Predicting Deep Mineral Deposits based on Surface level soil samples

CMPE 251 Final Project Report

Created By: Frank Siyung Cho

Student #: 20277846

Table of Contents

Introduction

Mineral deposits are one of the earth's most valuable resources being made up of several major raw materials that can be extracted by humans for industrial and non-industrial purposes (*Mineral Deposits*). Due to their properties and characteristics, these materials and ores are extracted from the mineral deposits and are used in everyday life and are present in all technological appliances used today. Minerals such as Bauxite, Graphite, Wolframite, Tantalite, and Chalcopyrite are only 5 of the hundreds of minerals used in mobile phones such as the iPhone (*General Information Product*).

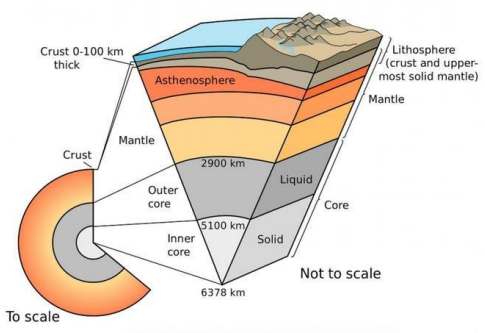
Mineral deposits themselves can be extremely difficult to find. Surface level mineral deposits can be easily extracted, however, due to their constant mining, they have become increasingly difficult to find (Ghorbani et al.). This raises an important task of finding new and innovative ways to extract mineral deposits from other regions of the world. Some possible alternative solutions are the retrieval of "mineral deposits that lie beneath the ocean and in rocks that form the continents" (*Mineral Deposit | Definition, Examples, & Facts | Britannica*). These are known as marine deposits and although potentially revolutionary, their retrieval poses a serious risk to hundreds of marine habitats and would destroy many ecosystems in deep ocean floor regions (*Deep-seaMiningFAQ*). Another alternative solution is the retrieval of deep mineral deposits. These can be abundant and rich in resources whilst also being a relatively preserved resource due to the exhaustive depletion of surface level mineral deposits and the overall difficulty in reaching deep layers of the earth's ground. Although these deep mineral deposits can be difficult to extract and find, by using focused techniques, mining companies can target certain areas which have high likelihoods of having a deep mineral deposit. One technique that heavily reduces the risk of mass land and habitat destruction is the use of geochemistry to find deeply buried deposits (Appleby). This technique utilizes the fact that deep mineral deposits often leave geochemical traces near the surface allowing for samples to be extracted and used to determine if deep mineral deposits are present in the area (Appleby). This paper looks further into this technique to find a correlation between the

geochemistry of a surface level soil sample and if there is a resulting mineral deposit. Furthermore, resulting correlations will be verified to see if an accurate prediction model or clustering algorithm can be used to predict future samples.

Background Information

Deep mineral deposits and their societal impact

As previously mentioned, due to the over mining and subsequent depletion of surface level mineral deposits, many organizations seek to dig deeper into the earth's crust in order to find more mineral deposits (Ghorbani et al.). As we can see in the figure below, the earth's crust is a small proportion (approximately 1%) of the overall volume of the earth.



However, it is still around 100km in depth which is very large when compared to the deepest mine in the world in 2012, having a depth to approximately 4km (*Deepest Mine | Guinness World Records*). Other depths have since been reached with the Kola Superdeep Borehole reaching an estimated depth of 12.2km. Whilst impressive this is still only around 10% of the earth's crust (*Why Did the Russians Seal Up the Deepest Hole in the World? | HowStuffWorks*). This demonstrates that there is still a vast space that is completely preserved and potentially contains large amounts of mineral deposits that can be extracted. This is one of the main reasons why deep underground mining and deep mineral deposits are heavily sought after. However, the viable constraints of deep mining restrict its large-scale production. In reality, there are many challenges that occur when mining at deep depths. Namely, the temperatures at around 12km deep in the earth's crust (where the Kola Superdeep Borehole reached) was approximately 250 °C (*Why Did the Russians Seal Up the Deepest Hole in the World? | HowStuffWorks*). This would only continue to rise exponentially as depths increased. Moreover, the stability of boreholes become increasingly unpredictable due to their dependency on stress, strain, drilling fluid composition and weight (*Why Did the Russians Seal Up the Deepest Hole in the World? | HowStuffWorks*). Although these constraints must be considered, the application of deep mining for mineral deposits could be viable when

used in a focused manner. The use of geochemistry allows for this as a solution to not having to deep mine at an industrial scale.

Geochemical technique and sampling

Geochemical exploration is the measurement process of chemical properties and trances through the area of a particular region with potential mineral deposits. This technique requires sampling of the nearby and inherit landscape and its elements. This includes samples of soil, bedrock, stream and lake sediments, plants/vegetation and many other constituents (*Learning Geology: Geochemical Sampling*). To be more specific, there are two main types of geochemical exploration sampling, conventional and non-conventional (Ghorbani et al.). Conventional geochemical sampling includes the sampling of lithogeochemical bedrock analysis, pedogeochemical/soil, stream sediments, hydrogeochemical, biogeochemical and homogeochemical analyses (Ghorbani et al.). This includes but is not limited to weathered bedrock, soil or transported overburden samples (Ghorbani et al.). Non-conventional sampling uses other media tools to make analyses (Ghorbani et al.). This includes vapor geochemistry, electro-chemistry and some isotope studies (Ghorbani et al.). This paper in particular analyses data that has used conventional soil sampling for their geochemical exploration.

Furthermore, the geochemical technique used in the data that will be further explored in this paper is derived from soil samples taken at different depths. These samples have been subjected to 5 different acids in order to extract the concentrations of the solute (minerals) present in the sample. The data was then recorded for several different elements; the five solvents that the samples were subjected to are ENZ, AAS, AA7, AQR and GDX.

Stakeholders

There are many stakeholders that are directly impacted by an increase in mineral deposit extraction and refinement. Firstly, many large corporations will be able to benefit economically by managing and focusing their efforts and resources to specific areas and regions with probable mineral deposits. Moreover, companies will also be able to decrease overall labor cost going towards industrial mining regions.

Another potential stakeholder that may be affected by increased deep mineral deposit extraction are the societies and individuals who are impacted by small scale mining companies that exploit impoverished people into labor and illegal mining. The table below demonstrates the 10 largest mines in the world as of 2019:

Name of Mine	Country
Mponeng Gold Mine	South Africa
TauTona Gold Mine	South Africa
Savuka Gold Mine	South Africa
Driefontein 5 shaft Gold Mine	South Africa
Kusasalethu Gold Mine	South Africa
Moab Khotsong Gold Mine	South Africa
South Deep Gold Mine	South Africa
Kidd Creek Copper and Zinc Mine	Canada
Great Noligwa Gold Mine	South Africa
Creighton Nickel Mine	Canada

As we can see in the table above, in 2019, 8 of the top 10 deepest mines in the world were in South Africa. Unfortunately, many of these mines do not adhere to human rights laws and subject their workers to severe and deathly conditions (*How Developing Countries Are Paying a High Price for the Global Mineral Boom | Global Development | The Guardian*). These operations are often funded and driven by large companies, and other countries such as Columbia, Philippines, Democratic Republic of Congo are also the target of exploitation due to their regions' vast resources and low economic state (*How Developing Countries Are Paying a High Price for the Global Mineral Boom | Global Development | The Guardian*). The increase in deep mineral deposit extraction could drive away companies from extracting minerals from 3rd world countries and aid the communities and people suffering from these abusive working conditions (Ghorbani et al.).

