5th African International Conference on Industrial Engineering and Operations Management, Johannesburg/Pretoria, South Africa, April 23 - 25, 2024
Publisher: IEOM Society International, USA
Published: April 23, 2024 DOI: 10.46254/AF05.20240204

Exploring Machine Learning on Geochemistry Data for Estimating Metal Concentrations in Copper Deposits

Lydia Joel and Richard Maliwatu
Faculty of Computing and Informatics
Department of Informatics
Namibia University of Science and Technology (NUST)
Windhoek, Namibia
lydiajoel91@gmail.com, rmaliwatu@nust.na

Abstract

Mining companies require metal concentration analysis for ore bodies, which can be costly, time-consuming, potentially negatively impacting production due to increased turnaround times. This research's aim is to explore machine learning on geochemical data and to evaluate how models perform in predicting metal concentrations in copper deposits. The research used geochemistry dataset comprised of 3282 samples from the Kombat Copper deposit area in Namibia to predict copper (Cu) concentrations from zinc (Zn) and lead (Pb) concentrations. In addition to the metal concentrations, the dataset had sample coordinates and grid names features. The four machine learning algorithms used were Random Forest (RF), K-Nearest Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM). These models were used because they were the commonly employed models for the similar purpose from the literature reviewed. Machine learning model's performances were assessed using the regression score (R-squared), which quantifies the model's ability to explain data variance. Other metrics like Mean Squared Error, Root Means Squared Error, Mean Absolute Error, Adjusted R-squared, and explained variance were also considered. The KNN model outperformed the other three models, predicting 57% of the relationship between the dependent and independent variables. Further optimization of the models improved their prediction accuracy, with KNN model still with a superior performance of R-squared at 70% (0.70) with n-estimators set at 4 and the test size set to 10%. Predicting metal contents from geochemistry data with machine learning can help mining companies reduce costs by supplementing lab-based analyses with model-based predictions in determining grades.

Keywords
Machine Learning, Metal Concentration, Mineralization and Geochemistry data

1. Introduction

Naturally occurring ore bodies like copper (Cu) often occur in association with other useful metals such as Silver (Ag), Lead (Pb) and Zinc (Zn). One of the major problems faced by the mining industry is determining the precise metal content in ore samples, which is a crucial step in assessing the viability of a mining operation from an economic standpoint. Typically, samples are collected from the field and sent to laboratories to undergo various analyses to determine the different metal concentrations in the sample. Depending on the number of metals to be analyzed in samples, the laboratory analysis can be costly for small-scale mining operations and exploration companies. Numerous tests are carried out and in Namibia for example, the cost of laboratory analysis for metals such as copper, zinc and lead can vary depending on the lab, the method used, and the number of metals being analyzed. At a local laboratory, analysis of copper alone in a sample cost N$550 (about US$29) using the Inductively Coupled Plasma –Mass Spectrometry (ICP-MS) analysis method. Double element analysis costs N$700 (about US$37) and N$900 (about US$47) when analyzing for 4 metal concentrations. These prices are for a local Anatech Laboratory. If samples are sent to other laboratories abroad, such as in South Africa or elsewhere, the cost per metal concentration analysis is increased by additional costs such as shipping and handling fees. As an example, according to ALS Global, a global

testing and analytical laboratory, the cost of analysis for copper, zinc, lead, and nickel can range from N$530 to N$830 (about US$28-44) per metal concentration in a sample, depending on the method used and the sample matrix (ALS, 2023). Additionally, the analysis of different metals in samples can take time, which increases the turnaround time of receiving results from the laboratory and negatively affects production. Depending on the workload of the lab and the complexity of the analysis, the turnaround time for laboratory analysis can also vary, but it normally takes a few days to several weeks (Gaudino et al. 2009). For instance, a study by Bortey-Sam et al. (2018), revealed that the typical turnaround time in Ghana for ICP-MS analysis of soil samples was almost three weeks. When samples are sent from Namibia to laboratories abroad, the process is further impacted by customs clearance, shipping, and other overheads which increases the turnaround times to about 2-3 months.

There have been attempts to address the issue, as evidenced by related work by Arslan et al. (2021) and Sun et al. (2020), who demonstrated the effectiveness of Machine Learning (ML) in predicting the presence of associated metals in ore deposits, but no published work in Namibia. Deposits are site-specific in that deposits from different areas differ based on their deposition modes, grades, minerals etc. Therefore, this research aims to develop a machine learning model that utilizes geochemistry data to predict the presence of copper using the associated metals in these copper deposits based on the geochemistry data in the Kombat area, Namibia. Figure 1 shows the map of the Kombat area with nine mining sites (Otasline, Otagross, Otainsel, Oaasisk, RL, Otaschn, Otastr, RP, Otakomso) situated in Grootfontein district of the Otjozondjupa region. It is situated in Namibia's Otavi Mountain land, which is well known for its copper deposits.

By accurately predicting different metal contents from the geochemistry data, the cost of analyzing multiple metals in ore samples can be reduced. Moreover, the mining companies can discover additional metals associated with the main commodity, which can be mined, processed, and sold as by-products leading to increased profitability. The model developed can be applied to geochemistry data from other sites and yield results that can be interpreted based on those sites' specific characteristics.

