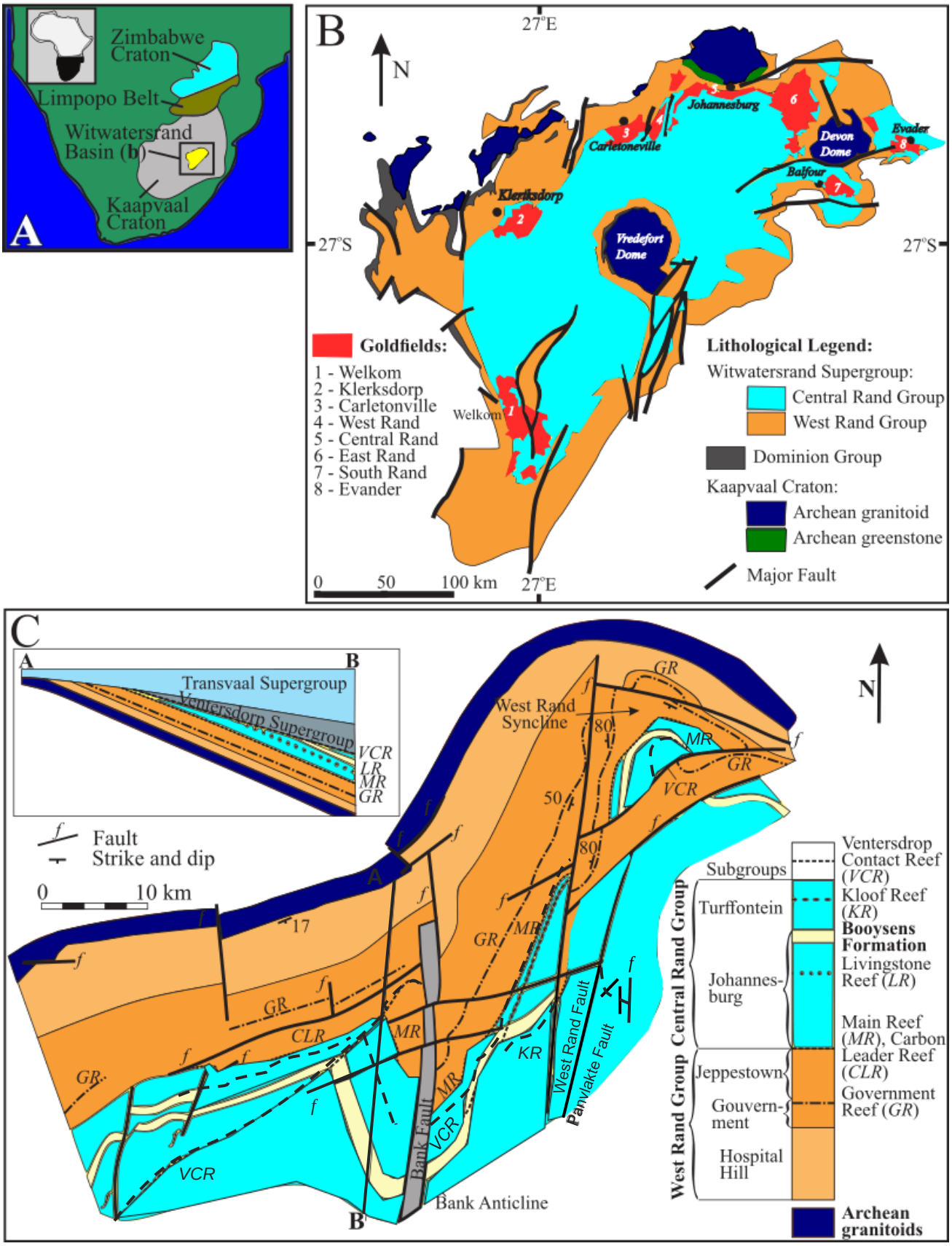


Figure 1.1: Locality map of the study area, included is the location of the Witwatersrand Basin (A); its goldfields (B); its cross-section (C); and its stratigraphic column with the position of the thin Kloof Reef outlined (adapted from Frimmel and Nwaila, 2020).

1.5. Research scope

This study is centred within the disciplines of geosciences, geostatistics, and data science; contextually, it explores the integration of ML in the field of economic geology and resource estimation. This research will assess geological sample data collected through traditional geological surveys from a gold-producing reef; these data will be compiled into a database, and the encompassed sedimentological properties will be used to predict gold grade. Various MLAs will be built using the Python programming language; the geological data will then be run through each of these algorithms with the ultimate goal of building a gold prediction model. This study will investigate the extent to which various sedimentological properties aid in predicting in-situ gold grade. Moreover, it will look at various measures that can be explored to improve prediction performance when predicting gold grade using sedimentological properties. This study will also explore the limitations of this approach to gold prediction. A summary of the scope of this project is illustrated in Fig 1.2:

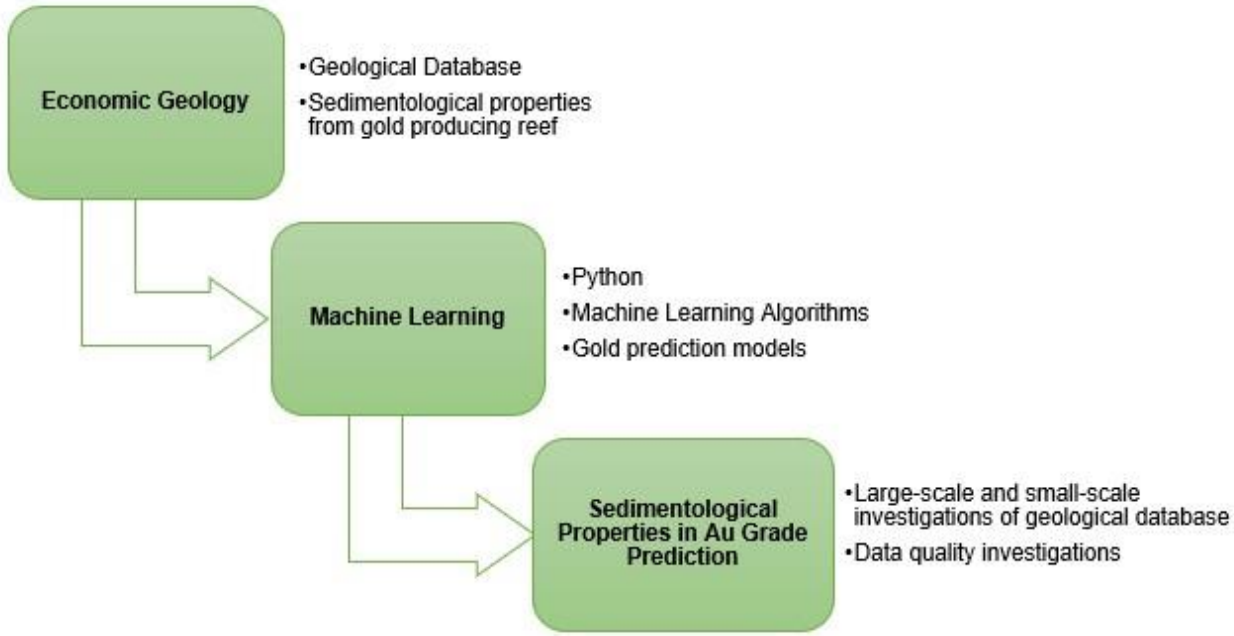


Figure 1.2: Research Scope

1.6. Thesis outline

This thesis comprises six chapters. Chapter 1 comprises the introduction, where the research background is provided. In this chapter, the aims and objectives are outlined, the study area is introduced, and the research scope is outlined. Chapter 2 contains the literature review, where the geological setting is discussed, and an evaluation of resource estimation and ML is provided. Chapter 3 includes the methodology section, outlining the steps taken during the experimental process. The results are outlined in Chapter 4; these are presented using various visualisations. Chapter 5 provides a discussion of the results attained while correlating them with information discussed in the previous

sections. Chapter 6 concludes the thesis and discusses the extent to which the aims and objectives were achieved.

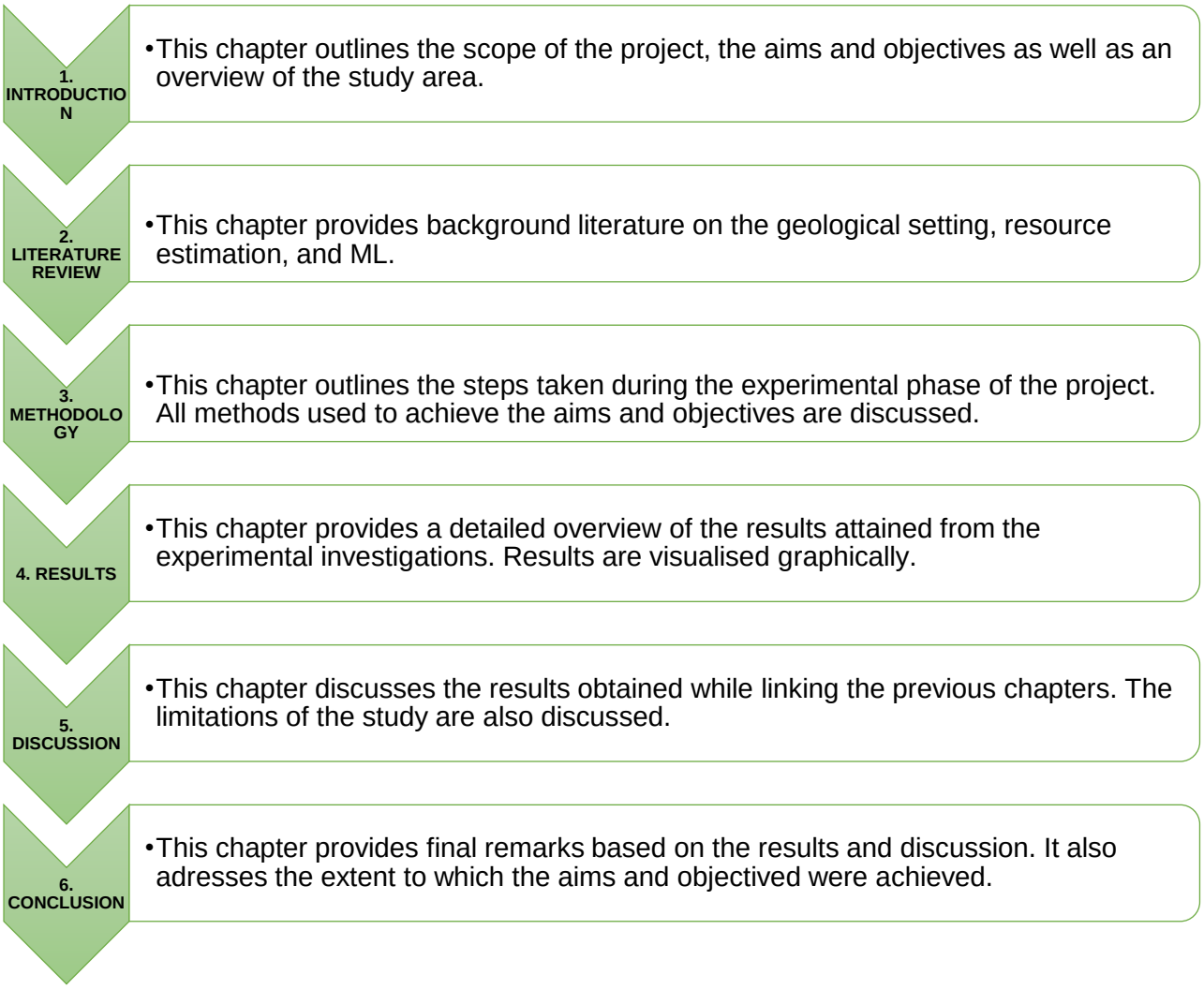


Figure 1.3: Thesis outline.

2. LITERATURE REVIEW

This chapter analyses literature relevant to the scope of this project. First, an overview of the geological setting is outlined; the Witwatersrand Basin's goldfields are discussed, and the geology of the Kloof operation is analysed. The subsequent section discusses resource estimation and then focuses on ML. An overview is provided, and thereafter a more detailed literature review of the ML modelling process is provided.

2.1. Geological setting

South Africa's Witwatersrand Basin (Fig. 2.1) is considered the largest goldfield province in the world; it accounts for over a third of the Earth's gold production, which amounts to about 53 000 tonnes of gold to date (Nwaila et al., 2020). The basin's gold deposits have been attributed to a series of unique interactions between a set of highly complex, coincidental geological events (Robb and Meyer, 1995; Dankert and Hein, 2009).

The ~2.9 Ga, intracratonic Witwatersrand Supergroup is located at the centre of the Archean Kaapvaal Craton (Fig. 2.1; Robb and Meyer, 1995; Frimmel et al., 2009). It is a sedimentary succession that has undergone several post-depositional changes, such as metamorphism and hydrothermal alteration. The supergroup includes the West Rand Group (~ 5150 m thick), which accounts for less than 5% of the basin's gold production and primarily comprises iron-rich shales and quartzites (Frimmel et al., 2009; Tucker et al., 2016). It is overlain by the Central Rand Group (~ 2550 m thick), which accounts for more than 85% of the basin's gold production and primarily comprises gold-bearing conglomerates (Frimmel et al., 2009; Tucker et al., 2016). The Witwatersrand sediments are underlain by the Dominion Group's meta-volcanics and minor clastic sediments, and overlain by the volcano-sedimentary succession of the Ventersdorp Supergroup and the quartzites and dolomites of the Transvaal Supergroup (Frimmel et al., 2009; Tucker et al., 2016). The Ventersdorp and Transvaal Supergroups account for the basin's remaining gold production (Taylor and Anderson, 2018).

The emplacement of the Bushveld Complex (~2.06 Ga) in conjunction with the Vredefort meteorite impact (~2.02 Ga) caused a series of tectonic-thermal events which are responsible for low-grade metamorphism (Greenschist facies) seen in the northern and western margins of the basin (Frimmel et al., 2009; Phillips and Powell, 2011). Metamorphic grade only reaches upper amphibolite facies in the deeper parts of the basin fill that is now upturned around the Vredefort Dome (Robb and Meyer, 1995; Frimmel et al., 2009; Phillips and Powell, 2011). Metamorphism in the basin can also be attributed to burial, along with syntectonic dynamic metamorphism that resulted from continental collision along the Kaapvaal Craton's northern margin (Robb and Meyer, 1995; Phillips and Powell, 2011). These metamorphic events prompted various degrees of fluid flow through the Witwatersrand

Basin units and consequently contributed to gold remobilisation (Safonov and Prokof'ev, 2006; Phillips and Powell, 2011).

Aside from gold, the Witwatersrand Basin also contains a mineral assemblage that includes uranium, iron sulphide, carbon, and detrital oxides (Hartnady, 2009; Tucker et al., 2016; Fuchs et al., 2016).

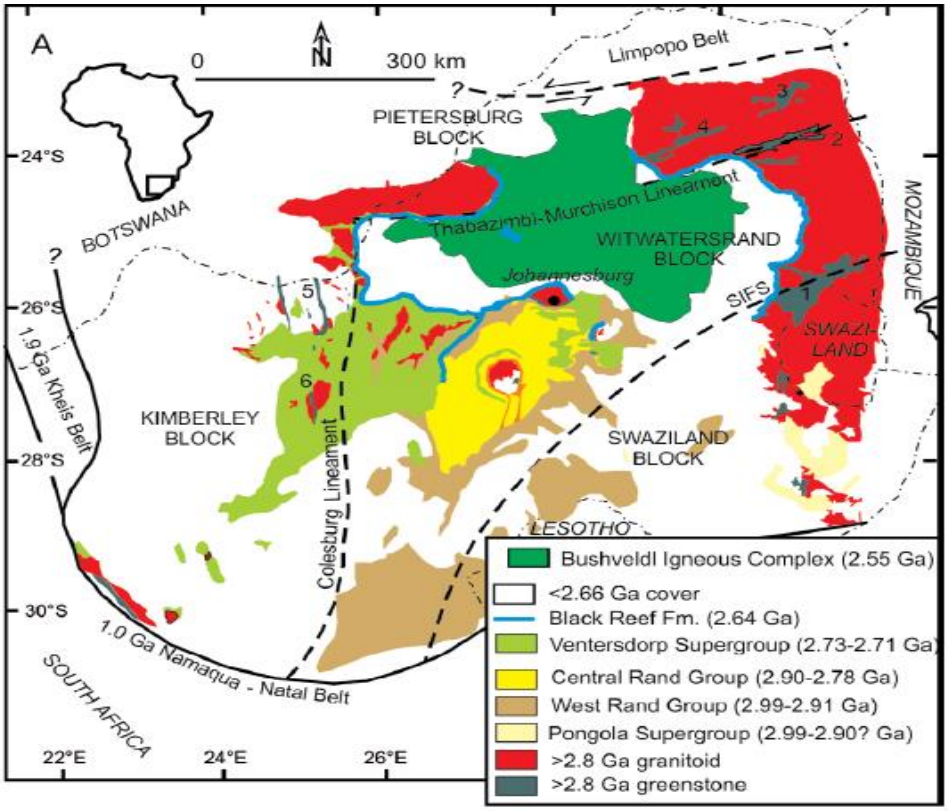


Figure 2.1: Locality map illustrating the geological setting of the Witwatersrand Basin (Frimmel and Nwaila, 2020).

2.1.1. Goldfields

The Witwatersrand Basin's gold-bearing conglomerate reefs are mined in 8 main goldfields (Fig. 1.1B) and a few minor occurrences across the basin. These are summarised below:

2.1.1.1. Welkom Goldfield

The Welkom Goldfield, situated at the south-western margin of the basin, is largely covered by younger Ventersdorp strata. Its orebodies – all of which are from the Central Rand Group – are contained in sedimentary packages comprising auriferous and uraniferous conglomerates found above multiple low-angle unconformities within the Central Rand Group (Hayward et al., 2005; Frimmel et al., 2009; Frimmel and Nwaila, 2020). The Beisa Reef (Fig. 2.2), its lowermost orebody, is a pebble layer located above a paleosurface that is distinguished by a thin kerogen layer (Hayward et al., 2005; Frimmel et al., 2009). The reef is auriferous but relatively low grade, and it only reaches mineable depths south of the goldfield (Tucker et al., 2016). The Basal and Steyn Reefs, located at the base of the Krugersdorp Formation (Fig. 2.2), are the Welkom Goldfield's most productive

orebodies (Frimmel et al., 2009; Frimmel and Nwaila, 2020). The Basal Reef constitutes an oligomictic placer pebble assemblage, while the Steyn Reef constitutes a polymictic placer pebble assemblage (Frimmel et al., 2009; Frimmel and Nwaila, 2020). Both reefs are highly uraniferous, with most uranium being hosted in semi-rounded detrital uraninite grains (Fuchs et al., 2016; Frimmel and Nwaila, 2020). The goldfield also comprises the Saaiplaas Reef, which contains sandy fluvial channel fills (Hayward et al., 2005; Frimmel et al., 2009). It is both auriferous and uraniferous, with Au and U being confined to trough crossbedding, foresets, and trough scour surfaces that are also enriched in rounded pyrite (Fuchs et al., 2016; Taylor and Anderson, 2018). Due to its lack of lateral continuity, the Saaiplaas Reef does not have high economic significance (Frimmel et al., 2009). The Leader Reef, found in the Booysens Formation (Fig. 2.2), is another placer orebody; however, its mineralization is confined to areas where it truncates older reefs (Hayward et al., 2005; Frimmel et al., 2009; Frimmel and Nwaila, 2020).

2.1.1.2. Klerksdorp Goldfield

The Klerksdorp goldfield, located on the north-western edge of the basin (Fig. 1.1B), is largely covered by younger Ventersdorp sequences. It contains auriferous and uraniferous deposits hosted in the Central Rand Group, the West Rand Group, and the Ventersdorp Supergroup, with the Central Rand Group hosting most of the orebodies (Frimmel et al., 2009; Tucker et al., 2016). The goldfield includes the Ada May Reef (Fig. 2.2), a conglomerate-kerogen layer (~1.5 m thick) with poor mineralisation located at the base of the Central Rand Group (Frimmel et al., 2009; Tucker et al., 2016; Frimmel and Nwaila, 2020). Its most significant orebody is the Vaal Reef (Fig. 2.2), which sits on an angular unconformity at the base of the Krugersdorp Formation (Frimmel and Nwaila, 2020). It is an auriferous oligomictic pebbly quartz arenite layer with noticeable kerogen seams (Hayward et al., 2005; Frimmel et al., 2009). The Ventersdorp Contact Reef (Fig. 2.2), found at the base of the Ventersdorp Supergroup, is the second most significant orebody in the goldfield (Tucker et al., 2016).

2.1.1.3. Carletonville and West Rand Goldfields

Along the basin's northern margin, the Carletonville and West Rand Gold fields can be located (Fig. 1.1B). The goldfields are separated by a fault known as the West Rand Fault and are covered by younger rocks from the Transvaal and Ventersdorp supergroups (Frimmel et al., 2009; Frimmel and Nwaila, 2020). Orebodies in these goldfields are hosted in conglomerate beds that overlie numerous low-angle unconformities (Frimmel and Nwaila, 2020). They vary from auriferous to uraniferous, with mineralisation being prominent where the underlying units are being truncated (Safonov and Prokof'ev, 2006; Frimmel et al., 2009; Dankert and Hein, 2009). The Main and Carbon Leader Reefs (Fig. 2.2) are the goldfields' most enriched orebodies (Frimmel et al., 2009; Frimmel and Nwaila, 2020). They are situated in the lower Central Rand Group on a disconformity at the base of the Main Formation. Today, the Carbon Leader Reef is completely mined out; however, it comprised a thin (10 cm) quartz pebble conglomerate layer underlain by a 1–30 mm thick kerogen layer that was highly

enriched in gold (Dankert and Hein, 2009; Frimmel and Nwaila, 2020). Orebodies with lag gravels characterised by thick, clast-supported conglomerate beds/lenses have proven to be highly enriched in gold and uranium (Fuchs et al., 2016; Taylor and Anderson, 2018). The goldfields also contain the Middelvlei, Livingstone and Elsburg Reefs (Fig. 2.2).

2.1.1.4. Central Rand, East Rand, South Rand, and Evander Goldfields

The Central Rand Goldfield (Fig. 1.1B), found immediately south of Johannesburg, is almost completely mined out and has produced more than 8 000 tonnes of gold to date (Frimmel and Nwaila, 2020). To its east is the East Rand goldfield, which has produced about 10 000 tonnes of gold to date and is also nearly depleted (Frimmel and Nwaila, 2020). The Main, Carbon Leader and South Reefs were the major gold-producing orebodies (Hayward et al., 2005; Frimmel et al., 2009; Frimmel and Nwaila, 2020). These were sitting on erosional unconformities (Frimmel et al., 2009; Dankert and Hein, 2009). Further south is the South Rand goldfield, where the South Reef constitutes the primary orebody (Fig. 2.2). The Evander goldfield, located north-east of the Central Rand Basin, is the youngest goldfield and is a different sub-basin to that of the other goldfields (Dankert and Hein, 2009; Tucker et al., 2016). The Central Rand Group is about 650 m thick here and is overlain by Ventersdorp and Transvaal Supergroup sequences, all of which are overlain by Karoo Supergroup rocks (Hayward et al., 2005; Frimmel et al., 2009). The Kimberley Reef, a 3 cm thick conglomerate layer, is the goldfield's only orebody (Frimmel et al., 2009; Frimmel and Nwaila, 2020).

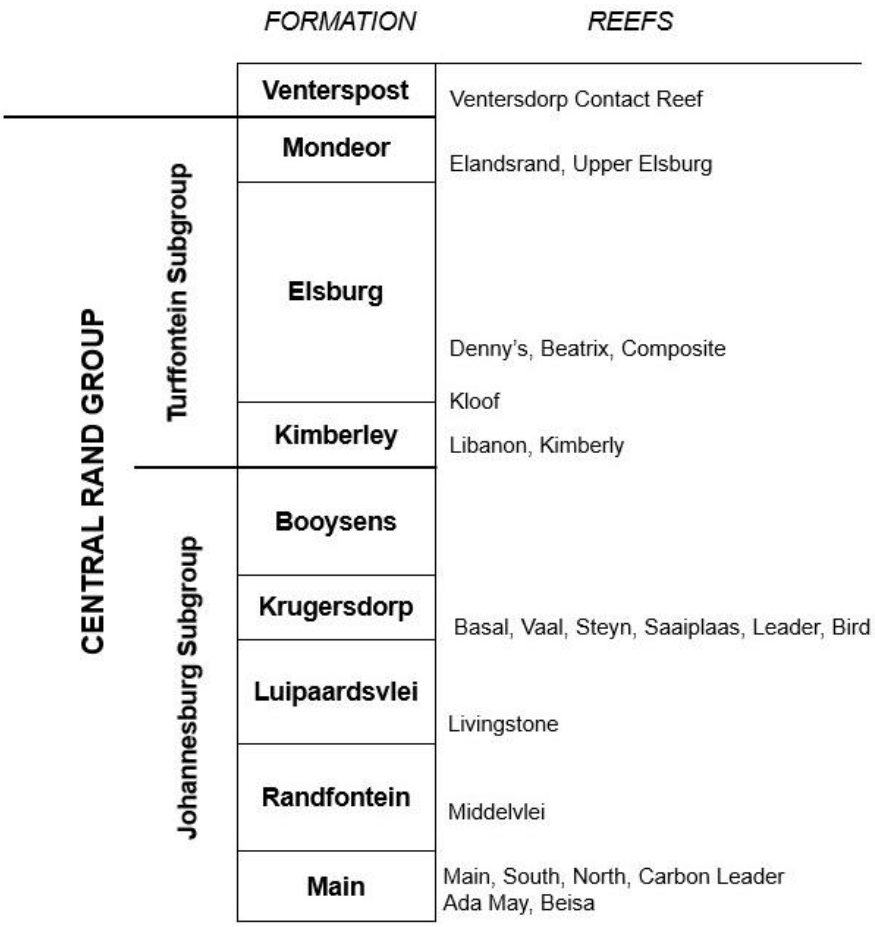


Figure 2.2: Generalised stratigraphic column of the Central Rand Group (adapted from Frimmel et al., 2009).

2.1.2. Kloof operation

Sibanye-Stillwater's Kloof operation exploits the Ventersdorp Contact Reef (71%), Middelvlei Reef (9%), Kloof Reef (18%) and the Leader Reef (2%) (Sibanye-Stillwater, 2020). The Kloof Reef can be located midway between the Ventersdorp Contact Reef (VCR) and the Middelvlei Reef, as illustrated in the stratigraphic column in Fig 2.2. The Kloof Reef is situated amidst the Central Rand Group sediments. These are primarily characterised by an arenaceous sequence that includes quartzites, quartz-pebble conglomerates, minor shales and quartzwackes (Frimmel et al., 2009; Tucker et al., 2016). The Central Rand Group comprises the bulk of the Witwatersrand Basin's gold mineralisation (Frimmel et al., 2009; Tucker et al., 2016). The Booysens Shale marker separates the group into the lower Johannesburg Subgroup and the upper Turffontein Subgroup, where the Kloof Reef can be located (Frimmel et al., 2009; Tucker et al., 2016). The bulk of the Central Rand Group reefs contain gold mineralisation described as submicroscopic crystalline grains that occur within detrital pyrite grains (Viljoen, 2009). Some of the gold grains have been described as rounded or flattened flakes (Viljoen, 2009). Furthermore, some degree of correlation between pebble sorting/packing has been suggested (Viljoen, 2009).

2.2. Resource Estimation

Mineral resources are the fundamental blocks of a modern society (Coates, 1985). They are an integral part of numerous industrial processes and play a significant role in manufacturing, socioeconomic development, and the construction of modern transport systems (Dubinski, 2013; Van Gosen et al., 2014). In addition, they drive the technology and energy sectors (Bardi, 2013; Calas, 2017). A mineral resource is a concentration of a naturally occurring material of economic interest in or on the crust of the Earth that is of a grade, quality, or quantity that allows for possible extraction of the commodity for commercial gain (Henckens et al., 2019; Zerzour et al., 2021). Estimating these resources is a vital and challenging task in mining and exploration projects, regardless of the size or type of deposit (Tokoglu, 2018). The mineral resource estimation process aims to extract information on the quantity and quality of a mineral deposit (Abzalov and Bower, 2014). Moreover, detailed consideration of the high spatial variability and uncertainty associated with geological formations is imperative in the estimation process (Abzalov and Bower, 2014; Rossi and Deutsch, 2014).

Mineral resources can be categorised as inferred, indicated and measured depending on the increasing order of geological knowledge and confidence (Fig. 2.3) (Zerzour et al., 2021). Mineral reserve estimation succeeds mineral resource estimation; this step aims to extract the tonnage and average grade of the economically minable portion of the mineral deposit (indicated/measured resource) (Hartman and Mutmansky, 2002). Furthermore, the mineral reserve estimate provides the foundation for economic analysis and consequently the future potential of a mine operation (Hartman and Mutmansky, 2002). The accuracy of a reserve estimation is vital to the quality of the subsequent geological interpretation (Yunsel, 2012). Mineral reserves are categorised as either probable or proven, and their estimation involves two underlying concepts: the first is extension, where attributes of a sample are extended to blocks to be estimated; the second is error estimation, where the associated error is used to assess the validity of an estimation method (Wellmer et al., 2007; Shurygin et al., 2019).

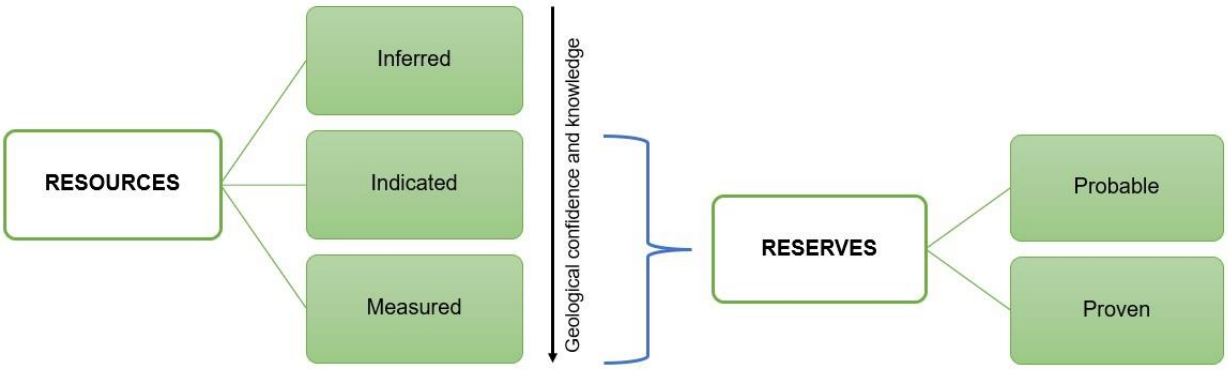


Figure 2.3: Generalised relationship between mineral resources and reserves (adapted from Dumakor-Dupey and Arya, 2021).

Several resource estimation methods have been suggested and applied in literature. These methods transform point-wise sample information into areal or volumetric information and are primarily categorised as either geometric or geostatistical (Hartman and Mutmansky, 2002; Rossi and Deutsch, 2014). Geometric techniques include methods such as random stratified grids, polygonal, triangular, and cross-sectional (Rossi and Deutsch, 2014). Moreover, they are simple, require minimal input parameters, and are commonly implemented in the initial stages of a mineral project; they can also be used to validate results attained from sophisticated estimation methods (Rossi and Deutsch, 2014). Geostatistical techniques include conditional simulation, kriging and inverse distance weighting (Jones et al., 2018). These are considered to be more sophisticated. However, traditional estimation methods tend to have several limitations: they require manual processing, tend to overestimate or underestimate resources, and underperform with highly heterogeneous datasets (Rossi and Deutsch, 2014; Jones et al., 2018). Two of the most widely applied geostatistical estimation methods are discussed below.

2.2.1. Inverse Distance Weighting (IDW)

IDW methods use attributes of a known point to interpolate the attributes of an unknown point/block with a weighing factor (Chen et al., 2018). This technique assumes that two points in proximity contain similar properties; the further they get, the less similar they become (Chen et al., 2018). IDW methods have been historically used to estimate the ore grades of various mineral deposits with success; presently, they are still used in some mining operations as part of their reserve estimation routine (Dominy and Hunt, 2001; Chen et al., 2018). Moreover, the method has been favoured due to its ease of use and low computational costs. According to Dominy and Hunt (2001), the IDW was the most common grade estimation method in Western Australia's Eastern goldfields in the late 90s to early 2000s.

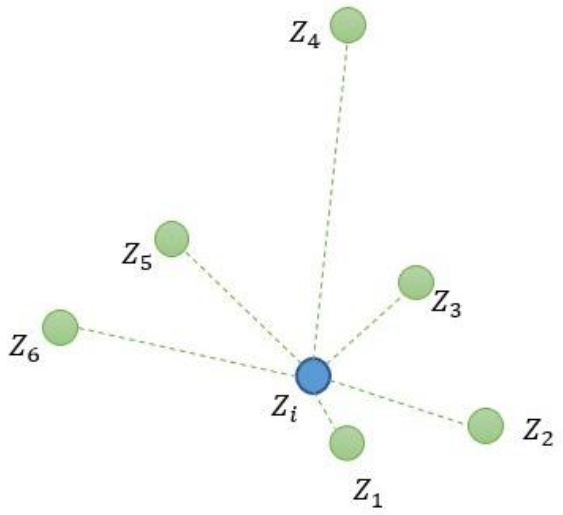


Figure 2.4: Generalised illustration of the IDW technique, where Z_i is the known point and the unknown points are in green. Z_1would be assigned the highest weight (adapted from Coleman, 2021).

IDW estimation uses an averaging approach, where each unknown sample point is allocated a weight equal to the inverse of their separation distance from the known point (Yasrebi et al., 2016). Consequently, closer points are allocated higher weights than those further away from the estimation point (Fig. 2.4) (Yasrebi et al., 2016; Coleman, 2021). The weights are rescaled to eliminate bias from the estimate – the inverse of the separation distances (weights) is summed to 1, ensuring there is no bias when the estimated grades are compared to the sample grades (Al-hassan and David, 2015; Coleman, 2021). The inverse distance estimation can be formulated as follows:

$$Estimate = \frac{\sum(sample\ value\ \times inverse\ distance\ weight)}{\sum(inverse\ distance\ weights)},$$

where:

$$inverse\ distance\ weights = \frac{1}{(sample\ to\ estimation\ point\ distance)^{power}}.$$

The inverse distance weights can be raised to a power. The user assigns the value of the power; the higher the power, the more weight assigned to closer sample points (Al-hassan and David, 2015; Coleman, 2021). The closer points are always assigned a weight higher than further points regardless of whether the power is high or low (Al-hassan and David, 2015; Coleman, 2021).

2.2.2. Kriging

Kriging is another popular mineral resource estimation method. Block grades are locally estimated using property data in or near the block; the process ensures the estimate has minimal error variance and bias (Ali Akbar, 2012; Rossi and Deutsch, 2014). Mining operations use multiple variants of kriging, with the most popular being: simple kriging, ordinary kriging, and indicator kriging (Ali Akbar, 2012; Dumakor-Dupey and Arya, 2021). Each variant has a different underlying assumption pertaining to the local domain mean (Rossi and Deutsch, 2014). Numerous authors have attributed ordinary kriging as the best estimator for ore reserve estimation (Yasrebi et al., 2009; Daya and Bejari, 2015). Kriging techniques not only take the distance between points into account; but also analyse the spatial structure of the data by comparing sample points separated by specific spatial distances two at a time (Yasrebi et al., 2009; Daya and Bejari, 2015). This process aims to comprehend the relationships between sample points separated by various lag distances – this information is thereafter accounted for in the variogram (Daya, 2015; Hekmatnejad et al., 2019). The variogram generates spatial weights for the sample points (Daya, 2015). Moreover, the original sample values are preserved in the interpolated map (Daya, 2015). The kriging weights are applied to the sample points as formulated:

$$\hat{Z}(X_0) = \sum_{i=1}^{N} \lambda_i Z(X_i),$$

where $\hat{Z}(X_0)$, the value of the predicted point is equal to the sum of each sampled point (X_i) multiplied by that predicted point's unique weight ($\lambda_i Z$) (Taboada et al., 2013; Hekmatnejad et al., 2019). The weights are derived from spatial variability modelling, either through a covariance function or a variogram model.

2.2.2.1. Variogram Modelling

The variogram describes a covariance as a function of variable separation (Rossi and Deutsch, 2014). When the separation distance is at zero, two points have maximum covariance; the covariance decreases as their separation increases to the range (Rossi and Deutsch, 2014; Daya and Bejari, 2015). The variogram is also termed the covariance function because it is an alternative way of conveying the degree of covariance between two points separated by distance (Rossi and Deutsch, 2014; Coleman, 2021). It is formulated as follows:

$$\gamma(h) = sill + nugget - C(h),$$

where $\gamma(h)$ is the variance of two points separated by distance h; the sill is the ceiling value for $\gamma(h)$ where the model variogram flattens out; the nugget refers to the value of $\gamma(h)$ when h is closest to 0; and $C(h)$ describes the covariance of two points separated by distance h. This relationship is illustrated in Fig. 2.5.

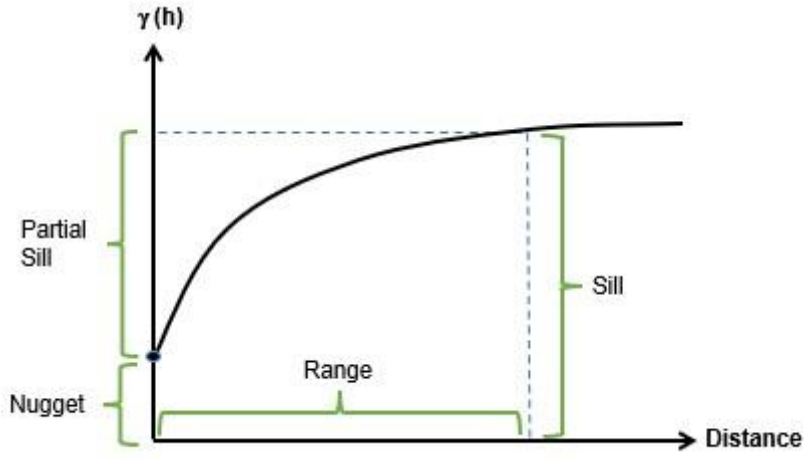


Figure 2.5: Simplified illustration of the semi-variogram (adapted from Rossi and Deutsch, 2014 and Coleman, 2021).