Dataset Properties

Dataset Attributes and Assumptions

The dataset used for this paper contains soil samples from a line and at different varying depths. The samples were subjected to 5 acidic solvents (AA5, ENZ, AA7, AQR, GDX), then the resulting concentrations of the extracted elements were measured and recorded. The dataset itself was formatted to contain only numeric data with no nominal values recorded. Certain assumptions were made as to which records in the data set were positively classified as having a mineral deposit. In particular, all records within the distance range from 186.2 to 214.6 inclusive were classified as positive and containing

a mineral deposit. All other records were identified as negative and assumed to not have a mineral deposit. A new column was appended to the dataset containing these nominal values.

Properties and Descriptive Statistics

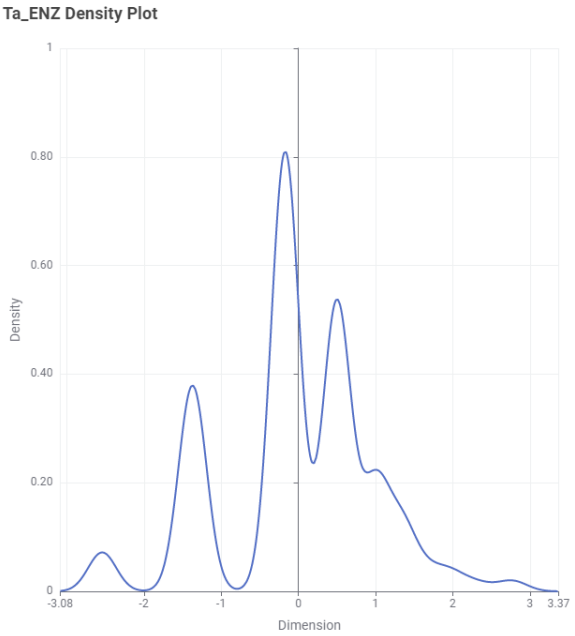
There were several properties of the dataset that were initially identified when analyzing the dataset. One property that was found was the range of the distance attribute being from 127.2 to 285.3 with the 33 different intervals of distance. The average number of samples per interval was 6 samples that varied in depth decrementing by 10 each time. To demonstrate this, a sample from the dataset is displayed in the table below:

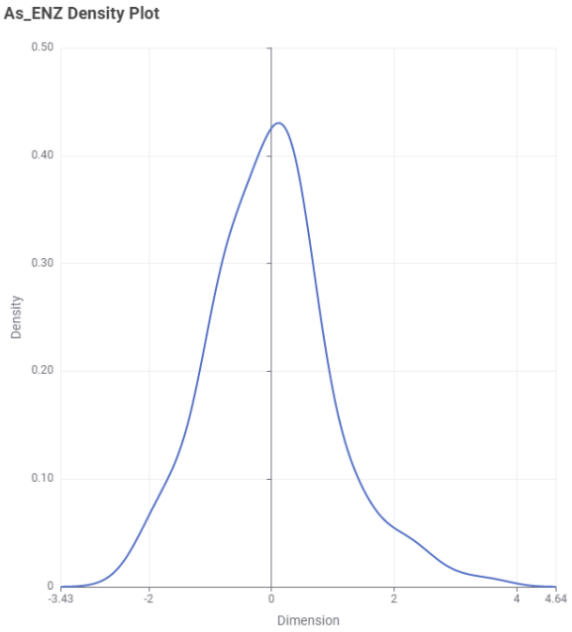
Dist	Dpth	As_ENZ	Au_ENZ	Ba_ENZ	Be_ENZ	Br_ENZ	Cd_ENZ
127.2	-5	2.09	-0.64	1.35	-0.57	0.68	1.62
127.2	-15	-1.02	0.49	-1.42	-0.57	-0.4	0.32
127.2	-25	-0.79	-0.64	-1.67	-0.57	0.53	-0.47
127.2	-35	-0.38	-0.64	-1.67	-0.57	-0.24	-1.09
127.2	-45	0.07	1.78	-0.79	-0.57	-0.13	-1.09
127.2	-55	0.56	-0.64	-0.79	-0.57	-0.82	-1.09

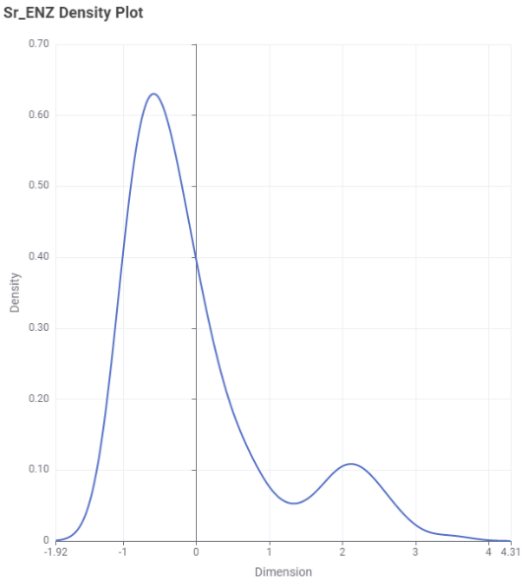
Furthermore, when analyzing the depth values, the range of depth values was from -5 to -105 in decreasing intervals of 10. However, throughout the data set it was evident that depths ranging from -5 to -65 dominated the dataset with about 93.96% of the data lying in this range, this means that only 6.04% of the data was sampled from depths less that -65. Due to this, all data points that were sampled at a depth less than -65 were removed from the dataset. Each sample from the dataset was also subjected to 5 different solvent acids (AA5, ENZ, AA7, AQR, GDX) and the resulting concentrations of different elements were recorded. The table in Appendix A demonstrates the different elements that were identified along with the unique solvents that were able to extract the element. This table was created using Python programming language and the Pandas library in order to extract the unique column names along with their solvents. From this table it was identified that 5 of the total 64 unique elements were extracted from only one solvent acid. These 5 elements were: Ag, Bi, Cl, pH and Re. It was found however, that one of these elements pH, is not an actual periodic element, but a measure of the acidity level of a concentration. Due to this, it was removed from the dataset. The resulting 4 elements that were found with single solvents were: Ag (silver), Bi (bismuth), Cl (chlorine) and Re (rhenium). Of these elements, rhenium and chlorine were found with the same solvent ‘ENZ’ and silver and bismuth were found with

the same solvent ‘AQR’. This demonstrates a potential correlation between the solvents ‘ENZ’ and ‘AQR’ in finding unique and harder to find elements from soil samples. Interestingly, silver and bismuth are both metallic elements and of the transition or post-transition metal family. This further demonstrates that the ‘AQR’ solvent may be correlated to these types of metals and could be an indicator for extracting them in further soil samples.