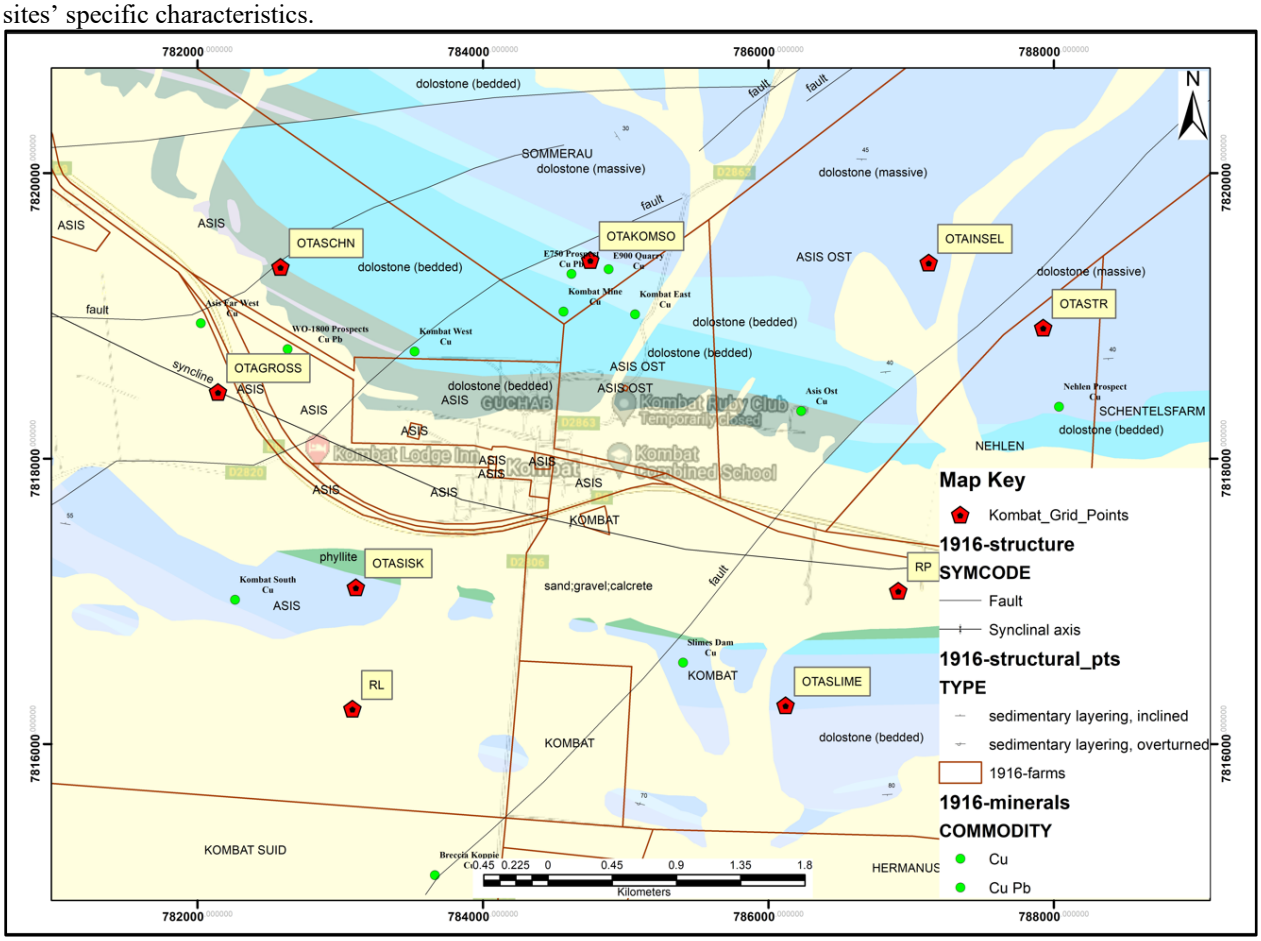

Figure 1: Area of study locality plan

1.1 Objectives

Main Objective: The study's main objective is to explore machine learning on geochemical data and to evaluate how machine learning methods perform in predicting metal concentrations in copper deposits.

Sub-Objectives

I. To explore the metal composition patterns in copper deposits.
II. To evaluate performance of the four commonly used machine learning techniques, namely Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) on geochemistry data.
III. To determine the most suitable technique based on the performance metrics and further fine tune parameters for performance improvement.

1.2 Paper Outline

Section 1 above served as an introduction to the research, covering the problem at hand as well as aims of the study. The paper's remaining sections are arranged as follows: Literature review is covered in section 2, and it covers the use of machine learning in mineralization and prediction. Section 3 describes the methodology used to accomplish the research objectives. Section 4 discusses the data collection. The outcomes and findings related to the objectives from the models are discussed in section 5 while section 6 concludes the work and discusses possible directions for future research.

2. Literature Review

This section examines studies that utilize various machine learning techniques, such as RF, SVM, DT and KNN to predict mineral deposits' metal content. This was done to better understand the problem domain and comprehend the machine learning methods that have been applied in other studies.

2.1 Mineralization in copper deposits

Mineralization refers to the process by which minerals, including metals, are deposited in rocks, creating concentrations of valuable elements such as copper. According to Santoro et al. (2018), this process can occur through various geological mechanisms, including hydrothermal activity, magmatic processes, sedimentary processes, and metamorphism. When it comes to copper deposits, mineralization typically involves the concentration of copper-bearing minerals within certain rock formations (Santoro et al. 2018). In many cases, a single rock may contain multiple metals with different metal concentrations. The rocks in copper deposits may contain lead, zinc, and nickel, among other metals. Different metal concentrations are unevenly distributed throughout the rock. Samples are sent to the labs for analysis to ascertain the amounts of metals in the rocks.

2.2 Geochemical data

Geochemical data include measurements of elements such as gold, copper, iron, zinc, and others that are commonly associated with different types of mineral deposits (Rollinson et al. 2020). The data may also include information about the abundance of certain minerals that are indicative of specific geological processes or mineralization events. According to Zuo et al. (2016), the geochemical data functions as the input features for the machine learning model within the framework of machine learning techniques. Numerical values that represent concentrations of various elements and minerals are assigned to each sample. The system uses these data points to discover patterns and connections between the different types of mineral deposits and the geochemical properties (Zuo et al.2016).

2.3 Laboratory analysis

Laboratory analysis of rocks is essential for several reasons: One of the primary objectives of analyzing copper-bearing rocks is to determine the ore grade (Jowitt and McNulty 2021). Ore grade refers to the concentration of copper or other metals within rock. Understanding the metal concentration helps in assessing the economic viability of mining and processing the deposit. Secondly, laboratory analysis provides valuable information about the mineralogical and chemical composition of the rocks. This data is crucial for conducting metallurgical studies and optimizing the extraction processes to achieve maximum metal recovery. Fourthly, the accurate laboratory analysis helps in estimating the overall metal concentrations present in a deposit. This information is vital for mine planning and long-term resource management. However, sending samples for laboratory analysis has its own challenges.