2.3. Machine Learning: Introduction

ML allows a computer to learn a specific task without being overly programmed (Dutta et al., 2010; Cate et al., 2017). A skeleton structure of the process followed is illustrated in Fig. 2.6:

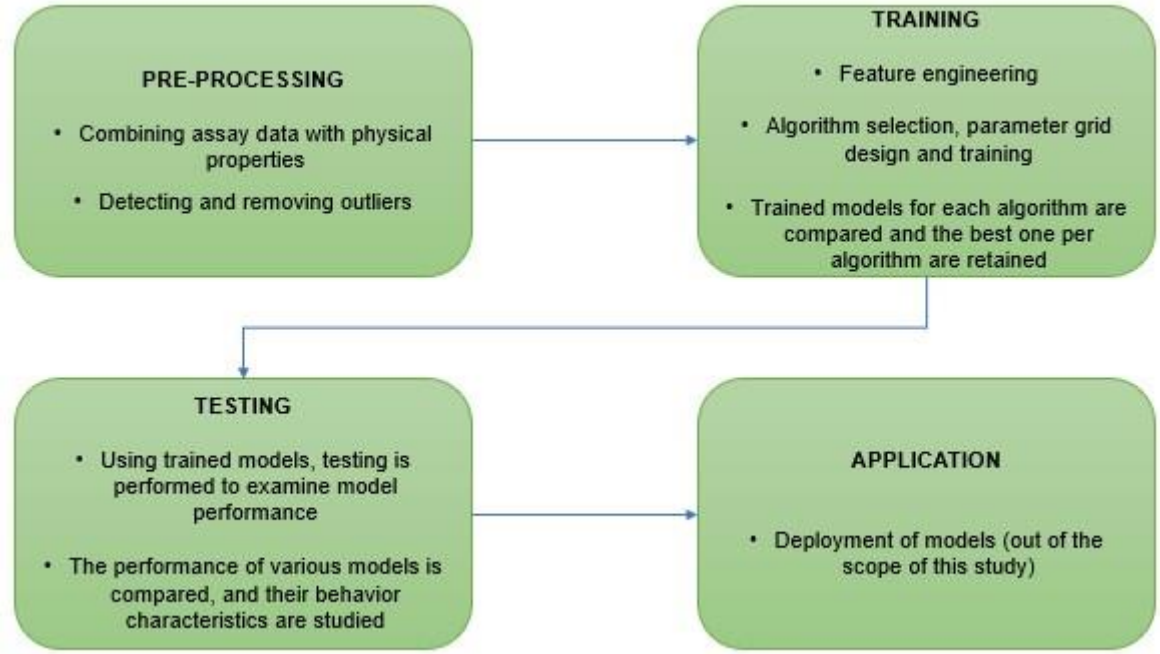


Figure 2.6: Simplified diagram illustrating the process followed in ML (adapted from Cate et al., 2017 and Samson, 2019).

Depending on the problem, ML tasks are solved using either supervised or unsupervised learning techniques. Both supervised and unsupervised MLAs follow a general methodology that entails data pre-processing, training, testing and application (Fig. 2.6).

In supervised ML, a dataset is provided with labels (Dutta et al., 2010; Cate et al., 2017). Relationships between the input features and data labels are assumed to exist and can be used for inference (Samson, 2019; Kaplan and Topal, 2020). Supervised learning problems can be categorised into regression and classification problems. In regression problems (Fig. 2.7), labels are continuous; consequently, predictions are also continuous (Samson, 2019; Kaplan and Topal, 2020). In classification problems (Fig. 2.7), labels are discrete; consequently, predictions are also discrete (Samson, 2019; Kaplan and Topal, 2020). In unsupervised learning, the data are not provided with labels and MLAs are applied to create labels (Samson, 2019; Kaplan and Topal, 2020). In this manner, a structure can be derived from the data using unsupervised learning techniques such as clustering and dimension reduction techniques (Fig. 2.8) (Dutta et al., 2010; Samson, 2019).

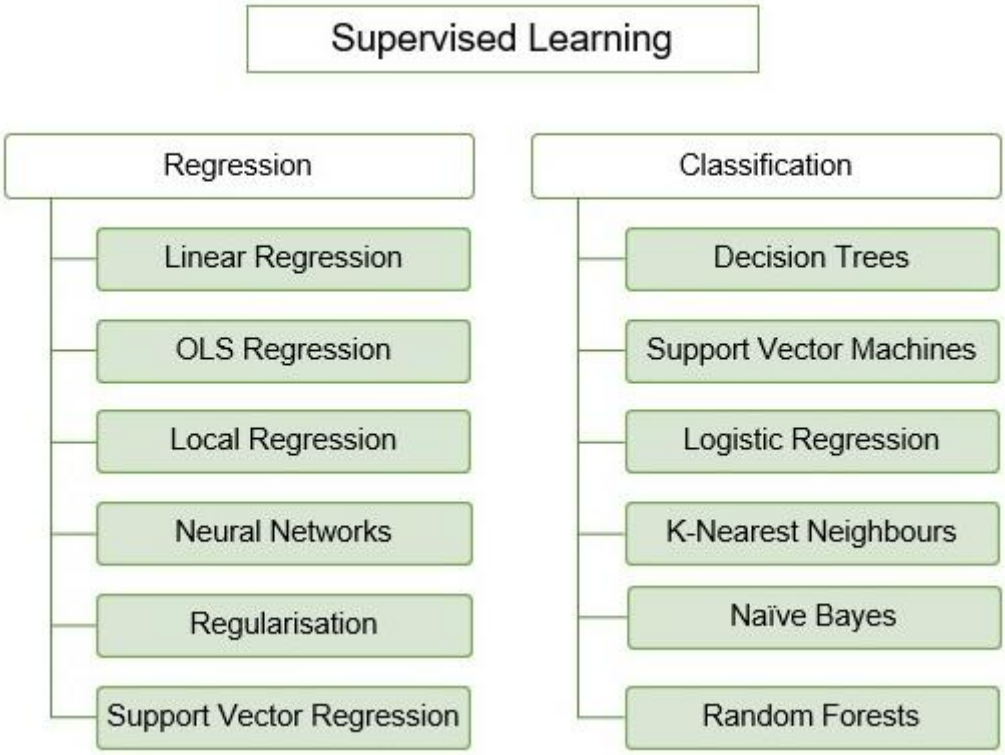


Figure 2.7: Common Supervised Learning algorithms (adapted from Samson, 2019 and Jung and Choi, 2021).

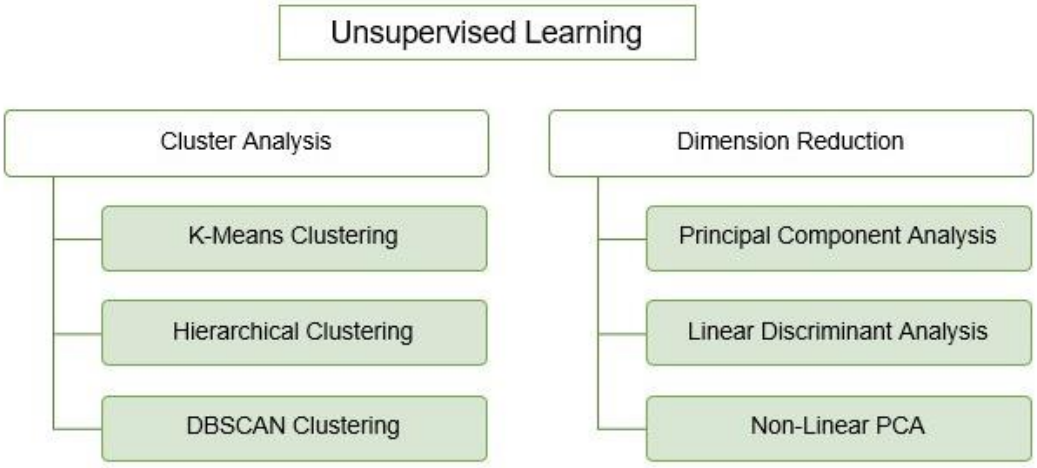


Figure 2.8: Common Unsupervised Learning algorithms (adapted from Samson, 2019 and Jung and Choi, 2021).

2.3.1. Terminology

An *algorithm* describes a system used by a machine to accomplish a certain goal. It is like a recipe in which both the inputs and outputs are defined, and all the steps required to generate the output from the inputs are outlined. In standard programming, the computer is given explicit instructions, while ML operates through indirect programming by feeding a computer with data (Lishchuk et al., 2019; Molnar, 2022).

A *MLA* refers to the algorithm that a machine uses to learn from data (Lishchuk et al., 2019; Molnar, 2022).

A *ML model* refers to a learned program (also referred to as the predictor) that maps features to labels (Lishchuk et al., 2019; Molnar, 2022).

Interpretable ML refers to models and methods that make the behaviour and predictions of ML systems understandable to humans (Molnar, 2022).

An *observation* refers to an individual datapoint. For tabular data, it can be visualised as a row in the dataset.

The *features* are the inputs used for prediction. For tabular data, it can be visualised as a column in the dataset. Features in the sense of traditional statistical data modelling can be thought of as independent variables.

The *target* refers to the information the machine learns to predict. In traditional statistical modelling, this is referred to as a dependent variable.

A *ML task* is the combination of a dataset with features and a target to achieve a goal (Molnar, 2022). The task can be classification, clustering, outlier detection or regression, depending on the type of target (Lishchuk et al., 2019; Molnar, 2022).

The *prediction* is a ML model inference of the target value based on provided features (Molnar, 2022).

2.4. Machine learning modelling: Application

All ML tasks, be it estimating house prices, fraud detection or product recommendations, are generally solved with the same approach. Firstly, data collection. The more data you collect, the better; moreover, the data should contain the desired outcome to be predicted (in this case, gold grades) as well as additional information that will aid prediction (Fisher et al., 2018; Molnar, 2022). For instance, if you want to build a house value estimator, you would need to collect data on previous house sales as well as related real estate data such as location, areal size etc. The next step would be to process the data and add the compiled information into a ML algorithm that will generate a desired model (Li et al., 2011; Fisher et al., 2018). The model will then be used with new data (Fisher et al., 2018; Molnar, 2022).

Machines exceed humans' performance in numerous tasks like weather prediction or chess. In instances where a machine is just as good as a human or perhaps slightly worse at a task, it remains superior in some regards and applications because it would still surpass a human in terms of scaling, availability, speed and reproducibility (Dutta et al., 2010; Apley and Zhu, 2020). Once a ML model is implemented, it will complete tasks faster and consistently deliver reliable results (Li et al., 2011). In addition, one can replicate a ML model on any machine at a low cost – the exception to this is with

deep learning which may entail millions of parameters; training may require a lot of electricity, and the overall cost could be quite high (Fisher et al., 2018; Apley and Zhu, 2020). This is in contrast to training a human for a task, which in the case of a professional geostatistician, can take years and can be quite costly. However, non-general purpose artificial intelligence in the form of ML algorithms still requires human intervention, monitoring and decision audits. Presently, artificial intelligence is an unsolved problem, meaning that for each task, a specific workflow would have to be formulated and no single best approach exists. This is particularly important to appreciate in the geosciences, where the widespread adoption and desirability of ML and artificial intelligence has been fairly recent, and most problems have not yet had demonstrably useful workflows created.

With complex ML models, there can be a lack of clear insight into the data and the task being solved by the machine, thus making it difficult to understand the model in its entirety (Fisher et al., 2018). Algorithms such as Random Forest, make use of a large ensemble of decision trees to arrive at an inference; to understand how each decision was made, one would have to evaluate the structure of each decision tree (Li et al., 2011; Apley and Zhu, 2020). This is not a practical human task. Multiple research papers (Breiman, 2001; Luckert and Schaefer-Kehnert, 2016; Wright and Ziegler, 2017) have attributed tree-based algorithms as highly suitable for regression ML tasks. The best performing ML models usually comprise an ensemble (blend) of several individual models that can be too complicated to interpret (even if they are interpretable as individual models) or arise from emergent behaviour of interactions of a large number of nodes, e.g., in neural networks (Li et al., 2011; Fisher et al., 2018; Apley and Zhu, 2020).

2.4.1. Interpretability

Interpretability can be described as the extent to which a human can comprehend the reasoning behind a decision (Ghorbani et al., 2019; Molnar, 2022). The higher the interpretability of a ML model, the easier it is for humans to understand the reason behind certain predictions or decisions made by a machine (Apley and Zhu, 2020; Molnar, 2022). In general, models with fewer parameters are more interpretable. Users of MLAs cannot just trust a model without understanding why it made a certain decision, even if it performs well in general. A single metric, like regression/classification accuracy, is not a complete description of most real-world tasks (Li et al., 2011; Ghorbani et al., 2019).

In predictive modelling, it is not just a matter of what is being predicted, e.g., the probability of a client procuring a certain product, but also a matter of why the prediction was made (Apley and Zhu, 2020; Molnar, 2022). This is of utmost importance in the sciences because the discipline originated with the human desire to understand the physical reality. Simply transferring the understanding to machines using MLAs is not the goal of most scientific endeavours. In practice, ML can be used to expand data modelling in the sciences for hypothesis testing (e.g., Nwaila et al., 2022); novelty discovery (e.g., Zhang et al., 2021b); mapping and surveying (e.g., Zhang et al., 2021c); etc. In other disciplines, such

as engineering, explainability may be less important because such disciplines focus on the applicative aspects of modelling instead of physical understanding of the data and physical reality. Investigations into the causation of outcomes and models may result in a drop in predictive performance through the use of simpler MLAs (Casey et al., 2019). With MLAs such as the ones we have used in this project, knowing the why or the ability to know is essential because, in addition to scientific requirements, models are expected to be implemented in a high-risk operational environment; thus, a mistake may incur severe consequences or liabilities (Li et al., 2011; Casey et al., 2019). Moreover, ML in geosciences is relatively new, so ensuring interpretability is essential for future researchers to assess the scientific merit of ML models and results. This is similar to other applications where ML models tackle real-world tasks that need testing and safety measures (Dutta et al., 2010; Casey et al., 2019); decisions should be traceable and human-understandable, e.g., in autonomous driving, it is important for humans to be able to investigate accidents by auditing the machine's decisions. The key constraint to avoid a lack of interpretability in this study is to avoid the use of deep learning methods. Although this does not ensure complete interpretability, shallow learning algorithms generally make use of fewer parameters per model.

2.4.2. Predictive modelling

To put it simply, predictive models are used to "guess" the values of a variable (label) of interest using its features (Biecek and Burzywowski, 2021; Molnar, 2022). As an example, you could consider predicting the risk of cancer based on a patient's genetics or predicting political attitudes based on Twitter comments. Humans have been using predictive models throughout their recorded history; for example, ancient Egyptians would predict the probability of the Nile flooding using the rising of Sirius (Barucci et al., 2021). The least squares method, developed more than 200 years ago by Andrien-Marie in 1805 and by Carl Friedrich Gauss in 1809, was the more meticulous model construction approach used back then (Dobson, 2002). As time progressed, numerous applications of predictive modelling evolved in the fields of economics, medicine, biology, and agriculture (Dobson, 2002; Efron and Hastie, 2016). Regression, a term devised by Francis Galton in 1886, was initially used in biological applications (Dobson, 2002). Presently, it is used in predictive modelling for continuous variables (Dobson, 2002). Alternatively, classification, a term speculated to have been coined by Ronal Fisher in 1936, is used to predict nominal variables (Dobson, 2002). In the sciences, it is commonly understood that more predictive models are more likely to be correct where more than one model exists. This is certainly the case of scientific abductionism, which is heavily used in geosciences (e.g., ore genesis models) (Kleinhans et al., 2005).

Numerous statistical models have been developed in the last century for the purpose of variable prediction; among these are regression trees, generalised linear models, rule-based models, classification and many more (Dobson, 2002; Efron and Hastie, 2016). The present 'big data" era we are in, accompanied by increased computational power and infinite data streams, has prompted

evolutionary developments in the mathematical foundations of predictive models. The increased demand for predictive models has sparked an interest in model properties like feature engineering, flexibility, and high-precision predictors (Ishikawa and Gulick, 2013; Efron and Hastie, 2016). As mentioned in the previous section, it had become a common practice to use ensembles of models to obtain an ultimate, high-performance model (Efron and Hastie, 2016; Molnar, 2022). Modelling techniques like model stacking, bagging or boosting use a combination of hundreds or perhaps thousands of simpler models to create a single "super" model (Breiman, 2001; Dobson, 2002; Efron and Hastie, 2016). More advanced deep neural models tend to comprise millions of nodes. Although this evolution in predictive modelling is quite impressive, it comes at a cost. It is often problematic and, at times, impossible to understand the effect of changes in the input data on a model's predictions (Liaw and Wiener, 2002; Molnar, 2022).

In typical statistical modelling, a good understanding of the application domain is necessary when building models (Fig. 2.9). Domain knowledge assists in creating and selecting the most important variables to be included in relatively simple models to attain predictive scores (Breiman, 2001). The goodness-of-fit and hypothesis testing are the prominent techniques used for model validation (Breiman, 2001; Hall et al., 2019). The statistical hypothesis is specified before data analysis; furthermore, the p-values obtained should not impede the data processing or interfere with the model construction process (Breiman, 2001, Dobson, 2002).

In ML, the trade-off between the availability of data and domain knowledge is exploited (Fig. 2.9) (Inglis et al., 2021). There is a change in emphasis from trying to attain a deep understanding of the application domain, and a more computationally heavy effort is made towards the construction and fitting of models (Biecek, 2019; Inglis et al., 2021). Flexible models use a large amount of data to isolate useful features and eliminate – or rather filter out – features deemed less useful (Biecek, 2019; Inglis et al., 2021). Much weight is placed on the validation process as it provides feedback on the modelling process (Inglis et al., 2021).

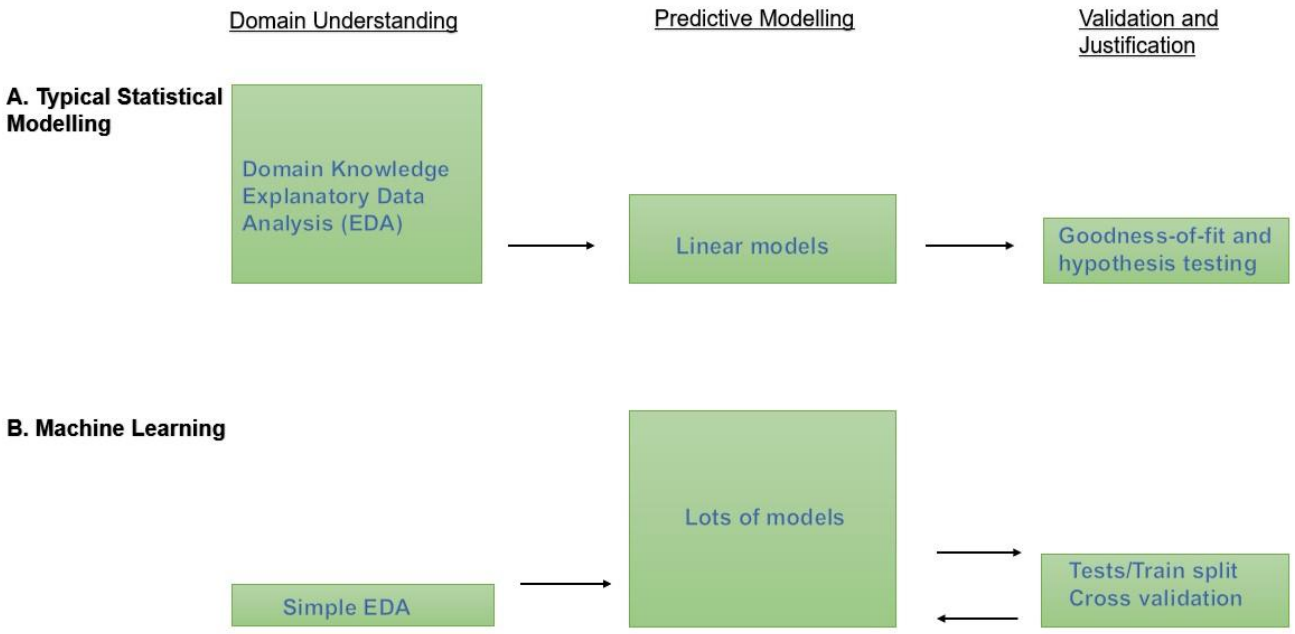


Figure 2.9: A simplified overview of the different predictive modelling processes in typical statistical modelling compared to machine learning. In typical modelling (A), much weight is placed on domain understanding, while in machine learning (B), much focus is placed on the predictive modelling process (adapted from Inglis et al., 2021).

2.4.3. Project management and model development

Numerous approaches have been suggested to manage data science-type projects, which proceed in a manner that may be entirely foreign to many traditional geoscientists. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is one of the most common approaches (Chapman et al., 1999; Hotz, 2022). However, awareness of CRISP-DM or similar processes, as well as particular management models suitable for the exploratory and high-risk nature of data-driven activities, seems to be generally lacking in the sciences. Researchers such as Biecek (2019) and Hall et al. (2019) have also introduced methodologies specific to predictive modelling. The common goal of model development approaches is to standardise the process and provide specific considerations to the tasks being executed or planned for (Biecek and Burzywowski, 2021). Standardisation ensures a clear layout of the resources required to develop and maintain a model and that no essential steps are omitted during the development process (Ishikawa and Gulick, 2013; Wickham and Grolemund, 2017).

The CRISP-DM tool-agnostic procedure sections the model development process into 5 phases: data understanding, data preparation, modelling, evaluation, and deployment (Wickham and Grolemund, 2017; Hall et al., 2019). These phases can be iterated. In scientific applications, data gathering requirements may be necessary to adhere to the practices of hypothesis testing, and additionally, the outcomes of the modelling process may have scientific implications. However, data gathering in geosciences is an expensive and prolonged endeavour where projects are expected to execute a full CRISP-DM process, as well as fulfil discipline-specific tasks of data generation (e.g., through

sampling and analysis); such projects are overloaded and unnecessarily add to the already high-risk nature of data-driven research. Furthermore, it generally cannot be determined a-priori whether the data intended to be gathered would be sufficient and of high enough quality for data-driven tasks (especially abundance) until exploratory data analysis and the evaluation phase of model development have already occurred. Consequently, projects that attempt to combine data gathering and CRISP-DM would be unacceptable for a typical graduate degree duration. In the case of this study, we chose to repurpose legacy data instead and focus on the application of CRISP-DM. The risk with data repurposing is that it is not generally possible to anticipate whether the data are relevant and of sufficient quality.

Fig. 2.10 describes a variant of the iterative process – the lifecycle of a predictive model sectioned into 5 phases. Data collection and preparation precedes any modelling (Wickham and Grolemund, 2017; Hall et al., 2019). However, data collection does not generally refer to the data generation activity (e.g., sampling and analysis) that occurs in the sciences, but to the collection of modern data (e.g., samples of big data). Good quality data are essential if one aims to build a good, high-performing model and therefore extract insights from the model or deploy it. However, exploration of data, as well as using it for predictive modelling and assessing the model performance, is necessary before conclusions can be made about the quality of the data. The collected data are then investigated to understand their structure (Wickham and Grolemund, 2017; Hall et al., 2019). A model can then be selected and fitted to the data. Once the model has been constructed, it must be validated (Wickham and Grolemund, 2017; Hall et al., 2019). The data understanding, model assembly and model audit phases can be iterated until a model with the best predictive performance is obtained (Wickham and Grolemund, 2017; Hall et al., 2019). This high-performance model can thereafter be deployed once testing is complete and the necessary documentation has been developed (Wickham and Grolemund, 2017; Hall et al., 2019). Therefore, it is clear that most data-driven activities are exploratory until a first pass of the CRISP-DM process has been entirely implemented. Any refinement of models for deployment would clearly come after much of the labour had already been conducted. This is why data-driven projects are high risk because failures are only detectable after much of the work has been commissioned and executed.

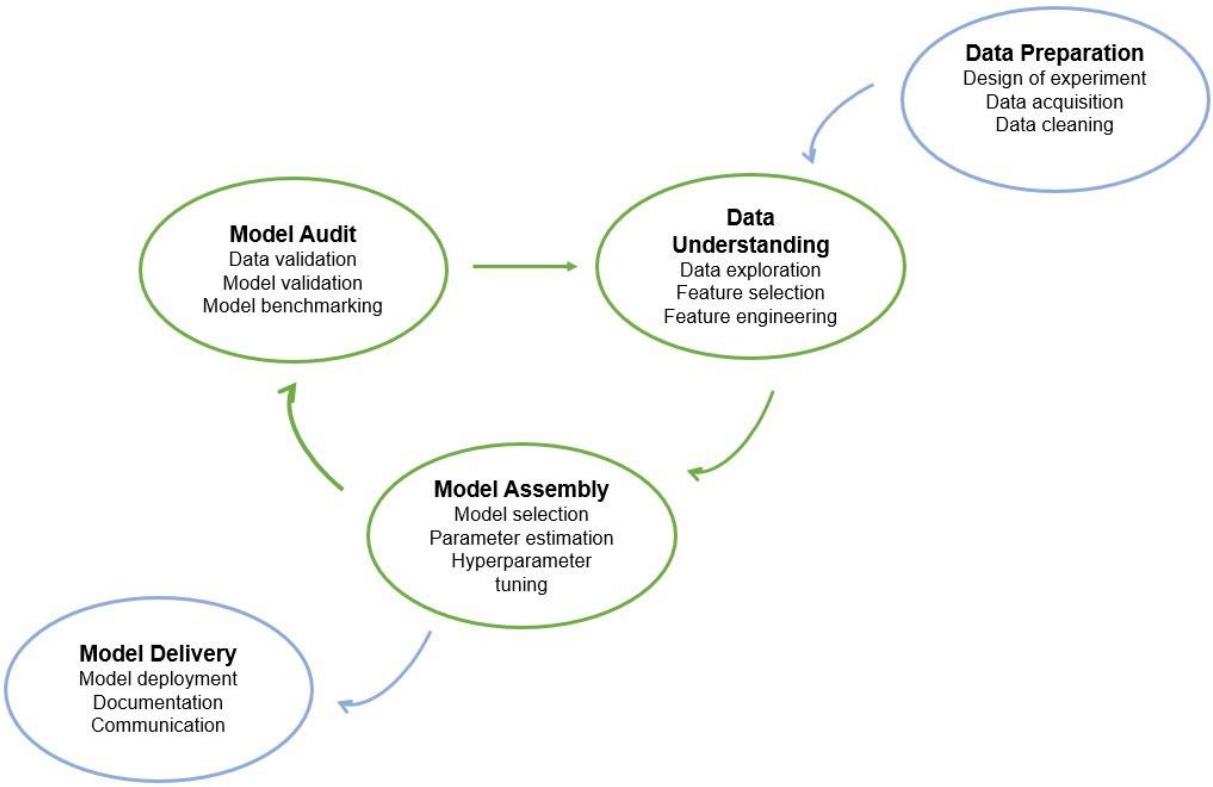


Figure 2.10: Flow chart illustrating an iterative variant of the CRISP-DM tool-agnostic predictive modelling procedure. The process is initiated by data preparation and followed by data understanding, model assembly and model audit. These three phases can be iterated. Then finally, the model delivery step concludes the process (adapted from Wickham and Grolemund, 2017 and Hall et al., 2019).

2.4.4. Data understanding

Prior to any model construction, the data has to be explored and understood. To achieve this, various data exploration tools such as statistical analysis, tabular summaries as well as various visualisations can be exploited. The type of exploration tool used is dependent on the characteristics of the dataset's features (Inglis et al., 2021). In the case of scientific data, domain-specific knowledge can be used to guide the understanding of data for ML purposes.

Exploration of features commonly investigates their distributions (Biecek and Burzywowski, 2021; Inglis et al., 2021). For a continuous target, investigations into how normal or symmetrical the distributions are are prioritised due to the prevalence of multiple methods and models that make use of the normality assumption (Biecek and Burzywowski, 2021; Inglis et al., 2021). Should a feature have an asymmetric distribution (skewness), the possibility of applying a transformation technique that could make the distribution more symmetric can be explored (Liaw and Wiener, 2002). In the case of a categorical target, it is essential to investigate the proportions of observations for each category to check for data bias (Biecek and Burzywowski, 2021; Inglis et al., 2021). Some classification methods may not produce desirable results should the categories contain a substantial imbalance in classes (Liaw and Wiener, 2002; Biecek and Burzywowski, 2021).

Investigation into the variability of features is essential because features with low variability may not be considered as valuable in model construction (Biecek and Burzywowski, 2021; Inglis et al., 2021). Predictive feature exploration should also include investigations into the relationships between the predictive features themselves as well as the relationships between the individual features and the target (Liaw and Wiener, 2002; Biecek and Burzywowski, 2021). The outcomes of such investigations can have significant implications for model construction and understanding the results of the final model (Liaw and Wiener, 2002). Should a feature appear to have no observable relationship with the target, it can be eliminated from the model construction process. This is known as feature selection or filtering (Biecek and Burzywowski, 2021; Inglis et al., 2021). Moreover, the feature can also be transformed with the aim of making its relationship with the target more linear. This is known as feature engineering (Biecek and Burzywowski, 2021; Inglis et al., 2021). Predictive feature analysis also takes note of strong correlations (Biecek and Burzywowski, 2021).

2.4.5. Model fitting

During model construction, it is common for the dataset to be split into a training set and a testing set (Wickham and Grolemund, 2017; Hall et al., 2019). The training set comprises the 'learning' dataset that is used to estimate the model coefficients, while the testing set comprises the 'validation' dataset that is used to validate the model (Ishikawa and Gulick, 2013; Hall et al., 2019). The train/test split process can be done repeatedly using k-fold cross-validation (Wickham and Grolemund, 2017; Hall et al., 2019). Estimating a model's parameters (i.e., fitting a model) is a process that is unique to the training data; thus, it varies with different algorithms (Wickham and Grolemund, 2017; Hall et al., 2019). However, this can be viewed as an optimisation problem in most instances (Biecek and Burzywowski, 2021). A model parameter refers to a configuration variable internal to the model, and its value can be predicted from the data (Wickham and Grolemund, 2017; Hall et al., 2019). Moreover, the model requires parameters to make predictions. A model with a fixed number of parameters is termed parametric, while models with a flexible number of parameters are termed non-parametric (Ishikawa and Gulick, 2013; Hall et al., 2019). A model hyperparameter refers to a configuration variable external to the model and requires user input (Biecek and Burzywowski, 2021). These can be tuned to aid the predictive modelling process; furthermore, they can also help with model parameter estimation (Biecek and Burzywowski, 2021). Hyperparameters must be optimised to adjust a model to a dataset – this is achieved by running a Grid Search technique that identifies the most ideal hyperparameters for the model (Ishikawa and Gulick, 2013).

All possible values of a model's coefficients are described in a space assigned as Θ (Hastie et al., 2009; Shmueli, 2010). During the process of fitting the model, a value of $\underline{\check{\theta}} \, \epsilon \, \Theta$ is selected to minimise a loss function $L\,(\,)$:

$$\check{\underline{\theta}} = arg \min_{\underline{\theta} \epsilon \Theta} L\left\{\underline{y}, f\left(\underline{\theta}; \underline{X}\right)\right\},$$

Where $\underline{y}$ describes the vector for the target feature's observations. $f\left(\underline{\theta}; \underline{X}\right)$ describes the resultant vector for the model's predictions calculated for model coefficients $\underline{\theta}$ and the matrix, $\underline{X}$ which comprises the predictive features' training set observation values (Hastie et al., 2009; Shmueli, 2010). Consequently, $f\left(\check{\underline{\theta}}; \underline{X}\right)$ would then describe the predicted version of the model (Hastie et al., 2009; Shmueli, 2010).

If we are to predict a new observation, the predictive feature's vector assumes a value of $\underline{x}_*$ (thus $f\left(\check{\underline{\theta}}; \underline{x}_*\right)$) (Hastie et al., 2009; Shmueli, 2010). If we assume that $E_{Y|\underline{x}_*}(Y) = f\left(\check{\underline{\theta}}; \underline{x}_*\right)$, the anticipated squared-error of prediction can be written as:

$$E_{(Y,\check{\underline{\theta}})|\underline{x}_*}\left\{Y - f\left(\check{\underline{\theta}}; \underline{x}_*\right)\right\}^2 = E_{Y|\underline{x}_*}\left\{Y - f\left(\underline{\theta}; \underline{x}_*\right)\right\}^2$$

$$+ [f\left(\underline{\theta}; \underline{x}_*\right) - E_{\check{\underline{\theta}}|\underline{x}_*}\{f\left(\check{\underline{\theta}}; \underline{x}_*\right)\}]^2$$

$$+ E_{\check{\underline{\theta}}|\underline{x}_*}[f\left(\check{\underline{\theta}}; \underline{x}_*\right) - E_{\check{\underline{\theta}}|\underline{x}_*}\{f\left(\check{\underline{\theta}}; \underline{x}_*\right)\}]^2$$

$$= Var_{Y|\underline{x}_*}(Y) + Bias^2 + Var_{\check{\underline{\theta}}|\underline{x}_*}\{\check{f}\left(\underline{x}_*\right).$$

The initial term on the right-hand side of the equation explains the variability of Y around a conditional expected value $f\left(\underline{\theta}; \underline{x}_*\right)$ that cannot be reduced (Hastie et al., 2009; Shmueli, 2010). This is the irreducible part of the prediction error. The squared differences between the expected value and its estimate are described by the second term (squared bias) (Hastie et al., 2009; Shmueli, 2010). Errors in model selection, such as opting for a simpler model, can result in bias (Hastie et al., 2009; Shmueli, 2010). The estimate's variance is described by the third term (Hastie et al., 2009; Shmueli, 2010). The equation above emphasises the importance of minimizing the sum of the squared bias and the variance estimation in predictive modelling (Hastie et al., 2009; Shmueli, 2010). That is, it ensures that the prediction error is minimised. In isolated cases, a bit of bias is accepted should it result in a lower prediction error – this would be in cases where a certain level of bias increases the estimator's precision significantly (Hastie et al., 2009; Shmueli, 2010).

2.4.6. Model audit

As illustrated in Fig. 2.10, the model development procedure is initiated by early CRISP-DM stages that are later fine-tuned in successive iterations (Hall et al., 2019). Numerous candidate models are then evaluated to attain a final model (Hall et al., 2019). This can be attained using various techniques. A useful technique may be one that breaks down the model's predictions into components that can be linked back to specific predictive features (Lary et al., 2016; Biecek and Burzywowski, 2021).

Another useful evaluation technique would be one that can assess sensitivity analysis for a model's predictions (Hall et al., 2019; Biecek and Burzywowski, 2021). The model audit component can also encompass summaries of the model's predictive performance – this can, in turn, give insight into the model development's progress (Hall et al., 2019). Moreover, assessments into the significance of various predictive features as well as their effect on the model's predictions will also be of use (Hall et al., 2019). These may highlight predictive features that can be excluded from the model's fine-tuning process (Goswami et al., 2016; Hall et al., 2019). Finally, thorough investigations into the model's large- and small-scale (observation level) performance are vital. In this instance, residual-diagnostic tools may shed light on the various motivations behind a model's performance (Liaw and Wiener, 2002).

The techniques described are used to assess the current model's performance. Results attained from these investigations can thus be used to improve the model – these improvements can then be implemented in successive iterative CRISP models or in a fine-tuning phase (Hall et al., 2019; Biecek and Burzywowski, 2021). One commonly used regression ML performance metric is the Mean Absolute Error (MAE), which is calculated by averaging the absolute distance between the actual values and the predicted values (Botchkarev, 2018):

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

where n is the total number of datapoints, y is the actual/true output value, and $\hat{y}$ is the predicted output value. A lower MAE value alludes to higher model performance –. for example, if we have an MAE of 12 g/t, it means that, on average, the model is predicting 12 g/t more or less than the actual value. Another frequently used metric is the Root Mean Square Error (RMSE), which represents the sample standard deviation of the disparities between the predicted and observed values (Alencar et al., 2017; Botchkarev, 2018). It is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y - \hat{y})^2}{n}}.$$

The RMSE has been deemed as a better performance metric than the MAE because it squares the errors before taking an average (Wang and King, 2010). Consequently, larger errors receive a more severe penalty. The Coefficient of Determination (CoD), also known as R^2, tells us how good a model is for a dataset by measuring the extent to which the independent variable adjusted with the variance (Alencar et al., 2017; Botchkarev, 2018). It is calculated as follows:

$$CoD = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

This equation can be explained graphically, as illustrated in Fig 2.11. The CoD value usually lies between 0 and 1, and it can be negative should the regression line have a worse fit than the average (Alencar et al., 2017; Botchkarev, 2018). The further the value is from 1, the poorer the model.

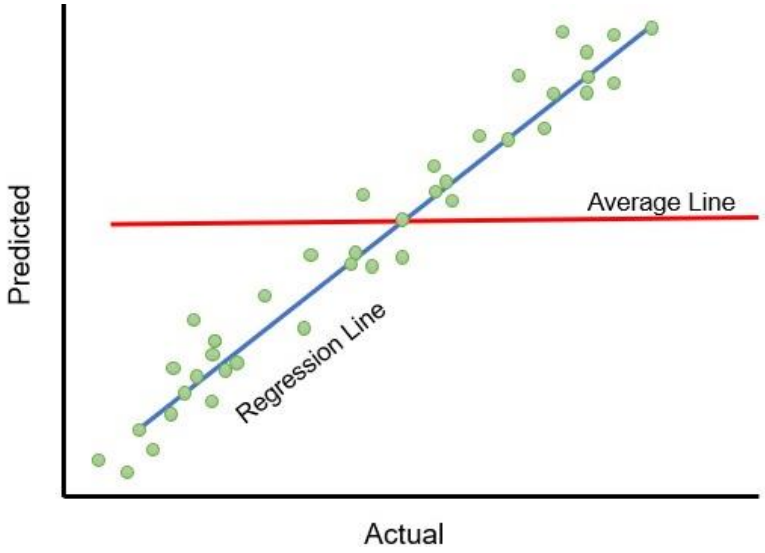


Figure 2.11: Graphical explanation of the CoD metric (adapted from Alencar et al., 2017 and Botchkarev, 2018).

The Mean Absolute Percentage Error (MAPE) metric also measures model performance using the disparity between the actual and predicted values (Kim and Kim, 2016). It can be regarded as a regression loss function to define the error termed by the model evaluation (Alencar et al., 2017). A lower MAPE value alludes to a better fitting model. It is calculated as follows:

$$MAPE = \frac{100}{n}\sum_{i}^{n}\frac{|y - \hat{y}|}{|y|}.$$

Lastly, the Relative Error (RE) evaluates the model performance by measuring how close the predicted value is to the true value (Botchkarev, 2018). The RE value is usually between 0 and 1; the closer it is to 1, the better the model. Multiplying the RE by 100 will return a percentage (Alencar et al., 2017; Botchkarev, 2018). It is calculated as follows:

$$RE = \left|\frac{\hat{y} - y}{y}\right|.$$

3. METHODOLOGY

The following chapter provides a step-by-step workflow of the experimental methodology followed for this research project.

3.1. Dataset

To minimise project risk, this study repurposes existing legacy data instead of seeking to generate any data. Sibanye-Stillwater provided a dataset comprising information derived from rock samples collected from their Kloof mine. The experimental process entailed coding and was conducted using the Python programming language. The software used was Jupyter Notebook.

- The experimental process was initiated by importing Python libraries that were to be used. These included directory libraries, pandas, NumPy and various sub-libraries in the Scikit-learn ML package that was used in the modelling process.
- The working directory was set in the next step. The dataset was loaded as a .csv file, and a folder that stores the images created in Jupyter notebook was formed.
- Once the dataset was loaded, it was analysed to determine the number of records, the number of features, and the nature of the features, i.e., whether they were categorical or qualitative. Similarly, the target feature was also analysed.
- Given the extent of the dataset, a subset block that contained a cluster of datapoints was isolated and extracted – subsetting ensured a consistent portion of the dataset was used for modelling. This was the set of data used in this project.
- Subsequently, the variable names were set. This was to ensure an easy way to extract specific data in the steps to follow. The spatial data were grouped under ‘meta_names’, the predictive features were grouped under ‘feature_names’, and the target feature was assigned as ‘target_name’. The dataset comprised four predictive features (sedimentological properties): Channel Width, Conglomerate Percentage, Internal Waste and Basal Contact. The physical relationship between the Internal Waste and Conglomerate Percentage is illustrated in Fig. 3.1.