Another property of the dataset is that all values were already normalized using Z-score normalization. This means that all solvents and their elements had a standard deviation near 1, a mean value near 0 with the elements in the AA5 solvent having the most variation with a standard deviation below 1 and a mean below 0. This demonstrates that the attributes could all be distributed normally and would therefore be well suited for further analyses such as distribution-based clustering algorithms. However, through further analysis it was found that not all elements were normally distributed, and many had other distinct distributions, or their distribution was undefinable. Pertaining to the ‘ENZ’ solvent below is a comparison of the density distributions for the ‘Ta_ENZ’, ‘As_ENZ’ and ‘Sr_ENZ’ elements.







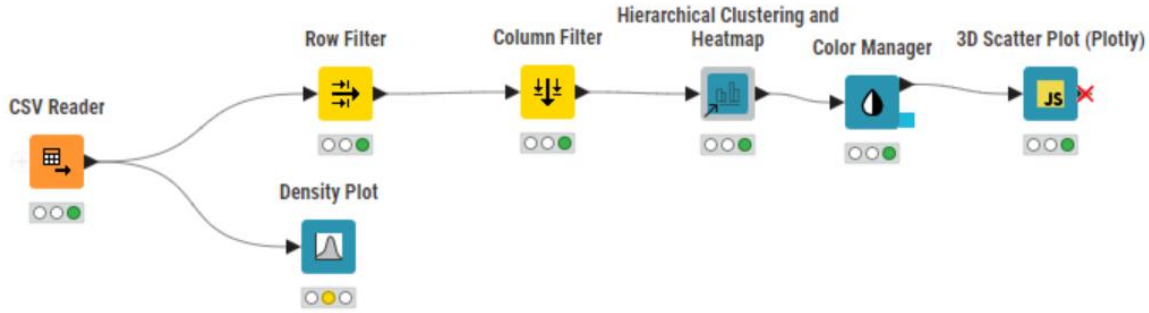
As we can see from the density distribution of the Ta_ENZ plot there is no formal understanding of its distribution as it is unclear. In contrast, the second plot of the As_ENZ distribution is a clear Gaussian distribution. Finally, the plot of the Sr_ENZ distribution is evidently a Poisson distribution. This property of the dataset demonstrates that distribution-based clustering algorithms will not be a viable solution for further analytical tasks due to its nature of assuming that all attributes are comprised of the same distribution.

Another analytical technique that has been deemed as unviable is the use of multi layer perceptrons or Neural Networks. Although neural networks could be implemented and could be used with this dataset due to the high amount of attributes available, there are too few actual records for it to be a viable option. The dataset itself contains approximately 50 positive records within a dataset of about 200 total records. This means that if a balanced dataset were to be used only about half of the total dataset would be used to train the neural network. This is not a viable execution as most neural networks are trained on thousands of records. If we were to train a neural network using this small dataset, the model would likely heavily under fit and most likely not generalize well to future data samples. The implementation of certain techniques such as oversampling could be a potential solution to aiding in the dataset's shortcoming however, it is believed that this would only marginally help.

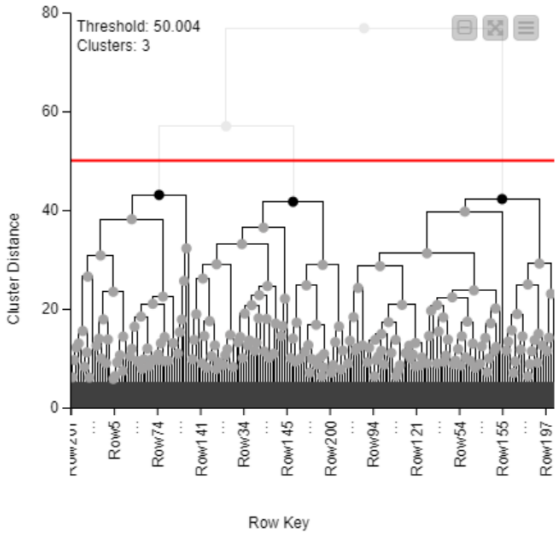
Clustering algorithm Implementation

Hierarchical Clustering Algorithm

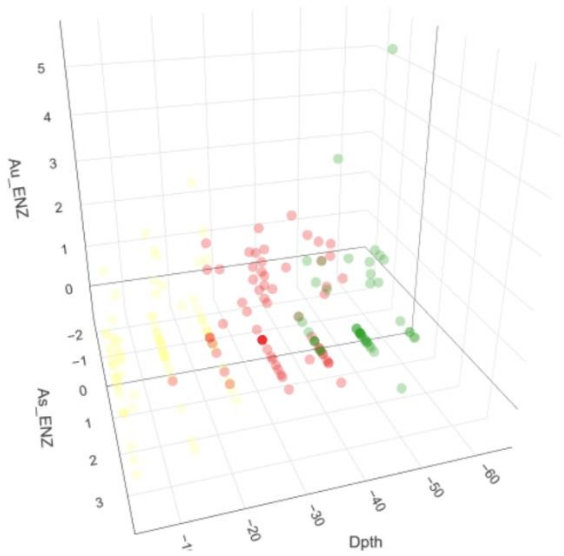
The clustering algorithm used with this dataset is the Hierarchical clustering technique. Hierarchical clustering aims to group similar records by creating a hierarchy based on distances between individual records. This technique groups clusters together using a dendrogram and can be implemented either divisively (Top-Down), where an initial cluster is created containing all records and then broken down into smaller clusters using distance measures, or agglomeratively (Bottom-Up), where all records are initially their own individual clusters and then are conjoined iteratively based on their distance. Furthermore, distance measures are used to calculate the distances between neighbouring clusters and used to determine the next hierarchy of joined clusters. In this implementation, the average linkage type along with the Euclidian distance measure was used to calculate the 'closeness' of two clusters, in order to determine which clusters should be joined. The Hierarchical clustering algorithm was implemented using the Knime software and the project workflow is demonstrated below:



Furthermore, a dendrogram was produced in order to demonstrate the different clusters as well as the threshold which would split the number of final clusters. Demonstrated below is the dendrogram produced as well as the threshold function which separates the records into 3 clusters. The range for the threshold function to split the records into 3 clusters is a distance of approximately 43.9 – 56.6. As we can see in this dendrogram, the dataset was split into relatively even partitions throughout all clusters and even though the root node is a distance of almost 80 units away from the bottom-most nodes there is consistency in the partitions created at each hierarchy level.