2.4 Machine Learning in metal concentration prediction

It is crucial to first determine the presence of the metal in the rocks before estimating the concentration of metals in rocks or ore samples. As previously stated, applying machine learning methods can assist in ascertaining the existence of related metals. Once the metal's existence has been confirmed, more research can be done to use machine learning to forecast the metal's concentration and related metals. Machine learning was defined by Antoine and Miranda (2017) as a technique that can recognize patterns and trends in datasets and then extrapolate predictions from those trends. SVM and RF are two popular machine learning techniques that have been utilized to forecast mineralization (Dumakor-Dupley and Ayra 2021).

Wenau et al. (2015), highlights that the ability of ML to continually improve the outcomes with the increase of input data into the system and not being limited by the mathematical calculations is one of the most recognized features of the tool. In addition, Cate et al. (2017), mentioned that the main attracting characteristic of Machine Learning for metal concentration prediction, is that ML requires minimal data pre-processing, secondly it can work with non-linear datasets, thirdly the approach is cheaper and faster.

2.5 Commonly used machine learning techniques for metal concentration prediction

The goal of Adebayo et al. (2019) study was to forecast the existence of related metals with copper in ore resources by means of RF algorithm. Their prediction model was fed with geochemical data that was extracted from the deposit. The study's findings proved that the random forest algorithm is capable of reliably and very accurately forecasting the presence of related metals, which offers important insight into the possible existence of different metal resources in the studied area.

Liu et al. (2019) looked at applying the neural network technique to forecast the existence of related metals with copper in ore resources in a different study. They used geochemical data gathered from the deposit as the input for their prediction model, just like in the earlier work. The neural network method showed exceptional efficacy in accurately predicting the existence of linked metals, offering a dependable indicator of the presence or absence of metals in the ore deposit. Sheng et al. (2015) employed RF with 100% prediction accuracy on iron ore samples. The study utlized silicon and tin emission spectral lines as input data and the ore class as the output data.

In a different study, Zaki et al. (2022) compared five machine learning algorithms, namely Gaussian Process Regression (GPR), Support Vector Regression (SVR), Decision Tree Ensemble (DTE), Fully Connected Neural Network (FCNN), and KNN, to predict highly askew gold data in a vein deposit. According to the ranking, krigging techniques are significantly outperformed by the GPR with logarithmic regularization as the most effective technique for predicting grades. The link between the independent and dependent variables in both procedures is indicated by the statistical parameter values of R-squared, which were found to be 0.4571 and 0.6889, respectively. The R-squared score changed to 0.8987 after fuzzy logic and neural networks were joined to form an adaptive neuro-fuzzy inference system. When testing data originating from a mixed or complicated distribution, this approach should result in a notable improvement (Zaki et al., 2022).

2.6 Models and Parameters

2.6.1 K-NN Model

K-Nearest Neighbors has several variants and extensions that address specific challenges or adapt the algorithm for different scenarios. These variants address different challenges and trade-offs associated with the original K-NN algorithm, making them suitable for specific use cases and types of datasets (Cunningham and Delany 2021). The dimensionality of the data, the quantity of the dataset, and the available computer power all play a role in the variant selection process. Cunningham and Delany (2021) reported several noteworthy variants, including Weighted K-NN, Radius Neighbors Classifier/Regressor, K-Dimensional Trees and Brute-force K-NN. The K-NN regression approach, which predicts the target variable for a new data point by averaging the target values of the five nearest neighbors in space, was applied in this work in its basic version. After this parameter was adjusted even more, the four closest neighbors produced an improved prediction accuracy of 70%.

In KNN, the value of K is an important parameter that greatly affects the algorithm's performance. When predicting a new data point, K represents number of closest neighbors considered (Bansal et al. 2022). According to Bansal et al (2022), a small K value (K=1, for example) means the model will be susceptible to data noise and outliers. Overfitting will result from this, when the model underperforms on fresh, untried data because it catches the noise in the training set. A big K value (e.g., K=10 or more) means the model smoothens over local patterns in the data yet becomes more resilient to noise. Too basic model to accurately represent the underlying structure of datasets, may result in underfitting.

The cross-validation technique is employed, according to Bansal et al. (2022), to ascertain the ideal value of K. By training and assessing the model with several values of K to determine which one works best on unseen data, the optimal K value is obtained through experimentation and validation. In this investigation, a K value of five was

utilized. After further fine-tuning this parameter, the K value of the four closest neighbours produced a prediction accuracy of 70%.

2.6.2 Random Forest

To decrease overfitting and increase overall accuracy, the Random Forest ensemble learning method constructs several decision trees and combines their predictions (Boateng et al. 2020). Here are some key parameters that users typically tune according to Boateng et al. (2020): n-estimaters, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap, random state, criterion, and n_jobs.

A value of 100 trees was selected for the n_estimators in this research project. Up to a certain point, performance can be enhanced by increasing number of trees. Starting point of 100 is frequently used because it offers a fair balance between computational efficiency and model accuracy. To guarantee reproducibility, a random state value of 42 was also employed. The same outcomes would be produced if the model were to be run several times using the same dataset and settings. This helps with debugging, code sharing, and maintaining consistency between model evaluations.

2.6.3 Decision Tree

Over time, numerous decision tree algorithms have been developed, each with unique advantages and disadvantages (Somvanshi et al. 2016). The ways in which these algorithms handle numerical and categorical features, partition data, and construct trees vary. The features of the dataset, the kind of problem (classification or regression), and the particulars of the current task all influence the choice of algorithm. Some of the primary decision tree algorithms described by Somvanshi et al. (2016) are Iterative Dichotomiser 3 (ID3), Classification and Regression Trees (CART), Multivariate Adaptive Regression Splines, Gradient Boosting Machines (GBM) and Decision Stump.

2.6.4 SVM

SVM is a flexible technique that may be used for a range of tasks, including regression, anomaly detection, and both linear and non-linear classification (Tanveer M, Rajani T, Rastogi R, Shao Y H, Ganaie M A, 2022). The type of data, the existence of outliers, and the selection of suitable kernel functions are some of the variables that affect the task selection and SVM's efficacy (Navada A., Ansari A., Patil S. & Sonkambe B. A, 2011). The following are some uses for SVM, as described by Navada A., et al. (2011): SVM for Linear Classification, SVM for Non-linear Classification, SVM for Regression, SVM for Anomaly Detection and SVM for Multiclass Classification.