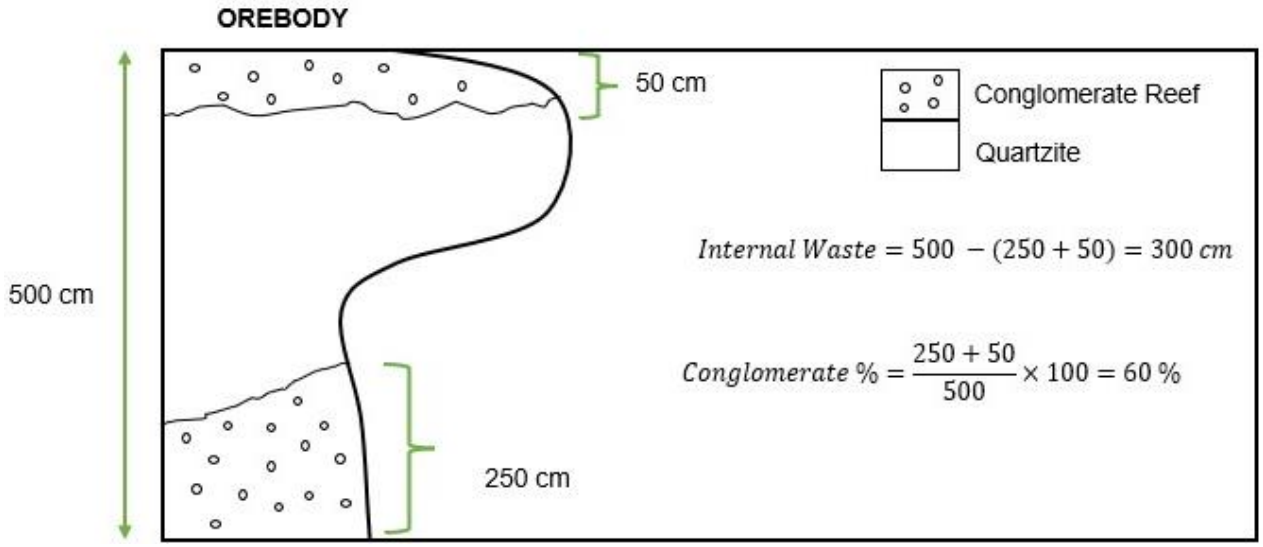


Figure 3.1. Simplified sketch of the physical relationship between the Internal Waste (cm) and Conglomerate Percentage.

3.2. Data pre-processing

Prior to any modelling, the dataset was analysed and cleaned to ensure it was in the most desirable form for predictive modelling:

- The subset was filtered to exclude observations that contained values equal to 0 for the target gold feature, as well as to exclude any observations with missing values for any of the predictive features.
- The 'MultiColumnLabelEncoder' was used to encode categorical features into qualitative integers in the next step.
- The data's distribution, as well as various statistical summaries, were investigated in the next step. This included the creation of visualisations like spatial maps, histogram plots and box and whisker plots that give us more information on the data from a large-scale and small-scale point of view.
- In the outlier detection step, the dataset was filtered to exclude data that fell outside a specified percentile. In this case, any datapoints that fell outside the upper and lower limits of two percentile-based thresholds were removed, while ensuring that > 90% of the original data was retained.
- Subsequently, a correlation matrix was created to analyse the relationships between the features and the target. In this step, the Pearson Correlation Coefficient was used.
- A prediction dataframe that included only the predictive features and the target was created in the subsequent step.
- The features were prepared for prediction by a scaling process. The Standard Scaler was used to ensure that all features would contribute equally to the model fitting. The predictive features were

standardised by removing the mean and scaling to unit variance. This process is formulated below:

$$y = \frac{(x - mean)}{std},$$

where y is the standard score, x is the sample, and std is the standard deviation. The target feature was not scaled. It is important to note that all predictive modelling steps were performed with scaled predictive features.

Once the pre-processing stage was complete, the predictive modelling phase could commence. This was run over two experiments, with the first making use of select MLAs in a straight-forward method. The second experimental phase repeated the prediction process with a dataset that had been run through a clustering algorithm first.

3.3. Experiment 1

- The dataset was split into a training and testing set using the 'train_test_split' library. The code was configured to split 30% of the dataset into the testing set and the remaining 70% into the training set.
- The ML algorithms were initially applied to the dataset using a straightforward model-fitting approach. A grid search specific to each algorithm was computed to calculate the hyperparameters that would return the most optimal prediction results. The ML algorithms used in this step are described in the following section:

3.3.1. Random Forest Algorithm

A Random Forest (RF) is a non-parametric algorithm that uses an ensemble of decision trees (Breiman, 2001). It can be used in both regression and classification ML tasks and makes use of aforementioned ensemble learning techniques. A forest of decision trees is built using the ensemble meta-algorithm known as bagging (Breiman, 2001; Wright and Ziegler, 2017). Bagging is incorporated to enhance the accuracy and performance of ML algorithms (Breiman, 2001). In the case of RF, trees are built using bootstrap sampled features, such that the trees are uncorrelated. Subsequently, an output is averaged from the results of the entire forest (Breiman, 2001). The more trees used in the prediction process, the more accurate the outcome (Breiman, 2001; Wright and Ziegler, 2017). A RF works around the constraints of a simple decision tree algorithm by reducing overfitting and increasing accuracy (Breiman, 2001). The RF model hyperparameters are outlined in Table 3.1.

Table 3.1: RF model hyperparameters

Hyperparameter	Function
number of trees	Sets the number of trees needed in the RF
max features	Sets the number of features at each node
max depth	Sets the depth of the tree (restricts how much the tree can grow)
min samples per split	Sets the minimum number of samples required to split a node
min samples per leaf	Sets the minimum number of samples in the leaf node

Decision Trees

A decision tree (DT), the foundational component of a RF algorithm, comprises three constituents: decision nodes, leaf nodes and a root node (Luckert and Schaefer-Kehnert, 2016). When the DT algorithm is implemented, the training (learning) dataset is separated into branches that are then sectioned into more branches (Goswami et al., 2011; Wright and Ziegler, 2017). The process is iterated until a leaf node that cannot be branched out any further is reached, which can be set as a model hyperparameter (Luckert and Schaefer-Kehnert, 2016). A DT's nodes are representative of features used to predict the outcome; moreover, its decision nodes link to the tree leaves (Fig. 3.2) (Luckert and Schaefer-Kehnert, 2016).

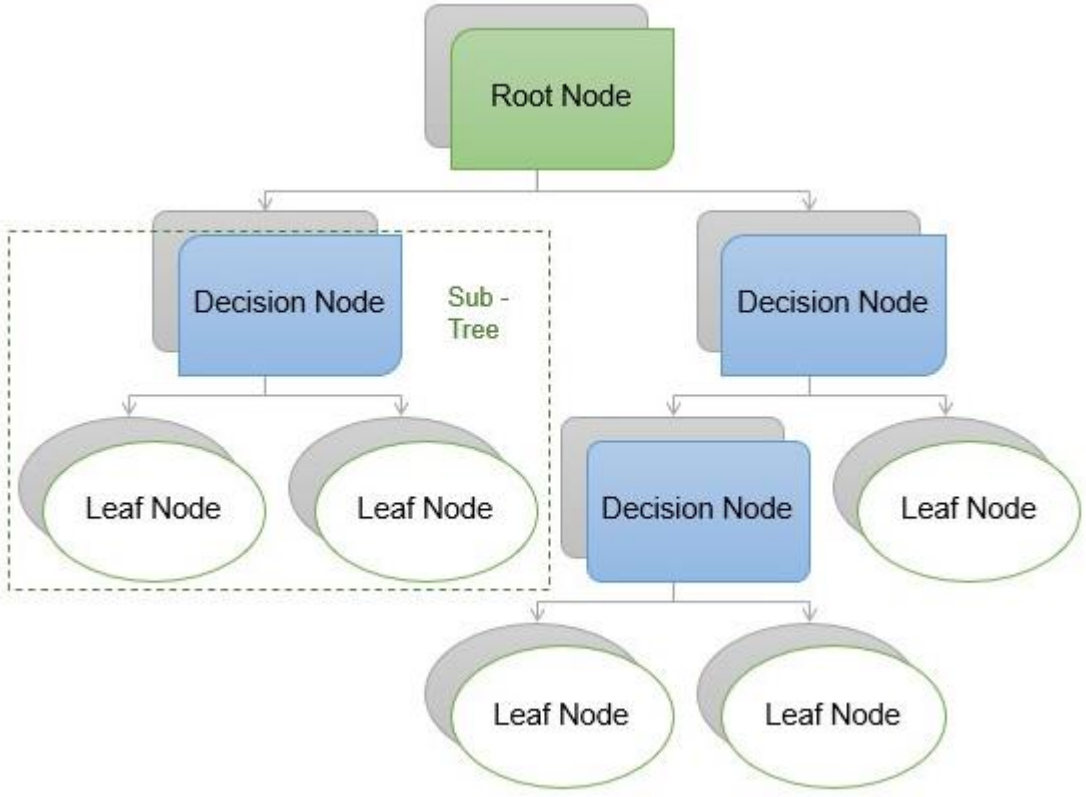


Figure 3.2: Illustration of a decision tree showing the Root, Decision and Leaf Nodes (adapted from Luckert and Schaefer-Kehnert, 2016).

Random Forest regression

RF regression follows a concept similar to simple regression. Each tree generates a unique prediction, and the mean of the individual trees is taken as the output of the regression (Wright and Ziegler, 2017). RF regression differs from linear regression (LR) in that its function is much more complex, in contrast to LR's simple $y = bx + c$ equation where the target y is predicted from the feature x, the estimation parameter b and the constant c.

Feature Importance

Each feature in the dataset can be analysed to determine how greatly it affects the predictive quality of the overall model. In other words, we can determine how relevant a certain property is to the output target (gold prediction). The Random Forest algorithm uses an information gain metric to split leaves, called the Gini importance. It can also be used to measure feature importance and can be given by (Saarela and Jauhiainen, 2021):

$$\text{Gini Importance} = p_1(1 - p_1) + p_2(1 - p_2),$$

where p_1 and p_2 represent the probabilities of classes 1 and 2. As the probabilities approach 0, the Gini index decreases (Saarela and Jauhiainen, 2021). The lower the Gini index, the more important the input feature in the sense that it contributes more to the outcome of the prediction (Saarela and Jauhiainen, 2021).

3.3.2. AdaBoost Algorithm

AdaBoost is a boosted (and also ensemble) non-parametric algorithm that compiles a number of weak learners, such as individual decision trees, and puts them together to form a strong learner (Koduri et al., 2019; Wang et al., 2021). It can be used for ML tasks involving both classification and regression. During the AdaBoost process, the training set's data are analysed, the first weak learner is identified, and incorrectly predicted samples are extracted (Jafrasteh et al., 2018; Wang et al., 2021). The extracted datapoints are combined with an untrained dataset to form a new training set for the algorithm to learn from (Koduri et al., 2019; Lishchuk et al., 2019). A second weak learner is identified from the new training set, and the process is repeated (Koduri et al., 2019; Wang et al., 2021). In this way, the subsequent learners attempt to correct the previous ones in a sequence. The boosting cycle is iterated until an optimal, high-performing learner is obtained.

In each boosting iteration, weights $w_1, w_2 \ldots, w_N$ are applied to the training data (Koduri et al., 2019; Wang et al., 2021). Prior to any iterations, the weights are all set to $w_i = \frac{1}{N}$ and the first weak learner is trained on the original data (Jafrasteh et al., 2018; Wang et al., 2021). The weights are independently modified in each successive iteration, and the AdaBoost algorithm is then reapplied to

the reweighted data (Koduri et al., 2019; Wang et al., 2021). In the case where the boosting model provides incorrect training sample predictions in a previous iteration, their weights are increased in the successive iteration (Jafrasteh et al., 2018). Alternatively, correctly predicted samples are assigned lower weights (Koduri et al., 2019; Wang et al., 2021). This ensures that the AdaBoost algorithm concentrates on incorrect predictions (Koduri et al., 2019; Wang et al., 2021). Essentially, the predictions attained from each step are combined over a weighed sum, and a final prediction is made. This process can be formulated as follows (Wang et al., 2021):

$$F_n(x) = F_{m-1}(x) + \arg\min_h \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + h(x_i)\big),$$

where $F_n(x)$ refers to the final model. $F_{m-1}(x)$ is the model obtained in the previous iterations. y_i refers to the i^{th} tree's prediction result and $h(x_i)$ refers to the newly added tree. Although the base learner could, in theory, be any weak learner, the original AdaBoost formulation uses decision trees as the base learner. All hyperparameters, except for the number of trees, are directly inherited from the base estimator (learner).

3.3.3. K-Nearest Neighbours Algorithm

K-Nearest neighbours (KNN) is a non-parametric algorithm that can be used in both classification and regression ML tasks. The KNN algorithm attempts to assign classes to the test dataset by computing the distance between the testing set's datapoints and all the training set's datapoints in the feature space (Imandoust and Bolandraftar, 2013; Novitasari et al., 2019). The 'K' number of points closest to the test data are then chosen; this is a model hyperparameter (number of neighbours). In classification tasks, the probability of the testing set's datapoints belonging to the classes of K training data are thereafter computed (Imandoust and Bolandraftar, 2013; Novitasari et al., 2019). In our regression task, the mean of the K training set's datapoints is computed to attain testing set values (Imandoust and Bolandraftar, 2013; Novitasari et al., 2019).

The algorithm's workflow is described as follows (Fig. 3.3): firstly, the distance between the training set datapoints and the new testing set data are computed. Based on the computed distances, the nearest K datapoints are selected. An average of these selected K datapoints is then taken as the final prediction value.

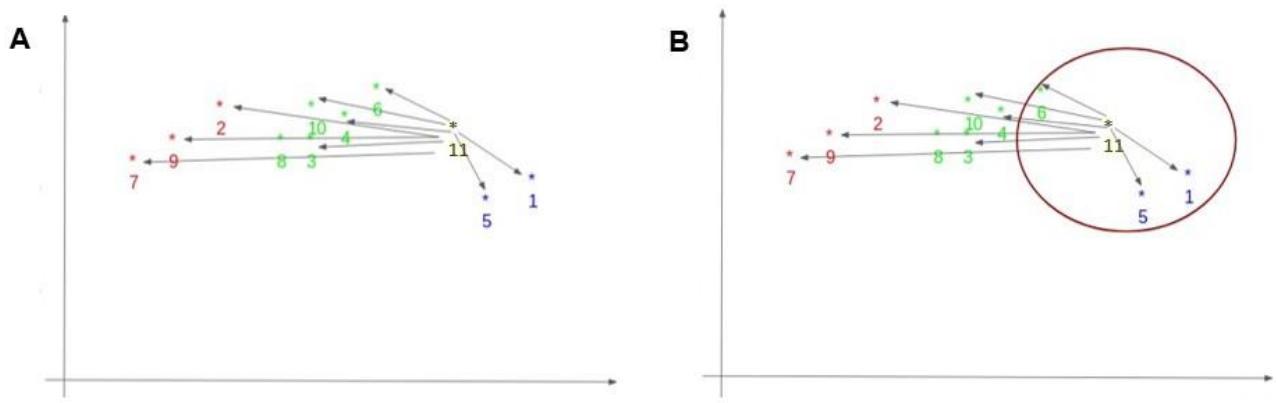


Figure 3.3: Simplified sketch illustrating the KNN algorithm's process. The training data's values are represented by red, green and blue, while black represents a value from the testing set. The coordinates are arbitrary feature dimensions. In subfigure A, the distance between the training and testing datapoints is computed. In subfigure B, we assume that k = 3, for example; in this case, 5,1 and 6 will be selected as our point's nearest neighbours based on their proximity (adapted from Kramer, 2013).

The distance between each training point and the testing points to be predicted is commonly computed using the Euclidean Distance function. The function calculates the square root of the sum of the squared differences between the point to be predicted (x) and the training point (y) (Imandoust and Bolandraftar, 2013; Novitasari et al., 2019). The formula is shown below:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}\,.$$

Alternatively, the Manhattan Distance function can also be used. This function computes the distance between real vectors by taking the sum of their absolute difference (Imandoust and Bolandraftar, 2013; Novitasari et al., 2019). The formula is shown below:

$$\sum_{i=1}^{k}|x_i - y_i|.$$

3.3.4. Elastic Net Algorithm

The Elastic Net (EN) regressor is a linear parametric algorithm that uses a penalised least-squares estimation technique (Liu and Li, 2017; Al-Jawarneh et al., 2020). The EN regressor uses a combination of the ridge (L_2 penalty) and the LASSO[1] (L_1 penalty) regression methods (Liu and Li, 2017; Al-Jawarneh et al., 2020). In EN regression, a generalised regression model is fitted to look for a relationship between the predictive features (x) and the target feature (y) (Liu and Li, 2017; Al-Jawarneh et al., 2020). This process can be formulated as follows:

[1] Least Absolute Shrinkage and Selection Operator.

$$y = \sum_{j=1}^{p} x_j \beta_j + \varepsilon,$$

where x_j is representative of the model's predictors and β_j is representative of the regression coefficient. Furthermore, ε represents the random observation error. For the sake of simplification, the x and y features are standardised (Liu and Li, 2017; Al-Jawarneh et al., 2020).

The EN regressor makes use of shrinkage and variable selection techniques. β_j, the regression coefficient, is shrunk towards 0 or equated to exactly 0 if possible (Liu and Li, 2017; Al-Jawarneh et al., 2020). This process seeks to eliminate or minimise the number of model parameters and the contribution of non-relevant features. The regression process ensures that the remaining coefficients have a non-zero value (Liu and Li, 2017; Al-Jawarneh et al., 2020). Once the shrinkage stage is complete, the remaining non-zero regression coefficients are selected during the model fitting (Liu and Li, 2017; Al-Jawarneh et al., 2020). Coefficient shrinkage is a method to deal with excessive model variance, especially in the case of feature multicollinearity. Feature multicollinearity implies that with changes in one feature, there would be correlated changes in other features such that a model naively fitted to this type of data would exhibit excessive output variance with respect to input. The EN's model hyperparameters include the L_1 ratio, the ratio of the L_1 penalty; alpha(α), the strength of regularisation; and max iterations, which outlines the number of iterations (Liu and Li, 2017; Al-Jawarneh et al., 2020). The EN algorithm addresses multicollinearity between the predictive features and the target variable by reducing model variance through penalising both the L_1 and L_2 norms of the model coefficients, thus reducing the prediction error. The EN algorithm minimises the following objective function where w is the sample weight:

$$\frac{1}{(2 \times n_{sample})} \times \|y - X_w\|_2^2 + \left(\lambda \times L_{1\,ratio} \times \|w\|\right) + \left(0.5\lambda\left(1 - L_{1_{ratio}}\right) \times \|w\|_2^2.\right.$$

The coefficient's (β) vector, which is unknown, is attained using the OLS (Ordinary Least Squares) estimator. The resulting vector regression coefficient is dependent on the minimised residual sum of squares (RSS) (Liu and Li, 2017; Al-Jawarneh et al., 2020). The RSS sums up the differences between the actual y values and the predicted $\check{y}$ values. This is formulated below (Al-Jawarneh et al., 2020):

$$RSS\,(\boldsymbol{\beta}) = \sum_{l=1}^{m} (y_l - \check{y}_l)^2 = \|y_l - \check{y}_l\|^2.$$

$\check{y} = X\check{\boldsymbol{\beta}}$ represents the estimated vector model (Al-Jawarneh et al., 2020). The OLS estimation of the coefficient's vector is formulated as follows (Al-Jawarneh et al., 2020):

$$\breve{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} \|y_l - \breve{y}_l\|^2.$$

3.3.5. Cross-Validation

Each algorithm used the K-fold cross-validation technique to measure model performance. The K-fold cross-validation technique uses a simple flow illustrated in Fig. 3.4. The subset is (usually randomly) split into K 'folds' of equal size following train-test splitting (Kapageridis, 2005). Each fold acts as the training set K – 1 times and the testing set 1 time. Taking an average of the testing results returns cross-validation scores. In this project, each algorithm was run over 10 K-fold cross-validations.

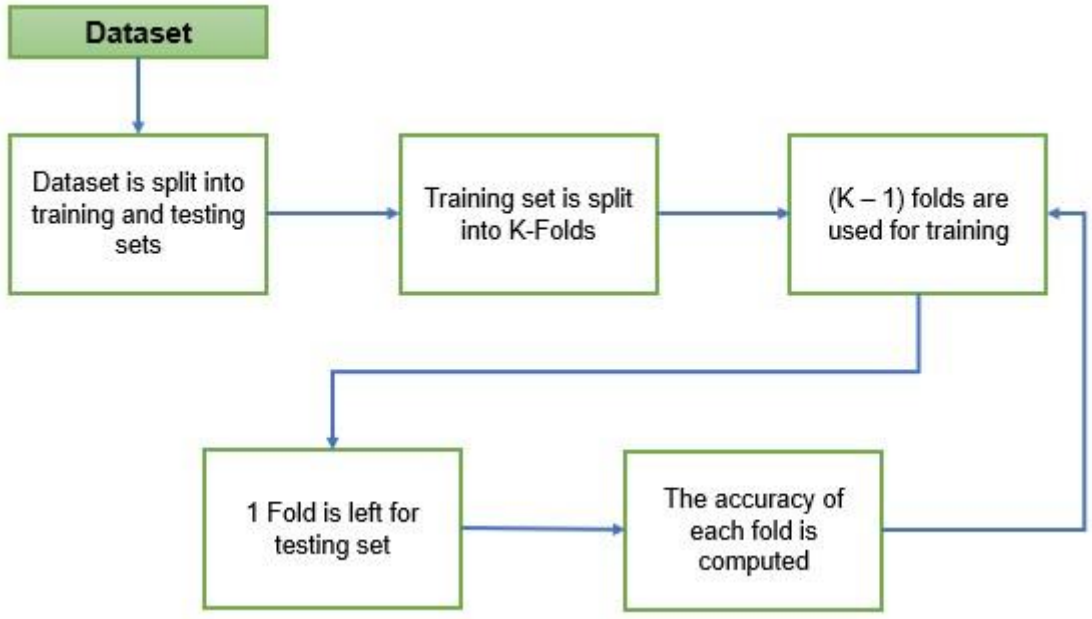


Figure 3.4: K-Fold Cross-Validation (adapted from Kapageridis, 2005).

3.4. Experiment 2

Once the first straightforward predictive modelling step was complete, a different modelling process was implemented to evaluate whether the prediction results would improve or change. This step introduced a clustering algorithm to classify our dataset into different clusters. The clustering algorithm essentially assigned each datapoint to a group (cluster); this information was then compiled in a new column and used as a feature (clustering can be used as a feature engineering step in supervised ML tasks). The predictive modelling step was thereafter repeated, and the predictions were performed per cluster for each algorithm. The K-Means clustering algorithm used is described below:

3.4.1. K-Means clustering

K-Means is a non-supervised MLA that clusters data by feature similarity. It works by calculating cluster centres (centroids) (Ahmar et al., 2018; Nainggolan et al., 2019). The process is then iterated up until the best centroid is attained. The 'K' in K-Means refers to the number of clusters, which is a model hyperparameter. The clustering algorithm allocates datapoints to a cluster by ensuring that the

sum of the squared distances between the datapoints and the cluster centre is at its most minimal (Dutta et al., 2010; Nainggolan et al., 2019).

The process followed by the algorithm is described (Ahmar et al., 2018; Nainggolan et al., 2019):

- The number of clusters (K) to be created from the dataset is provided to initiate the process.
- Datapoints are then randomly chosen and allocated to a specific cluster.
- The algorithm then generates the cluster centroids.
- The following stages are repeated up until the best centroid is attained:
    - Firstly, the algorithm computes the sum of squared distances between the datapoints and centroids.
    - Each datapoint is then assigned to a cluster based on the other datapoints it is in proximity with.
    - The cluster's centroid is thereafter computed by taking an average of all the datapoints within the cluster.

The procedure followed by the K-Means clustering algorithm makes use of the Expectation-Maximisation technique (Dutta et al., 2010; Ahmar et al., 2018). The Expectation stage manages the allocation of datapoints to the closest cluster, while the Maximisation stage manages the computation of each cluster's centre (Ahmar et al., 2018; Nainggolan et al., 2019).

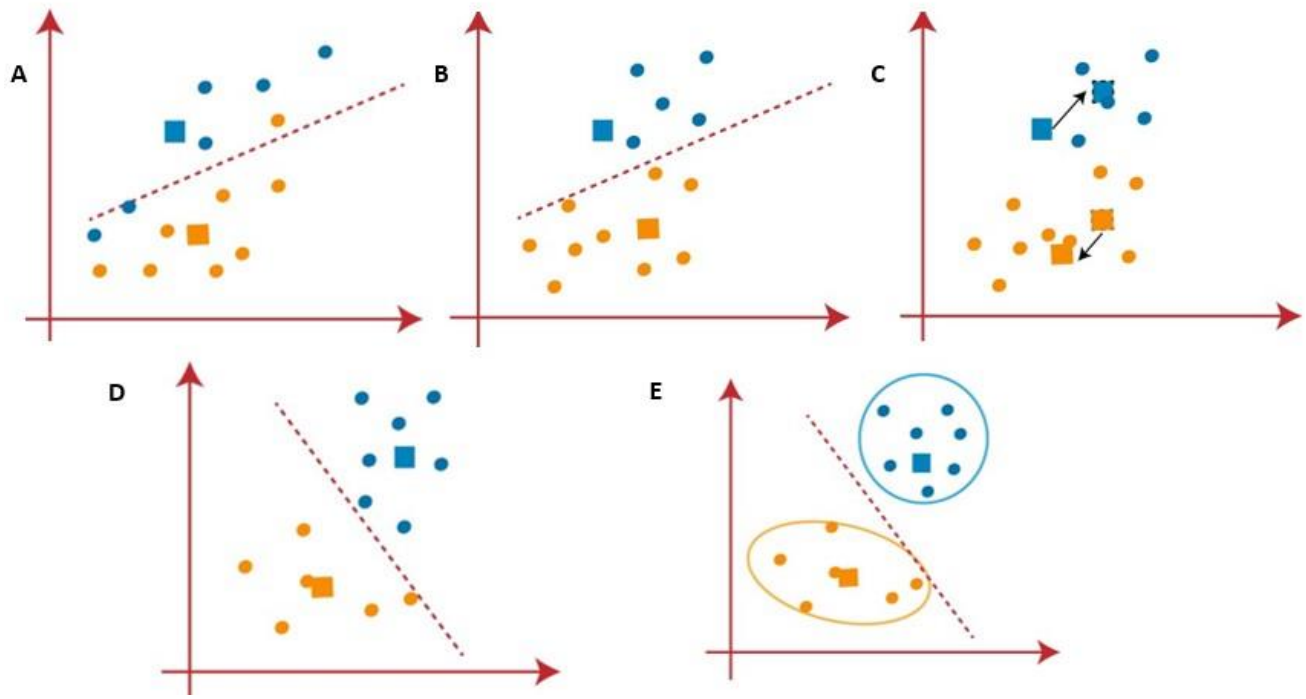


Figure 3.5: Graphical representation of the K-Means algorithm's clustering process. A K of 2 is assigned, and the clustered datapoints are represented in blue and orange. The median line (dashed) is generated by computing the median between the two centroids (adapted from Tan et al., 2021).

Consider an example where a K of 2 has been allocated (Fig.3.5): In A, two random datapoints are chosen as cluster centroids, and the surrounding datapoints are then clustered based on the centroid they are closest to. This is computed by determining the median between both centroids (represented as the dashed median line). As illustrated in B, the process is iterated with different centroids selected, and the datapoints are then reallocated to their appropriate clusters. Following successful reallocation, the centroids' centre of gravity is computed as illustrated in C. Once again, new centroids are computed, and a median is generated to reallocate the datapoints to the closest cluster (D). Finally, the datapoints are grouped based on their proximity to the median line (E). The final clusters are distinct, with no overlapping points in either group. In this project, the Silhouette Algorithm (Shitaywi and Kachouie, 2021) was used to determine the value of K.

3.5. Post-prediction investigations

Once the first and second prediction attempts were completed, further attempts to investigate the data and attain improved prediction results were conducted.

3.5.1. Residuals Analysis

The dataset's residuals were calculated by subtracting the predicted target (gold grade) values from the original target values:

$$Residuals = \hat{y} - y.$$

The relative errors (RE) were thereafter computed using the following formula (as outlined in section 2.4.6.):

$$RE = \left|\frac{residuals}{y}\right|.$$

Once the relative errors were computed, the dataset was then evaluated in sets according to its relative errors – that is, observations with relative errors that fell between 0 and 0.1 were grouped, and those between 0.1 and 0.2 were grouped together and so on. Subsequently, the predictive modelling processes described above were run on each sub-dataset, and the phased prediction results were collected. This process was thereafter used to isolate datapoints that improve the prediction performance and those that do not.

3.5.2. Obtaining predictions from the entire 'good' dataset

Once the well-predicting portion of the dataset was isolated, the entire 'good' set was run through the prediction process using randomised shuffling. The 'dataset was initially predicted using a test size of 30% and training size of 70%, and a data frame to store all the testing results was created. The entire set was thereafter predicted through a randomisation function where each datapoint was repeatedly removed, and then returned into the dataset and predicted again. The resulting scores were shuffled

N = 100 times over 5-fold cross-validations, and an average was taken to give a final model where all the datapoints had been predicted.

3.6. Assessing the efficiency of sedimentological features in gold grade prediction

3.6.1. Feature Selection

In this step, the predictive features were removed one by one, and the predictive modelling process was repeated with the remaining features. This was to assess if the prediction scores/accuracy improved when a certain sedimentological feature was missing or whether its absence lowered the scores.

3.6.2. Feature Importance

The feature importance statistics were computed using the tree-based algorithms' best parameters. These statistics were visualised graphically and assessed to understand which sedimentological features were considered most important for gold grade prediction. Feature importance statistics were generated on a larger scale for the entire subset and on a smaller scale for well-predicting and poor-predicting portions of the data.

3.7. Kriging Block Model

Once the ML predictive modelling process was completed, 2D block models were generated using the traditional kriging estimation method described in Chapter 2.2.1, specifically, Ordinary Kriging (OK). Block models are spatially georeferenced and divided into fixed-size blocks (Ansah, 2018). The block model's geometry can be dependent on the deposit's characteristics, mine planning requirements, and various geological properties (Erarslan, 2012). In this study, we used block models to demonstrate observable changes in the resource model with changes in data quality. Unfortunately, we are not able to compare our estimates with actual metal yields.

3.7.1. Ordinary Kriging

OK and the IDW (Inverse Distance Weighting) estimation method are alike in that they apply weights to explain grade similarities according to distance. OK, however, differs in that a variogram model is used to select the weights; this is in accordance with the sample's direction and distance from the estimation point (Rossi and Deutsch, 2014; Coleman, 2021). Consequently, the weights take the spatial correlation between the samples and estimation point into consideration (Fig. 3.6).

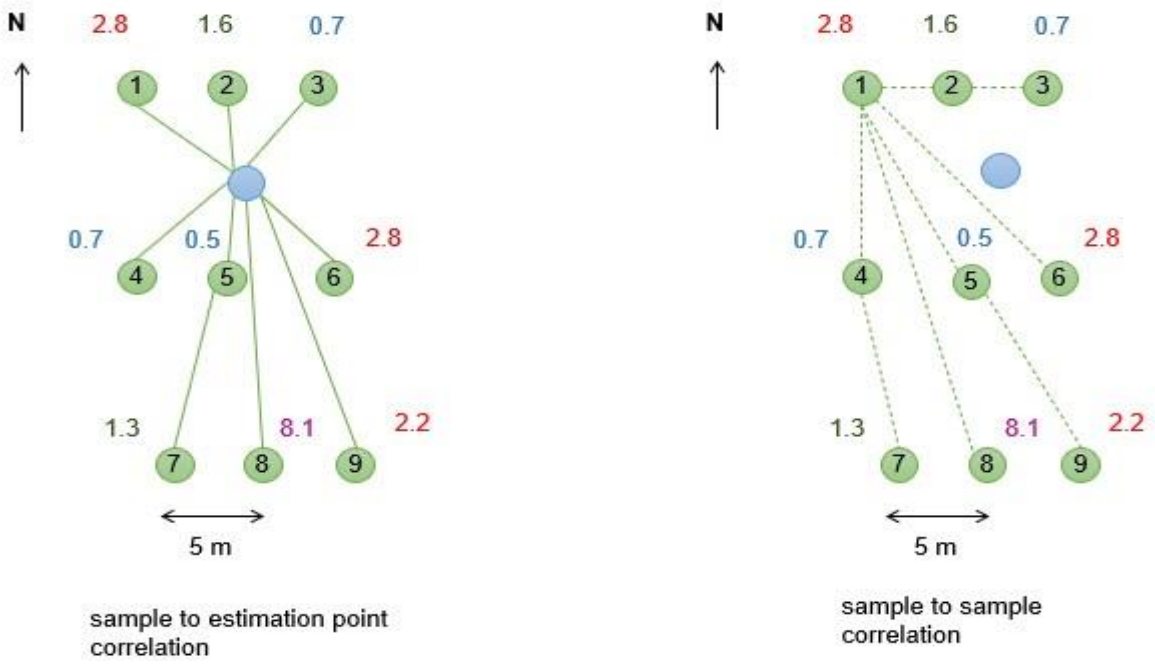


Figure 3.6: Samples weighed by variogram in OK estimation (adapted from Coleman, 2021).

- The kriging estimation equation can be formulated as follows:

$$Estimate = \sum sample\ value \times kriging\ weight.$$

- The kriging estimate derived should have the least overall errors and would be unbiased (Yasrebi et al., 2009; Rossi and Deutsch, 2014). A least overall error means the overall variance between the estimate and the true values must be at a minimum:

$$error = (estimate - true\ value)^2.$$

- The LaGrange multiplier, an adjustment factor, is applied to ensure the weights sum to one, as illustrated in the following updated equation (Yasrebi et al., 2009; Rossi and Deutsch, 2014):

$$error = \left[\sum (sample\ value \times kriging\ weight) - true\ value\right]^2 + [LaGrange\ multiplier \times (\sum weights - 1)].$$

- The equation above can be differentiated, resulting in a kriging system of equations used to determine kriging weights (Yasrebi et al., 2009; Rossi and Deutsch, 2014). The kriging system, which can be summarised in three matrices, is formulated below: where μ is the LaGrange multiplier, γ is the variogram, and λ is the kriging weight. Matrix A summarises the variogram values between all the samples used for estimation, B is the matrix of weights the system attempts to estimate, and C is the matrix of variogram values for the distances and directions between the samples and the estimation point.

Matrix A		**Matrix B**	**Matrix C**
Sample to sample correlation	x	*weight*	= *sample to estimation point correlation*

$$\begin{pmatrix} \gamma(1,1) & \gamma(1,2) & \cdots & \gamma(1,9) & 1 \\ \gamma(2,1) & \gamma(2,2) & \cdots & \gamma(2,9) & 1 \\ \vdots & \vdots & \cdots & \vdots & 1 \\ \vdots & \vdots & \cdots & \vdots & 1 \\ \vdots & \vdots & \cdots & \vdots & 1 \\ \vdots & \vdots & \cdots & \vdots & 1 \\ \vdots & \vdots & \cdots & \vdots & 1 \\ \vdots & \vdots & \cdots & \vdots & 1 \\ \gamma(9,1) & \gamma(9,2) & \cdots & \gamma(9,9) & 1 \\ 1 & \vdots & \cdots & 1 & 0 \end{pmatrix} \times \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \\ \lambda_7 \\ \lambda_8 \\ \lambda_9 \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(\bullet,1) \\ \gamma(\bullet,2) \\ \gamma(\bullet,3) \\ \gamma(\bullet,4) \\ \gamma(\bullet,5) \\ \gamma(\bullet,6) \\ \gamma(\bullet,7) \\ \gamma(\bullet,8) \\ \gamma(\bullet,9) \\ 1 \end{pmatrix}.$$

3.7.2. Validation

The quality of the estimates was assessed using quality check measures for kriging, specifically the Kriging Slope of Regression (SLOR) and Kriging Efficiency (KEFF).

Kriging Slope of Regression (SLOR)

The SLOR refers to the linear regression of the estimated block values against true values and is used to approximate the kriging estimate's conditional bias (Kentwell, 2022). It is a function of the variogram model and the local block/sample geometry and is calculated at each block. It is described in the equation below (Deutsch et al., 2014):

$$b = \frac{Cov\{Z_V, Z_V^*\}}{\sigma_{Z_V^*}^2} = \frac{\sum_{i=1}^{n} \lambda_i \bar{C}(v_j, V)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(v_i, v_j)},$$

where b is the slope of regression; Z is the unknown true value of a random variable; Z^* is the known estimate of the random variable; V is the estimation volume and $C(v_i, v_j)$ is the covariance between two data locations v_i and v_j. In OK, the SLOR is usually less than one, thus indicating a conditional bias (Deutsch et al., 2014; Deutsch and Deutsch, 2015). The greater the deviation from the SLOR, the greater the conditional bias, as illustrated in Fig. 3.7.

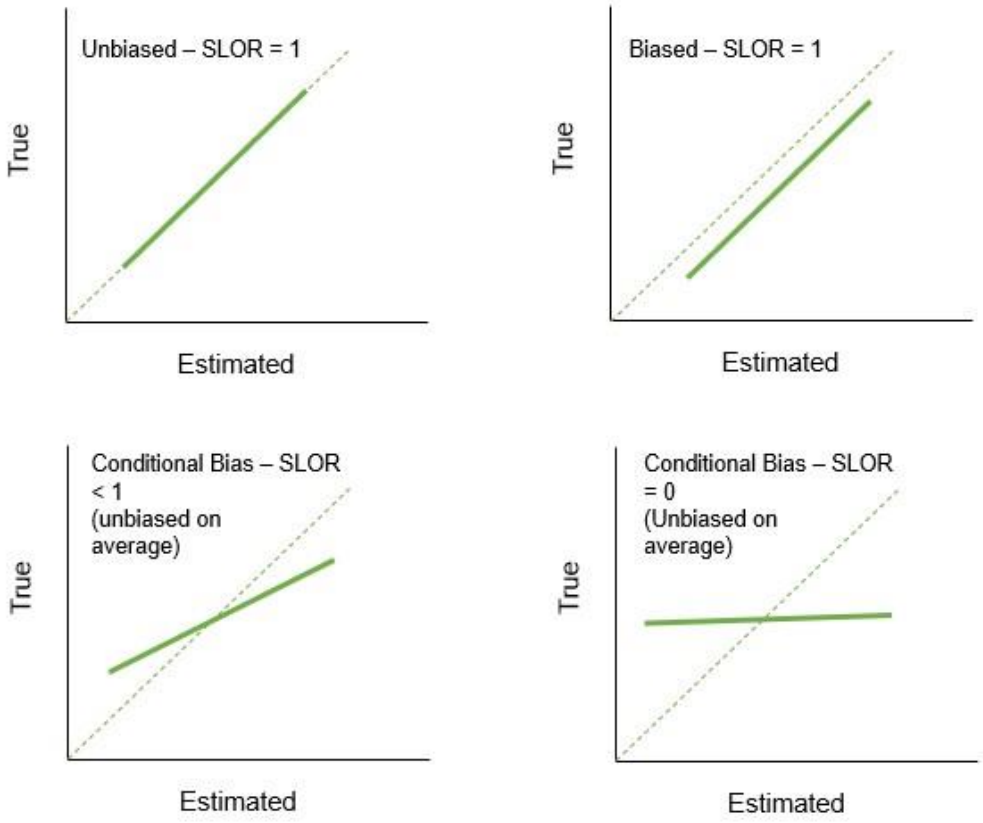


Figure 3.7: SLOR and bias (adapted from Kentwell, 2022).