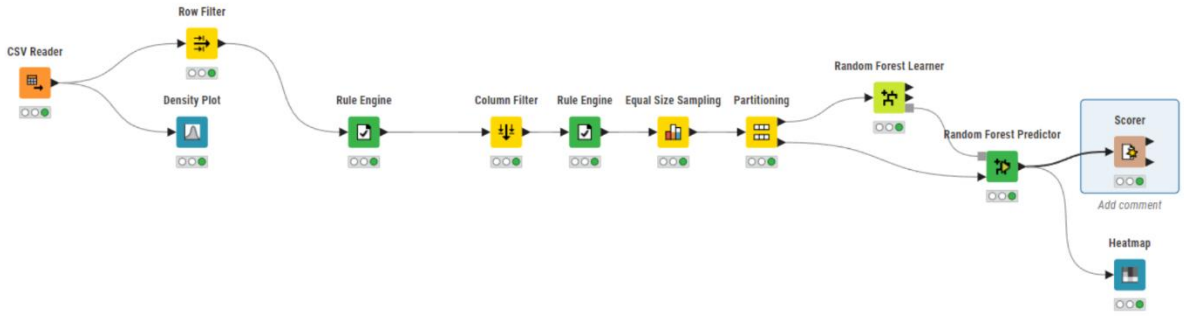
Furthermore, in the image below we can see the 3-dimensional plotting of 3 attributes: Dpth, As_ENZ and Au_ENZ. We can see that there are 3 distinct clusters formed subjecting the data to hierarchical clustering. This demonstrates that the dataset is able to be clustered. One important caveat however, is the use of the Dpth attribute in this graph. As we know the dpth attribute only contains values in intervals of -10 starting from -5 and ending at -65. This quality of the dpth variable may contribute to the separation of the clusters in the graph below.



Prediction model Implementation

Random Forest Model

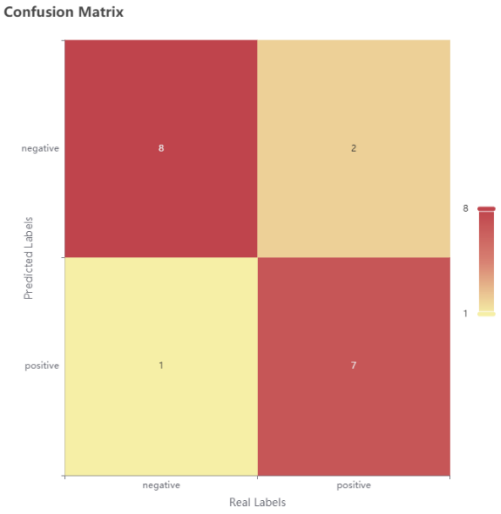
The Random Forest model was the first prediction model developed and implemented with this dataset. Random Forests are an ensemble technique which combines sometimes thousands of decision trees during training in order to collect multiple different predictions and choose the average between all choices. Random forests can be effective with noisy data where not all features affect the final prediction to the same magnitude. This feature of random forests may be useful in this case due to the high amount of attributes present and the unlikelihood of all concentrations and elements contributing to the final predictions to the same degree. In addition, random forests are useful due to their bagging and bootstrapping technique which allows for diversity among all trees by random sampling with replacement. The random forest model will also provide insights into the importance of certain features which will allow us to see which attributes the model believes contributes the most to its final prediction. The random forest model was implemented using the Knime software and the project workflow can be demonstrated below:



The dataset was changed using specifications mentioned previously in this report. The dataset was then equally sampled between positive and negative classes and partitioned into a training and test set with an 80% training and 20% testing split. The training set was then passed into the Random Forest Model with the baseline Gini Index splitting criteria. The model then made its own predictions using the test set and the performance of the model was measured.

Below, we can see in the confusion matrix that the model was often able to accurately predict real negative cases. This means that whenever the model was tasked with classifying a record that did not contain a mineral deposit it was often succesful in doing so. This is a desirable outcome because the model is not predicting that there is a mineral deposit when there actually is not. If the model were to

perform poorly in this, it could lead to wasted resources and materials spent on mining deposits that are not there. The model was also able to accuractely predict positive cases. This means that whenever the model was tasked with classifying a record that did contain a mineral deposit it was accurately able in doing so. This is a desirable outcome as the model rarely missed finding a mineral deposit. The least desirable outcome is when this happens because we are missing the chance to collect as many mineral deposits as possible.



Furthermore, in the table below we can see several metrics of performance such as Recall, Precision, Sensitivity and F-score as well as the model's overall accuracy on the testing set.

Metric	Value
Accuracy	0.833
Recall	0.778
Precision	0.875
Specificity	0.889
F-measure	0.824

Inspecting the accuracy value in the table above, the model performed moderately well on the testing data, reaching an overall accuracy of 83.3%. However, through investigation of further metrics we can verify the legitimacy of this accuracy value. As previously mentioned, the model was able to accurately predict positive cases within the test dataset. The Recall value exemplifies this and as we can see, the model is able to identify positive cases around 77.8% of the time. Furthermore, the Precision value

demonstrates something similar, that is, the overall acccuracy of positive predictions by the model. The Precision value of 87.5% demonstrates that 7 out of 8 positive predictions made by the model are correct. Moreover, we previously identified that the model was performing well on negatively classed cases in the test data set. The Specificity value also demonstrates the same findings as the model is able to correctly identify 88.9% of negative cases.

Looking further into the random forest model itself, 50 decision trees were created to comprise the overall random forest. Out of all the mineral elements, Calcium (Ca) was deemed to be the most important attribute when splitting the decision trees. Overall, there were 12 attributes that were used for the top splitting criteria more than once in the random forest. Of these 12 attributes, 3 of them had the mineral element containing calcium: Ca_AA5, Ca_AA7 and Ca_GDX. This demonstrates that calcium is a strong indicator which can be used in determining if mineral deposits are present. Another important feature that proves that this is the amount of times each of these 3 calcium concentrations were considered for the top splitting criteria and how they compare to the number of times they were actually chosen. As we can see in the table below, there were several attributes that were chosen as candidates multiple times but were not often chosen as splitting attributes. For example, attributres such as Cr_ENZ were not deemed to be strong indicator, as even though they were chosen as the top splitting attribute twice they were also chosen as a candidate 7 times. Moreover, other attributes such as Ag_AQR and Ba_GDX could be considered as good indicators. However, they were not chosen as candidates very often, meaning they may have a smaller contribution to the maintaining purity inside the random forest. Calcium, however, displays all these qualities as overall, the element was chosen as a candidate 9 times and was used as the top splitting criteria 7 times. This means that it was often considered by the random forest model to increase the purity across the decision trees and was then likely chosen as the attribute to do so.