3. Methods

This research's methodology draws from Saunders et al. (2017)'s research onion model. This research used a quantitative method research design, which involved numerical data on metal concentrations. The approach used in this study blended case study with design science research. Design science research strategy, according to Saunders et al. (2009), enhances research by assessing elements like models that address operations' problems. Case Study strategy was used along in this research because it is an in – depth inquiry used along design research to establish rich knowledge about an aspect (Saunders et al. 2019). A case –study was used to refine work and use data of one area to understand and get more in-depth knowledge on the data for the area -Kombat Area, Namibia. Due to the nature of the problem domain, the study employed the cross-sectional time horizon, where data is collected at one point in time.

3.1 Data Preparation

To further prepare the data for the models, the following steps were undertaken. Removing the rows with missing data was considered more appropriate because imputation was going to give unrealistic metal concentrations and affect the prediction results. Furthermore, the missing values were distributed across a smaller portion of the dataset, and it was not going to lead to significant data loss. Furthermore, an analysis on feature importance was done to determine most important features for predicting Cu metal concentration. Feature weighting was done and served as criteria for

determining irrelevant features not required for the predictions. Features below the weighting of 0.20 were considered non-important and the dataset was vertically scaled by dropping these features.
Performance of four ML methods (RF, KNN, DT and SVM) shown to function well and frequently employed in related work, were assessed on the task of metal concentration prediction and were evaluated using six indicators, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared (Coefficient of Determination), Mean Absolute Error (MAE), Adjusted R-Squared, and Explained Variance Score.

4. Data Collection

The analytical/geochemical data of the area of study were obtained from Earth Data Namibia (EDN) database, which is a comprehensive database of geological data, including mineral deposits, exploration and mining licenses, drilling data, geochemistry, maps, and reports. To store and manage this factual, geometrical, and unstructured information the database uses ORACLE and ARCVIEW as platforms. The principal source of geological data for Namibia is the Ministry of Mines and Energy's Geological Survey of Namibia (GSN), which is also in charge of maintaining this database. Professionals at GSN can use this secure server-hosted database, which provides data to interested clients at no cost to students and at a nominal fee to non-student researchers. The database is up to date (new data is added on a regular basis), accurate (gathered by trained geologists and subjected to stringent quality control procedures). This study focuses on 3282 available samples comprising Cu, Pb and Zn from the Kombat region. Table 1 below lists the eight fields from the dataset.

Table 1: Fields in the dataset.

Field	**Description**	**Data Type**	**Example in the dataset (Row 1)**	**Unit**
Sample Number	A unique identifier for the sample.	Numerical	1	-
Northing	northing geographic coordinate in UTM.	Numerical	786553.1	-
Southing	southing geographic coordinate in UTM.	Numerical	7816087.5	-
Grid Name	The location where the sample was collected.	Categorical	OTASLIME	-
Type	Type of the sample	Categorical	soil	-
Zn	Concentration of zinc in the sample	Numerical	174	ppm
Cu	Concentration of copper in the sample	Numerical	39	ppm
Pb	Concentration of lead in the sample	Numerical	89.0	ppm

5. Results and Discussion

5.1 Numerical Results

Table 2 below summarizes the ML model performances on predicting Copper metal contents from Zinc and Lead metal contents in samples. A higher number in the R-squared score is desired, as it quantifies the percentage of variance in the target variable that the model explains. ML tasks tend to have different objectives and the models may perform differently in different contexts. Therefore, it is imperative to consider a variety of performance evaluation metrics, including explained variance score, RMSE, MAE, adjusted R-squared, and explained variance.
The KNN model scores the highest R-squared score of 0.57, meaning that the features account for around 0.57% of variance in the target variable. Furthermore, the KNN model exhibits the lowest RMSE (104.24) with a moderate MAE (32.83), indicating that, on average, the predictions have less errors. Additionally, the KNN model shows greater explained variance and modified R-squared scores, demonstrating improved model fitting and the capacity to account for variance in the target variable. Based on these metrics overall, the KNN model outperforms the other three models in predicting Cu concentration in ore samples.

Table 2: Summary of ML model performance.

Metrics	**SVM Model**	**KNN Model**	**DT Model**	**RF Model**
Mean Squared Error:	13922.84	10866.86	12905.74	11428.92
R-squared Score:	0.45	0.57	0.49	0.55

Root Mean Squared Error (RMSE):	118.00	104.24	113.60	106.91
Mean Absolute Error (MAE):	33.69	32.83	38.36	32.20
Adjusted R-squared:	0.45	0.57	0.49	0.55
Explained Variance Score:	0.46	0.57	0.49	0.55

5.2 Graphical Results

5.2.1 Trends in the dataset

Figure 2 shows the distribution of Cu concentrations across the nine mining sites. The box plot below shows that high Cu concentrations are more common at Otasline, Otagross and Otainsel, while Lower Cu concentrations are more prevalent at Oaasisk, RL, Otaschn, Otastr and RP.