Kriging Efficiency (KEFF)

The KEFF is explained as the proportion of kriging variance (Deutsch and Deutch, 2015). It provides a measure of the efficiency of block estimates and can be formulated as follows (Deutsch et al., 2014):

$$KE_{DK}(\%) = \frac{\sigma^2 - \sigma_K^2}{\sigma^2} = \frac{\bar{C}(V,V) - \sigma_K^2}{\bar{C}(V,V)},$$

where is KE_{DK} the kriging efficiency; σ_K^2 is the kriging variance; σ^2 is the variance of the true blocks. In block kriging, the estimates are created on some volume V; thus $\bar{C}(V,V)$ refers to the average variance between the points within the blocks. A high KEFF alludes to a low kriging variance, meaning the variances of the block estimates and the true block values are near equal (Deutsch et al., 2014; Deutsch and Deutsch, 2015). A low KEFF indicates a lower block variance relative to the kriging variance. Since the kriging variance differs from block to block, the KEFF varies as well (Deutsch et al., 2014).

4. RESULTS

The following chapter analyses the results obtained through the experiments conducted in the previous methodology chapter. It is initiated by the experimental data analysis (EDA), which provides a detailed analysis of the dataset. Subsequently, the results of the predictive modelling steps are outlined, followed by the post-prediction investigations conducted. The next section then outlines the results obtained when assessing the efficiency of sedimentological features for gold grade prediction. Finally, the 2D block modelling results are outlined.

4.1. Explanatory Data Analysis

The raw data comprises 55 304 samples and nine sample properties (Table 4.1). These properties include *BHID* (borehole ID), a unique identification column comprising a serial number with a combination of letters and numbers. *X* is a spatial coordinate indicating Southing coordinates in the Gauss Conform geographical coordinate system, while the Y contains Westing coordinates. The *Channel Width* and *Internal Waste* properties are described in centimetres (cm). The *Percentage Conglomerate* property is expressed as a percentage of the volume of the sample that is estimated to be of the conglomerate type of lithology. The *Basal Contact* is a categorical property that describes the nature of the contact between the sample and the underlying lithology. *Au (cm. g/t)* describes gold accumulation in cm grams per ton, and *Au (g/t)* describes the target property in grams per ton (gold grade) as determined through chemical assay.

Table 4.1: Selection of raw dataset illustrating the identification column, spatial coordinates, the predictive properties, and the target property.

BHID	X	Y	Channel Width	Internal Waste	Conglomerate Percentage	Basal Contact	Au (cm. g/t)	Au (g/t)
420SE0000128363	11225.37	-18483.9	126	0	100	Sharp	405	3.2
420SE0000132231	11236.59	-18467.7	87	27	69	Sharp	687	7.9
420SE0000132232	11237.54	-18465.8	70	11	84	Sharp	342	4.9
420SE0000132233	11238.96	-18464.5	55	0	100	Sharp	293	5.3
420SE0000132234	11239.81	-18462.8	59	0	100	Sharp	336	5.7
420SE0000135873	7583.76	-20848.8	67	0	100	Sharp	1122	16.7
420SE0000138606	6050.55	-21408.1	129	21	84	Sharp	1792	13.9
420SE0000139793	7207.308	-21156.4	24	0	100	Sharp	274	11.4
420SE0000140841	7577.563	-20832.7		0		Faulted		
420SE0000144464	6705.238	-21331.8	9	0	100		214	23.8
420SE0000144989	8329.257	-20729	101	0	100	Sharp	300	3
420SE0000146672	8117.969	-20239.3	185	0	100	Sharp	500	2.7
420SE0000146673	8120.777	-20243.6	113	0	100	Sharp	312	2.8
420SE0000146674	8122.805	-20248.4	92	44	52	Sharp	1763	19.2
420SE0000146675	8125.752	-20252.4	111	60	46	Sharp	1602	14.4

For the purpose of this project, the data was subdivided to isolate a subset study area between -22 000 and -20 000 Westing (Y) and between 5000 and 9000 Southing (X). This portion of the data were

selected because it contained a spatially dense cluster of samples and would thus be ideal for estimation. A spatial map of the raw data are illustrated in Fig. 4.1.

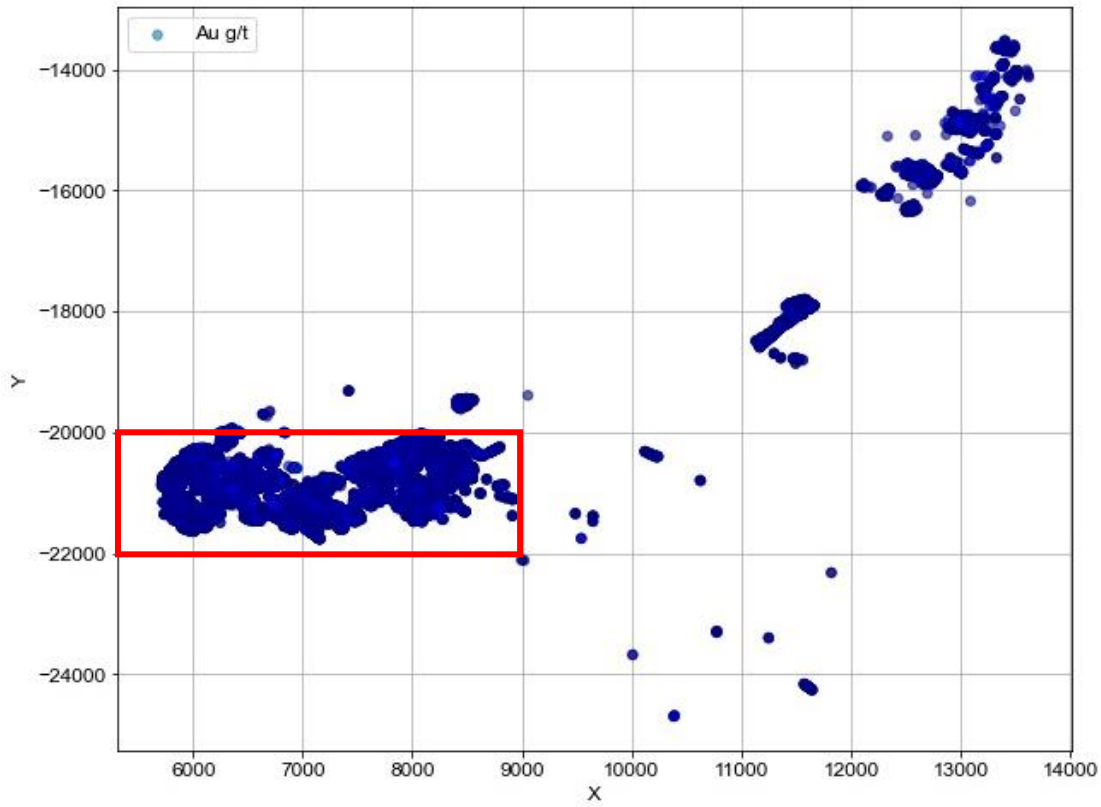


Figure 4.1: Spatial map illustrating the distribution of the original/raw dataset. Datapoints are colour-coded by their gold grade. The red box marks the subset area that this project will focus on.

A spatial map of the subset block highlighted in Fig. 4.1 is illustrated in the following figure:

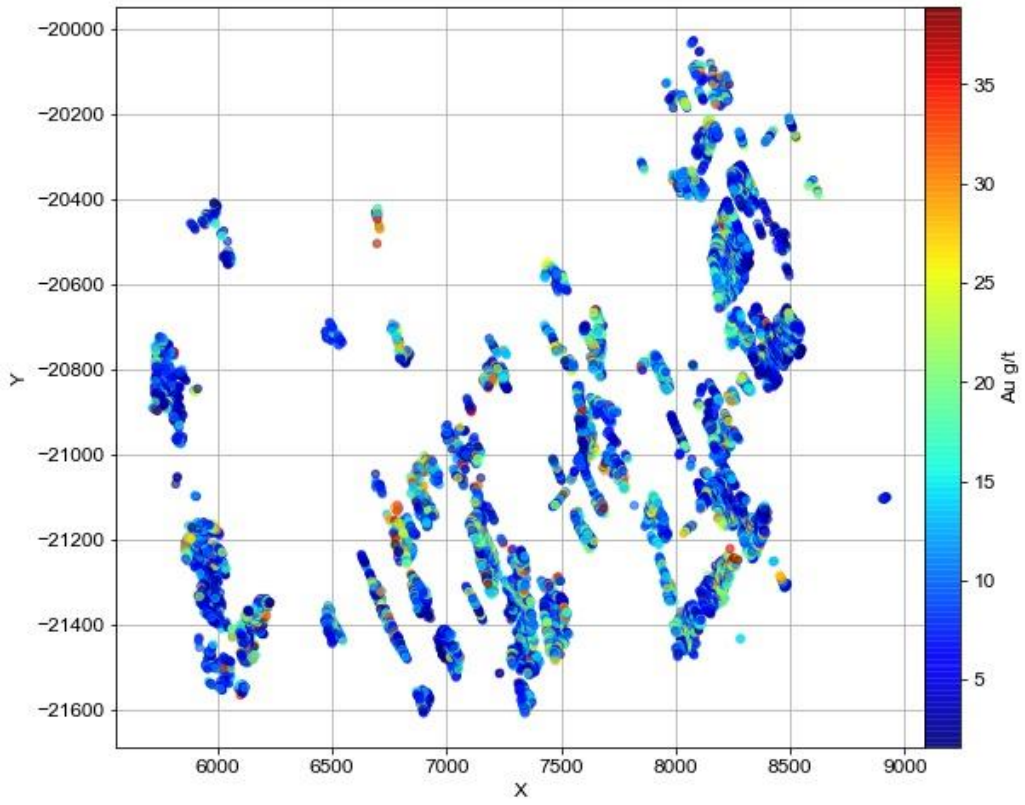


Figure 4.2: Spatial map illustrating the distribution of the selected subset block. Datapoints are colour-coded by their gold grade.

The subsetted data were filtered to exclude observations that contained gold grade values equal to 0, as well as to exclude any observations with missing values for any of the features. The resulting dataset contained a total of 10682 observations after outlier removal. The gold accumulation (Au cm. g/t) feature was removed because it is just another variant of the target feature and would be highly correlated with the target gold grade. As a result, the prediction model would return false, highly exaggerated values. Consequently, our subset contains a total of 8 features. Prediction requires all features to be numerical; to ensure this, the basal contact feature, comprising three categories – namely Gradational, Sharp and Faulted – was assigned numerical values (integers) using a multiple label encoder function. Upon analysis of the properties of the cleaned dataset (Table 4.2), we see that there are no null instances, and the features entail six float data types and one integer data type. Float point values are recognised as numerical values with decimals.

Table 4.2: Python extract evaluating the features' data types and the occurrence of non-null values.

#	Column	Non-Null Count	Data Type
0	X	10682 non-null	float
1	Y	10682 non-null	float
2	Channel Width	10682 non-null	float
3	Internal Waste	10682 non-null	float
4	Conglomerate Percentage	10682 non-null	float
5	Basal Contact	10682 non-null	integer
6	Gold Grade (Au g/t)	10682 non-null	float

4.1.1. Feature Analysis

The following figure illustrates spatial maps similar to Fig. 4.2, but colour coded by each predictive feature.

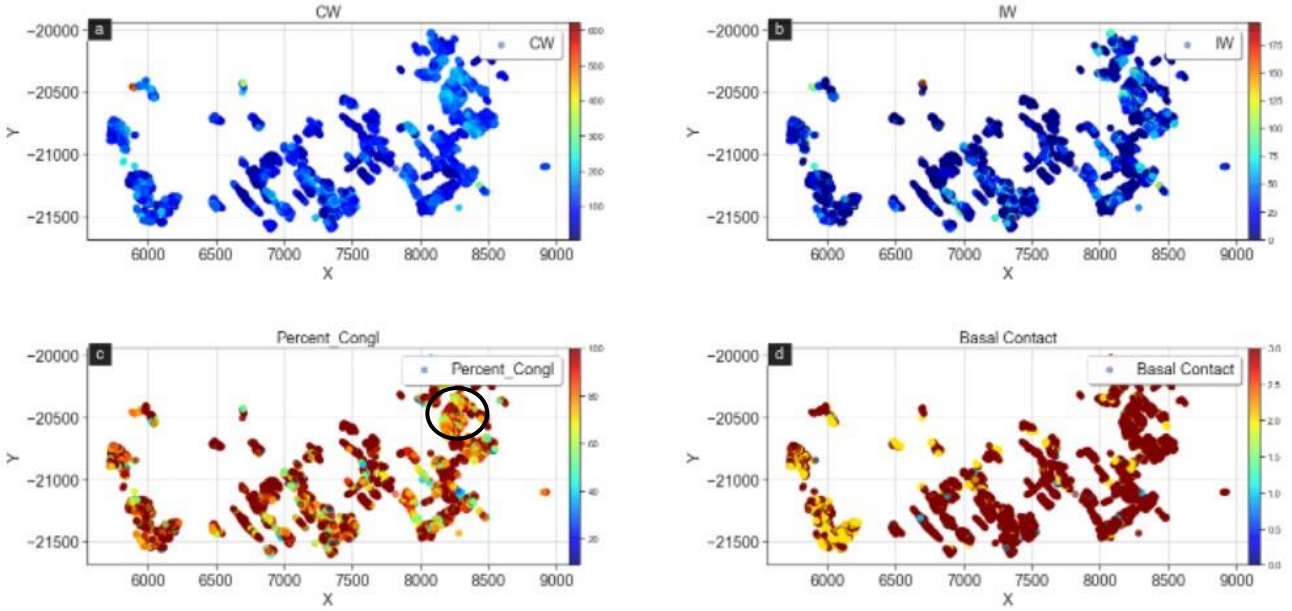


Figure 4.3: Spatial maps illustrating the distribution of the datapoints, colour-coded by individual features. The black circle on the conglomerate percentage map marks an area where some sort of cluster has been observed.

4.1.1.1. Channel Width

Channel Width is a quantitative feature comprising continuous values with a minimum of 6 cm and a maximum of 662 cm. Its histogram and box and whisker plots (Fig. 4.4) reveal that the dataset is almost symmetrical with some outliers. A statistical analysis (Table 4.3) of the data reveals a standard deviation (std) of 56.57 cm and a mean of 112.5 cm. As shown in Fig. 4.3a, its spatial distribution is consistent with results seen in the histogram. Values range between 0 and 200 cm, with most concentrations being centred around 100 cm. There is no major clustering of specific Channel Width values observed.

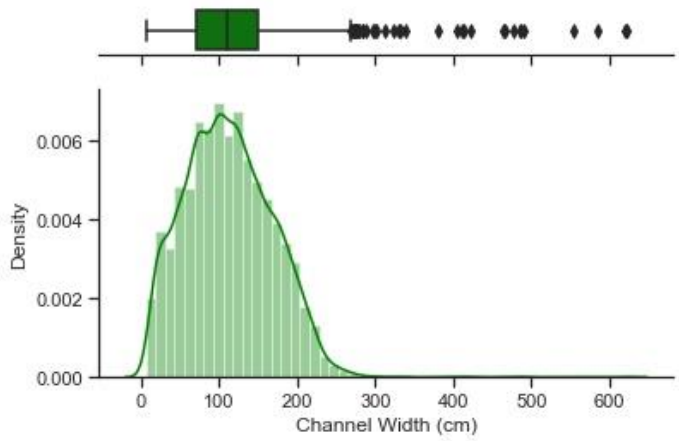


Figure 4.4: Boxplot and histogram illustrating the distribution of Channel Width data.

4.1.1.2. Internal Waste

Internal Waste is a quantitative feature comprising continuous values with a minimum of 0 cm and a maximum of 195 cm. A statistical analysis (Table 4.3) of the data reveals a standard deviation of 25.6 cm and a mean of 19.9 cm. Its histogram and box and whisker plots (Fig. 4.5) reveal that the data are skewed to the right with numerous outliers. Its spatial distribution map (Fig. 4.3b) and histogram plot reveal that a large number of the observations are concentrated below 15 cm.

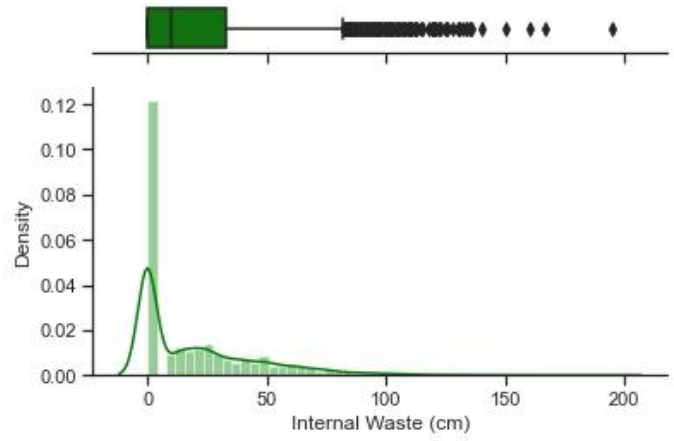


Figure 4.5: Boxplot and histogram illustrating the distribution of the Internal Waste data.

4.1.1.3. Percentage Conglomerate

Percentage Conglomerate is a quantitative feature comprising continuous values with a minimum of 17.1 and a maximum of 100%. A statistical analysis (Table 4.3) of the data reveals a standard deviation of 17.1% and a mean of 85.6%. Its histogram and box and whisker plots reveal that the data are skewed to the left with a number of outliers. Its spatial distribution map and histogram plot (Fig. 4.6) reveal that most values are clustered between 90 and 100%. The area surrounding 8520 X and -20 500 Y (marked by a black circle) appears to have a concentrated cluster of values between 60 and 80% (Fig. 4.3c).

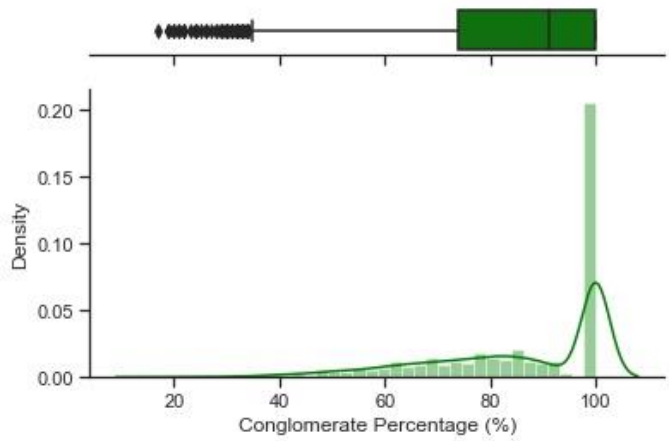


Figure 4.6: Boxplot and histogram illustrating the distribution of the Percentage Conglomerate data.

4.1.1.4. Basal Contact

Basal Contact is a qualitative feature that has been transformed into a discrete quantitative feature with a minimum of 0 and a maximum of 3. Its descriptive statistics (Table 4.3) reveal that over 80% of the observations contain a value of 3 (Sharp). Its spatial distribution map (Fig. 4.3d) also affirms this, as most of the observations are marked by the same red hue.

4.1.2. Target Analysis (Au g/t)

Gold grade, the target, is quantitative and continuous, with minimum and maximum values of 1.6 and 38.9 g/t respectively. Statistical analysis reveals a standard deviation of 17 g/t and a mean of 13.8 g/t (Table 4.3). Its histogram and box and whisker plots (Fig. 4.7) reveal an asymmetrical distribution that is skewed to the right with many outliers. Over 90% of the observations contain values between 0 and 20 g/t; this is also affirmed by its spatial distribution plot (Fig. 4.2), which shows that most of the data are essentially a single colour throughout.

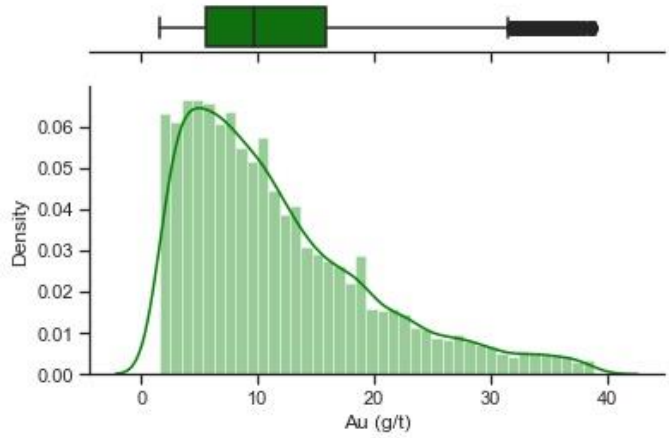


Figure 4.7: Boxplot and histogram illustrating the distribution of the gold grade data.

The filtered subset's summary statistics are presented in the following table.

Table 4.3: Descriptive statistics for the full subset.

	X	Y	Channel Width	Internal Waste	Conglomerate Percentage	Basal Contact	Gold Grade (Au g/t)
count	10682.0	10682.0	10682.0	10682.0	10682.0	10682.0	10682.0
mean	7500.9	-20998.7	112.5	19.9	85.6	2.8	11.7
std	836.1	367.6	56.7	25.6	17.1	0.4	8.1
min	5717.6	-21607.8	6.0	0.0	17.0	0.0	1.6
25%	6980.7	-21319.6	71.0	0.0	74.0	3.0	5.5
50%	7647.0	-21090.5	109.0	10.0	91.0	3.0	9.7
75%	8219.9	-20717.8	150.0	33.0	100.0	3.0	15.9
max	8924.2	-20027.4	622.0	195.0	100.0	3.0	38.9

4.1.3. Correlation Matrix

The correlation matrix (Fig. 4.8), generated using the Pearson correlation metric, gives insight into the relationships between the features and the target. Channel Width has a strong positive correlation with the Internal Waste and a strong negative correlation with Conglomerate Percentage; this implies a near-linear relationship between these three features. The Basal Contact is poorly correlated with all features. Relative to the target, the Channel Width, Internal Waste, and Conglomerate Percentage all have low correlations – this may imply a non-linear relationship.

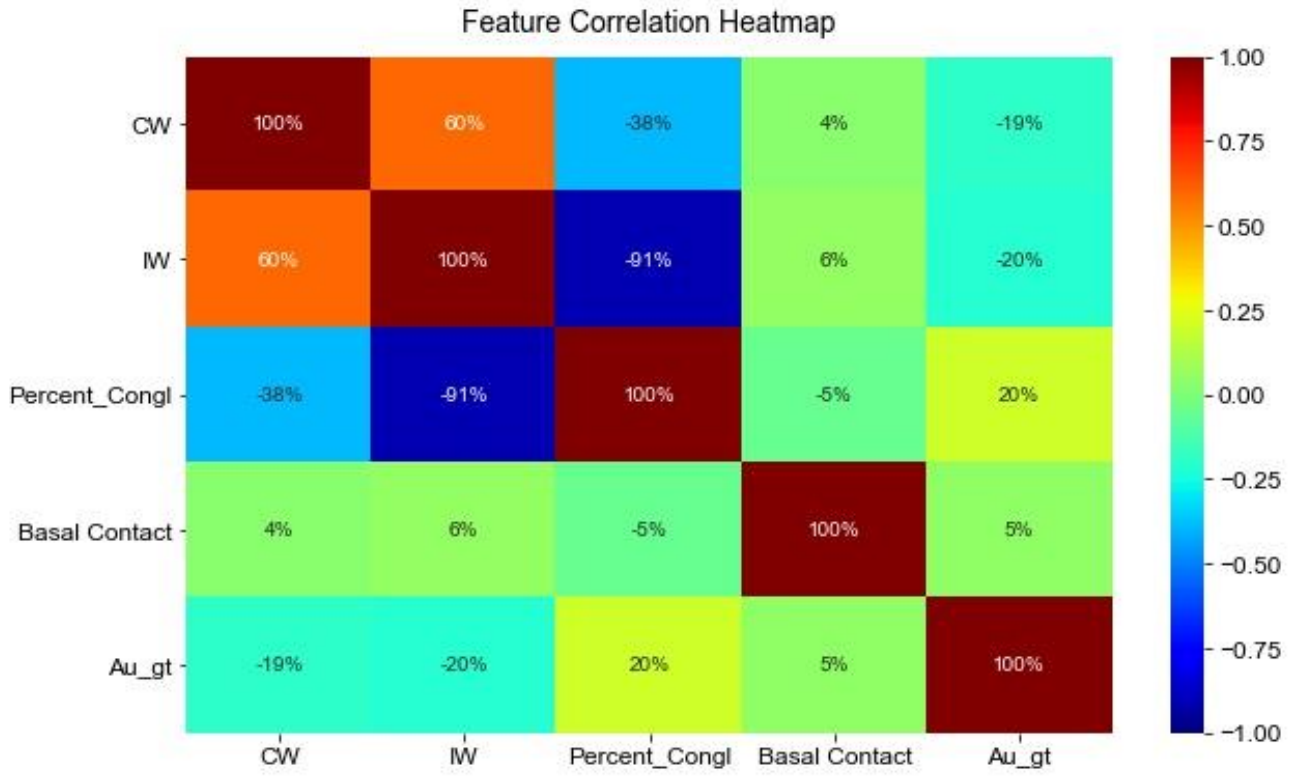


Figure 4.8: Feature correlation heatmap.

4.2. Full Subset

4.2.1. Experiment 1

The first round of predictive modelling included training 70% of the data and testing with 30% of the dataset using a simplified implementation of our MLAs. The results, obtained with hyper-tuned parameters and run over 10 K-fold cross-validations, are presented in the following sections. Prior to segmenting the data, the full subset comprising 10 682 observations is predicted. The testing results are presented in the following table:

Table 4.4: First round of test scores for the full subset.

	RF	KNN	EN	AdaBoost
30% Test Size	0.01	-0.05	0.02	0.03

The prediction accuracies are 1% for the RF algorithm, -5% for the KNN algorithm, 2% for the EN algorithm and 3% for the AdaBoost Algorithm. A graphical representation of these results is presented in Fig. 4.9. The scatter plots do not show a distinct trend but rather display irregular clusters.

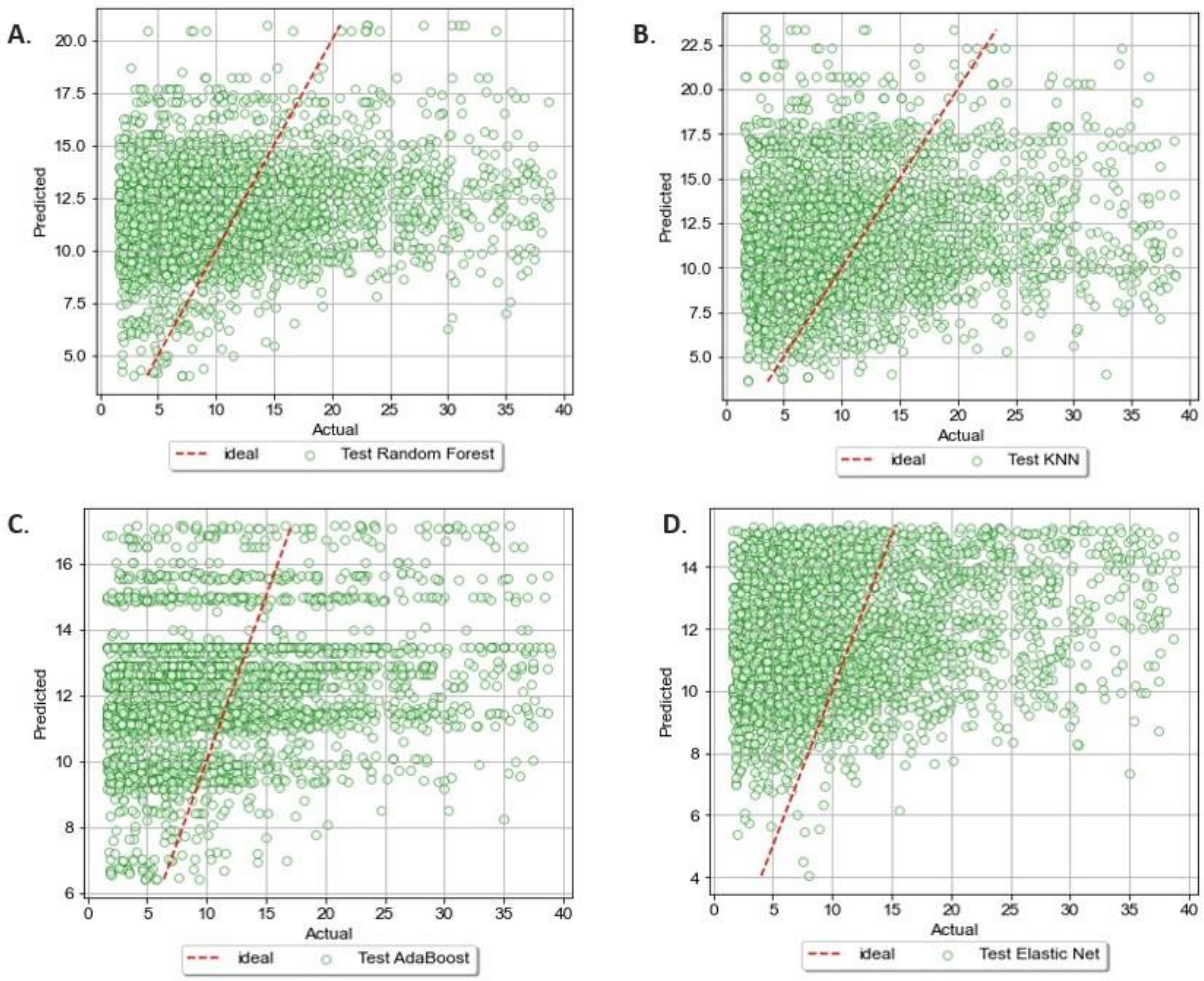


Figure 4.9: Graphical illustrations of the full subset's first round of testing for each of the four algorithms.

The final hyperparameters used for the full subset's predictions are presented in the following table:

Table 4.5: Full subset's final hyperparameters.

Algorithm	Hyperparameters
RF	max depth = 10, max features = 4, min samples per leaf = 4, min samples per split = 2, number of trees = 400
EN	max iteration = 10, α = 0.0001, L_1 ratio = 0.4
KNN	number of neighbours = 9
AdaBoost	base estimator = decision tree, base estimator max depth = none, base estimator max features = 2, base estimator min samples per leaf = 3, base estimator min samples per split = 3, number of trees = 500

4.2.2. Experiment 2

In the next round of predictive modelling, the dataset was clustered using the K-Means method. The results are presented in the following sections.

4.2.2.1. Prediction with K-Means clustering

The silhouette algorithm (Appendix A) suggested 3 clusters for the full dataset. The K-Means algorithm then allocated each datapoint to one of 3 groups. This information was compiled into a new column and used as a feature. The spatial map produced is presented in Fig 4.10.

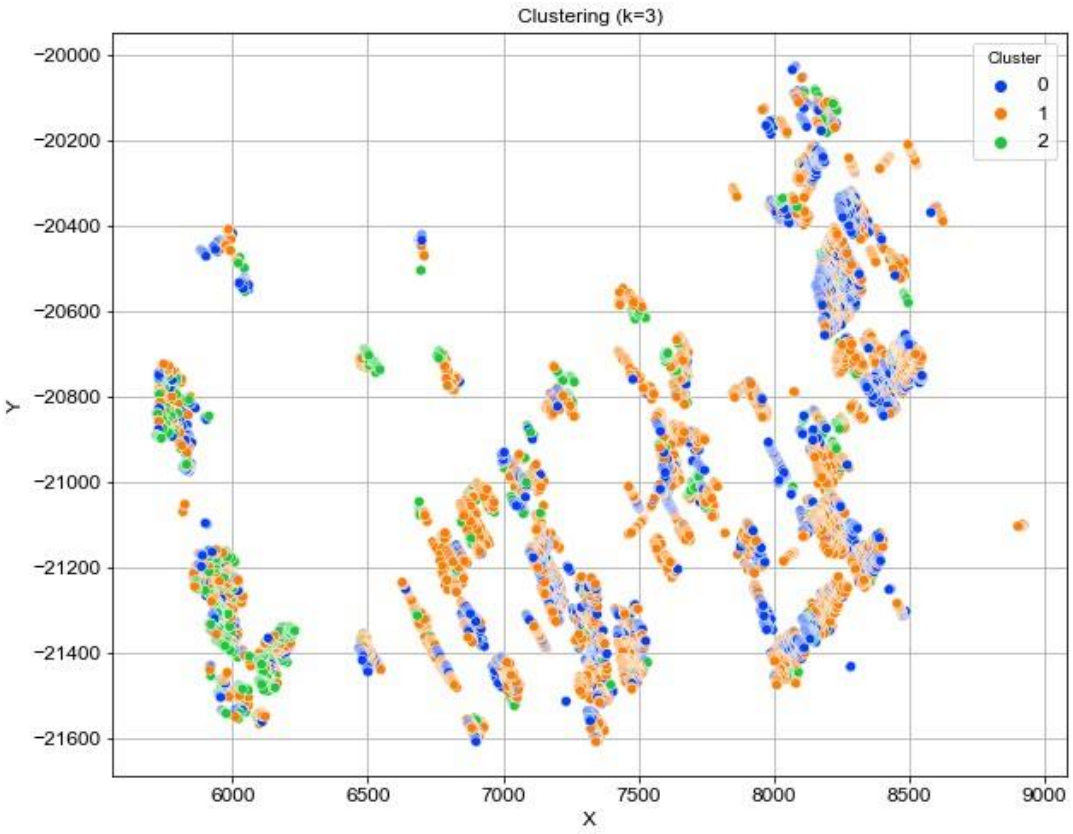


Figure 4.10: Spatial map visualising the full subset's clustered datapoints when K = 3.

The prediction results, run over 10 K-fold cross-validations, are presented in table 4.6. Included are the CoD (Coefficient of Determination) scores for cross-validation obtained when the test size was increased to 50%. CoD values are expressed as decimals but can be converted to percentages (see Chapter 2.4.6).

Table 4.6: Second round of test scores for the full dataset.

CoD scores (K=3)				
	RF	**KNN**	**EN**	**AdaBoost**
30% Test Size	0.02	-0.02	0.04	0.04
50% Test Size	0	-0.04	0.03	0.02

The prediction accuracies attained are 2% for the RF algorithm, - 2% for the KNN algorithm, 4% for the EN algorithm and 4% for the AdaBoost Algorithm. The prediction accuracies have increased by 1%, 3%, 2%, and 1% for the RF, KNN, EN and AdaBoost algorithms, respectively, compared to the previous result obtained in Table 4.4. Increasing the test size decreases the prediction accuracies on average. A graphical representation of the prediction accuracies with a 30% testing set is presented

in Fig 4.11. The scatter plots do not have a linear trend but rather display irregular clusters; furthermore, the AdaBoost algorithm's plot is governed by horizontal linear clusters.

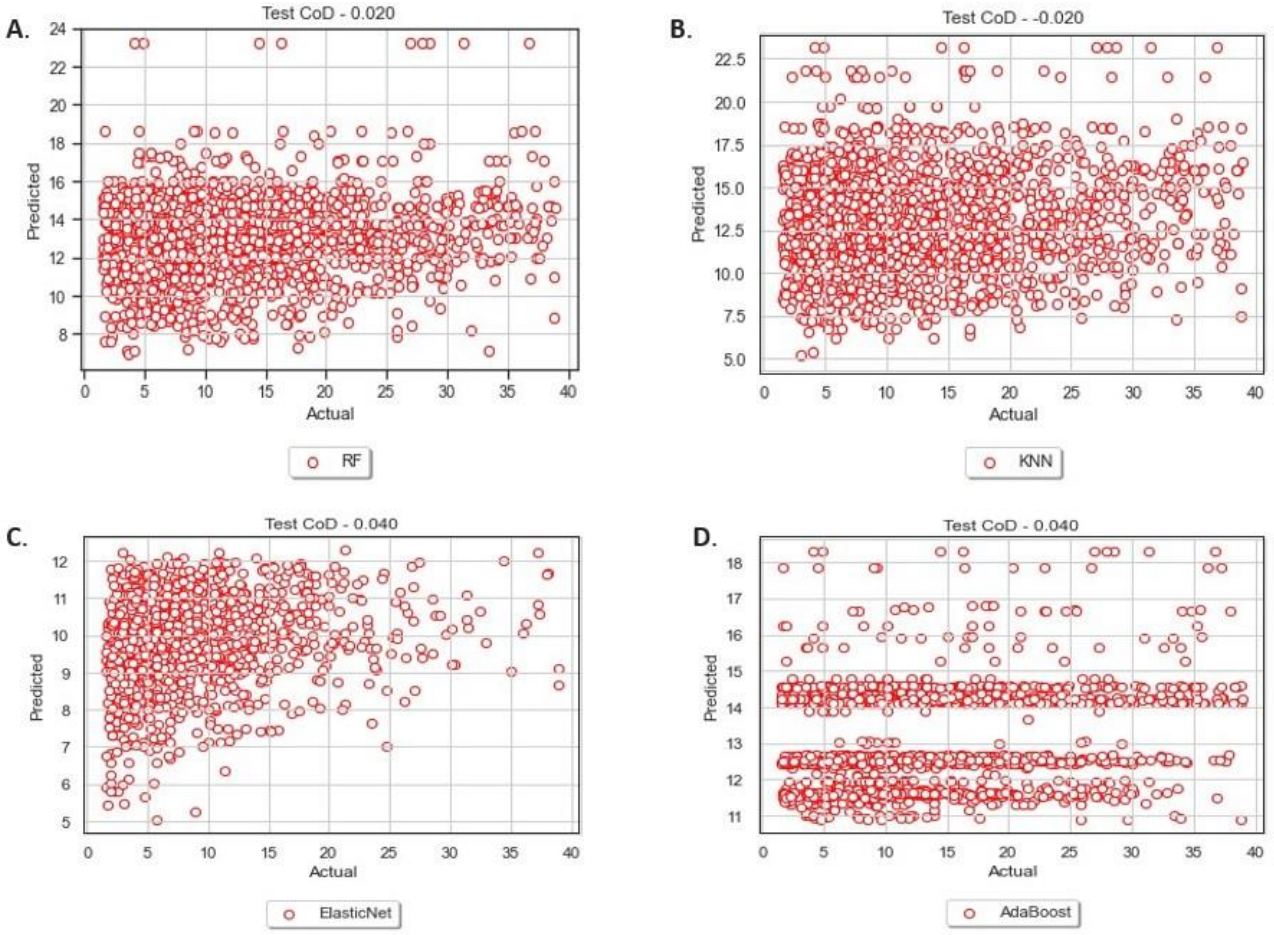


Figure 4.11: Graphical illustrations of the full subset's second round of testing, visualising each algorithm's CoD scores.

4.2.2.2. K-Means clustering with spatial data

In an effort to assess the prediction performance, the K-Means algorithm was configured to include the spatial features when assigning clusters and in the prediction process. The spatial map produced is presented in Fig 4.12, where it can be observed that the datapoints have been assigned three neatly defined clusters.