Attribute	Split (top level)	Candidate (top level)	Split/Candidate
CA_AA5	3	4	0.75
Cr_ENZ	2	7	0.29
Mn_ENZ	2	4	0.50
Sr_AA5	2	3	0.67
V_AA5	2	5	0.40
Ca_AA7	2	2	1.00
Sr_AA7	2	4	0.50

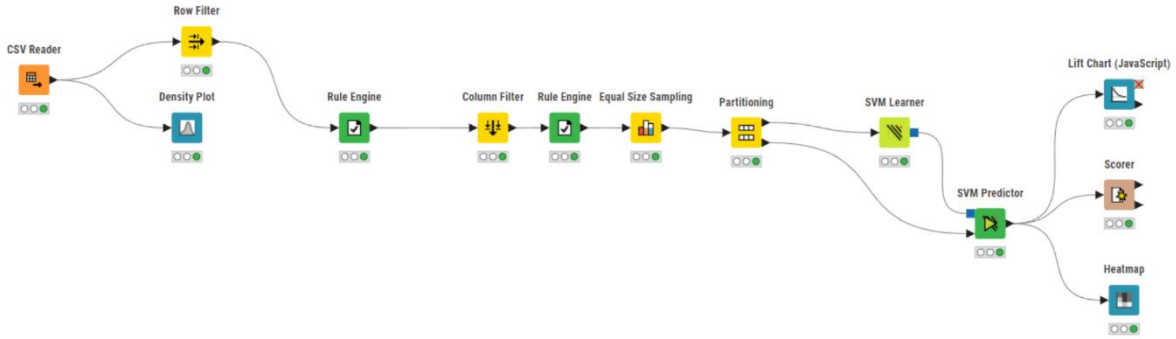
Tm_AA7	2	4	0.50
Ag_AQR	2	2	1.00
Ba_GDX	2	2	1.00
Ca_GDX	2	3	0.67
Rb_GDX	2	3	0.67

Overall, the model is able to perform moderately well on both positive and negative cases which is beneficial in identifying mineral deposits and ensuring that resources and materials are not wasted in regions with no mineral deposits. Moreover, the random forest model has high interpretability, so it was found that attributes containing the calcium element had a high contribution to predicting mineral deposits by maintaining purity in the model. However, there is still lots of room for improvements with the model for this initial test and further techniques will be implemented to improve the models performance such as cross validation and over sampling. These techniques and their relevance will be discussed further in the paper.

Support Vector Machine Model

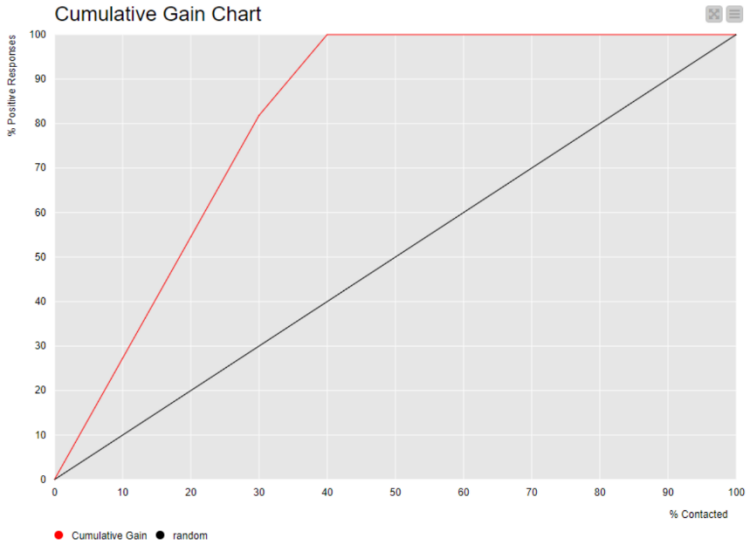
The next prediction model that was developed and implemented with this dataset is the Support Vector Machine Model. SVM models are geometric induced supervised models that separate classes using hyperplanes. This is done by finding the optimal seperator between records placed in multidimensional spaces. SVM models can be utilized in higher dimensional spaces by finding the optimal seperator that is also of a higher dimension. For example, marignal seperators for a 2 dimensional space could be a line but could also be a plane or some 3 dimensional function. SVM models use linear functions such as the dot product to find the distances (which can also mean similarity) between records in higher dimensional spaces. This allows all records to be mapped to the same space and for a seperator to be iteratively adjusted when new training records are added. These seperators split the data into various clusters, but they can also be used for regression tasks. In this context and dataset, the binary classification of the dataset is well suited for SVM models as well as the small amount of records available. In contrast to a logistic regression approach, which also use hyperplanes to classify data and is well suited for binary classification tasks; the SVM is founded upon statistical methods whereas logistic regression models are founded upon probabilistic methods, making the SWM more suited for this dataset. SVM models also implement a penalty parameter to penalize records that are near the seperator and therefore reward

distinct and seperated records. The SVM model was implemented using the Knime software and the project workflow can be viewed below:

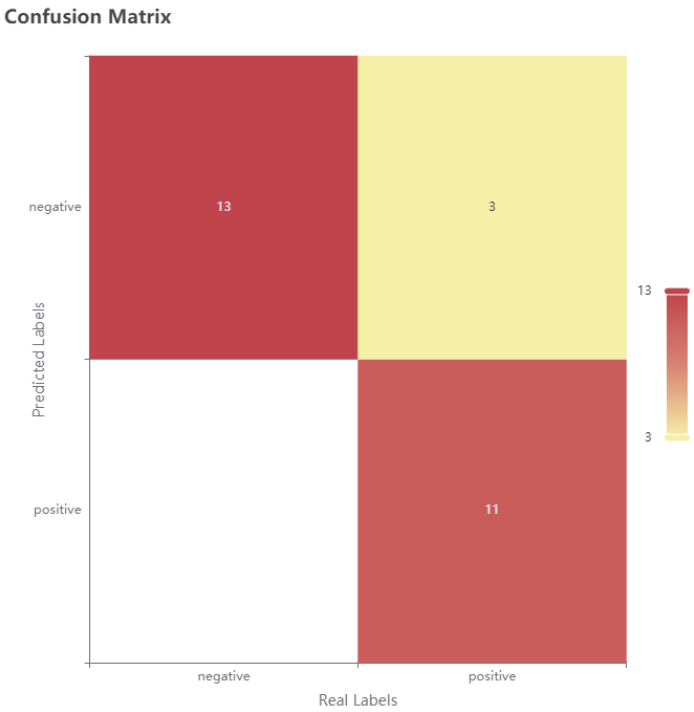


The dataset was changed using specifications mentioned previously in this report. The dataset was then equally sampled between positive and negative classes and partitioned into a training and test set with a 80% training and 20% testing split. The training set was then passed into the SVM Model with all baseline parameters (all values set to 1) using the polynomial kernel. The model then made its own predictions using the test set and the performance of the model was measured.