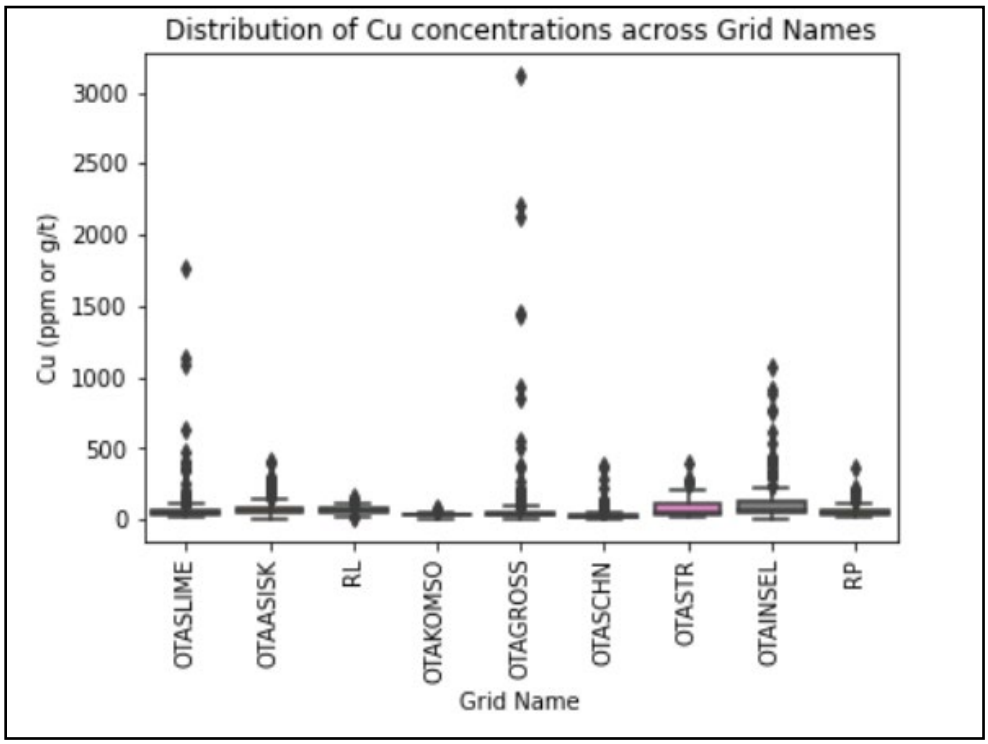

Figure 2: Box plot of the dataset showing distribution of metal across the area of study.

A scatter plot in Figure 3 below was plotted to visualize the relationship between Cu, Zn, and Pb concentrations and help observe any potential correlations or patterns between these variables. The plot shows that most of the concentrations are below 500ppm. The scatter plot further indicates that for high Copper (Cu) concentrations the Lead (Pb) concentrations are in similar ranges whereas the Zinc (Zn) concentrations are much lower.

Additional analyses were conducted to understand the individual range and distributions of Cu, Zn, and Pb concentrations. This aided in determining the data's central tendency, dispersion and skewness, and see the frequency or count of data points that fall into specific intervals. The histogram (not shown in this section) showed that within the limit of 500 ppm, most Pb and Cu concentrations are in the lower ranges whereas some of the Zn concentrations are close to 500ppm.

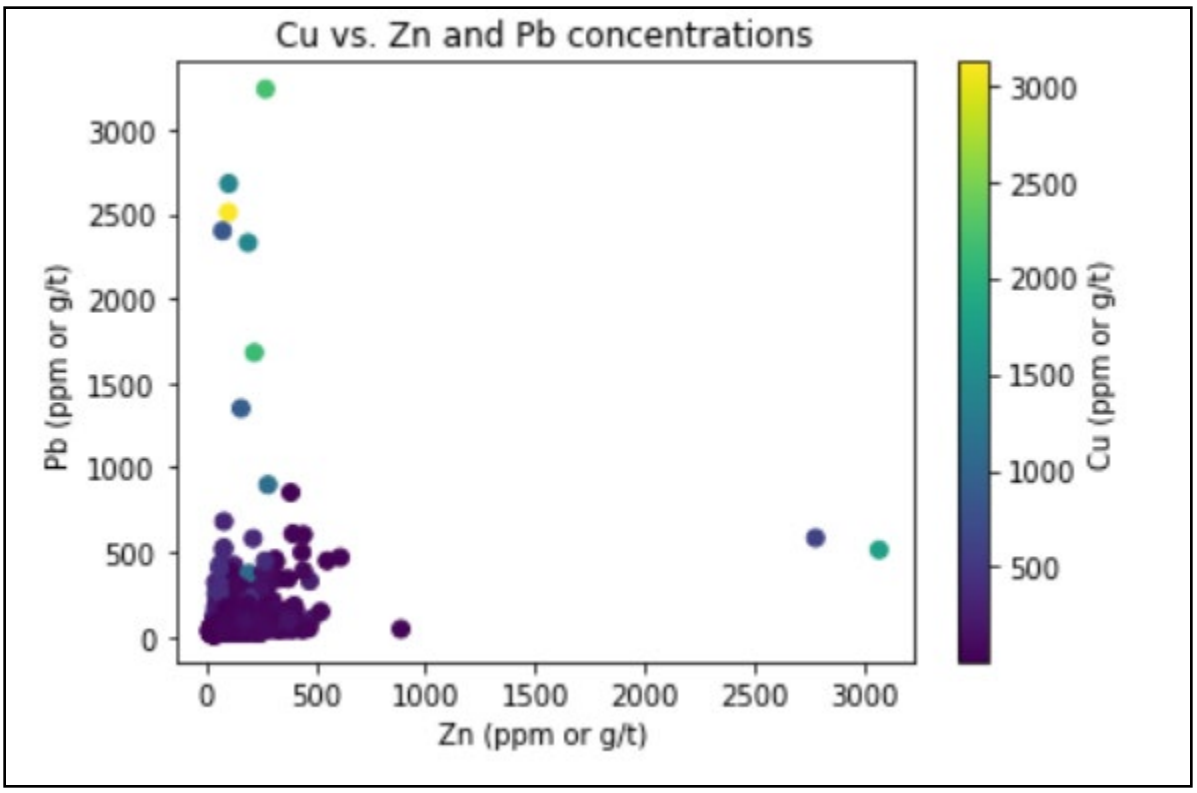

Figure 3: Relationship between Cu, Zn and Pb concentrations.

A heatmap illustrating the pairwise correlations between Cu, Zn, and Pb concentrations is displayed in Figure 4 below. This type of analysis aids in determining the direction and intensity of the interactions between the variables. The results show that there is a stronger correlation of 0.76 between Cu and Pb than that of 0.28 between Cu and Zn. The graph further shows that Zn is more correlated to Cu than it is to Pb as the correlation coefficients of 0.28 and 0.2 show. On the other hand, Pb is more strongly correlated to Cu than it is to Zn as the correlation coefficients of 0.76 and 0.22 show.