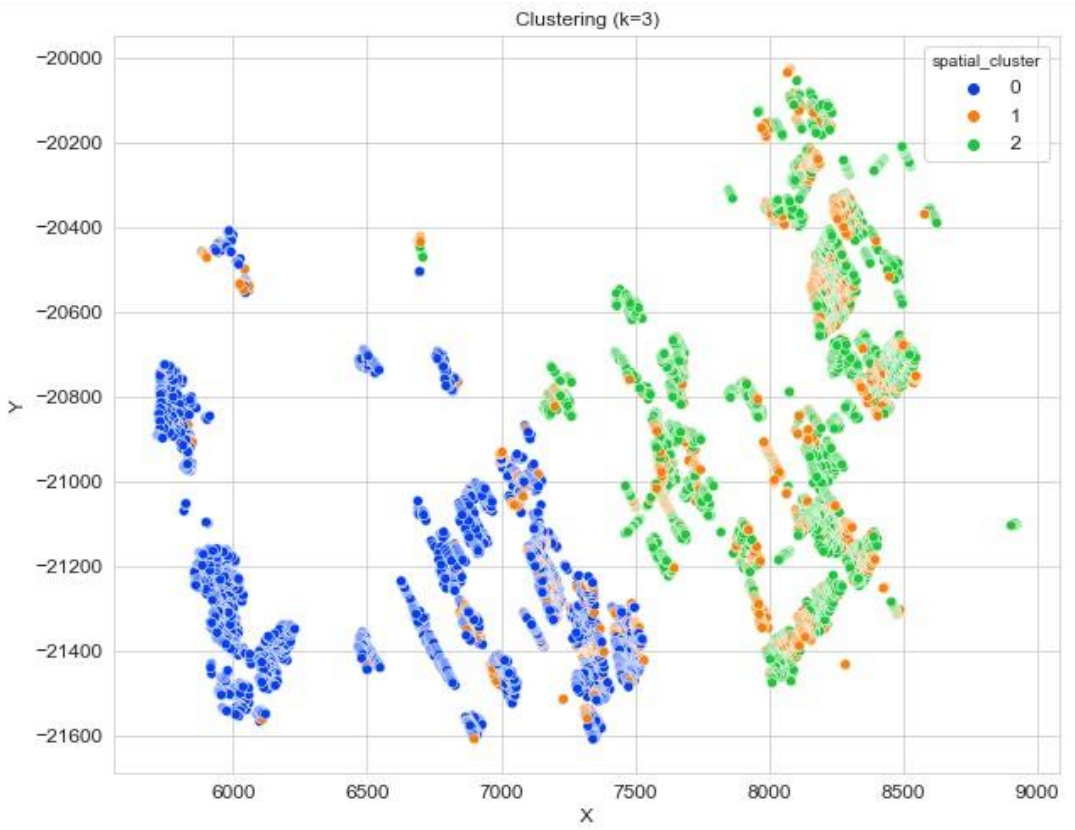


Figure 4.12: Spatial map visualising the full subset's clustered datapoints when spatial features are incorporated.

The prediction results, run over 10 K-fold cross-validations, are presented in the following table.

Table 4.7: Second round of test scores for the full dataset when spatial features are incorporated.

CoD values (K = 3)				
	RF	**KNN**	**EN**	**AdaBoost**
30% Test Size	0.17	0.08	0.05	0.1
50% Test Size	0.14	0.05	0.06	0.08

The prediction accuracies attained are 17% for the RF algorithm, 8% for the KNN algorithm, 5% for the EN algorithm and 10% for the AdaBoost Algorithm. There is a 15%, 6% and 1% increase for the RF, KNN and EN algorithms, respectively, compared to the previous results (Table 4.6), where no spatial features were considered. Increasing the test size decreases the prediction accuracies on average.

4.2.3. Feature Importance

The RF and AdaBoost algorithms contain feature importance attributes that were computed to assess which features took precedence over others for this full dataset. The results are presented in Fig. 4.13. The RF algorithm categorised the Channel Width as most important, while the AdaBoost

algorithm categorised the Conglomerate Percentage as most important. The AdaBoost algorithm recognised the Internal Waste to be quite important, while the RF algorithm recognised it to a lesser extent. The Basal Contact is the least important feature for both algorithms.



Figure 4.13: Graphical illustrations of the RF and AdaBoost algorithms' feature importance statistics for the full subset.

Feature importance statistics were also computed when the spatial features were included in the modelling process (Fig. 4.14). Both algorithms considered coordinates to be important and relatively more so than other features. The AdaBoost algorithm barely took other features into account (except Internal Waste), while the RF algorithm considered them to a much lesser extent.

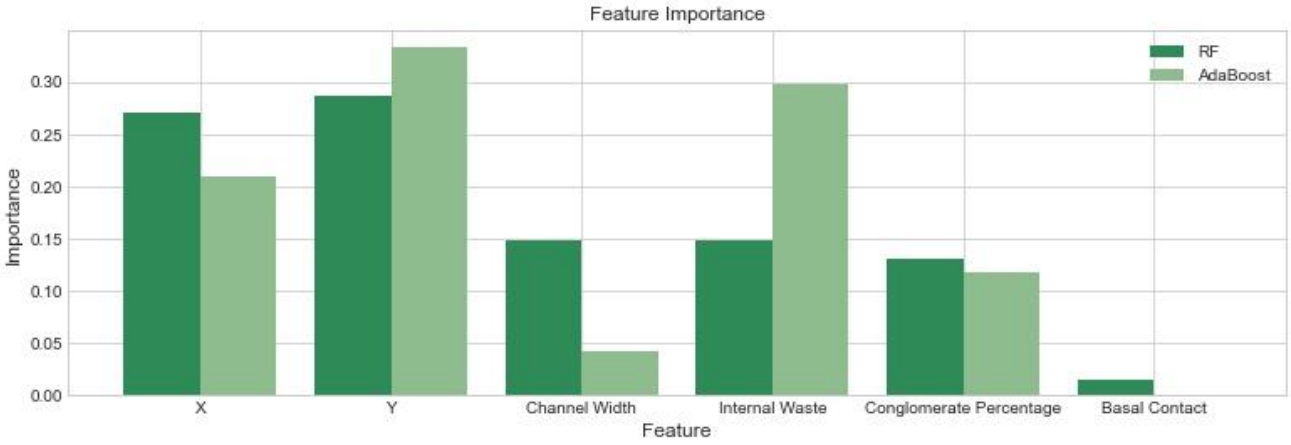


Figure 4.14: Graphical illustrations of the RF and AdaBoost algorithms' feature importance statistics for the full subset (including spatial features).

4.3. Post-prediction Investigations

Training and testing using the full dataset returned poor results, as seen in the previous section. In an effort to investigate the dataset and troubleshoot the cause of poor performance of predictive modelling, the REs were computed for each sample in the dataset. The dataset was then segmented

according to various RE intervals and predicted at a smaller scale; the results are presented in the following table:

Table 4.8: CoD test scores according to Relative Error (RE) intervals.

RE subset statistics											
RE Interval	**0 to 0.1**	**0.1 to 0.2**	**0. to 0.3**	**0.3 to 0.4**	**0.4 to 0.5**	**0.5 to 0.6**	**0.6 to 0.7**	**0.7 to 0.8**	**0.8 to 0.9**	**0.9 to 1**	**>1.0**
count	1199	1126	1146	1101	949	794	483	314	224	236	3110
% Of full dataset	11.22	10.54	10.73	10.31	8.88	7.43	4.52	2.94	2.10	2.21	29.11
Test CoD values for each RE subset											
RF	0.87	0.61	0.48	0.29	0.12	0.02	-0.05	0.62	0.72	0.52	0.28
KNN	0.87	0.58	0.35	0.37	0.01	0.09	-0.08	0.58	0.77	0.41	0.27
Elastic Net	0.72	0.43	0.32	0.31	0.14	0.04	0	0.28	0.64	0.62	0.16
AdaBoost	0.8	0.55	0.41	0.41	0.09	0.03	-0.07	0.75	0.86	0.65	0.32

The predictions were conducted in intervals, i.e., in Table 4.8, the RE = 0 to 0.1 column refers to datapoints with RE between 0 and 0.1. Upon analysis of these results, the datasets with the highest prediction accuracies were grouped and predicted as a unit. These, however, returned poor results. Subsequently, a decision was made to isolate datapoints that fell within the proximity of the group with the highest CoD (RE = 0 to 0.1). Ultimately, the final dataset grouped datapoints with RE that fell between 0 and 0.2 (21.8% of the full dataset). This dataset was found to produce high prediction accuracies; from this point forth, this dataset will be referred to as the ‘good’ dataset, while the remainder will be referred to as the ‘bad’ dataset due to its poor prediction performance. The spatial distribution of these two datasets is illustrated in Fig. 4.15, where the ‘bad’ datapoints seem to occur at the edge of ‘good’ clusters.

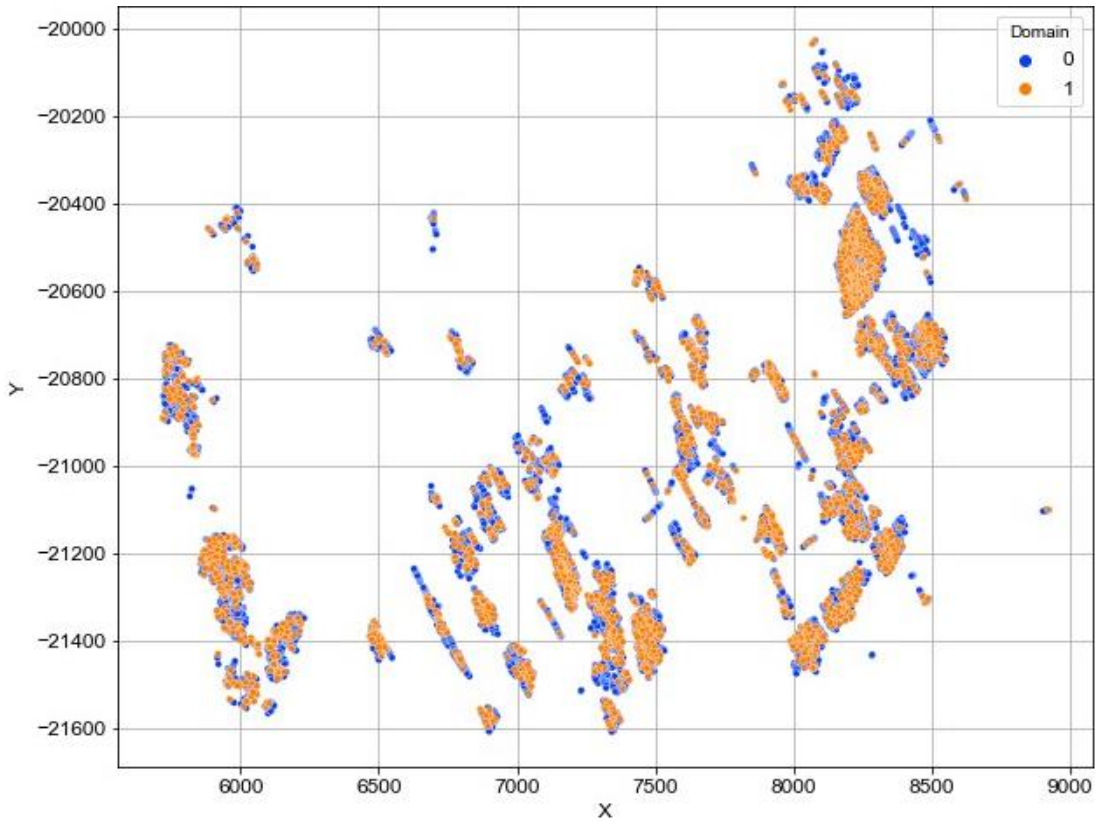


Figure 4.15: Spatial distribution of the ‘good’ vs ‘bad’ datasets, represented in orange and blue respectively.

4.3.1. ‘Good’ Dataset

This dataset comprises 2 325 datapoints. Its descriptive statistics are presented in the following table:

Table 4.9: Descriptive statistics for the ‘good’ subset.

	X	Y	Channel Width	Internal Waste	Conglomerate Percentage	Basal Contact	Gold Grade (Au g/t)
count	2325.0	2325.0	2325.0	2325.0	2325.0	2325.0	2325.0
mean	7515.9	-20996.1	115.8	20.9	85.0	2.8	13.4
std	810.4	367.6	56.4	26.0	17.3	0.5	5.0
min	5718.7	-21599.3	8.0	0.0	20.0	0.0	2.0
25%	7048.2	-21317.1	74.0	0.0	74.0	3.0	10.2
50%	7666.5	-21099.0	111.0	12.0	90.0	3.0	12.6
75%	8214.0	-20707.1	154.0	35.0	100.0	3.0	15.6
max	8921.2	-20030.0	622.0	195.0	100.0	3.0	38.9

4.3.1.1. Experiment 1

The test results of the first round of predictive modelling are presented in the following table:

Table 4.10: First round of test scores for the ‘good’ dataset.

	RF	KNN	EN	AdaBoost
30% Test Size	0.79	0.79	0.54	0.71

The prediction accuracies are 79% for the RF algorithm, 79% for the KNN algorithm, 54% for the EN algorithm and 71% for the AdaBoost Algorithm. A graphical representation of these results is presented in Fig. 4.16. A linear trend can be traced on the scatter plots for all algorithms except EN, which presents a more curved trend.

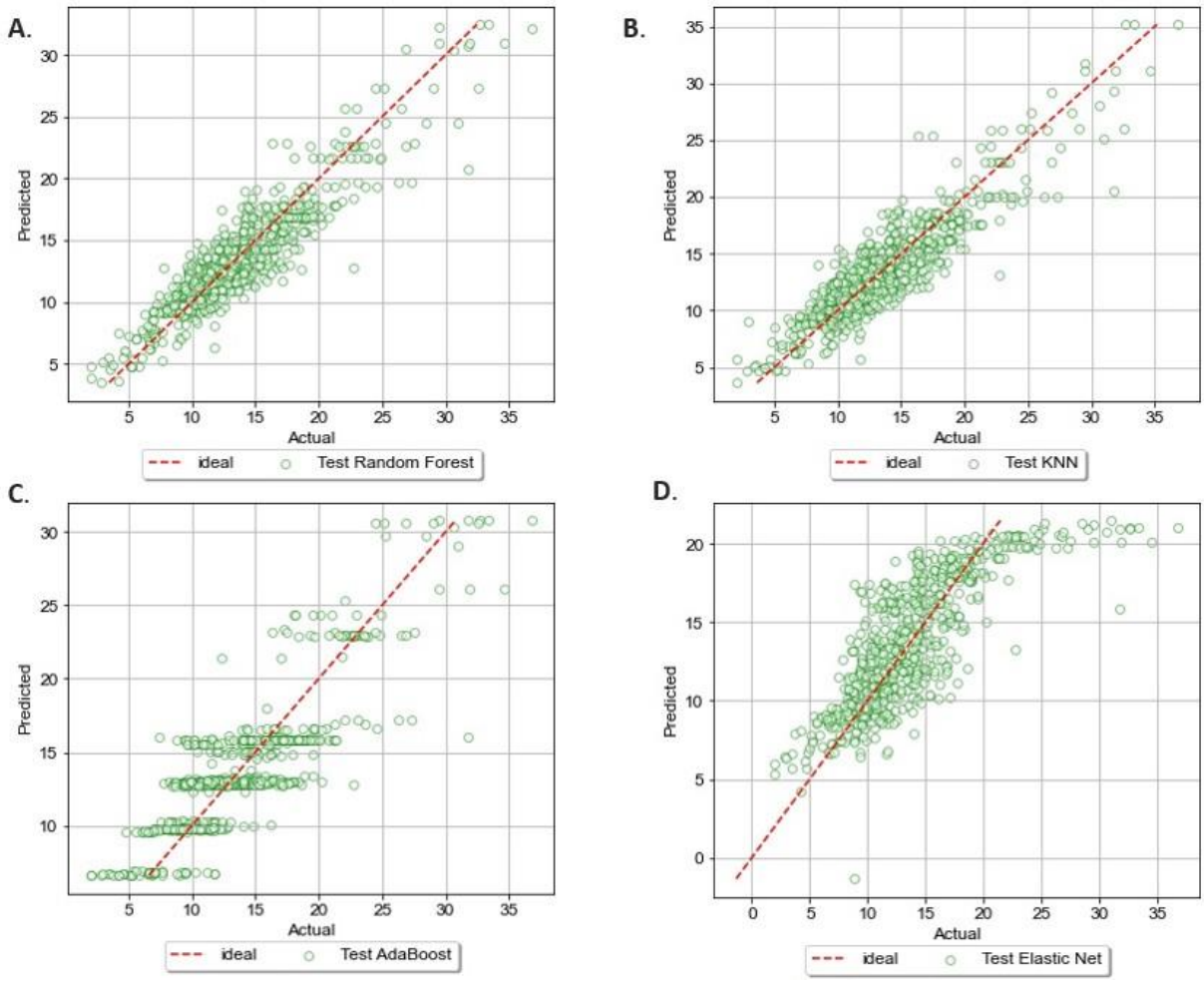


Figure 4.16: Graphical illustrations of the good subset’s first round of testing for each of the four algorithms.

The hyperparameters used for the full subset’s predictions are presented in the following table:

Table 4.11: 'Good' dataset's final hyperparameters

Algorithm	Hyperparameters
RF	max depth = 10, max features = 4, min samples per leaf = 1, min samples per split = 5, number of trees = 2000
EN	max iteration = 10, α = 0.0001, L1 ratio = 0.8
KNN	number of neighbours = 5
AdaBoost	base estimator = decision tree, base estimator max depth = none, base estimator max features = 2, base estimator min samples per leaf = 2, base estimator min samples per split = 5, number of trees = 500

4.3.1.2. Experiment 2

The silhouette algorithm (Appendix A) assigned 5 clusters for the 'good' dataset. The K-Means algorithm then allocated each datapoint to one of 5 groups. This information was compiled into a new column and used as a feature. The spatial map produced is presented in Fig 4.17.

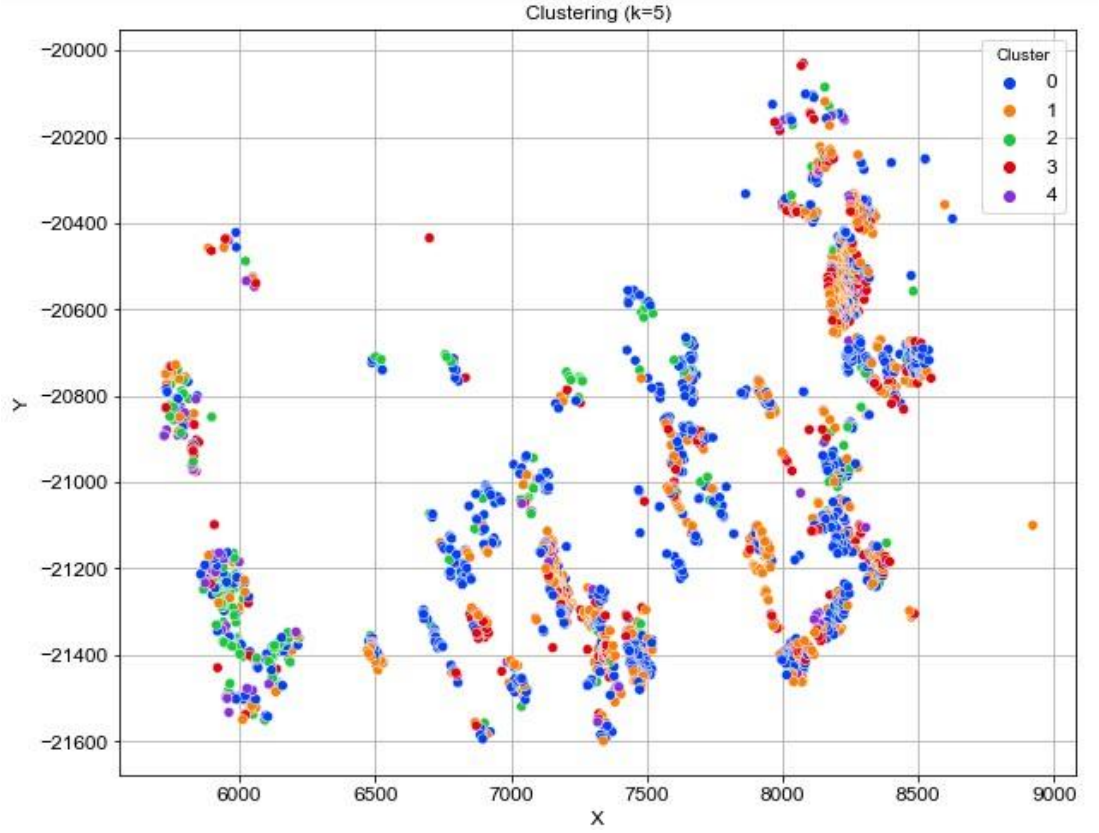


Figure 4.17: Spatial map visualising the 'good' dataset's clustered datapoints when K = 5.

The prediction results, averaged over 10-fold cross-validations, are presented in the following table.

Table 4.12: Second round of test scores for the ‘good’ dataset.

Test CoD (K = 5)				
	RF	**KNN**	**EN**	**AdaBoost**
30% Test Size	0.84	0.81	0.59	0.81
50% Test Size	0.8	0.78	0.51	0.79

The prediction accuracies attained are 84% for the RF algorithm, 81%% for the KNN algorithm, 59% for the EN algorithm and 81% for the AdaBoost Algorithm. The prediction accuracies have increased by 5%, 2%, 5%, and 10% for the RF, KNN, EN and AdaBoost algorithms, respectively, compared to the previous results obtained in table 4.10. A graphical representation of the prediction accuracies with a 30% testing set is presented in Fig. 4.18. A linear trend can be traced on the scatter plots for all algorithms except EN, which presents a more sparsely scattered trend.

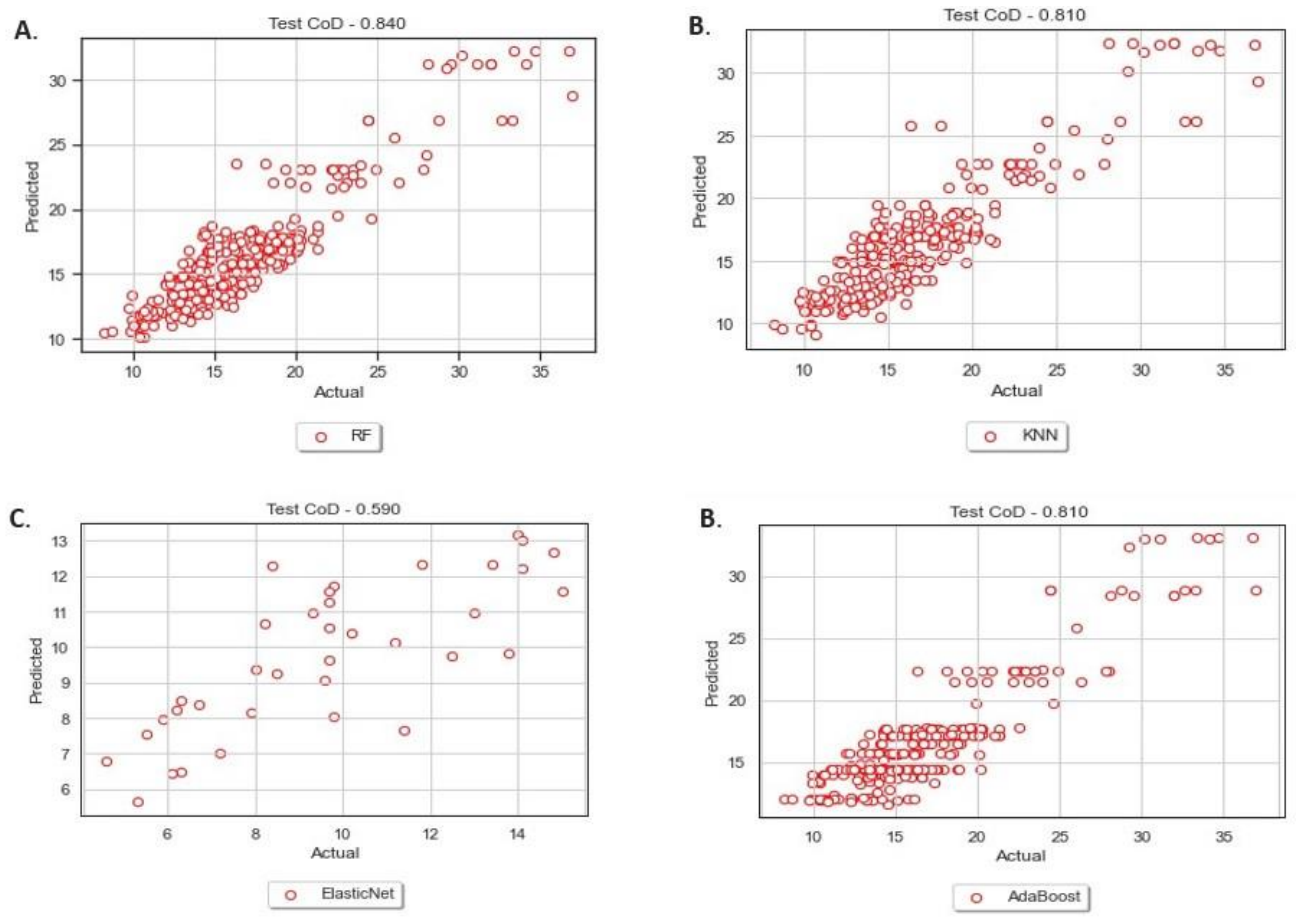


Figure 4.18: Graphical illustrations of the ‘good’ dataset’s second round of testing, visualising each algorithm's CoD scores.

K-Means clustering with spatial data

Incorporating the spatial coordinates during per-cluster prediction returns a spatial map with neatly defined clusters, as presented in Fig. 4.19.

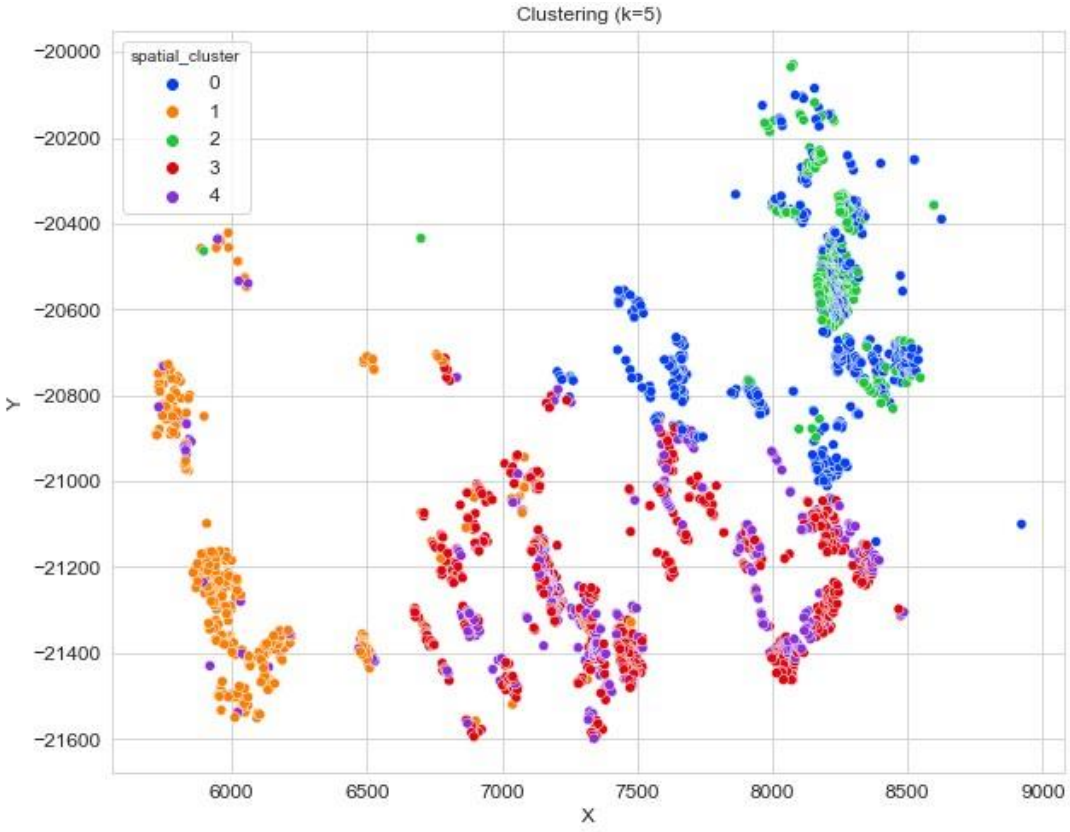


Figure 4.19: Spatial map visualising the 'good' dataset's clustered datapoints when spatial features are incorporated.

The prediction results, averaged over 10-fold cross-validations, are presented in the following table:

Table 4.13: Second round of test scores for the 'good' dataset when spatial features are incorporated.

Test CoD (K = 5)				
	RF	**KNN**	**EN**	**AdaBoost**
30% Test Size	0.75	0.63	0.56	0.76
50% Test Size	0.75	0.6	0.56	0.75

The prediction accuracies attained are 75% for the RF algorithm, 63% for KNN, 56% for the EN algorithm and 76% for the AdaBoost Algorithm. The prediction accuracies have decreased on average for all algorithms when compared to the previous results (Table 4.12), where no spatial features were considered. Moreover, increasing the test size decreases the prediction accuracies on average.

4.3.1.3. Predicting the entire 'good' dataset

Upon observing that the training scores are often higher than the test scores (Appendix B), the entire 'good' dataset was predicted using randomised k-fold cross-validation. The dataset was shuffled 100 times over 5-fold cross-validations, and an average of the results was taken to profile model performance (Fig. 4.20). The final model yielded the following results: a cross-validation score of 81%

was attained for the RF algorithm, while the AdaBoost algorithm attained a score of 82%. The prediction accuracy improved by 1%% for the AdaBoost algorithm when compared to the results in Table 4.12. The scatter diagrams produced (Fig. 4.20) show a linear trend consistent with the high prediction results observed in Fig. 4.18.

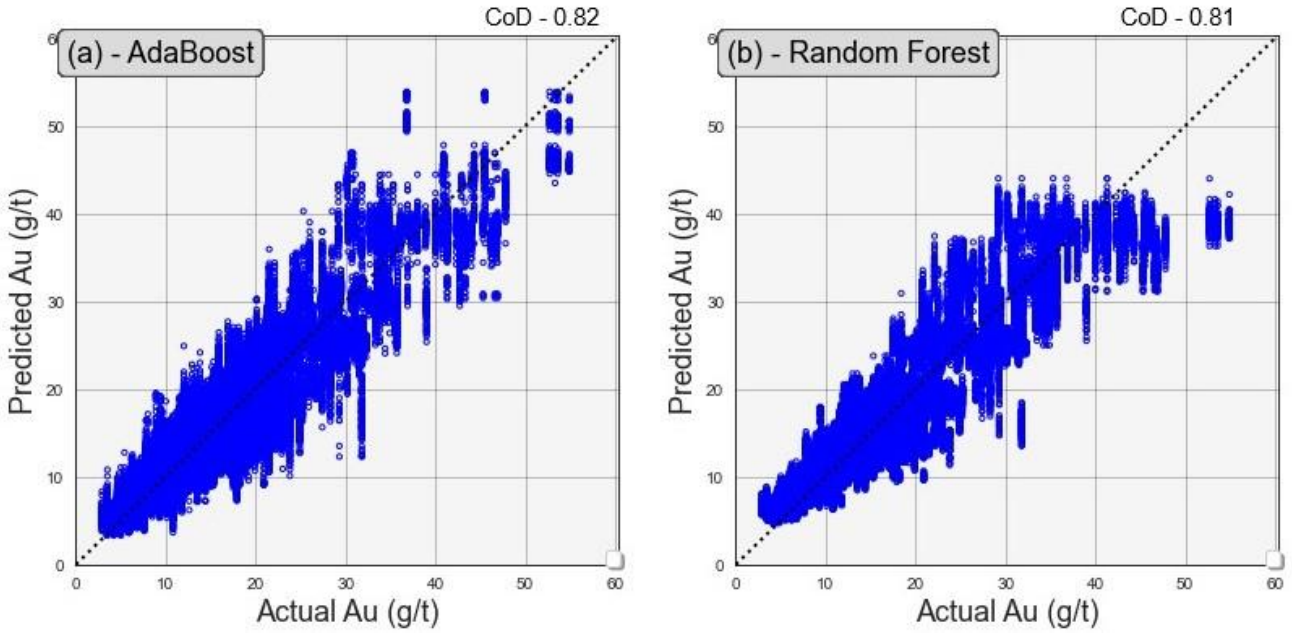


Figure 4.20: Graphical illustration of the 'good' dataset's final randomised prediction model.

The final hyperparameters used during prediction are presented in the following table:

Table 4.14: Final parameters for the randomised prediction of the 'good' dataset.

Algorithm	Hyperparameters
RF	max depth = none, max features = 2, min samples per leaf = 3, min samples per split = 3, number of trees = 1000
AdaBoost	base estimator = decision tree, base estimator max depth = none, base estimator max features = 2, base estimator min samples per leaf = 3, base estimator min samples per split = 3, number of trees = 100

4.3.2. 'Bad' Dataset

This dataset comprises 8 357 datapoints. Its descriptive statistics are presented in the following table:

Table 4.15: Descriptive statistics for the ‘bad’ dataset.

	X	Y	Channel Width	Internal Waste	Conglomerate Percentage	Basal Contact	Gold Grade (Au g/t)
count	8357.0	8357.0	8357.0	8357.0	8357.0	8357.0	8357.0
mean	7496.7	-20999.4	111.5	19.7	85.7	2.8	11.2
std	843.2	367.6	56.7	25.5	17.0	0.4	8.7
min	5717.6	-21607.8	6.0	0.0	17.0	0.0	1.6
25%	6946.6	-21320.5	70.0	0.0	74.0	3.0	4.7
50%	7640.8	-21089.5	109.0	10.0	92.0	3.0	8.0
75%	8221.6	-20720.7	150.0	33.0	100.0	3.0	16.1
max	8924.2	-20027.4	621.0	167.0	100.0	3.0	38.9

4.3.2.1. Experiment 1

The test results of the first round of predictive modelling are presented in the following table:

Table 4.16: First round of test scores for the ‘bad’ dataset.

	RF	KNN	EN	AdaBoost
30% Test Size	-0.02	-0.08	0.01	0.01

The prediction accuracies are -2% for the RF algorithm, -8% for the KNN algorithm, 1% for the EN algorithm and 1% for the AdaBoost Algorithm. These results are presented in Fig. 4.21, where the scatter diagrams display an irregular, clustered trend similar to that observed in the full dataset’s test score diagram (Fig. 4.11).

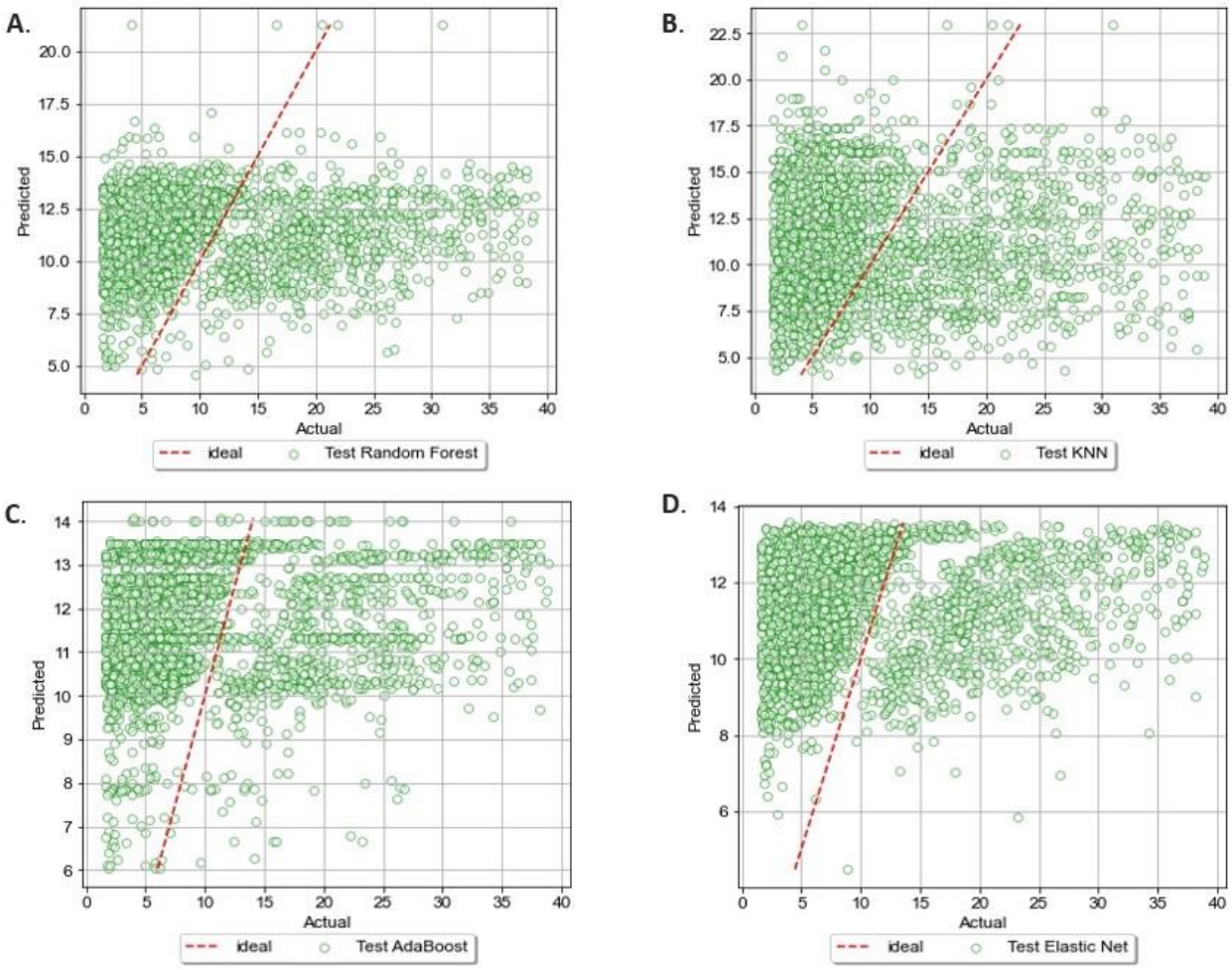


Figure 4.21: Graphical illustrations of the bad dataset's first round of testing for each of the four algorithms.

The final hyperparameters used for the 'bad' dataset's predictions are presented in the following table:

Table 4.17: 'Bad' dataset's final hyperparameters.