Below, we can see the results of a Gain chart which can be a useful representation of how our model is performing. The Cumulative Gain chart is effective for binary classification tasks and demonstrates the ratio between cumulative positive events up to the total number of possible positive events. The Gain chart is often viewed in intervals of deciles and can be used to view how much of the total target population is being reached after a certain amount of target records are predicted by the model. This is one reason why the cumulative gain chart is a useful representation for how the model is performing specifically in identifying positive events. The red curve on the graph below demonstrates the SVM model's performance on the test set (known as lift curve) and the black curve is the baseline if the model were to be classifying at random. The cumulative gain chart is useful with the case of deep mineral deposits because it can help us understand how many mineral deposits will be reached after a certain amount of total samples are checked. In the most optimal scenario the cumulative gain chart should demonstrate that the model is able to reach the total amount of positive samples as quickly as possible. In this case there were 14 records with a positive class in the test set, and looking at the cumulative gain chart, the SVM model was able to reach 100% of positively identified cases after about 40% of the records were seen. This demonstrates that the model is performing well and can reach positively classed records quickly.



Furthermore, below, we can see in the confusion matrix, that the model was often able to accurately predict real negative cases. This means that whenever the model was tasked with classifying a record that did not contain a mineral deposit it was often succesful in doing so. This is a desirable outcome as it is often going to be the case as mineral deposits are rare occurences. Additionally, the model was also able to accurately predict real positive cases only making minimal errors on the small testing set. This is a desirable outcome as the model is able to identify mineral deposits accurately without missing any mineral deposits. This would be the least desirable outcome as we are missing the opportunity to collect these mineral deposits because the model is not able to identify it.



Furthermore, in the table below we can see several metrics of performance such as Recall, Precision, Sensitivity and F-score as well as the model's overall accuracy on the testing set.

Metric	Value
Accuracy	0.889
Recall	0.786
Precision	1
Specificity	1
F-measure	0.880

Inspecting the accuracy value in the table above, the model performed very well on the testing data reaching an overall accuracy of 88.9%. However, through investigation of further metrics we can see verify the legitimacy of this accuracy value. As mentioned previously, the model was able to accurately predict positive cases within the test dataset. The Recall value demonstrates that our findings were not as good as expected and we can see this because the value is 0.786 which is approximately 10% worse then the overall accuracy of the model. This demonstrates that even though the model is fairly good at predicting positive cases there is still significant room for improvement as errors are still being made. Furthermore, the Precision and Specificity values demonstrates our previous findings that the model does not predict

records as postive when in reality they are negative. Both metric's value of 1 show that the model made no errors in this kind of prediction. These metrics depict a stronger view of the models performance, however there are still several caveats. The main issue is the low amount of testing samples used to measure these metrics. In the case of the False Positive predictions where the model made no errors of this kind, it is highly unlikely that the model will do this in a real setting and this would likely be demonstrated if a larger testing set were available. Further implementation of techniques such as Cross validation will help mitigate this and hopefully demonstrate an even more in depth view of the model's performance.

Model Tuning and further evaluation

Optimization techniques

There are several different optimization techniques are deployed to increase model performance. However, it is extremely important to identify shortcomings within all facets of the project workflow in order to deploy the right optimization techniques. In this case the dataset itself has several shortcomings that could be focused on to optimize the performance of our models.

Firstly, the dataset itself is small and only has around 200 records in total. This is not ideal for many predictive models as the dataset may not be a true representation of the total space of possibilities. On top of this the data set is not a balanced set meaning that the occurrences of positive and negative classes in the dataset are not equal. This is not ideal because several prediction models, even those used in this paper, perform better and sometimes require a balanced dataset in order to be make accurate predictions. For example, if a neural network is trained on a dataset containing 9990 negative records and 10 positive records, performing a binary classification, then the model would likely gain an overall accuracy of 99.9%. However, this is not a true representation of the model's performance, if the model were to guess only ever false on every single piece of training record shown to it, then it would gain an accuracy of 99.9%. This means that the model is never learning any true representations and features present in the dataset, it is only ever learning to predict false.

K-Fold Cross Validation

One technique implemented to optimize the performance of the prediction models used in this paper is K-Fold Cross validation. This technique partitions the dataset into, K, amount of equally sized folds/partitions and then trains the model K amount of iterations with each iteration using one of the partitions as a validation set. The subsequent average performance of each metric through all iterations are used as an estimate of the model's overall performance. This technique is extremely useful for several

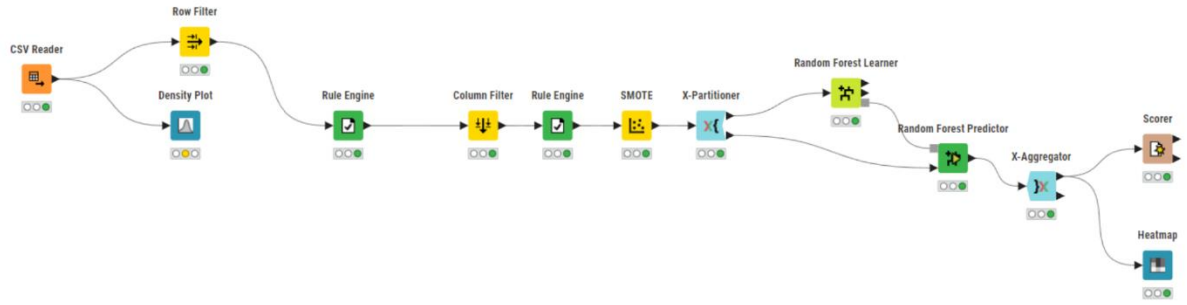
reasons however, in this case it aids with the datasets imbalanced nature as well as the fact that certain models such as the Random Forest model used in this paper has several hyperparameters that can impact the performance of the model. This technique helps mitigate this by helping in this tuning process and finding optimal performing parameters whilst showing the model several variations of the training set.

Oversampling

Another technique implemented to optimize the performance of the prediction models used in this paper is oversampling. Oversampling is the main technique used to mitigate unbalanced datasets and is deployed by increasing the number of cases with the minority class. There are various ways to implement this technique and the one used in this paper is Synthetic Minority Oversampling (SMOTE). SMOTE is able to generate new cases of the minority class by creating 'synthetic' cases which is done by interpolating existing cases of the minority class. To be more precise, SMOTE determines the nearest neighbor of all cases in the minority class, then takes the difference between the nearest neighbor and a new sample it creates, then uses the difference between these two multiplied by a random value between 0 and 1 and added onto the new sample in order to generate new synthetic cases inside the same space (Maklin). This oversampling technique will be useful with our random forest model in order to improve against bias of an imbalanced dataset as well as determining more attributes that may contribute greatly in classifying mineral deposits.

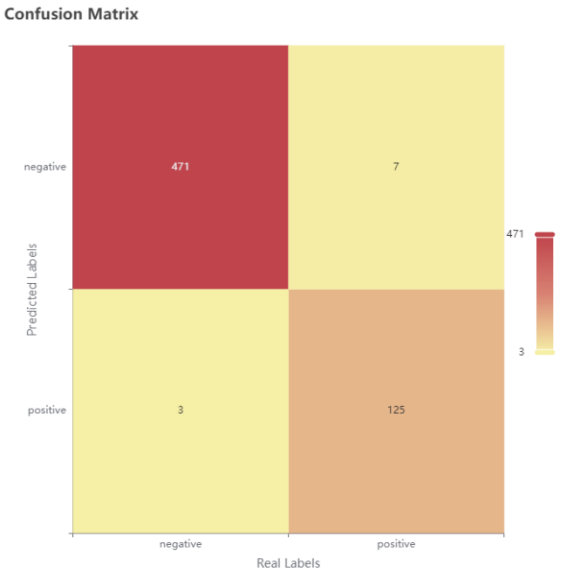
Random Forest Final Results

As mentioned previously, the two techniques used to help improve our Random Forest model's performance is K-Fold cross validation and oversampling. Injected into the project workflow with Knime, the new workflow is demonstrated below.