Furthermore, distributions of the variables were studied using density plots (not shown in this section). Density plots are used to show the probability densities of the variables and can be used to uncover errors in the data. Right-skewed (aka positively skewed) density plots were observed for the concentrations of all three metals, which confirmed that the majority of values were to the left side i.e. less than 500ppm. There are some samples with metal concentrations exceeding 500ppm, but only very few. The type of copper ore for the samples collected are predominantly sulfide-based, with occurrences of oxide ores as well. Determining the metal concentration is a key step in affirming the metal ore grade, which is critical in assessing the economic viability of the planned mining venture. For instance, studies have shown that the unitary energy cost of extracting metals such as copper increases as the metal concentration decreases (Fizaine & Court, 2015). Existing mining operations can similarly use information regarding predicted metal concentrations to examine the effectiveness of the mineral extraction processes that are currently in place by comparing the estimated metal output with the actual outputs.

The geological setting, mineralization type, mineral associations, geochemical processes, and sampling characteristics can all contribute to the observed correlations between different metal concentrations in the samples. The dataset is from volcanic-hosted massive sulfide deposits which contain minerals that are rich in both lead and copper. This means in these ore formations, lead and copper minerals occur together in significant quantities due to similar geological processes. In addition, the strong correlation between Pb and Cu can also be due to mineral association. Lead and copper minerals often occur together in association with each other within the same mineral assemblages or ore veins. This association can lead to a strong correlation between their concentrations in samples. These minerals might share similar geochemical behaviors, such as solubility, transport, and precipitation processes, which can result in their co-occurrence. It is also possible that the observed correlation between Pb and Cu concentrations compared to Zn and Cu concentrations could be influenced by sampling bias or the specific characteristics of the samples collected.

The samples were collected from areas with known Pb-Cu mineralization or geological settings favoring the formation of Pb-Cu ores, this could skew the correlation analysis.

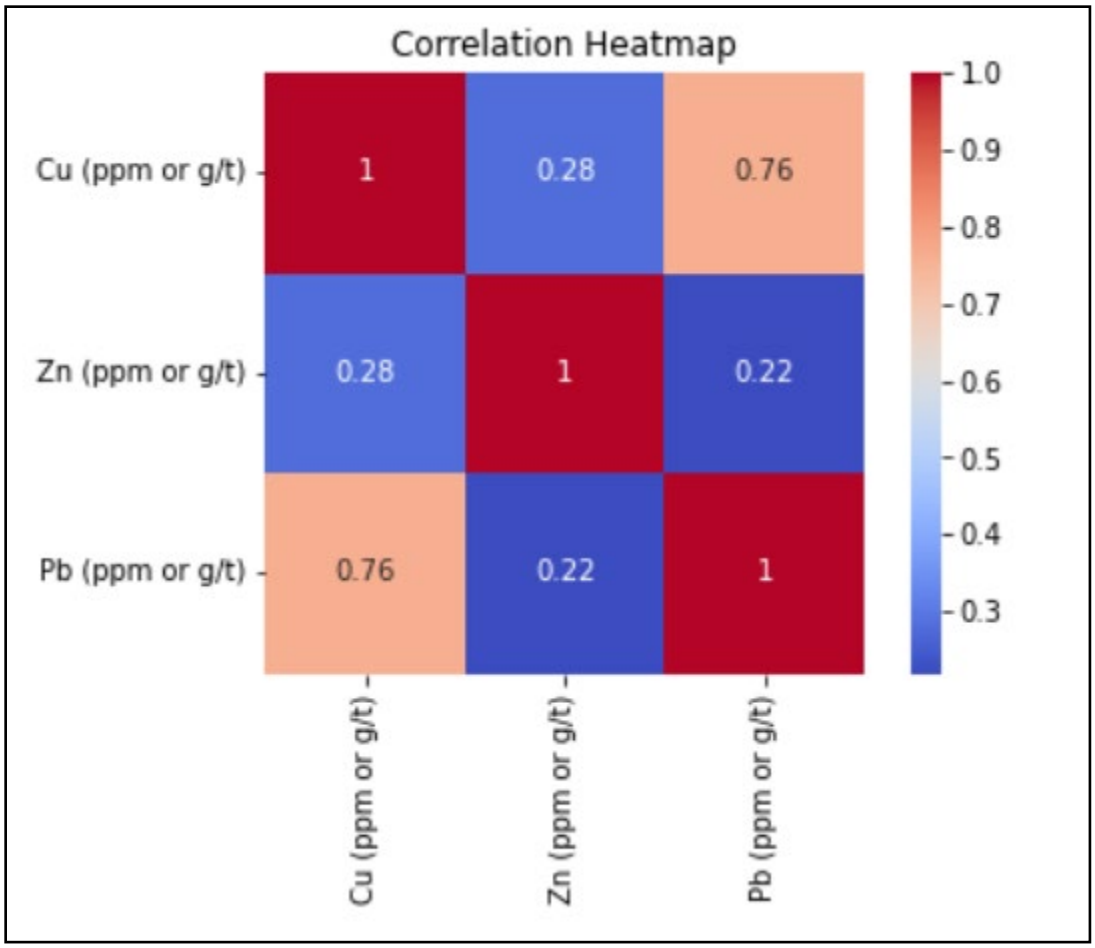

Figure 4: A heat map of pairwise Seaborn's correlations between Cu, Zn and Pb concentrations of the dataset.

5.2.2 Predicted versus actual values of the best performing model K-NN

A comparison of the actual Cu concentrations and the KNN predicted values is presented in Figure 5. The results demonstrate the ML model's ability to predict metal concentrations with marginal error.

Perfect prediction is marked by the red line in Figure 6 and indicates where the dots would fall if all the predictions and actual values were exactly equal. The graph compares the output to the perfect prediction line/target to show how the projected values closely match the actual values. The result shows that more than 90% of points are dispersed tightly near the ideal prediction line, indicating that the actual values and predictions correspond marginally. However, there are a few outlier points that are widely dispersed, indicating a greater variation between predicted and actual values as seen in Figure 6 for samples 6, 21, 25 and 29. Despite scaling done, there could still be issues with the model's sensitivity to feature magnitudes between Cu, Pb, and Zn concentrations, contributing to discrepancies between actual and predicted values for these samples.