Algorithm	Hyperparameters
RF	max depth = 10, max features = 4, min samples per leaf = 2, min samples per split = 2, number of trees = 400
EN	max iteration = 10, α = 0.0001, L1 ratio = 0.4
KNN	number of neighbours = 9
AdaBoost	base estimator = decision tree, base estimator max depth = none, base estimator max features = 2, base estimator min samples per leaf = 2, base estimator min samples per split = 3, number of trees = 500

4.3.2.2. Experiment 2

The silhouette algorithm (Appendix A) assigned 5 clusters for the 'bad' dataset. The K-Means algorithm then allocated each datapoint to one of 5 groups. This information was compiled into a new column and used as a feature. The spatial map produced is presented in Fig 4.22:

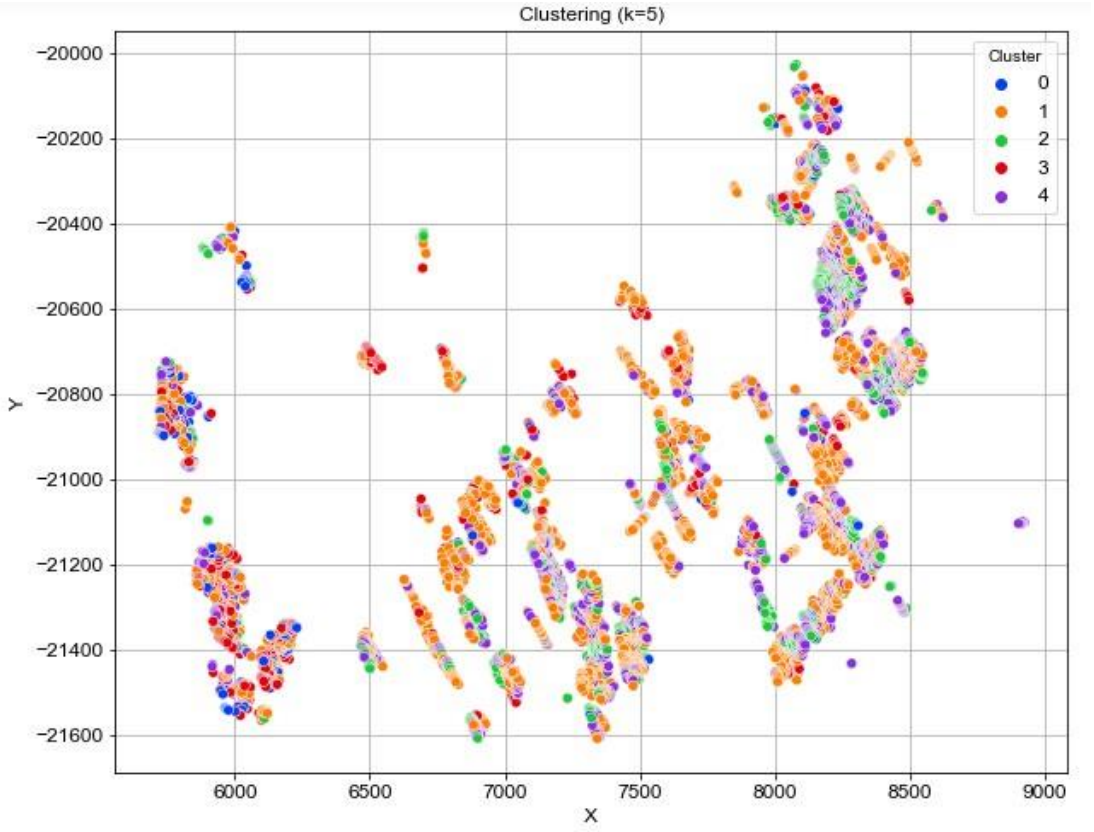


Figure 4.22: Spatial map visualising the 'bad' dataset's clustered datapoints when K = 5.

The prediction results, run over 10-fold cross-validations, are presented in the following table:

Table 4.18: Second round of test scores for the 'bad' dataset.

Test CoD (K = 5)				
	RF	**KNN**	**EN**	**AdaBoost**
30% Test Size	-0.01	-0.01	0.02	0.01
50% Test Size	-0.02	0	0.01	0

The prediction accuracies attained are -1% for the RF algorithm, -1% for the KNN algorithm, 2% for the EN algorithm and 1% for the AdaBoost Algorithm. The prediction accuracies have increased by 1%, 7% and 1% for the RF, KNN and EN algorithms, respectively, compared to the previous results obtained in Table 4.16. A graphical representation of the prediction accuracies with a 30% testing set is presented in Fig. 4.23. The scatter plots do not have a linear trend but rather display irregular clusters, with KNN's cluster being more sparsely distributed. The AdaBoost algorithm's plot once again displays discrete horizontal linear clusters.

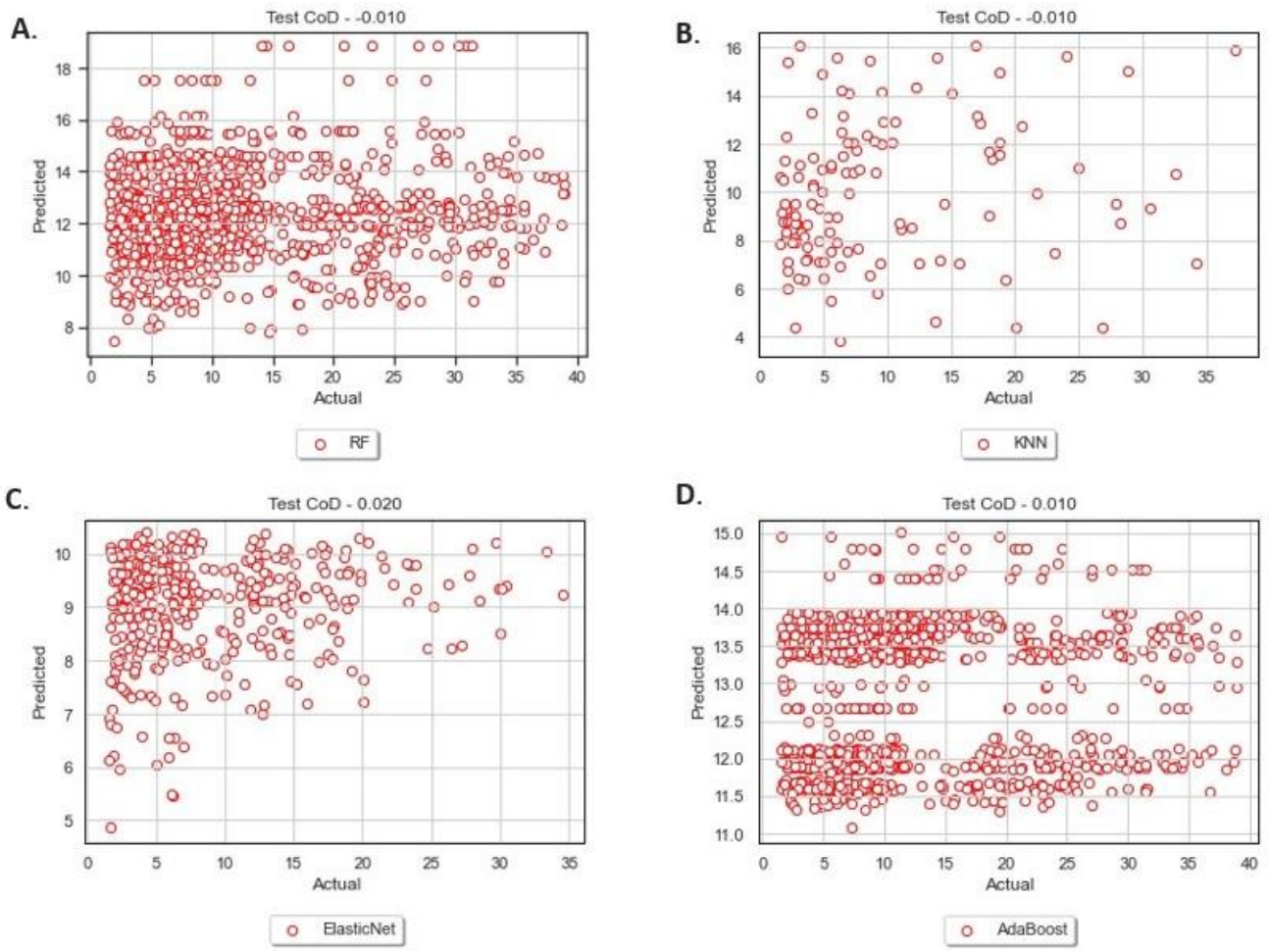


Figure 4.23: Graphical illustrations of the 'bad' dataset's second round of testing, visualising each algorithm's CoD scores.

K-Means clustering with spatial data

Once again, incorporating the spatial coordinates during per-cluster prediction returns a spatial map with more neatly defined clusters, as presented in Fig 4.24:

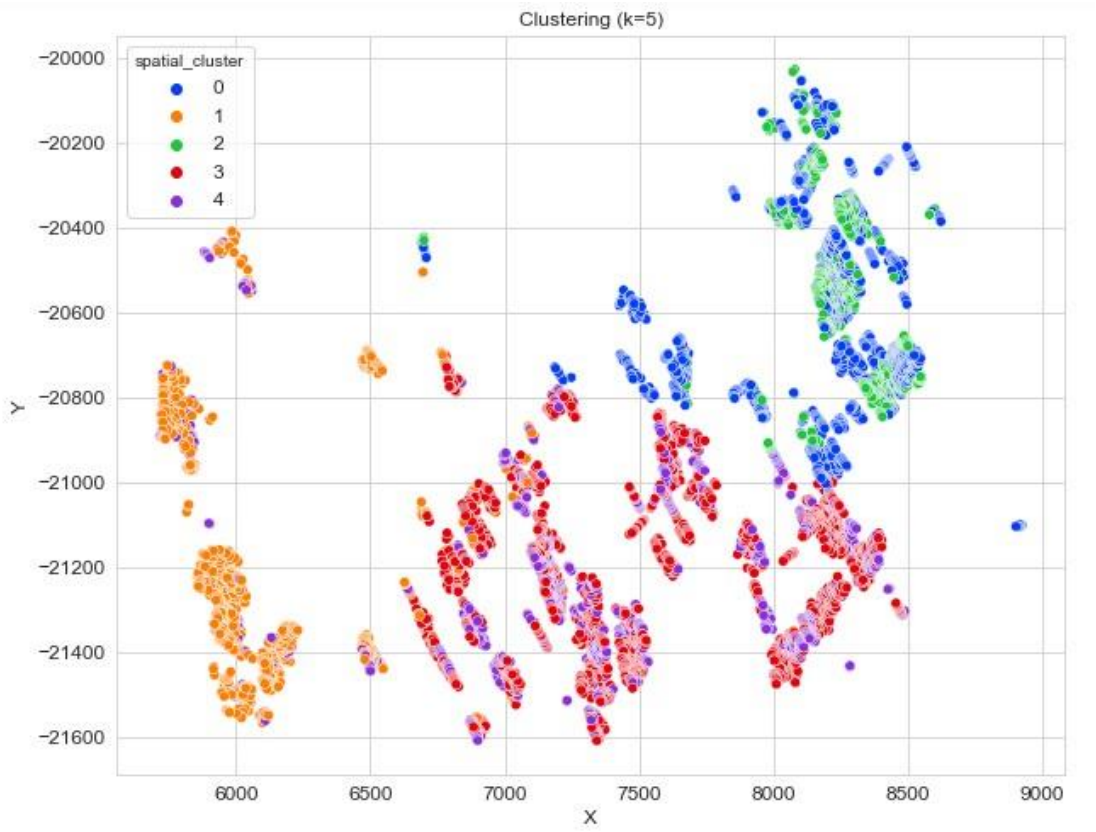


Figure 4.24: Spatial map visualising the 'bad' dataset's clustered datapoints when spatial features are incorporated.

The prediction results, averaged over 10-fold cross-validations, are presented in the following table.

Table 4.19: Second round of test scores for the 'bad' dataset when spatial features are incorporated.

Test CoD (K = 5)				
	RF	**KNN**	**EN**	**AdaBoost**
30% Test Size	0.18	0.05	0.08	0.14
50% Test Size	0.13	0.03	0.09	0.13

The prediction accuracies attained are 18% for the RF algorithm, 5% for the KNN algorithm, 8% for the EN algorithm and 14% for the AdaBoost Algorithm. The prediction accuracies have increased by 16%, 3%, 6%, and 13% for the RF, KNN, EN and AdaBoost algorithms, respectively, compared to the previous result obtained in table 4.18, where no spatial features were considered. Moreover, increasing the test size decreases the prediction accuracies on average.

4.3.3. RE Investigations

The generated RE values were also analysed visually in relation to the other features. A first line of inquiry is in the spatial domain. It is possible to examine the spatial distribution of datapoints with respect to their RE values (Fig. 4.25). Datapoints with a high RE seem to occur as discrete and linear

patterns at the outer edges of large clusters. Furthermore, smaller, isolated linear clusters appear to be dominated by high RE points.

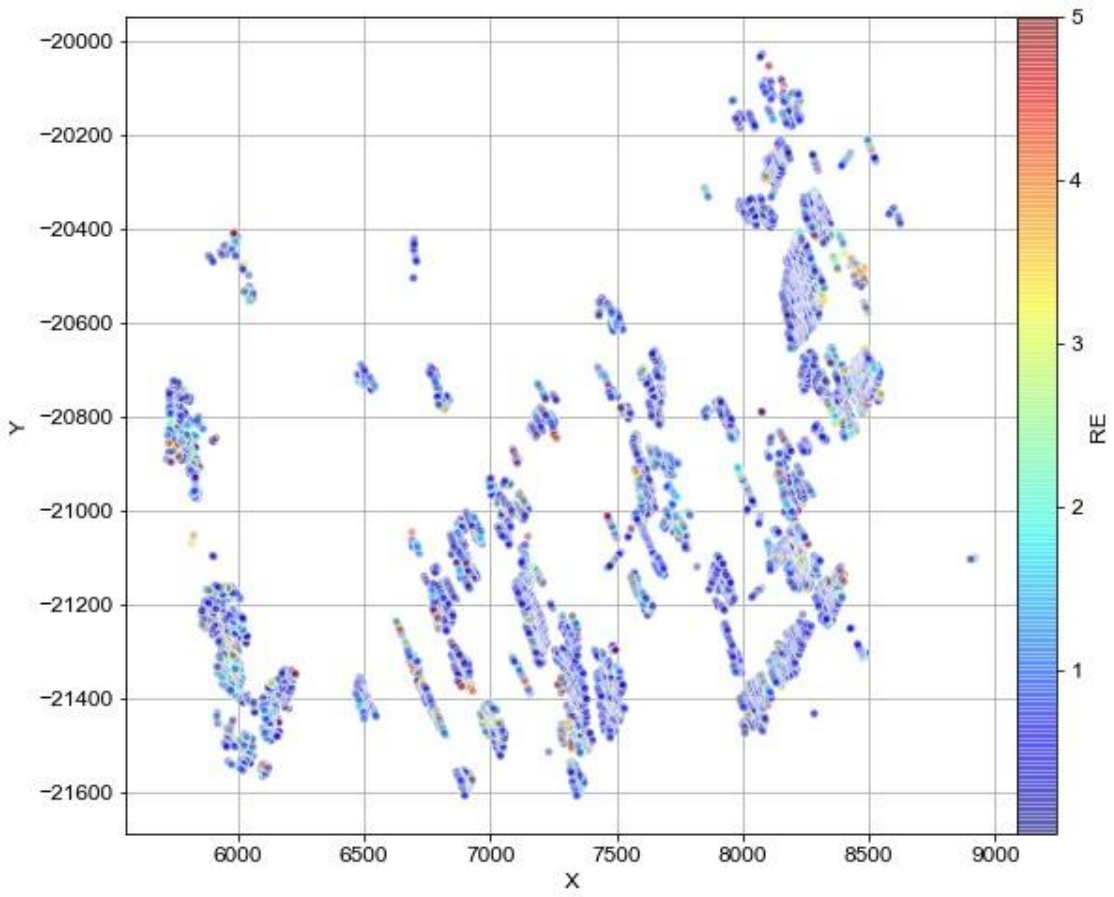


Figure 4.25: Spatial Map of the full subset colour coded by the RE.

A second line of inquiry is in the variable domain. It is possible to examine the behaviour of RE with respect to various characteristics of the data, including its targets and features. On a plot of RE versus the gold grade (Figure 4.26), there appear to be two divergent trends: the first shows a cluster of data where the RE increases with decreasing gold grade, while the other shows a cluster of data with lower RE values and higher gold grade values.

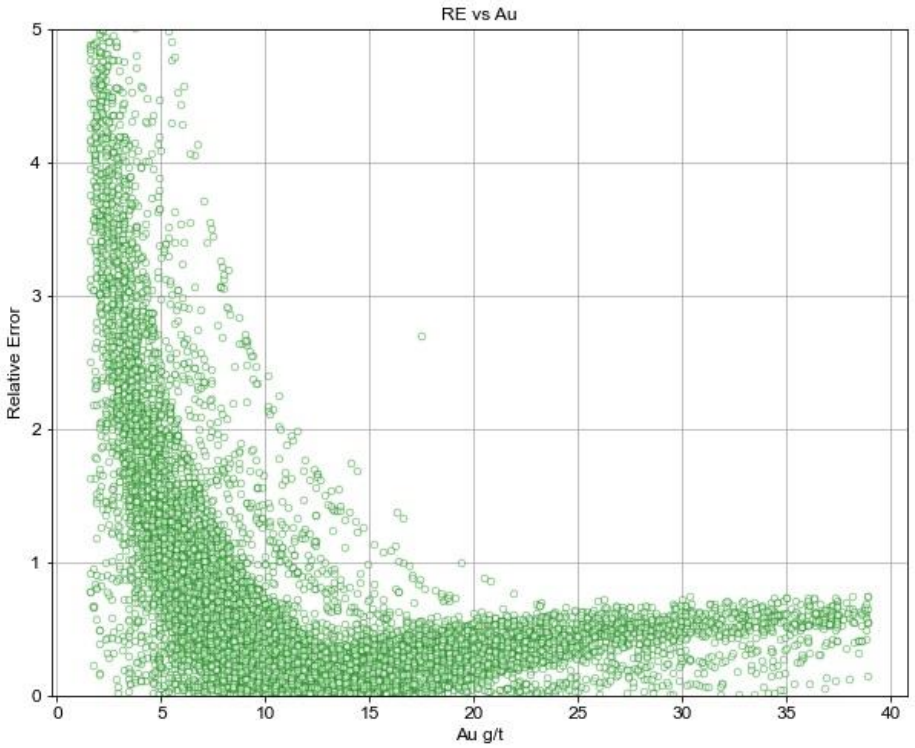


Figure 4.26: Scatter plot investigating the relationship between the RE and the gold grade (Au g/t).

Similarly, it is possible to examine the RE versus a few key features. On a plot of RE versus Conglomerate Percentage (Fig. 4.27), there appears to be no district relationship except the linear cluster observed at 100% – this is generally due to most Conglomerate Percentage values being 100% as described in the Explanatory Data Analysis section (Chapter 4.1.1). Below the maximum value of 100%, there is a weak association of higher RE with higher Conglomerate Percentage. When Conglomerate Percentage is lower than 40%, it is uncommon to find RE values higher than 1. However, this could be attributable to data bias since there are systematically fewer datapoints with Conglomerate Percentage below 40% compared to other bins.

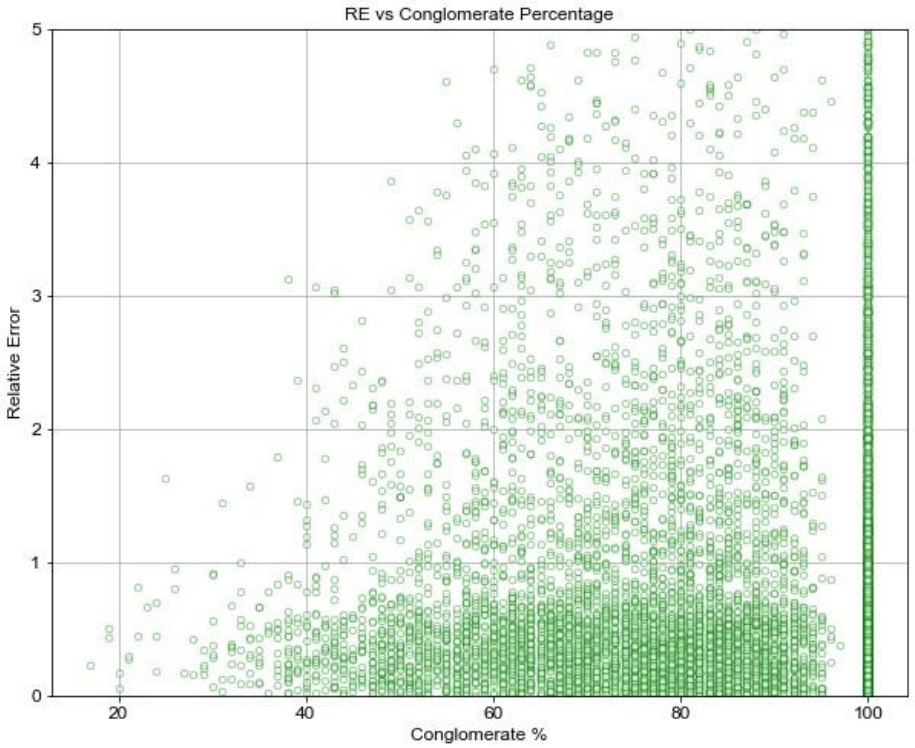


Figure 4.27: Scatter plot investigating the relationship between the RE and the Conglomerate Percentage.

There appears to be no district relationship between the RE and Internal Waste (Fig. 4.28) except the linear cluster observed at 0 cm – this is generally due to most Internal Waste values being 0 cm, as described in the Explanatory Data Analysis section. A large portion of datapoints are clustered below RE = 1 and between 10 to 75 cm.

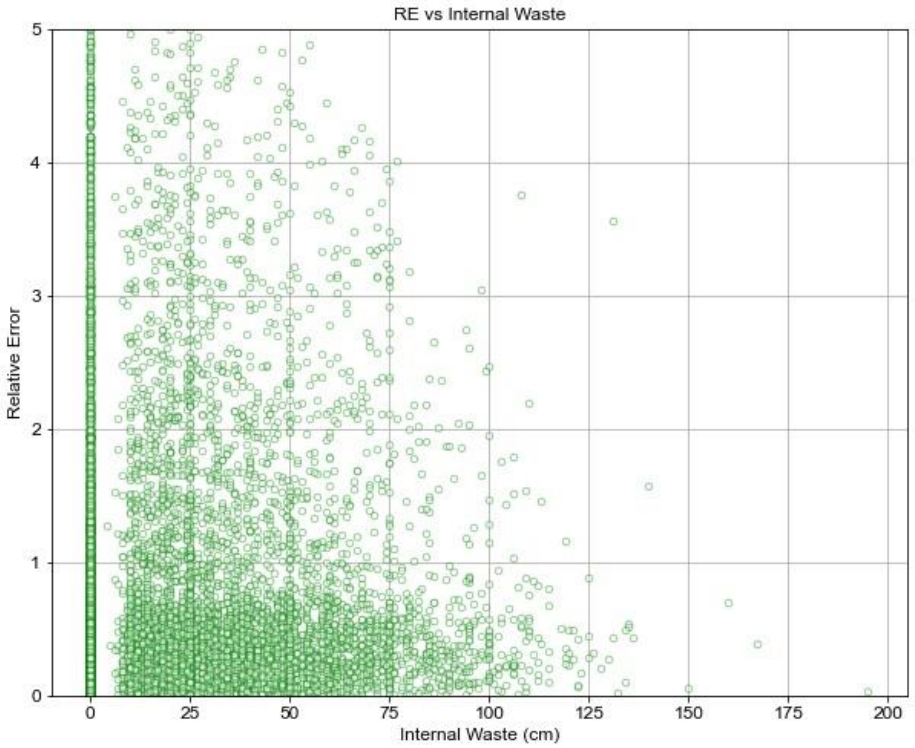


Figure 4.28: Scatter plot investigating the relationship between the RE and the Internal Waste.

There appears to be a symmetrically distributed cluster consistent with the symmetrical distribution observed in the Explanatory Data Analysis section (Fig. 4.29). Here, a large proportion of data values were clustered between 0 and 200 cm.

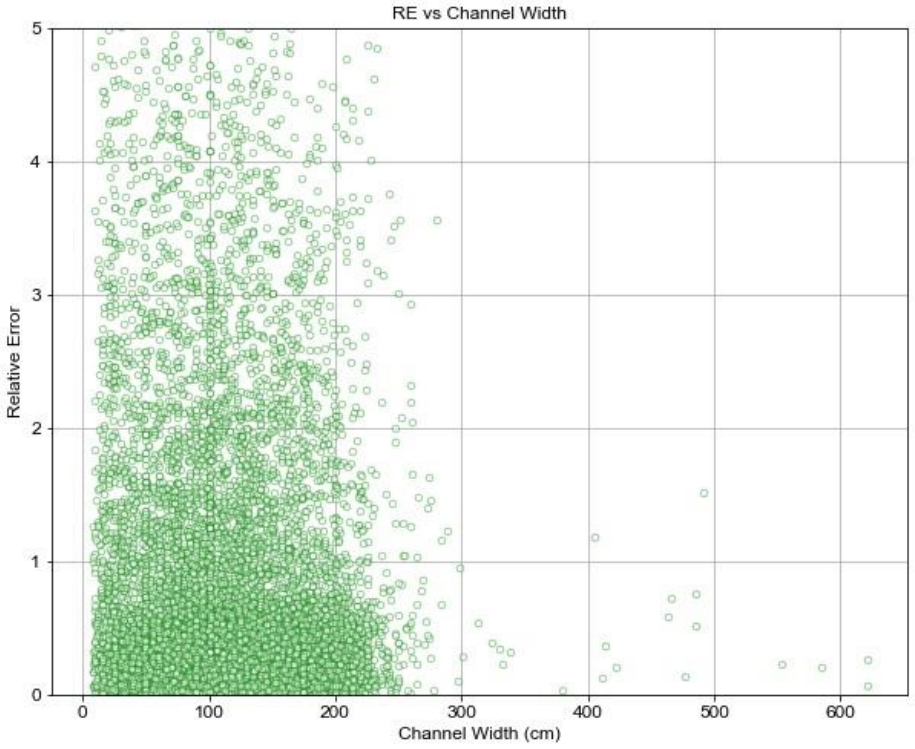


Figure 4.29: Scatter plot investigating the relationship between the RE and the Channel Width.

4.4. Assessing the importance of sedimentological features in gold grade prediction

4.4.1. Feature Selection

4.4.1.1. 'Good' Dataset

Once the 'good' dataset was identified and isolated, efforts could be made to assess which of the sedimentological features improved/worsened the prediction accuracy. Each feature was dropped one by one, and the second prediction round was repeated with the remaining features. The results are presented in the following table:

Table 4.20: 'Good' dataset's feature selection testing scores.

	Test CoD (30% Test Size)			
Dropped Feature	**RF**	**KNN**	**EN**	**AdaBoost**
Basal Contact	0.69	0.68	0.46	0.65
Conglomerate Percentage	0.85	0.84	0.52	0.81
Internal Waste	0.85	0.82	0.48	0.81
Channel Width	0.59	0.52	0.45	0.61
Internal Waste and Conglomerate%	0.58	0.56	0.41	0.55

The prediction accuracies improve on average when either the Conglomerate Percentage or the Internal Waste features are removed. The RF score improved by 1% when either of the two features were dropped, while the KNN score improved by 3% and 1% respectively, when the Conglomerate Percentage and Internal Waste features were removed. Removing the Basal Contact feature reduced the prediction accuracies by about 10% on average, while removing the Channel Width reduced them by about 20% on average. This analysis is in comparison to the results obtained in Table 4.12. Removing both the Internal Waste and Conglomerate Percentage features resulted in a significant drop in model performance compared to when only one was missing.

4.4.1.2. 'Bad' Dataset

The feature selection step was also repeated on the 'bad' dataset to assess the effect this would have on its prediction accuracies. The results are summarised in the following table.

Table 4.21: 'Bad' dataset's feature selection testing scores.

	Test CoD (30% Test Size)

Dropped Feature	RF	KNN	EN	AdaBoost
Basal Contact	-0.02	-0.05	0.03	0.03
Conglomerate Percentage	-0.03	-0.09	0.03	0.03
Internal Waste	-0.03	-0.08	0.02	0.02
Channel Width	0	-0.02	0	0.05
Internal Waste and Conglomerate%	-0.04	-0.09	0	0.02

Removing certain features did not improve or reduce the prediction scores significantly compared to the results in table 4.18. The differences are $\pm$ 1 or 2% on average for most of the algorithms except for KNN, which has a 6% decrease on average. Similarly, removing both the Internal Waste and Conglomerate Percentage features resulted in a significant drop in model performance.

4.4.2. Feature Importance

In an effort to analyse the role the sedimentological features may have played in determining which set of data predicts well and which does not, the feature importance statistics were computed for both the 'good' and 'bad' datasets.

RF Algorithm

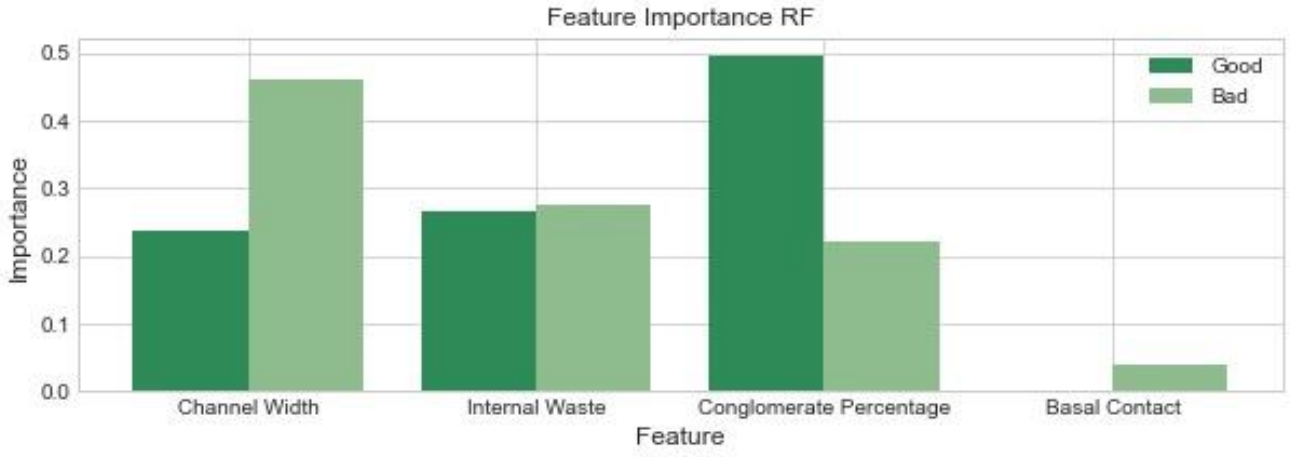


Figure 4.30: Graphical illustrations of the RF algorithm's feature importance statistics for the 'good' and 'bad' datasets.

When the 'good' dataset (Fig. 4.30) was modelled using the RF algorithm, it recognised the Conglomerate Percentage as most important; the Channel Width and Internal Waste features were roughly equal in importance, and the Basal Contact feature was considered insignificant. In contrast,

the Channel Width was most important for the ‘bad’ dataset, while the Internal Waste and Conglomerate Percentage had lower importance (Fig. 4.30).

AdaBoost Algorithm



Figure 4.31: Graphical illustrations of the AdaBoost algorithm’s feature importance statistics for the ‘good’ and ‘bad’ datasets.

The AdaBoost algorithm’s feature importance statistics show a similar trend to the RF algorithm (Fig. 4.31). For the ‘good’ dataset, the Conglomerate Percentage feature was considered most important, followed by the Internal Waste. Meanwhile, the Basal Contact feature was not considered at all for both datasets. The Channel Width was considered as most important for the ‘bad’ dataset, and the Internal Waste came second.

4.4.2.1. Influence of Spatial Data

The RF algorithm considered the Internal Waste as most important in the ‘good’ dataset; however, the presence of the spatial features significantly reduced the importance of the Conglomerate Percentage and Channel Width (Fig. 4.32). The RF algorithm considered the spatial features most important in the ‘bad’ dataset.

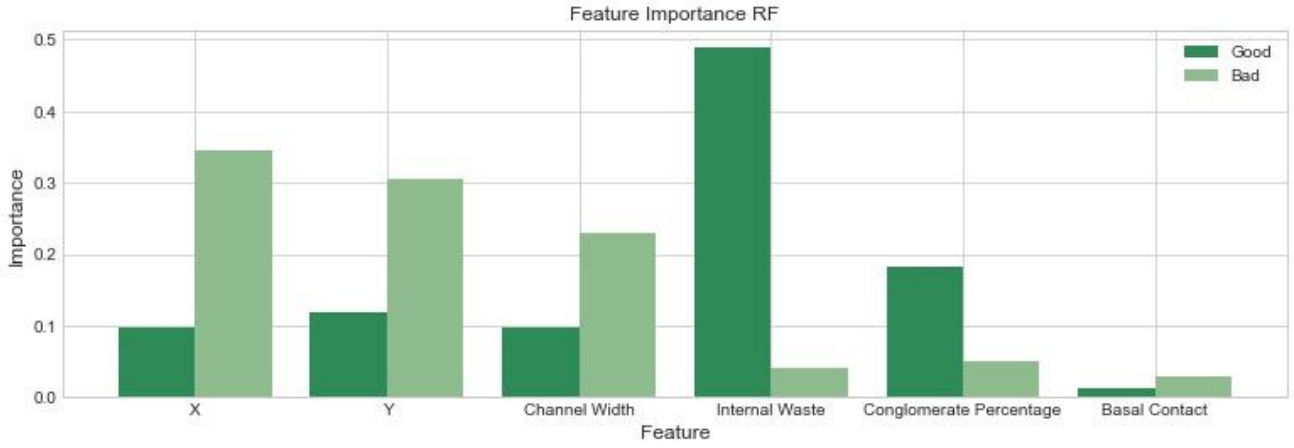


Figure 4.32: Graphical illustration of the RF's feature importance statistics for the 'good' and 'bad' datasets when spatial features are incorporated.

The AdaBoost algorithm still considered the Internal Waste and Conglomerate Percentage features as most important for the 'good' dataset (Fig. 4.33). However, the spatial features contributed more than the Basal Contact, which was again barely considered. The Channel Width feature still dominates among the 'bad' dataset's predictive features, but once again, the spatial features surpass all the predictive features.

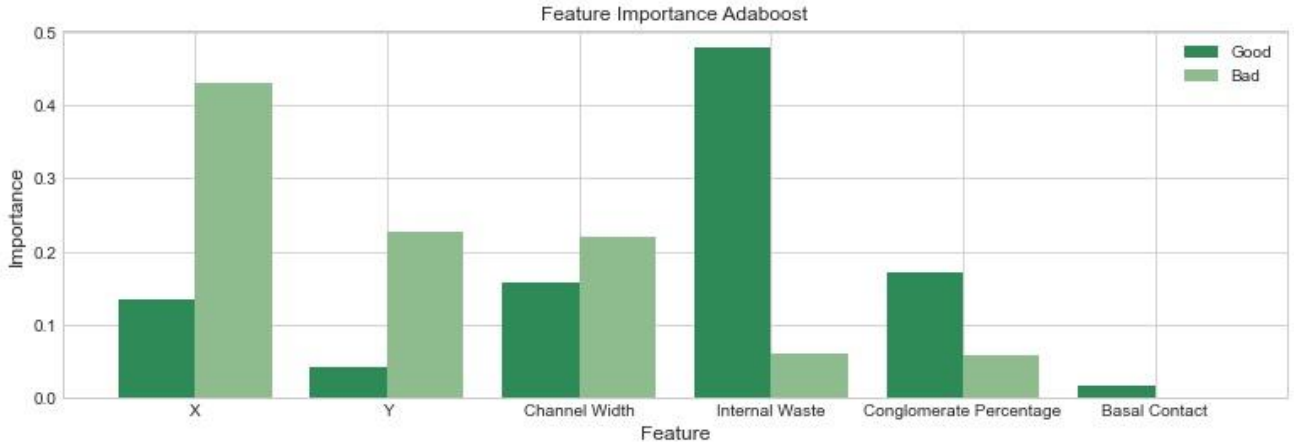


Figure 4.33: Graphical illustration of the AdaBoost's feature importance statistics for the 'good' and 'bad' datasets when spatial features are incorporated.

4.5. Kriging block model

The summary statistics of the kriging block model estimates are outlined in Table 4.22. The block model estimated a mean grade of 14.01 g/t for the 'good' dataset and a lower 11.39 g/t for the full dataset. The full dataset has a greater number of low-grade blocks (minimum grade = 2.93 g/t) compared to the 'good' dataset (minimum grade = 7.72 g/t).

Table 4.22: Summary statistics of kriging estimates for the full and 'good' datasets

Dataset	Mean Grade	std	Skew	Kurtosis	Coefficient of Variation	Max Grade	Min Grade	No. of Blocks	Datapoint count
Full	11.39	3.95	0.46	-0.08	0.35	26.74	2.93	4348	10862
'Good'	14.01	3.49	1.05	0.89	0.25	27.02	7.72	4265	2325

Variography

The variogram model produced is illustrated in Fig. 4.34. The block semi-variogram estimates were fitted with a three-component spherical model. The nugget threshold is 0.35 with a sill range of 0.01 to 0.38. The major and minor azimuths are 60° and 150° respectively.

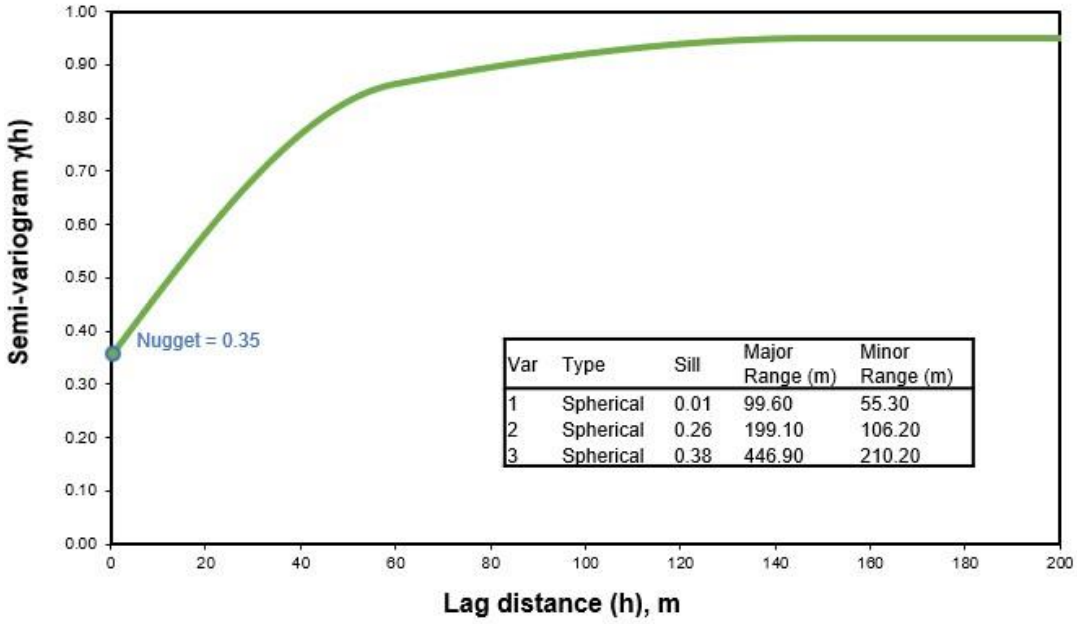


Figure 4.34: Kriging block semi-variogram model, including parameters.