Furthermore, below, we can see the results of the new model with its resulting confusion matrix. As we can see the model performed extremely well with the new optimizations and was able to accurately predict real and negative cases. This is similar to the original random forest model's confusion matrix properties however, this case differs due to the magnitude of test cases the model was able to predict. In

the original implementation of the random forest the model was only able to test on 18 cases whereas with the new implementation it is able to test on 606 cases.



Furthermore, in the table below we can see the metrics of performance for the new implementation of the random forest model.

Metric	Value
Accuracy	0.983
Recall	0.994
Precision	0.985
Specificity	0.994
F-measure	0.989

Inspecting the accuracy value in the table above, the model performed very well on the testing data reaching an overall accuracy of 98.3%, around a 15% increase then the initial trial. Through investigation of further metrics, we can verify the legitimacy of this accuracy value. As mentioned previously, the model was able to accurately predict positive cases within the test dataset. The Recall value exemplifies this and as we can see the model is able to identify positive cases around 99.4% of the time. Furthermore, the Precision value demonstrates something similar, with a high value of 0.985. We previously identified that the model was performing very well on negatively classed cases in the test data set. The Specificity value also demonstrates the same findings as the model is able to correctly identify 99.4% of negative cases.

Looking further into the new random forest model, 50 decision trees models were created to comprise the overall random forest. Similar with the initial random forest model, calcium was found to be the most important attribute when splitting the decision trees. Overall, there were 11 attributes that were used for the top splitting criteria more than once in the new random forest model. Of these 11 attributes, 3 of them contained the element calcium: Ca_AA5, Ca_AA7 and Ca_AQR. Below we can see the new results of the 11 attributes used for top level splitting.

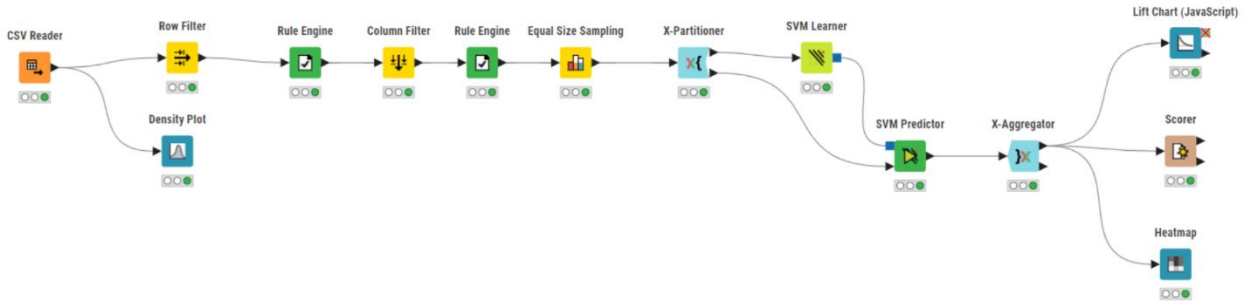
Attribute	Split (top level)	Candidate (top level)	Split/Candidate
Ca_AQR	5	5	1.00
Nb_ENZ	4	5	0.80
Ti_AA5	4	5	0.80
Ca_AA7	3	3	1.00
Sr_AQR	3	4	0.75
B_GDX	3	3	1.00
Be_GDX	3	3	1.00
Be_ENZ	2	3	0.67
Mo_ENZ	2	4	0.50
B_AA5	2	6	0.33
Ca_AA5	2	2	1.00

With the new implementation calcium is chosen as the splitting criteria 10 times, this is 5 more times than the next leading element whilst also being chosen as a candidate 10 times. This demonstrates that the calcium element is extremely pure and was not only considered by the random forest often but chosen as the splitting criteria every time it was considered.

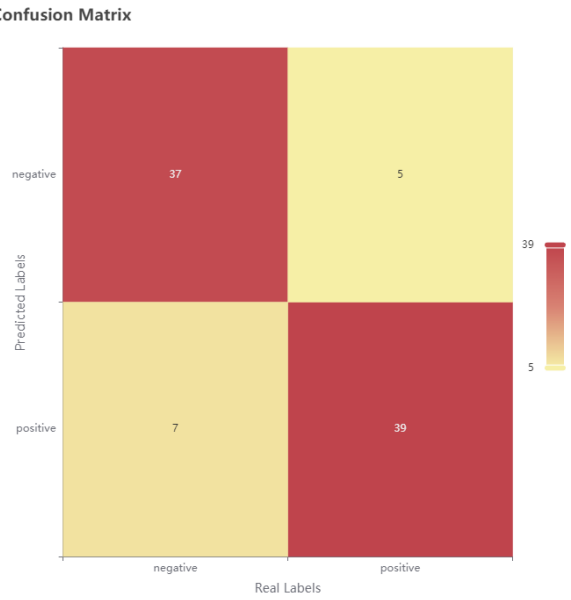
Overall, we can see that the new model is able to perform extemely well on both positive and negative cases which is beneficial in identifying mineral deposits and ensuring that resources and materials are not wasted in regions with no mineral deposits. Moreover, the random forest model has high interpretability, and it was found that attributes containing the Calcium element had a high contribution to predicting mineral deposits by maintaining purity in the model. It was also shown that the model was also able to significantly improve the overall accuracy and performance whilst also displaying a high level of interpretability than the original model using oversampling and k-fold cross validation.

Support Vector Machine Final Results

As mentioned previously, the technique used to help improve our Support Vector Machine model's performance is K-Fold cross validation. Injected into the project workflow with Knime the new workflow is demonstrated below.



Furthermore, below, we can see the results of the new model with its resulting confusion matrix. As we can see the model performed extremely well with the new optimizations and was able to accurately predict real and negative cases. This is similar to the original SVM's confusion matrix properties however, this case differs due to the magnitude of test cases the model was able to predict. In the original implementation of the SVM the model was only able to test on 25 cases whereas with the new implementation it is able to test on 88 cases.



Furthermore, in the table below we can see the metrics of performance for the new implementation of the SVM model.