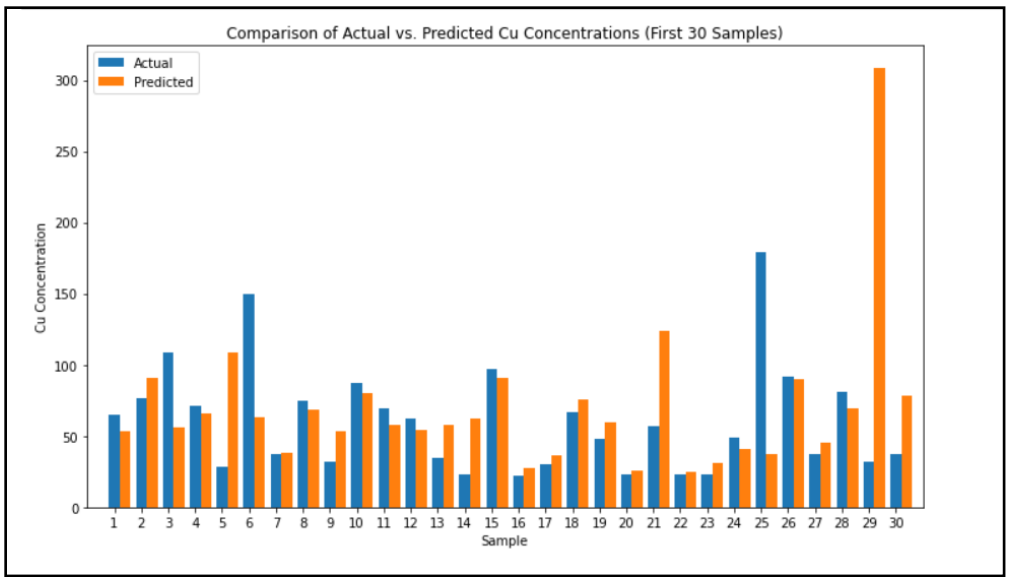

Figure 5: Comparison of actual Cu concentration and the KNN model predicted value.

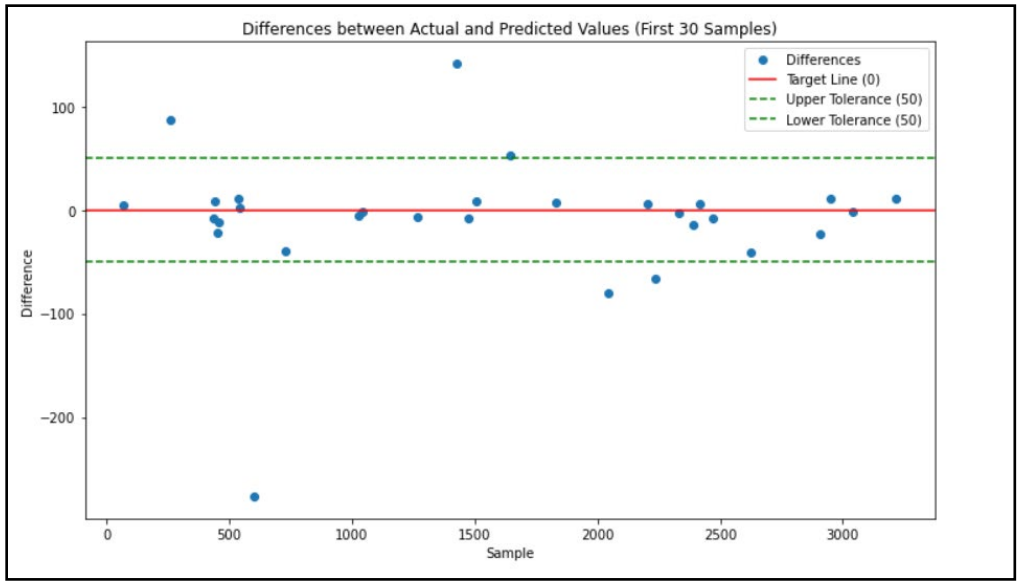

Figure 6: Variations between the expected and actual copper values.

5.3 Proposed Improvements

The KNN model performed comparatively well in terms of prediction accuracy. Tweaking the testing size and the nearest neighbor estimator, from 0.2 (20%) and 5 respectively to 0.1 (10%) and 4 yielded a R-squared score of 0.70 (70%). On the testing size of 10% other models' prediction accuracies have also improved. New R^2 values were 0.54 (54%), 0.51 (51%) and 0.50 (68%) for Decision Tree, SVM and RF respectively. Additional hyperparameter tuning can be applied for further performance improvement. There is a need to draw up specific industry standards on acceptable ML model performance margins in different application domains. The superior performance of KNN than other models in this study could be attributed to its ability to capture complex relationships in the data, its robustness to outliers which are in the dataset and its suitability for small datasets. Furthermore, the dataset exhibits clear patterns of where concentrations of Cu, Pb, and Zn tend to cluster together. For these types of datasets, KNN can effectively capture these patterns without making strong assumptions about the underlying data distribution.

6. Conclusion

The first sub-objective was to explore the metal composition patterns in copper deposits. The observation is that 99% of the Zn, Cu, and Pb metal concentrations are below 500ppm. Correlations between metal concentrations were observed among the metals in ore samples. The scatter plot further shows that for high Cu concentrations the Pb concentrations are proportionately high whereas the Zn concentrations are much lower. Furthermore, analysis shows that higher Cu concentrations are more common at Otasline, Otagross and Otainsel while the lower Cu concentrations were more common at Oaasisk, RL, Otaschn, Otastr and RP locations.

The second sub-objective was to evaluate performance of four leading ML techniques (RF, DT, KNN and SVM) in predicting metal concentrations on the geochemistry data. Six performance evaluation metrics were considered, and overall KNN outperformed the other three models. Adjusting KNN's testing size and nearest neighbour estimator improved the R-squared score from 57% to 70%.