Kriging block model

The block models generated for the full and 'good' datasets are illustrated in Fig. 4.34. It can be observed that the 'good' dataset contains a greater number of high gold grade domains, while the full dataset contains a mixture, with a noticeable number of very low gold grade domains. This is consistent with the block model statistics in Table 4.22. The difference in the number of datapoints between the full dataset (Fig. 4.35 (a)) and the 'good' dataset (Fig. 4.35 (b)) resulted in a difference in the spatial resolution of the block models. The blank patches observed represent areas without datapoints and where the semi-variogram search radius did not reach.

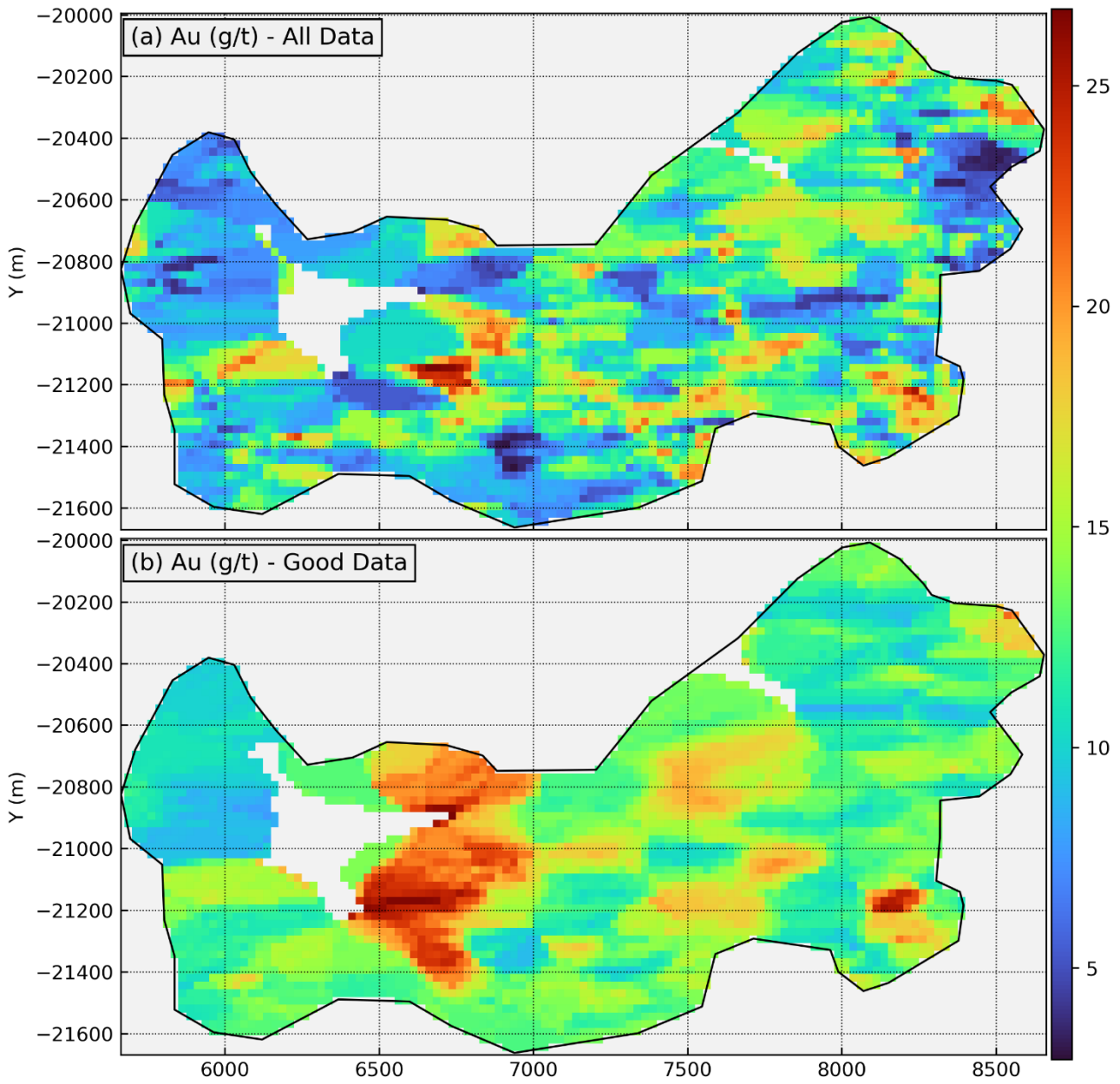


Figure 4.35: Block model estimates for the full (a) and 'good' (b) datasets.

4.5.1. Model Validation

The kriging estimates were verified using the SLOR and KEFF kriging quality measures. The analysis results are presented in Table 4.23 and reveal the same high accuracy averages for both datasets.

Table 4.23: Average SLOR and KEFF results for the full and 'good' datasets

Dataset	Avg. SLOR	Avg. KEFF
Full	0.93	0.89
'Good'	0.93	0.89

A visualisation of the KEFF results for both datasets is presented in Fig. 4.36. The ‘good’ dataset has a much smoother gradient despite the validation scores (Table 4.23) being the same. The blank patches observed represent areas without datapoints and where the semi-variogram search radius did not reach.

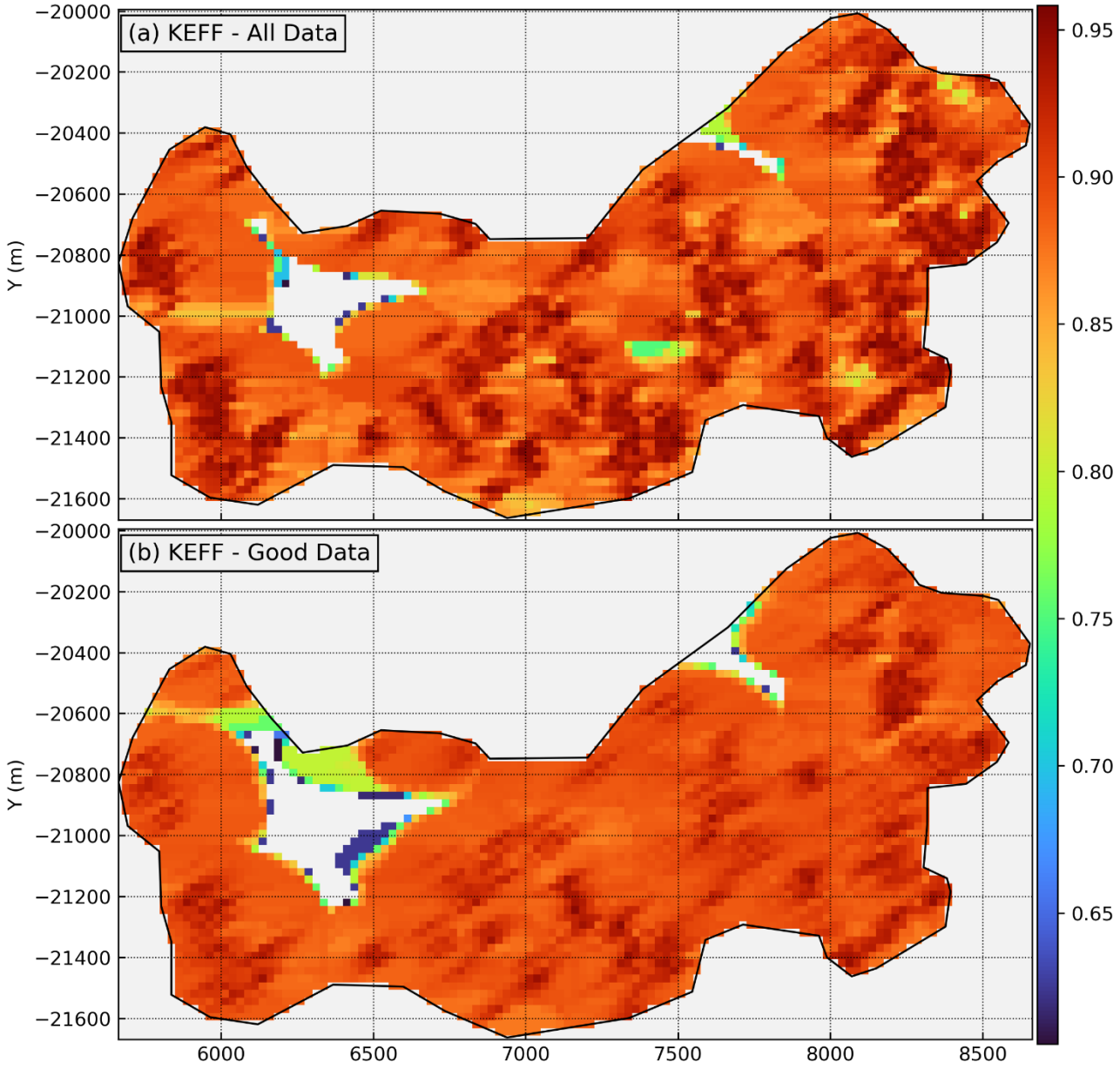


Figure 4.36: KEFF results for the full (a) and ‘good’ (b) datasets.

A visualisation of the SLOR results for both datasets is presented in Fig. 4.37. The ‘good’ dataset has a much smoother gradient despite the accuracy scores (Table 4.23) being the same.

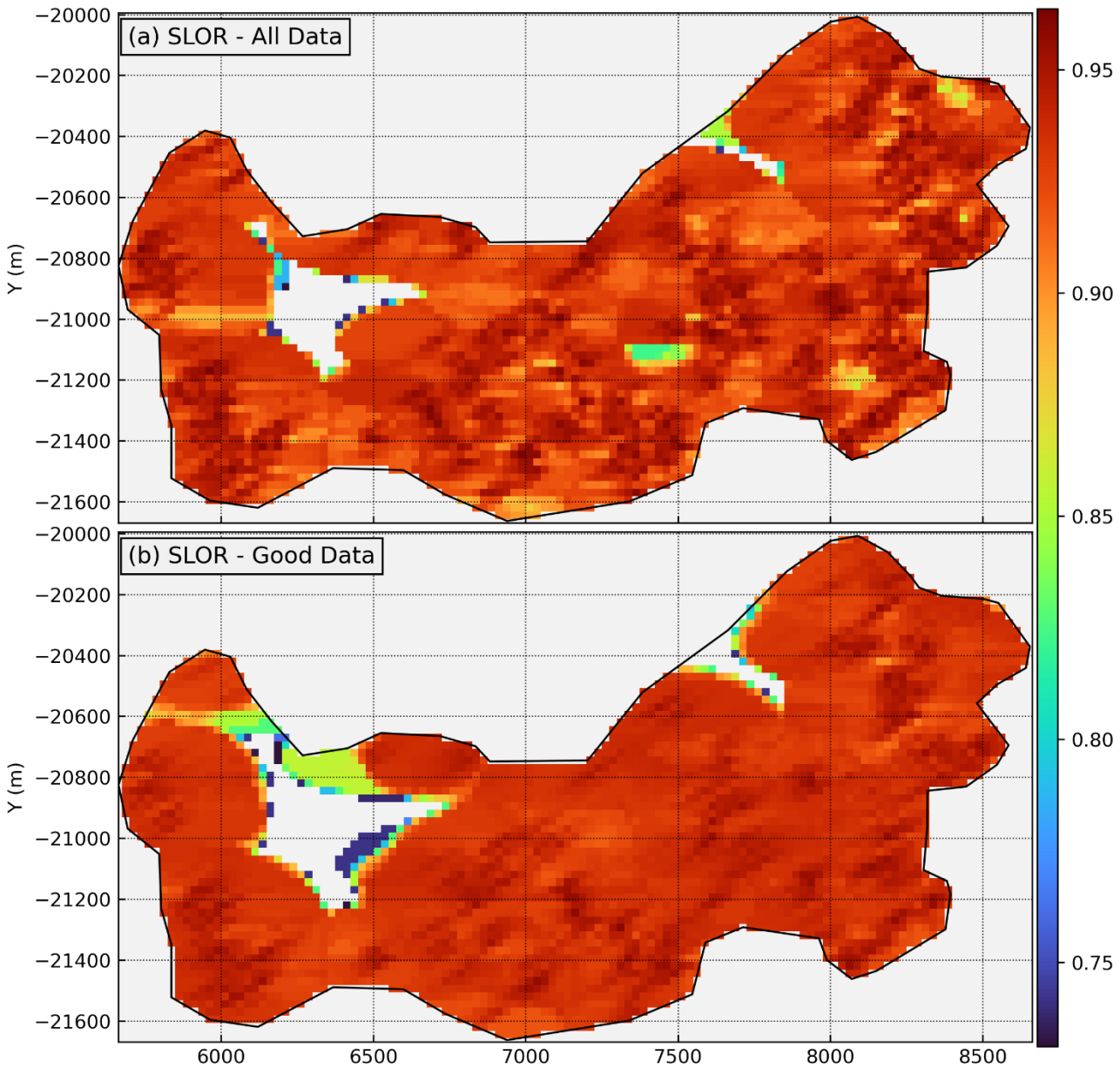


Figure 4.37: SLOR results for the full (a) and ‘good’ (b) datasets.

5. DISCUSSION

The following chapter discusses the findings of the experimental methodology process (chapters 4 and 5).

5.1. Evaluation of the results of experiments 1 and 2

Cross-validation of the full dataset produced very low accuracy scores, with the highest being 3% (AdaBoost). Per-cluster prediction introduced in experiment 2 did not have a significant effect on the cross-validation score, with the highest accuracy score being 4% (AdaBoost). Filtering the dataset using the RE metric proved to be productive because it was able to delineate between the subsets of the data by predictive modelling performance (the 'good' and 'bad' datasets, characterised by prediction performance). The 'good' dataset produced high accuracy scores (Tables 4.10 and 4.12), with the highest being 79% (RF) prior to clustering. Further predicting the 'good' and 'bad' datasets using clustering proved beneficial because it boosted the 'good' dataset's accuracy score to 84% (RF). However, adding spatial features in the second phase of experiment 2 reduced the highest accuracy score to 76% (AdaBoost). The poor-predicting ('bad') portion of the dataset shares similar results with the full dataset, which is expected given that the 'bad' data constitutes 78% of the full dataset, hence the full dataset's accuracy scores were strongly influenced by poorly predicting data. Per-cluster prediction did not have much of an influence on the 'bad' dataset's prediction scores; however, the scores increased significantly from 1% (AdaBoost) to 18% (RF) when spatial features were incorporated. Predicting the entire 'good' dataset proved to have a positive effect on the overall prediction results, with scores of 83% and 84% for the RF and AdaBoost algorithms respectively. The tree-based algorithms (RF and AdaBoost) seem to produce the best results for the dataset used in this research.

5.2. Importance of sedimentological data for gold grade prediction

Sedimentological properties have proved significant for gold grade prediction, judging from the results obtained in this study. The feature importance testing illustrated in Fig. 4.30 and 4.31 show that the most important features for the 'good' dataset are the Conglomerate Percentage and the Internal Waste. In contrast, the 'good' dataset's feature selection results (Table 4.20) demonstrate that the Conglomerate Percentage and Internal Waste features are also responsible for poor predictions since prediction accuracies remain high or increase when either feature is removed. In general, this could be the result of interactions with other features. In this case, the Conglomerate Percentage and Internal Waste are by design, complementary features in the sense that the portion of the orebody that is not conglomerate is considered Internal Waste. This was observed through analysing the data (Fig. 3.1). Hence, at least in principle, having one or the other of Internal Waste and Conglomerate Percentage should be sufficient for predictive modelling. This reduces feature space dimensionality and therefore can improve predictive modelling performance. However, the removal of both resulted in a loss of performance. Nevertheless, it seems that points which predicted poorly within the 'good'

dataset were a result of un-useful information recorded in the Conglomerate Percentage and Internal Waste. Meanwhile, the Channel Width was most important in the 'bad' dataset, while the other features contributed a lot less. There is a consistent difference between the feature importance statistics for the ‘good’ and ‘bad’ datasets from both the RF and AdaBoost algorithms. However, the AdaBoost algorithm consistently categorised Internal Waste as less important; this may mean that this feature is noisy or irrelevant. The boosting function in the AdaBoost algorithm magnifies prediction errors and uses the subsequent prediction to correct the previous one; this makes the AdaBoost algorithm very sensitive to noise. This implies that there is a structural difference in the data between the ‘good’ and ‘bad’ datasets, which itself could be a recording and data entry consistency issue, or a sample issue (including sample characteristics and sampling methods). It seems that at least some of the difference must be at the sample level, since there is a clear spatial association and gold grade difference between the ‘good’ and ‘bad’ datasets. The ’bad’ dataset’s samples generally contained lower gold grades and were distributed more on the periphery of the clusters of datapoints (Fig 4.15).

The spatial coordinates are not considered to be useful in the ‘good’ dataset, as illustrated in Fig. 4.32 and 4.33; this is also supported by the results in table 4.13, where the prediction scores decrease when spatial features are incorporated. This confirms that the geological characteristics of these samples more strongly control the gold distribution than their spatial locations. In contrast, spatial features in the ‘bad’ dataset’ were important – this tells us that the spatial data are less noisy in the ‘bad’ dataset compared to the sedimentological attributes, which were used as ML features. Consequently, it can be assumed that the ‘good’ dataset does not have a spatial pattern, while the ‘bad’ dataset does, although weakly – this is to such an extent that the other features are substantially less significant (Table 4.19). These results tell us that the location of the ‘bad’ datapoints makes a difference; furthermore, the non-spatial features are less important than the spatial features. The spatial distribution map of the ‘good’ and ‘bad’ datasets (Fig. 4.15) can support the aforementioned results because there is no clear spatial pattern observed in the ‘good’ data, while the ‘bad’ data tends to occur at the edge of ‘good’ clusters. The Basal Contact feature is categorised as least important in all feature importance statistics; however, when it was removed in the feature selection step, the prediction scores significantly lowered. It can therefore be assumed that the Basal Contact has an inconsistent control on gold occurrence, maybe through interactions with other sedimentological attributes. If this behaviour is not an artefact of sampling, then it is a physical property of the samples. In this case, the lithological contacts would play a role in the gold deposition, although it would be difficult to determine the causal relationship through this dataset. Yang et al. (2021) conducted research that found that gold mineralisation was significantly concentrated on lithological contacts, while Jolley et al. (2004) encountered gold mineralisation associated with contacts in various parts of the Witwatersrand Basin.

5.3. Importance and reliability of sedimentological data in resource estimation

This research has reiterated that sedimentological properties can be quite valuable for resource estimation. The feature importance analysis suggests that the properties are useful for gold grade prediction, but not uniformly throughout the dataset. The reliability of different sedimentological attributes varies between the 'good' and 'bad' datasets. Moreover, the feature selection step revealed that excluding certain features (i.e., Channel Width) significantly lowers prediction accuracies across several MLAs. Subsequently, it can be assumed that a larger set of sedimentological properties would provide a strong database for resource estimation and most likely result in more precise prediction accuracies. The inclusion of spatial features in this research was however purely investigative. Spatial coordinates are not transportable from one area to another; thus, models generated using spatial features are area-specific and can only be incorporated locally. If spatial co-ordinates improve the results, it means there are properties of the rocks that have not been described or adequately captured by the data; for example, a 13% improvement in the prediction score (as measured using the CoD metric) would suggest that there is 13% variability to the rock that is spatially dependent and has not been captured in the database (although not necessarily linearly variable, either). The reliability of using sedimentological data for resource estimation is once again highly dependent on the quality of the geological database. Although sedimentological properties improve model predictions on average, the extensive QA/QC (Quality Assurance and Quality Control) process that goes into cleaning the dataset reduces its size significantly, as seen in this research.

5.4. The effect of data quality on gold prediction

This research has demonstrated the importance of high-quality data for predictive modelling. Unlike chemical assays, the sedimentological descriptions seem to be of highly variable quality and are poorly predictive of gold grade at low concentrations. Predicting the full dataset returned very poor prediction accuracies, and given the scale of the data, attempts had to be made to filter the dataset and separate the well-predicting and poorly predicting datapoints. Investigations into their spatial distribution (Fig. 4.15) also reveal that these bad datapoints were sampled from isolated clusters and possibly at the outer edges of larger clusters. The datapoints in the small, isolated clusters may have been from trial sample areas that were abandoned due to low gold grades – this can be correlated with the graph of gold grade versus RE (Fig. 4.25), where we see increasing RE with low gold grade values. Furthermore, since it has been determined that the occurrence of 'bad' data are spatially coherent, it can be suggested that some of these datapoints may have been sampled off-reef or were marginal. Due to the overwhelming amount of poorly predicted datapoints, only about 22% (Fig. 5.1) of the full dataset contained well-predicting datapoints. The discrepancies between the sample locations (i.e., on-reef or off-reef) would also affect the prediction accuracies due to poor interactions between such sets. Table 4.8 shows that there are datapoints that predicted well within the 'bad' dataset (within the RE interval of between 0.8 and 0.9). However, when these were added to the

'good' dataset, prediction scores plummeted – suggesting that the two sets may not be from the same reef. These observations indicate that, for those willing to adopt predictive modelling into their business operations, it is imperative that the data generation processes be engineered for downstream uses, which would change if predictive modelling were to be employed. This is an often-overlooked aspect as the most visible aspect of predictive modelling seems to be the compelling results (where successful), but the journey starts with data generation. Repurposing data implies that although the data may meet or exceed traditional expectations, it may not meet expectations under repurposing.

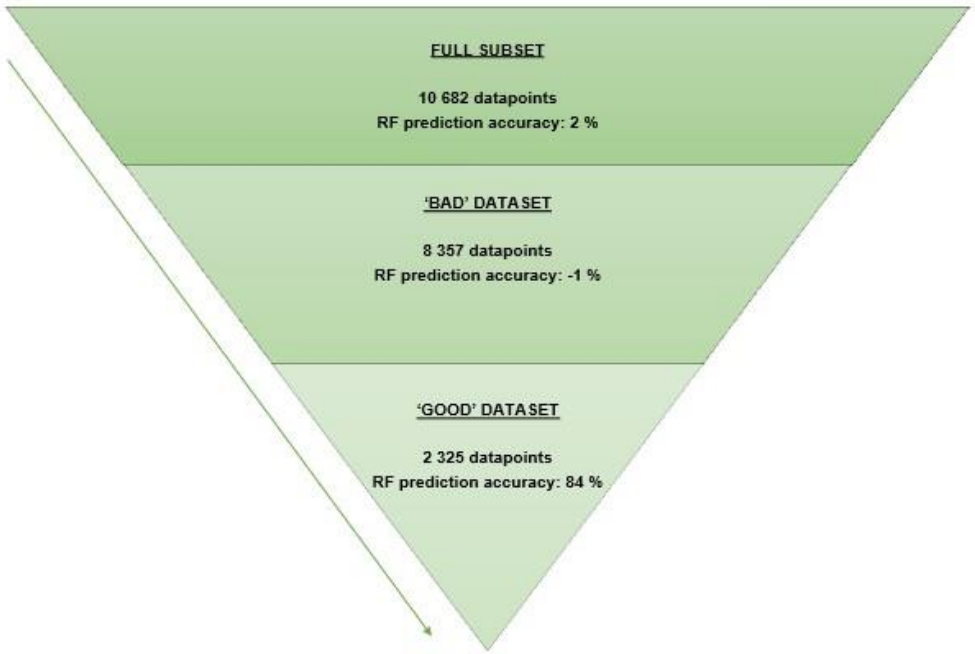


Figure 5.1: Figure illustrating the sample distribution between the full, 'bad' and 'good' datasets. The RF prediction score for each set is included.

5.5. Comparison of geostatistical and ML techniques

The use of ML is a recent phenomenon in geoscience; however, geoscientists have attributed MLAs as being useful for complex geospatial data with high spatial variations (Kanevski, 2009; Lary et al., 2016). Moreover, MLAs have produced favourable results in regression and classification tasks when compared to geostatistical techniques – particularly for tasks with non-linear relationships. In this study, MLAs have proven to be robust, comprehensive, adaptive, non-linear, and effective in several applications. Furthermore, they enabled the discovery of new information from the dataset, and a more in-depth understanding of existing geoscientific data, which is evidenced in this research by how a sedimentological database initially assembled for qualitative assessment of gold mineralisation could be contextualised and used meaningfully. Gold grade estimations made with MLAs used a combination of several geospatial parameters. MLAs accommodated the use of spatial coordinates and unlike geostatistical techniques, the MLAs did not assume an underlying spatial distribution and

could handle incomplete data. Furthermore, despite the non-linear relationships between the predictive features and the target, the MLAs managed to produce highly accurate models at the sample scale. However, at present, the use of ML in general in the mineral industry is experimental, as much of the methodology development is occurring within the research realm.

Geostatistical techniques remain more popular for mineral resource estimation despite ML's strong potential. Presently, ML techniques are primarily used as a validation tool for geostatistical methods and not solely for resource estimation. Goswami et al. (2016) noted that the overwhelming preference for traditional geostatistical techniques is mainly attributable to stakeholders having more confidence in them since they have been sufficiently accurate over the years (e.g., trust due to familiarity). In addition, a large number of geostatistical algorithms have been modernised and standardised in resource estimation software packages such as Datamine and Surpac, thus ensuring greater ease of use. Another reason for the aforementioned preference can be attributed to geostatistical techniques being widely recognised by ore reserve estimation committees like (SAMREC) the South African Code for the Reporting of Mineral Resources and Mineral Reserves, or Europe's Pan-European Reserves and Resources Reporting Committee (PERC). Despite being an estimation standard, predictions made with geostatistical techniques have to rely on second-order statistics like variograms because geostatistical methods are mostly linear estimators (Goswami et al., 2016). Second-order statistics produce reliable results for data following a Gaussian distribution but are less reliable with complex data with high variability (Goswami et al., 2011).

Studies conducted to compare geostatistical and ML techniques have returned mixed results; however, most agree that ML techniques have higher accuracies (Kanevski, 2009; Lary et al., 2016). One limitation of ML techniques that has been referenced pertains to how they are only effective when the testing and training sets are of similar distribution (Karami and Afzal, 2015; Jafrasteh et al., 2018); however, in this research, having a bigger testing size reduced the prediction accuracies significantly. Kapageridis (2005) found that the physical dimensions of the predictive features (sedimentological properties) significantly affect ML estimates after examining a geological dataset with 2D and 3D properties. This observation is in line with the results observed in this study: combing 1-dimensional features (Channel Width, Internal Waste) with 2D features (Conglomerate Percentage, Basal Contact) may be a contributor to the poor predictions initially encountered. Ultimately with ML techniques, there is generally no single configuration that works for all tasks; it depends on the characteristics of the data and the nature of the mineral deposit (Fig. 5.2) (Dumakor-Dupey and Arya, 2021). However, the MLAs discussed in this research have proven to be efficient for gold prediction, with RF and AdaBoost being the most effective.

With continued adoption of ML techniques in resource estimation, it may be necessary to create specific codes of conduct for ML-based estimation of resources. SAMREC, for example, does not

address the specifics of ML-based resource estimation. It would seem at this time that the onus to demonstrate accountability is a big and unknown risk with the adoption of ML-based resource estimation. This risk is currently unmitigable because there are no known frameworks that account for resource accounting and auditing in existence that cater to ML-based estimation aside from performance assessment frameworks within ML. In particular, the situation becomes even more unclear where estimations differ between a traditional geostatistical approach and a ML-based approach, and risk-aversive businesses may decide to adhere to the traditional approach, given the unmitigated risks associated with the ML-based approach. Unless a framework becomes available, it would be difficult to imagine that there would be systematic uptake of ML-based resource estimation within profitable businesses.

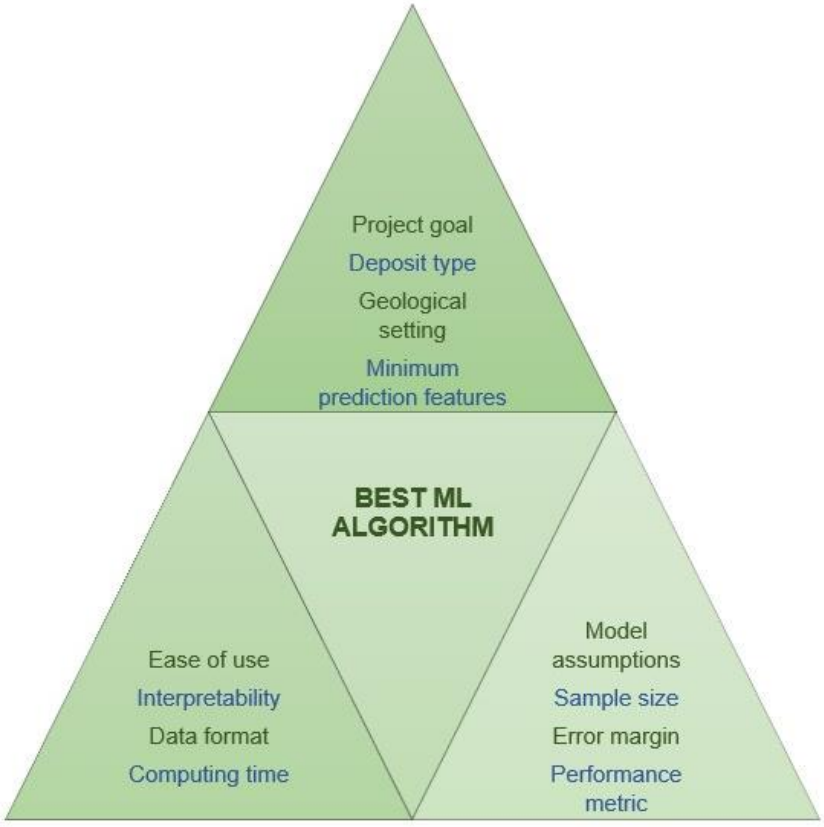


Figure 5.2: Factors affecting the choice of algorithm (adapted from Dumakor-Dupey and Arya, 2021).

5.6. Implications of this study for the exploration of Witwatersrand-type gold deposits

This study has generated a methodology that can be replicated and used on any gold deposit in the Witwatersrand Basin, where similar types of data exist. Consequently, similar deposits in other locations can also use the same methodology. The QA/QC processes and limitations addressed in this study can aid mining operations in understanding the importance of data quality and how it has/will influence their resource estimates, thus inciting a possible investigation into on-ground pre- and post-sampling methodologies. The ability to automate the ML techniques used in this study would mean that mining operations working on Witwatersrand-type gold deposits have first-hand access to a more

accurate and novel estimation method; this could consequently reduce expenses. In addition, the use of such technology would put operations working on these deposits at the forefront of the digital transformation.

5.7. Limitations and recommendations

The primary challenge experienced during the course of this research was the quality of the data provided for predictive modelling. The poor nature of the data prompted the development of a complex QA/QC workflow to extract well-predicting datapoints. However, this proved to be a very productive tangential adventure, which often happens in data-driven research. As we have demonstrated, data reconstruction using ML is a very effective method to detect poor-quality data, which in this case, exhibited clear and consistent trends (e.g., that they are all of low grade). Consequently, only 22% of the data produced good results – this was another limitation because only a small portion of the data could be salvaged, and the gold grades predicted may not be an accurate reflection of the larger deposit because of the loss of data coverage in the spatial domain. In terms of MLAs, a limitation to consider is the high computation time experienced with the tree-based algorithms, especially RF. In addition, computations attempted with other MLAs like SVM (Support Vector Machines) had even higher computation time – these issues can however be solved with higher computational power. To aid the integration of ML with geoscience, ML techniques should be integrated into modern mining software, ensuring accessibility and ease of use. However, because of the number of artificial intelligence algorithms that may be suitable for mining applications, the more practical approach would be for mining companies to adopt trans-disciplinary expertise alongside their traditional teams. Studies conducted by scholars like Kapageridis et al. (1999) have already attempted to build ML estimation software that only requires an assay database; the software then automatically performs analysis and makes predictions. Ideally, once resource estimation is complete, ML techniques should also be able to use the generated model/estimates to create block models, determine cut-off grades and perhaps go as far as to design mining sequences. A benefit of this approach would be the potential to completely utilise the data, which in our case, involved the use of sedimentological descriptions. Such descriptions are of less or no value in traditional geostatistical workflows.

6. CONCLUSIONS

This research aimed to study the properties of sedimentological data and their importance as predictive features for in-situ gold prediction using MLAs in the Witwatersrand Basin. However, the original intent of this study was to solely explore algorithms that are capable of high-performance predictions in g/t units. The tangential nature of the findings of this study is a common occurrence in data-driven projects, especially in the research and development realm. This is because much of the procedures under CRISP-DM (or other similar development models) have to be completed before a feasibility assessment can occur. Clearly, should the original intent of this study have been the only goal, this study would not have been possible. Indeed, this is an under-appreciated phenomenon outside of data sciences. This study took upon a tangential approach to instead understand the nature of the data, its quality, and impact on predictive modelling. As it turned out, this was a fruitful investigation that yielded not only insights into the data itself, but also potential business operations. In addition, this study highlighted the importance of data engineering, which seems to be necessary in the mining environment, where downstream uses adopt data-driven methods, such as ML. It is naïve for any project proponent or business to expect that their data would be useful for predictive modelling, as unless much of the CRISP-DM procedures had already been executed, such expectation is not grounded in reality, and subsequent projects could easily fail to achieve their purpose.

This research extracted high-quality samples from the original dataset by developing a methodology to differentiate between good- and poor-quality data and managed to generate highly accurate ML models. I can conclude that sedimentological properties have been effective in predicting gold grade in the Witwatersrand Basin and that prediction in g/t units is possible, but only with high-quality data. This study has revealed that data quality plays a significant role in the prediction performance and, consequently emphasised the desirability of using data-driven methods in any project, including for business purposes. The sedimentological properties did not necessarily have a strong inferential relationship with the gold grade within the entire dataset. However, the feature importance and selection stages proved that some sedimentological properties, especially Channel Width, are highly effective in predicting in-situ gold grade, while the Conglomerate Percentage/Internal Waste are negligible. The strongest association, as proxied by the highest prediction performance, is associated with samples that are seemingly more central in each spatial sample cluster. This research has deduced that the difference in prediction performance attained for the ‘good’ and ‘bad’ datasets can be attributed to their spatial distribution (as proxied by their gold grades), where well-performing datapoints were sampled on-reef, and the low predicting samples were largely sampled off-reef. In addition, this study found that the RF is the best MLA for gold grade prediction in the Witwatersrand Basin, followed by the AdaBoost algorithm; and the EN is the least suitable. Estimations produced by the block model were less viable, considering only gold grade statistics were used, and no other

sedimentological properties were included. Through literature reviews and experiments conducted in this study, it is evident that ML techniques produce ore grade estimates at the point-wise level that are generally more accurate than traditional geostatistical methods. ML still has a long way to go in geosciences; however, it is quickly gaining momentum. Its potential for automation and data-driven objectivity makes it a robust and progressive prospect that exploits new technologies, thus moving the mining industry into the new high-tech world.

7. REFERENCE LIST

Abzalov, M.Z. and Bower, J. (2014). Geology of bauxite deposits and their resource estimation practices. *Applied Earth Sci*ences 123, 118–134.

Ahmar, A.S., Napitupulu, D., Rahim, R., Hidayat, R., Sonatha, Y., and Azmi, M. (2018). Using K-Means clustering to cluster provinces in Indonesia. *Journal of Physics: Conference Series*. 1028, 1-7. DOI :10.1088/1742-6596/1028/1/012006

Alencar, D., Carvalho, D., Koenders, E., Mourao, F., and Rocha, L. (2017). Devising a computational model based on data mining techniques to predict concrete compressive strength. *Procedia Computer Science*. 108, 455-464. https://doi.org/10.1016/j.procs.2017.05.018

Al-hassan, S. and David, A. (2015). Competitiveness of Inverse Distance Weighting Method for the Evaluation of Gold Resources in Fluvial Sedimentary Deposits: A Case Study. *Journal of Geosciences and Geomatics*. 3, 122–127.

Ali Akbar, D. (2012). Reserve estimation of central part of Choghart north anomaly iron ore deposit through ordinary kriging method. *International Journal of Mining Science and Technology*. 22, 573–577.

Al-Jawarneh, S.A., Ismail, M.T., Awajan, A.M., and Alsayed, A.R.M. (2020). Improving accuracy models using elastic net regression approach based on empirical mode decomposition. *Communications in Statistics - Simulation and Computation*, 1-21. DOI:10.1080/03610918.2020.1728319

Ansah, S.K. (2018). *Geostatistical estimation of a paleoplacer deposit with hard geological boundary: case study at Tarkwa gold mine, Ghana*. Master's Thesis, Memorial University of Newfoundland, Canada.

Apley, D.W., and Zhu, J. (2020). Visualising the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society*. 82(4), 1059-1086. DOI: 10.1111/rssb.12377

Bardi, U. (2013). The mineral question: how energy and technology will determine the future of mining. *Frontiers in Energy Research*. 1, 1-11. https://doi.org/10.3389/fenrg.2013.00009

Barucci, A., Cucci, C., Franci, M., Loschiavo, M., and Argenti, F. (2021). A deep learning approach to ancient Egyptian hieroglyphs classification. *IEEE Access*. 9, 123438-123447. DOI: 10.1109/ACCESS.2021.3110082.

Biecek, P. (2019). Model development process. *Warsaw University of Technology & University of Warsaw*. 1, 1-6. https://doi.org/10.48550/arXiv.1907.04461

Biecek, P. and Burzywowski, T. (2020). *Explanatory model analysis: explore, explain, and examine predictive models. With examples in R and Python*. Chapman and Hall.

REFERENCES

Botchkarev, A. (2018). Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. *SSRN Electronical Journal*. 1-17. DOI: 10.2139/ssrn.3177507

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001). Statistical modelling: The two cultures. *Statistical Science*. 16(3), 199–231.

Calas, G. (2017). Mineral Resources and Sustainable Development. *Elements*. 13, 301-306. DOI: 10.2138/gselements.13.5.301

Casey, B., Farhangi, A. and Vogl, R. (2019). Rethinking explainable machines: The GDPR's Right to Explanation debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*. 34(1), 143–88. DOI: 10.15779/Z38M32N986

Cate, A., Perozzi, L., Gloaguen, E., and Blouin, M. (2017). Machine learning as a tool for geologists. *The Leading Edge*, 36(3), 215-219. https://doi.org/10.1190/tle36030215.1

Chapman, P., Kerber, R., Clinton, J., Khabaza, T., Reinartz, T., and Wirth, R. (1999). The CRISP-DM process model. *The CRISP–DM Consortium*. 91, 1-310.

Chen, Y., Jiang, X., Wang, Y., and Zhuang, D. (2018). Spatial characteristics of heavy metal pollution and the potential ecological risk of a typical mining area: a case study in China. *Process Safety and Environmental Protection*. 113, 204–219.

Coates, D.R. (1985). *Mineral Resources*. In: Geology and Society. Environmental Resource Management Series. Springer, Boston, MA. https://doi.org/10.1007/978-1-4613-2543-7_2

Coleman, J.R. (2021). *Implicit Resources Modelling with vector indicator kriging*. Master's Thesis, The Polytechnic University of Turin, Italy.

Dankert, B.T. and Hein, K.A.A. (2009). Evaluating the structural character and tectonic history of the Witwatersrand Basin. *Precambrian Research*. 177(1), 1-22. DOI: 10.1016/j.precamres.2009.10.007

Daya, A. (2015). Ordinary kriging for the estimation of vein type copper deposit: A case study of the Chelkureh. *Iranian Journal of Mining and Metallurgy*. 51, 1–14.

Daya, A.A. and Bejari, H. (2015). A comparative study between simple kriging and ordinary kriging for estimating and modeling the Cu concentration in Chehlkureh deposit, SE Iran. *Arabian Journal of Geoscience*. 8, 6003–6020.