Metric	Value
Accuracy	0.864
Recall	0.886
Precision	0.881
Specificity	0.841
F-measure	0.867

Inspecting the accuracy value in the table above, the model performed very well on the testing data reaching an overall accuracy of 86.4%. Although this may seem like the model performed poorly due to a 2.5% decrease in accuracy this model's performance is a better representation of the overall performance of the SVM model. This is mainly due to the fact that the SVM in the original model trained on a significantly smaller test set and managed to never predict a False positive case. This is where the model predicts a case as being positive when in reality it was negative. In the context of our data this means that the model predicted there was a mineral deposit when in reality there was not. This affected the initial model's accuracy and further metrics because values for Precision and Specificity were calculated to be 1, the highest they can be. This, however, is highly unlikely in a real deployment of the model is not a precise demonstration of the true model's performance. By increasing the number of test cases the model was able to predict we can gain a true representation of the model's performance demonstrated in the new implementation. Due to this, using k-fold cross validation actually did improve our model as we are now able to see a precise and truthful interpretation of the metrics calculated and we are able to see the model's performance on a larger sample space. As mentioned previously, the model was able to accurately predict positive cases within the test dataset. The Recall value exemplifies this and as we can see the model is able to identify positive cases around 88.6% of the time. We also previously identified that the model was performing very well on negatively classed cases in the test data set. The Specificity value also demonstrates the same findings as the model is able to correctly identify 84.1% of negative cases.

Overall, we can see that the new model is able to perform extemely well on both positive and negative cases which is beneficial in identifying mineral deposits and ensuring that resources and materials are not wasted in regions with no mineral deposits. The SVM model was also able to improve with the implementation of k-fold cross validation with a more precise view of the metrics calculated.

Final Conclusions

In conclusion, this paper demonstrated that it was possible to identify, predict and cluster mineral deposits using the concentrations of elements extracted from soil samples. This paper used several techniques to do so including hierarchical clustering to create several clusters of the dataset as well as random forest prediction models and support vector machine models to classify if a mineral deposit was present from a sample. The final result demonstrate that the random forest model was the best and was able to accurately predict if mineral deposits were present with approximately 98.3% accuracy.

Appendix A:

Element	Solvents
Ag	['_AQR']
Al	['_AA5', '_AA7', '_AQR', '_GDX']
As	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Au	['_ENZ', '_AQR']
B	['_AA5', '_AA7', '_AQR', '_GDX']
Ba	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Be	['_ENZ', '_AA5', '_AA7', '_GDX']
Bi	['_AQR']
Br	['_ENZ', '_AA5', '_AA7', '_GDX']
Ca	['_AA5', '_AA7', '_AQR', '_GDX']
Cd	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Ce	['_ENZ', '_AA5', '_AA7', '_GDX']
Cl	['_ENZ']
Co	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Cr	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Cs	['_ENZ', '_AA5', '_AA7', '_GDX']
Cu	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Dy	['_ENZ', '_AA5', '_AA7', '_GDX']
Er	['_ENZ', '_AA5', '_AA7']
Eu	['_ENZ', '_AA5', '_AA7', '_GDX']
Fe	['_AA5', '_AA7', '_AQR', '_GDX']
Ga	['_ENZ', '_AQR', '_GDX', '_AA5', '_AA7']
Gd	['_ENZ', '_AA5', '_AA7', '_GDX']
Ge	['_ENZ', '_GDX', '_AA5']
Hf	['_ENZ', '_GDX', '_AA5']
Hg	['_ENZ', '_AQR']
Ho	['_ENZ', '_AA5', '_AA7', '_GDX']
I	['_ENZ', '_AA5', '_AA7']
K	['_AA5', '_AA7', '_AQR', '_GDX']
La	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Li	['_ENZ', '_AA5', '_AA7', '_GDX']
Lu	['_ENZ', '_AA5', '_AA7', '_GDX']
Mg	['_AA5', '_AA7', '_AQR', '_GDX']
Mn	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Mo	['_ENZ', '_AA7', '_AQR', '_GDX']
Na	['_AA5', '_AA7', '_AQR', '_GDX']
Nb	['_ENZ', '_AA5']
Nd	['_ENZ', '_AA5', '_AA7', '_GDX']
Ni	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
P	['_AA5', '_AQR', '_GDX']
Pb	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
pH	['_AA7']
Pr	['_ENZ', '_AA5', '_AA7', '_GDX']
Rb	['_ENZ', '_AA5', '_AA7', '_GDX']

Re	['_ENZ']
S	['_AA5', '_AA7']
Sb	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Sc	['_AA5', '_GDX']
Se	['_ENZ', '_AQR', '_GDX']
Si	['_AA5', '_AA7']
Sm	['_ENZ', '_AA5', '_AA7']
Sr	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Ta	['_ENZ', '_GDX']
Tb	['_ENZ', '_AA5', '_AA7', '_GDX']
Th	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Ti	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Tl	['_ENZ', '_AA5', '_AA7', '_AQR']
Tm	['_ENZ', '_AA5', '_AA7']
U	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
V	['_ENZ', '_AA5', '_AQR', '_GDX']
Y	['_ENZ', '_AA5', '_AA7', '_GDX']
Yb	['_ENZ', '_AA5', '_AA7', '_GDX']
Zn	['_ENZ', '_AA5', '_AA7', '_AQR', '_GDX']
Zr	['_ENZ', '_GDX', '_AA5', '_AA7']

Appleby, Candice. “Scratching the Surface Before Digging Deep.” *Geoscience BC*, 9 June 2020, https://www.geosciencebc.com/scratching-the-surface-before-digging-deep/.

Deepest Mine | Guinness World Records. https://www.guinnessworldrecords.com/world-records/66169-deepest-mine. Accessed 20 Nov. 2023.

Deep-seaMiningFAQ.Pdf. https://www.biologicaldiversity.org/campaigns/deep-sea_mining/pdfs/Deep-seaMiningFAQ.pdf. Accessed 20 Nov. 2023.

General Information Product. General Information Product, 2016.

Ghorbani, Yousef, et al. “Moving towards Deep Underground Mineral Resources: Drivers, Challenges and Potential Solutions.” *Resources Policy*, vol. 80, Jan. 2023, p. 103222. *ScienceDirect*, https://doi.org/10.1016/j.resourpol.2022.103222.

How Developing Countries Are Paying a High Price for the Global Mineral Boom | Global Development | The Guardian. https://www.theguardian.com/global-development/2015/aug/15/developing-countries-high-price-global-mineral-boom. Accessed 20 Nov. 2023.

Learning Geology: Geochemical Sampling. https://geologylearn.blogspot.com/2014/09/geochemical-sampling.html. Accessed 20 Nov. 2023.

Maklin, Cory. “Synthetic Minority Over-Sampling TEchnique (SMOTE).” *Medium*, 14 May 2022, https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c.

Mineral Deposit | Definition, Examples, & Facts | Britannica. https://www.britannica.com/science/mineral-deposit. Accessed 20 Nov. 2023.

Mineral Deposits. 15 Jan. 2007, https://www.geo.fu-berlin.de/en/v/geolearning/gondwana/special_topics/mineraldeposits/index.html.

Why Did the Russians Seal Up the Deepest Hole in the World? | HowStuffWorks. https://science.howstuffworks.com/engineering/civil/kola-superdeep-borehole.htm. Accessed 20 Nov. 2023.