This study demonstrates the potential of machine learning to predict metal concentrations given some geochemical data, which could save exploration costs, boost productivity, and enhance metal concentration estimation accuracy. The R^2 value of 0.70 is promising but is not sufficient for an ML model to be deployed for the metal prediction task. Other variables like the textures, colour, lithologies of samples could be captured to improve the predictive power of the model. Using grid search techniques, gamma, kernel type, regularization parameters could influence the model's performance. For future work, further hyperparameter tuning could be done to improve the models. In addition, assessing the effectiveness of other machine learning methods such as Artificial Neural Networks can also be carried out. Furthermore, having seen the promising results of ML on geochemical data, investigating the integration of ML-based ore grade estimates into automated mineral extraction process optimization would be the next step.

References

ALS, G., Geochemistry. Available: https://www..alsglobal.com/en/services-and-products/geochemistry/, May 2023.

Antoine, A. and Eduardo R., Musical Acoustics, Timbre, and Computer–Aided Orchestration Challenges, *Proceedings of the 2017 International Symposium on Musical Acoustics (ISMA)*, pp. 151-154, Montreal, Canada, June 18-22, 2017.

Arslan, H., Aslan, N., & Demirci, S., Mineral prospectivity mapping using machine learning techniques in the Murgul area, northeastern Turkey. pp.557-577, 2021.

Bansal, M., Goyal, A. and Choudhary, A., A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, vol.*3*, p.100071, 2022.

Boateng, E.Y., Otoo, J. and Abaye, D.A., 2020. Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. *Journal of Data Analysis and Information Processing*, vol.*8*, no.4, pp.341-35, 2020.

Bortey-Sam, N., Nakayama, S. M., Ikenaya, Y., Akoto, O., Baidoo, E. and Mizukawa, H., Human health risk assessment of heavy metals in soil and food crops from farms in Ghana receiving irrigation with polluted urban wastewater, 2018.

Caté, A., Perozzi, L., Gloaguen, E. and Blouin, M., Machine learning as a tool for geologists. *The Leading Edge*, vol. 36, no.3, pp.215-219, 2017.

Cunningham, P. and Delany, S.J., k-Nearest neighbour classifiers-A Tutorial. *ACM computing surveys (CSUR)*, vol.54 no.6, pp.1-25, 2021.

Dumakor-Dupey, N.K. and Arya, S., Machine Learning—A Review of Applications in Mineral Resource Estimation. *Energies*, vol.14, no.14, p.4079, 2021.

Gaudino, S., Galas, C., Belli, M., Barbizzi, S., de Zorzi, P., Jaćimović, R., Jeran, Z., Pati, A. and Sansone, U., 2007. The role of different soil sample digestion methods on trace elements analysis: a comparison of ICP-MS and INAA measurement results. *Accreditation and quality assurance*, 12, pp.84-93.

Jowitt, S.M. and McNulty, B.A., Geology and mining: mineral resources and reserves: their estimation, use, and abuse. *SEG Newsletter*, vol.125, pp.27-36, 2021.

Liu, X., Ma, Y. and Ma, X. The application of a convolutional neural network to the mapping of mineral distributions using remote sensing data. *International Journal of Remote Sensing*, pp.157-173., 2019

Navada, A., Ansari, A.N., Patil, S. and Sonkamble, B.A., Overview of use of decision tree algorithms in machine learning. In *2011 IEEE control and system graduate research colloquium*, pp. 37-42, Alam, Malaysia, June 27-28, 2011.

Rollinson, H.R., Rollinson, H. and Pease, V., *Using geochemical data: to understand geological processes.* Cambridge University Press, Cambridge, United Kingdom, pp. 1-10, 2021.

Santoro, L., Yav, S.T., Pirard, E., Kaniki, A., Arfè, G., Mondillo, N., Boni, M., Joachimski, M., Balassone, G., Mormone, A. and Cauceglia, A., 2018. Abstracts from the 2017–2018 Mineral Deposits Studies Group meeting. *Applied Earth Science*, vol.127, no. 2, pp.46-79, 2018.

Saunders, M., Lewis, P.H.I.L.I.P. and Thornhill, A.D.R.I.A.N., *Research methods*. Business Students 4th Edition, Pearson Education Limited, England, pp.1-268, 2017.

Somvanshi, M., Chavan, P., Tambade, S. and Shinde, S.V., A review of machine learning techniques using decision tree and support vector machine. In *2016 International conference on computing communication control and automation (ICCUBEA)*, IEEE, pp. 1-7, Pune, India, August 12-13, 2016.

Sun, Q.F., Wang, Y.C., Wang, K.Y., Sun, F.Y., Lai, C.K., Zhao, C.G. and Sun, L.X., Multistage metallogeny and tectonic evolution in eastern NE China and adjacent Russian Far East: geochronology, geochemistry, and Sr-Nd-Hf isotope perspectives. *International Geology Review*, vol. 65, no.11, pp.1800-1831, 2023.

Tanveer, M., Rajani, T., Rastogi, R., Shao, Y.H. and Ganaie, M.A., Comprehensive review on twin support vector machines. *Annals of Operations Research*, pp.1-46, 2022.

Wenau, S., Spiess, V., Pape, T. and Fekete, N., Cold seeps at the salt front in the Lower Congo Basin II: The impact of spatial and temporal evolution of salt-tectonics on hydrocarbon seepage. *Marine and Petroleum Geology*, vol. 67, pp.880-893, 2015.

Zaki, M.M., Chen, S., Zhang, J., Feng, F., Khoreshok, A.A., Mahdy, M.A. and Salim, K.M., 2022. A Novel Approach for Resource Estimation of Highly Skewed Gold Using Machine Learning Algorithms. *Minerals*, vol.12, no.7, p.12, 2022.

Zuo, R., Carranza, E.J.M. and Wang, J., Spatial analysis and visualization of exploration geochemical data. *Earth-Science Reviews*, vol.158, pp.9-18, 2016.

Biographies

Lydia Joel is a production geologist with 10 years' experience. She holds an Honours of Geology and Master of Business Administration from the University of Namibia. She is currently studying towards her Master of Data Science at Namibia University of Science and Technology.

Richard Maliwatu is a lecturer at Namibia University of Science and Technology in the Faculty of Computing and Informatics. He holds a Master's degree in Computer Science and a PhD from the University of Cape Town. His research interests span data science and analytics in a range of applications, including mining, tourism, governance, energy, financial transaction anomaly detection and computer networks.