Deutsch, C.V. and Deutsch, J. L. (2015). Introduction to choosing a Kriging Plan. *Geostatistics Lessons*. 1-7.

Deutsch, J.L., Szymanski, J., and Deutsch, C.V. (2014). Checks and measures of performance for kriging estimates. *The Journal of The Southern African Institute of Mining and Metallurgy*. 114. 223-230.

Dobson, A. J. (2002). *Introduction to Generalized Linear Models (2nd Ed.)*. Chapman and Hall/CRC.

Dominy, S.C. and Hunt, S.P. (2001). Evaluation of gold deposits-Part 2: Results of a survey of estimation methodologies applied in the Eastern Goldfields of Western Australia. *Transactions of the Institutions of Mining and Metallurgy*. 10, 167–175.

Dubinski, J. (2013). Sustainable Development of Mining Mineral Resources. *Journal of Sustainable Metallurgy.* 12, 1–6.

Dumakor-Dupey, N.K. and Arya, S. (2021). Machine learning - A review of applications in Mineral Resource Estimation. *Energies*. 14, 4079.

Dutta, S., Bandopadhyay, S., Ganguli, R., and Misra, D. (2010). Machine learning algorithms and their application to ore reserve estimation of sparse and imprecise data. *Journal of Intelligent Learning Systems and Applications*. 2(02), pp. 86-96.

Dutta, S., Bandopadhyay, S., Ganguli, R., and Misra, D. (2010). Machine learning algorithms and their application to ore reserve estimation of sparse and imprecise data. *Journal of Intelligent Learning Systems*. 2, 86.

Efron, B. and Hastie, T. (2016). *Computer age statistical inference: algorithms, evidence, and data science (1st Ed.)*. NY: Cambridge University Press.

Els, B.G. (1991). Placer formation during progradational fluvial degradation: the late Archean Middelvlei gold placer, Witwatersrand South Africa. *Economic Geology*. 86(2), 261-277. https://doi.org/10.2113/gsecongeo.86.2.261

Erarslan. K. (2012). Computer aided ore body modelling and mine valuation. *Earth Sciences*. 325-372. DOI: 10.5772/26020

Fisher, A., Rudin, c., and Dominici, F. (2018). All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *Statistics.* 20(1), 1-81. https://doi.org/10.48550/arXiv.1801.01489

for classification models. *SN Applied Sciences*. 3(272), 0-12. https://doi.org/10.1007/s42452-021-04148-9

forecasts. *International Journal of Forecasting*. 32(3), 669-67. DOI: 10.1016/j.ijforecast.2015.12.003

Frimmel, H.E. and Nwaila, G.T. (2020). Geologic evidence of syngenetic gold in the Witwatersrand goldfields, South Africa. *Society of Economic Geologists*. 23, 645-668. DOI: https://doi.org/10.5382/SP.23.31

Frimmel, H.E., Lehrmann, A.Z.B., Hallbauer, D., and Frank, W. (2009). Geochemical and geochronological constraints on the nature of the immediate basement next to the Mesoarchean auriferous Witwatersrand Basin, South Africa. *Journal of Petrology*. 50 (12), 2187-2220. https://doi.org/10.1093/petrology/egp073

Fuchs, S., Williams-Jones, A.E., and Przybylowicz, W.J. (2016) The origin of the gold and uranium ores of the Black Reef Formation, Transvaal Supergroup, South Africa. *Ore Geology Reviews*. 72(1), 149-164. https://doi.org/10.1016/j.oregeorev.2015.07.010

Ghorbani, A., Wexler, J., Zou, J. and Kim, B. (2019). Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*. 832, 9277–9286.

Gonzalez-Blanks, A., Bridgewater, J.M., and Yates, T.M. (2020). Statistical approaches for highly skewed data: evaluating relations between maltreatment and young adults' non-suicidal self-injury. *Journal of Clinical Child and Adolescent Psychology*. 49(2), 147–161. https://doi.org/10.1080/15374416.2020.1724543

Goodman, B. and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a 'right to explanation'. *AI Magazine*. 38(3), 50–57. https://doi.org/10.48550/arXiv.1606.08813

Goswami, A.D., Mishra, M.K., and Patra, D. (2011). Investigation of general regression neural network architecture for grade estimation of an Indian iron ore deposit. *Arabian Journal of Geoscience*. 10.

Goswami, A.D., Mishra, M.K., and Patra, D. (2016). *Adapting pattern recognition approach for uncertainty assessment in the geologic resource estimation for Indian iron ore mines*. In Proceedings of the International Conference on Signal Processing, Communication, Power and Embedded System, Germany.

Granek, J. (2016). *Application of machine learning algorithms to mineral prospectivity mapping*. PhD Thesis. The University of British Columbia, Vancouver.

Gray, N.K., Tucker, R.F. and Kershaw, D.J. (1994). The Sun Project – 1, Discovery of a major new Witwatersrand goldfield. *Proc. XVth CMMI Congress*. 3, 95-102.

Hall, P., Gill, N., and Schmidt, N. (2019). Proposed guidelines for the responsible use of explainable machine learning. *33rd Conference on Neural Information Processing Systems*. 3, 1-11.

Hammouri, H.M., Sabo, R.T., Alsaadawi, R., and Kheirallah, K.A. (2020). Handling skewed data: a comparison of two popular methods. *Applied Sciences*. 10, 1-14. doi:10.3390/app10186247

Hartman, H.L. and Mutmansky, J.M. (2002). *Introductory Mining Engineering, 2nd ed*. John Wiley & Sons.

Hartnady, C.J.H. (2009). South Africa's gold production and reserves. *South African Journal of Science*. 105(9), 328-329. DOI: 10.4102/sajs.v105i9/10.101

Hastie, T., Tibshirani, R and Friedman, J. (2009*). The elements of statistical learning: data mining, inference, and prediction (2nd Ed.)*. NY: Springer.

Hayward, C.L., Reimold, W.U., Gibson, R.L., and Robb, L.J. (2005). Gold mineralization within the Witwatersrand Band, South Africa: evidence for a modified placer origin, and the role of the Vredefort impact event. *Geological Society London*. 248(1), 31-58. DOI: 10.1144/GSL.SP.2005.248.01.02

Hekmatnejad, A., Emery, X., and Alipour-Shahsavari, M. (2019). Comparing linear and non-linear kriging for grade prediction and ore/waste classification in mineral deposits. *International Journal of Mining, Reclamation and Environment*. 33, 247–264.

Henckens, M.L.C.M., Biermann, F.H.B., and Driessen, P.P.J. (2019). Mineral resources governance: A call for the establishment of an International Competence Center on Mineral Resources Management. *Resources, Conservation and Recycling*. 141, 255–263.

Hotz, N. (2022). *What is CRISP DM?* Available at: https://www.datascience-pm.com/crisp-dm-2/ (Accessed: 18 November 2022).

Imandoust, S.B. and Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) approach for predicting economic events: theoretical background*. International Journal of Engineering Research and Applications*. 3(5), 605-610.

Inglis, A., Parnell, A., and Hurley, C.B. (2021). Visualising variable importance and variable interaction effects in machine learning models. *Statistics*. 2, 1-31. https://doi.org/10.48550/arXiv.2108.04310

Ishikawa, S.T. and Gulick, V.C. (2013). An automated mineral classifier using Raman spectra. *Computational Geosciences*. 54, 259–268.

Jafrasteh, B., Fathianpour, N., and Suárez, A. (2018). Comparison of machine learning methods for copper ore grade estimation. *Computational Geosciences*. 2018, 22, 1371–1388.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An introduction to statistical learning: with applications in R (Second Edition)*. NY: Springer.

Jolley, A.J., Freeman, A.R., Barnicoat, A.C., Phillips, G.M., Knipe, R.J., Fox, N.P.C., Strydom, D., Birch, M.T.G., and Henderson, I.H.C. (2004). Structural controls on Witwatersrand gold mineralisation. *Journal of Structural Geology*. 24, 1067-1086.

Jones, O., Lilford, E., and Chan, F. (2018). *The Business of Mining: Mineral Project Valuation*. CRC Press: Boca Raton.

Jung, D. and Choi, Y. (2021). Systematic review of machine learning applications in mining: exploration, exploitation, and reclamation. *Minerals*. 11(2), 1-22. https://doi.org/10.3390/min11020148

Kanevski, M. (2009). *Machine Learning for Spatial Environmental Data: Theory, Applications, and Software*. EPFL Press: Lausanne, Switzerland.

Kapageridis, I.K. (2005). Input space configuration effects in neural network-based grade estimation. *Computational Geosciences*. 31, 704–717.

Kapageridis, I.K. and Denby, B. and Hunter, G. (1999). *Integration of a neural ore grade estimation tool in a 3D resource modeling package*. In proceedings of the International Joint Conference on neural networks. Washington DC, USA.

Kaplan, U.E. and Topal, E. (2020). A new ore grade estimation using combine machine learning algorithms. *Minerals*. 10(10), 1-17. DOI: 10.3390/min10100847

Karami, R. and Afzal, P. (2015). Estimation of elemental distributions by combining artificial neural network and Inverse Distance Weighted (IDW) based on lithogeochemical data in Kahang porphry deposit, Central Iran. *Universal Journal of Geoscience*. 3, 59–65.

Kentwell, D. (2022). Empirical Geostatistics #1: Kriging slope of regression – sensitivities and impacts on estimation, classification, and final selection. *International Mining Geology Conference*. 1-21.

Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*. 32(3), 669-679.

Kleinhans, M.G., Buskes, C.J.J., and de Regt, H.W. (2005). Terra Incognita: Explanation and reduction in Earth Science. *International Studies in the Philosophy of Science*. 19 (3), 289-317. https://doi.org/10.1080/02698590500462356

Koduri, S.B., Gunisetti, L., Ramesh, C.R., Mutyalu, K.V., and Ganesh, G. (2019) Prediction of crop production using Adaboost regression method. *International Conference on Computer Vision and Machine Learning*. 1228, 1-11. doi:10.1088/1742-6596/1228/1/012005

Kramer, O. (2013). K-nearest neighbors. In: dimensionality reduction with unsupervised nearest neighbors. *Intelligent Systems Reference Library*. 51, 13-23. https://doi.org/10.1007/978-3-642-38652-7_2

Krapez, B. (1979*). Facies variations in the Ventersdorp Contact Reef at East Driefontein Gold Mine, Carletonville Goldfield*. MSc thesis. University of the Witwatersrand,, Johannesburg.

Lary, D.J., Alavi, A.H., Gandomi, A.H., and Walker, A.L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*. 7, 3–10.

Li, J., Heap, A.D., Potter, A., and Daniell, J.J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modeling and Software*. 26, 1647–1659.

Liaw, A. and Wiener M. (2002). Classification and regression by random forest. *R News*. 2, 18–22.

Lindsay, J.J., Hughes, S.R.H., Yeomans, C.M., Andersen, J.C.Ø., and McDonald, I. (2021). A machine learning approach for regional geochemical data: Platinum-group element geochemistry vs

geodynamic settings of the North Atlantic Igneous Province. *Geoscience Frontiers*. 12(3), 1-18. https://doi.org/10.1016/j.gsf.2020.10.005

Lishchuk, V., Lund, C., and Ghorbani, Y. (2019). Evaluation and comparison of different machine-learning methods to integrate sparse process data into a spatial model in geometallurgy. *Minerals Engineering*. 134, 156–165.

Liu, W. and Li, Q. (2017). An efficient Elastic Net with regression coefficients method for variable selection of spectrum data. *PLoS ONE*. 12(2), 1-13. e0171122. doi:10.1371/journal.pone.0171122

Luckert, M. and Schaefer-Kehnert, M. (2016). *Using Machine Learning Methods for Evaluating the Quality of Technical Documents*. Master's Thesis. The University of Stuttgart.

Marwanza, I., Nas, C., Badaruddin, S.P., and Azizi, M.A. (2018). Copper resource estimation in PT X Batu Hijau, Regency of West Sumbawa, West Nusa Tenggara Province using geostatistical method. *IOP Conference Series: Earth and Environmental Science*. 212, 0-10. doi:10.1088/1755-1315/212/1/012024

McMillan, M., Fohring, J., Haber, E., and Granek, J. (2019). Orogenic gold prospectivity mapping using machine learning. *Australian Society of Exploration geophysicists*. 2019(1), 1-4. DOI: 10.1080/22020586.2019.12073020

Molnar, C. (2022). *Interpretable machine learning: a guide for making black box models explainable*. Creative Commons Attribution.

Mpanza, M. (2015). *A comparison of ordinary and simple kriging on a PGE resource in the eastern limb of the bushveld complex*. MSc Thesis, University of the Witwatersrand, Johannesburg.

Myers, R.E., Zhou, T., and Phillips, N. (1993). Sulphidation in the Witwatersrand Goldfields: evidence from the Middelvlei Reef. *Mineralogical Magazine*. 57, 395-405.

Nainggolan, R., Perangin-angin, R., Simarmata, E., and Tarigan, F.A. (2019). Improved the performance of the K-Means cluster using the sum of squared error (SSE) optimized by using the elbow method. *Journal of Physics: Conference Series*. 1361, 1-7. doi:10.1088/1742-6596/1361/1/012015

Novitasari, H.B., Hadianto, N., Sfenrianto, M., Rahmawati, A., Prasetyo, R., Miharja, J., and Gata, W. (2019). K-nearest neighbor analysis to predict the accuracy of product delivery using administration of raw material model in the cosmetic industry (PT Cedefindo). *Journal of Physics: Conference Series*. 1367, 1-7. doi:10.1088/1742-6596/1367/1/012008

Nwaila, G.T., Zhang, S.E., Tolmay, L.C.K., and Frimmel, H.E. (2020). Algorithmic optimization of an underground Witwatersrand-type gold mine plan. *International Association for Mathematical Geosciences*. 1-23.

Nwaila, T. N. Zhang, S.E., Bourdeau, J.E., Negwagwatini, E., Rose, D.H., Burnett, M., and Ghorbani, Y. (2022). Data-driven predictive modeling of lithofacies and Fe in-situ grade in the Assen Fe Ore deposit of the Transvaal Supergroup (South Africa) and implications on the genesis of Banded Iron Formations. *Natural Resources Research*. 1-27. https://doi.org/10.1007/s11053-022-10105-z

Phillips, G.N. and Powell, R. (2011). Origin of Witwatersrand gold: a metamorphic devolatilization–hydrothermal replacement model. *Applied Earth Science*. 120(3), 112-129. DOI: 10.1179/1743275812Y.0000000005

Rahimi, H., Asghari, O., and Meysami, F. (2018). Investigation of linear and non-linear estimation methods in highly-skewed gold distribution. *Journal of Mining and Environment*. 9(4), 967-979. DOI: 10.22044/jme.2018.7023.1544

Robb, L.J. and Meyer, F.M. (1995). The Witwatersrand Basin, South Africa: Geological framework and mineralization processes. *Ore Geology Reviews*. 10(2), 67-94. https://doi.org/10.1016/0169-1368(95)00011-9

Rossi, M.E. and Deutsch, C.V. (2014). *Mineral Resource Estimation*. Springer Science & Business Media.

Saarela, M. and Jauhiainen, S. (2021). Comparison of feature importance measures as explanations

SACS (South African Committee for Stratigraphy). (1980). Lithostratigraphy of the Republic of South Africa, South West Africa/Namibia, and the Republics of Bophuthatswana, Transkei and Venda. *Geological Survey Southern Africa*. 8(1), 1-690.

Safonov, Y.G. and Prokof'ev, V.Y. (2006). Gold-bearing reefs of the Witwatersrand Basin: A model of synsedimentation hydrothermal formation. *Geology of Ore Deposits*. 48(6). 415-447. DOI: 10.1134/S1075701506060018

Samson, M. (2019). *Mineral resource estimates with machine learning and geostatistics*. MSc Thesis, University of Alberta, Edmonton.

Shmueli, G. (2010). To explain or to predict. *Statistical Science*. 25(3), 289–310. DOI: 10.1214/10-STS330

Shurygin, D.N., Vlasenko, S.V., and Shutkova, V.V. (2019). Estimation of the error in the calculation of mineral reserves taking into account the heterogeneity of the geological space. *IOP Publishing: Bristol*. 272, 22139.

Shutaywi, M. and Kachouie, N.N. (2021). Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*. 23(759), 0-17. https://doi.org/10.3390/e23060759

Taboada, J., Saavedra, Á., Iglesias, C., and Giráldez, E. (2013). Estimating quartz reserves using compositional kriging. *Abstract and Applied Analysis*. 2013, 1–6.

Tan, P., Steinbach, M., Karpatne, A., and Kumar, V. (2021). *Introduction to data mining (second edition)*. Pearson.

Taylor, R.D. and Anderson, E.D. (2018). Quartz-pebble-conglomerate gold deposits: Scientific Investigations Report 2010–5070–P. *U.S. Geological Survey*. 1-34. DOI: 10.3133/sir20105070P

Tokoglu, M. (2018). *Comparative analysis of 3D domain modelling alternatives: implications for mineral resource estimates*. Colorado School of Mines.

Tucker, R.F., Richard, P.V., and Viljoen, M.J. (2016). A review of the Witwatersrand Basin - The world's greatest goldfield. *Episodes*. 39(2), 105-133. DOI: 10.18814/epiiugs/2016/v39i2/95771

Van Gosen, B.S., Verplanck, P.L., Long, K.R., Gambogi, J., and Seal, R.R. (2014). *The Rare-Earth Elements: Vital to Modern Technologies and Lifestyles*. US Geological Survey.

Viljoen, M. (2009). The life, death and revival of the central rand goldfield. *World Gold Conference 2009: The Southern African Institute of Mining and Metallurgy*. 1-8.

Wang, C., Xu, S., and Yang, J. (2021). Adaboost Algorithm in Artificial Intelligence for Optimizing the IRI Prediction Accuracy of Asphalt Concrete Pavement. *Sensors*. 21(17). 1-16. https://doi.org/10.3390/ s21175682

Wang, D. and King, I. (2010). An enhanced semi-supervised recommendation model based on green's function. *International Conference on Neural Information Processing*. 397-404.

Wellmer, F.W., Dalheimer, M., and Wagner, M. (2007). *Economic Evaluations in Exploration*. Springer Science & Business Media.

Wickham, H, and Grolemund, G. (2017). *R for Data Science: import, tidy, transform, visualise, and model data*. O'Reilly Media.

Wright, M.N. and Ziegler, A. (2017). Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1-17. http://dx.doi:10.18637/jss.v077.i01

Yang, Y., Li, Y., Deng, X., and Yan, T. (2021). Structural controls on the gold mineralization at the eastern margin of the North China Craton: Constraints from gravity and magnetic data from the Liaodong and Jiaodong Peninsulae. *Ore Geology Reviews*. 139, 104522.

Yasrebi, J., Saffari, M., Fathi, H., Karimian, N., Moazallahi, M., and Gazni, R. (2009). Evaluation and comparison of ordinary kriging and inverse distance weighting methods for prediction of spatial variability of some chemical parameters. *International Journal of Biological Sciences*. 4, 93–102.

Yasrebi, A.B., Afzal, P., Wetherelt, A., Foster, P., Madani, N., and Javadi, A. (2016). Application of an inverse distance weighted anisotropic method (IDWAM) to estimate elemental distribution in Eastern

Kahang Cu-Mo porphyry deposit. *Central Iranian International Journal of Mining and Mineral Engineering*. 7, 340.

Young, D.R. (2008). The effect of ignoring the sample support on the global and local mean grade estimates, mineral resource classification and project valuation of variable width Merensky and UG2 Reef orebodies. *The Southern African Institute of Mining and Metallurgy*. 63-76.

Yunsel, T.Y. (2012). A practical application of geostatistical methods to quality and mineral reserve modelling of cement raw materials. *Journal of the Southern African Institute of Mining and Metallurgy*. 112, 239–249.

Zerzour, O., Gadri, L., Hadji, R., Mebrouk, F., and Hamed, Y. (2021). Geostatistics-Based Method for Irregular Mineral Resource Estimation, in Ouenza Iron Mine, Northeastern Algeria. *Geotechnical and Geological Engineering*. 39, 3337–3346.

Zhang, S.E., Nwaila, G.T., Tolmay, L., Frimmel, H.E., and Bourdeau, J.E. (2021a). Integration of machine learning algorithms with gompertz curves and kriging to estimate resources in gold deposit. *National Resources Research*. 30(3), 1-18. DOI: 10.1007/s11053-020-09750-z

Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Frimmel, H.E., Ghorbani, Y., and Elhabyan, R. (2021b). Application of machine-learning algorithms to the stratigraphic correlation of Archean shale units based on lithogeochemistry. Journal of Geology. 129(6), 647-672. https://doi.org/10.1086/717847

Zhang, S.E., Bourdeau, J.E., Nwaila, G.T., and Corrigan, D. (2021c). Towards a fully data-driven prospectivity mapping methodology: A case study of the Southeastern Churchill Province, Québec and Labrador. *Artificial Intelligence in Geosciences*. 2, 128-147.https://doi.org/10.1016/j.aiig.2022.02.002

8. APPENDICES

APPENDIX A: SILHOUETTE SCORE PLOTS

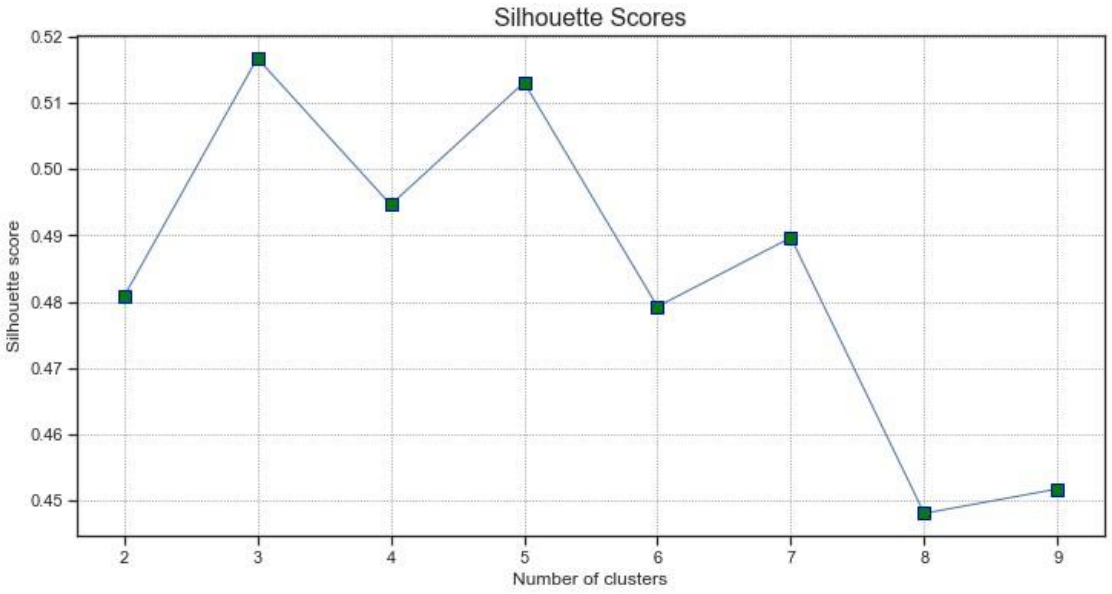


Figure A1: Graph visualising the full dataset's silhouette scores.

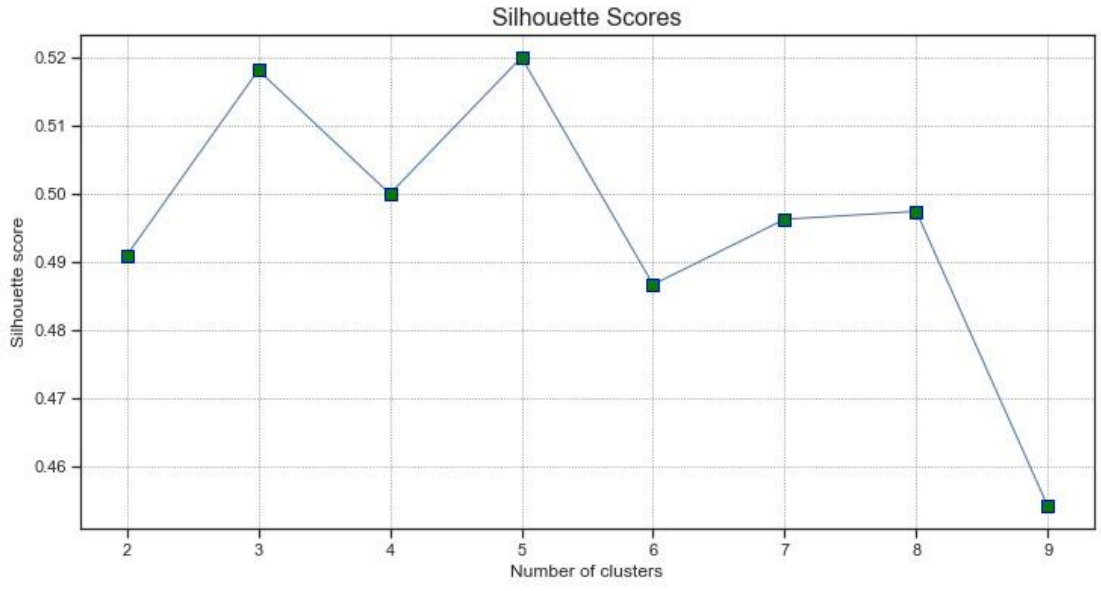


Figure A2: Graph visualising the 'bad' dataset's silhouette scores.

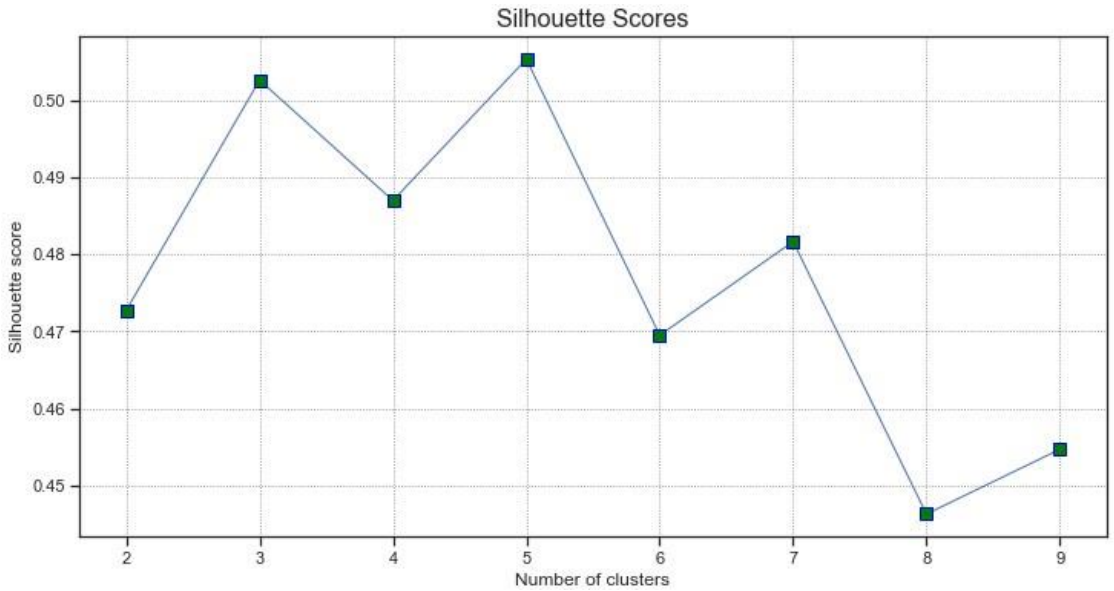


Figure A3: Graph visualising the ‘good’ dataset’s silhouette scores.

APPENDIX B: TRAINING SCORE PLOTS (EXPERIMENT 2)

Full Dataset

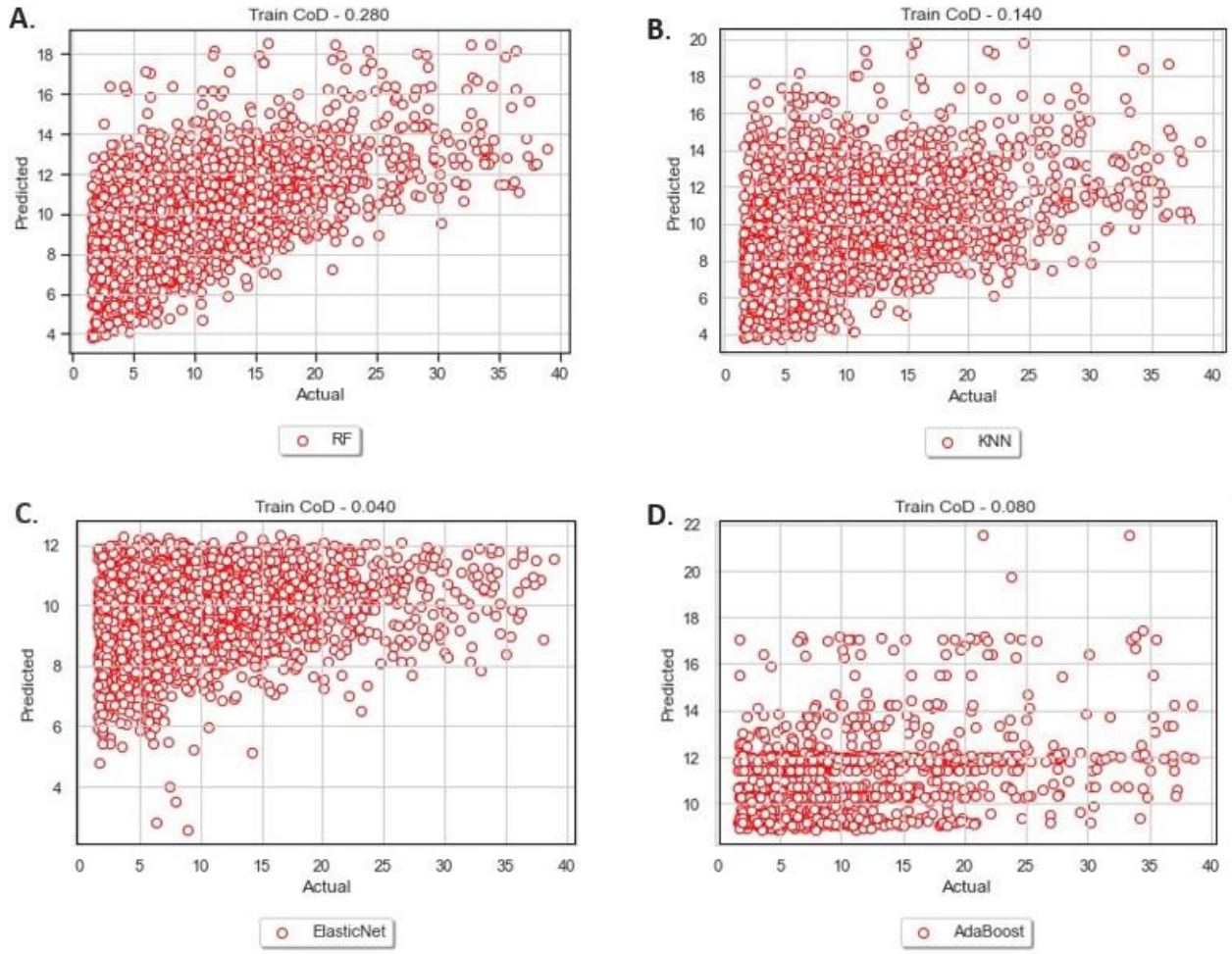


Figure B1: Graphical illustrations of the full dataset's second round of training, visualising each algorithm's CoD scores.

‘Bad’ Dataset

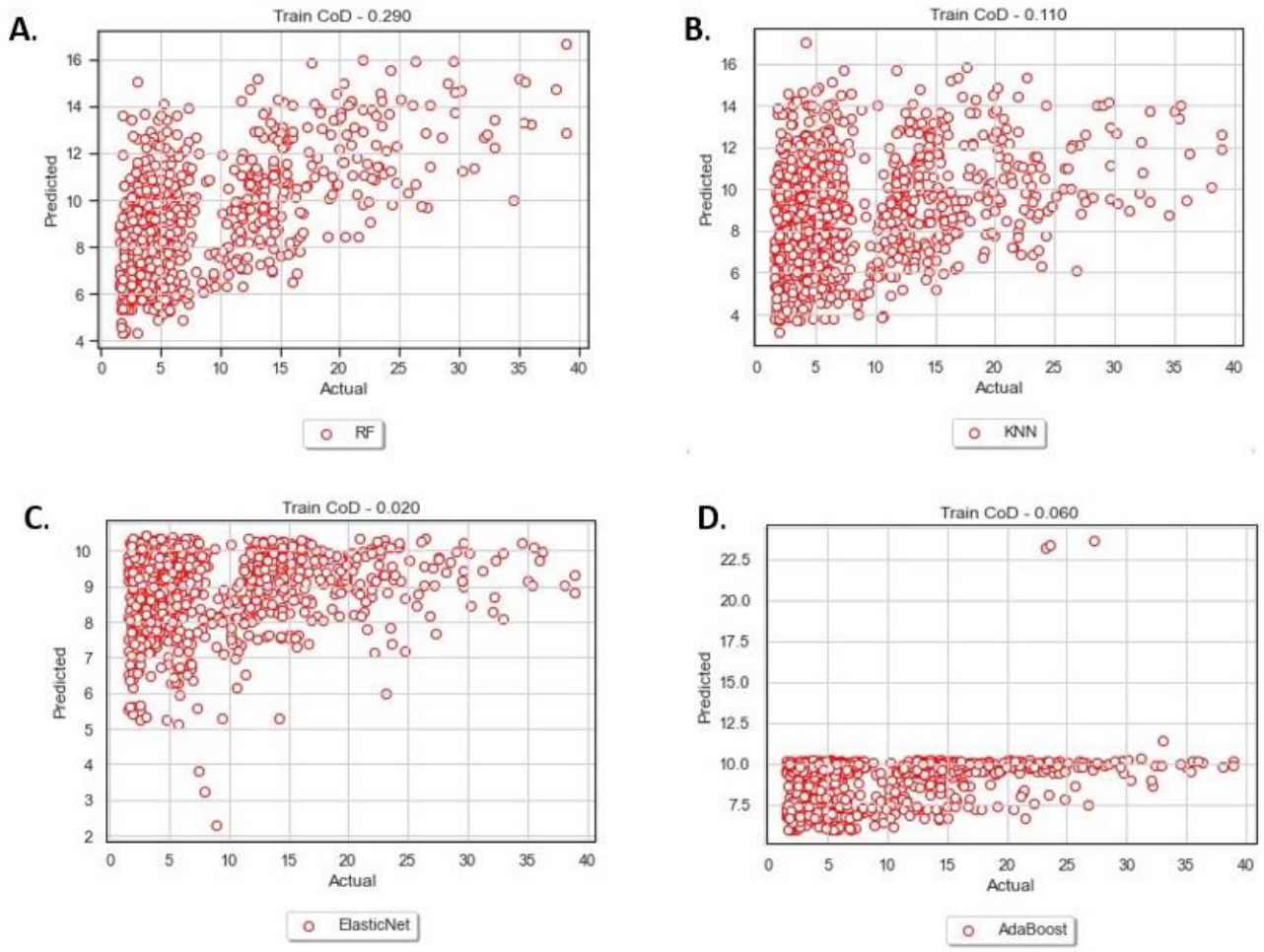


Figure B2: Graphical illustrations of the ‘bad’ dataset’s second round of training, visualising each algorithm's CoD scores.

‘Good’ Dataset

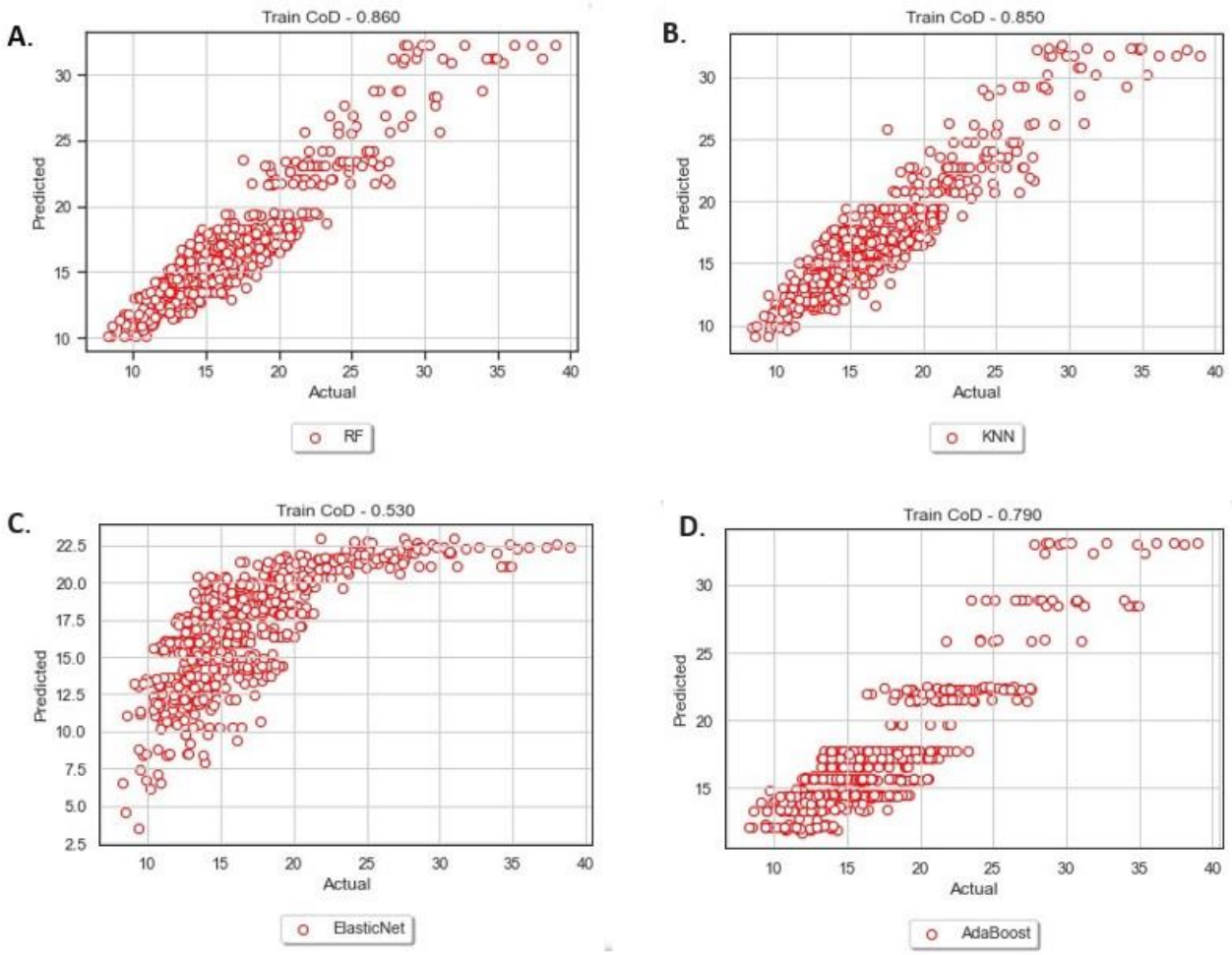


Figure B3: Graphical illustrations of the ‘good’ dataset’s second round of training, visualising each algorithm's CoD scores.