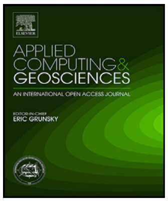
Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences



Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task

Antonella S. Antonini [a,c,*], Juan Tanzola [b,d], Lucía Asiain [b,d], Gabriela R. Ferracutti [b,d], Silvia M. Castro [a,c], Ernesto A. Bjerg [b,d], María Luján Ganuza [a,c]

[a] *VyGLab Research Laboratory (UNS-CICPBA), Department of Computer Science & Engineering, Universidad Nacional del Sur, San Andrés 800, Bahía Blanca, 8000, Argentina*
[b] *Geology Department, Universidad Nacional del Sur, San Juan 670, Bahía Blanca, 8000, Argentina*
[c] *ICIC, UNS-CONICET, Universidad Nacional del Sur, San Andrés 800, Bahia Blanca, 8000, Argentina*
[d] *INGEOSUR, UNS-CONICET, Universidad Nacional del Sur, San Juan 670, Bahia Blanca, 8000, Argentina*

ARTICLE INFO

Keywords:
Mafic–ultramafic rocks
Geochemical discrimination
Random forest classification
SHAP values
Explainable artificial intelligence

ABSTRACT

El Fierro intrusive body is one of the bodies that compose the La Jovita–Las Aguilas mafic–ultramafic belt, located in the Sierra Grande de San Luis, Argentina. The units of this belt carry a base metal sulfide (BMS) mineralization and platinum group minerals (PGM). The macroscopic description of mafic and ultramafic rocks, as is usually done by the mining exploration companies, leads to an imprecise modal classification of the rocks. In this study, we develop a random forest-based prediction model, which uses geochemical parameters to classify mafic and ultramafic rocks intercepted by drill cores. This model showed an accuracy of between 86% and 94%, and an f1_score of 96%. Random forest classification is a widely adopted Machine Learning approach to construct predictive models across various research domains. However, as models become more complex, their interpretation can be considerably difficult. To interpret the model results, we use both global and local perspectives, incorporating the SHAP (SHapley Additive exPlanations) method. The SHAP technique allows us to analyze individual samples using force plots, and provides a measure of the importance of each geochemical input attribute in the model output. As a result of analyzing the contribution of each input feature to the model, the three variables with the highest contributions were identified in the following order: Al_2O_3, MgO, and Sr.

Contents

[*] Corresponding author at: ICIC, UNS-CONICET, Universidad Nacional del Sur, San Andrés 800, Bahia Blanca, 8000, Argentina.
E-mail address: antonella.antonini@cs.uns.edu.ar (A.S. Antonini).

https://doi.org/10.1016/j.acags.2024.100178
Received 1 February 2024; Received in revised form 8 July 2024; Accepted 11 July 2024
Available online 22 July 2024

1. Introduction

The study area of this work comprises La Jovita–Las Águilas mafic–ultramafic belt (Kilmurray and Villar, 1981), which is located on the eastern slope of the Sierra Grande in San Luis, northeast of the capital city of San Luis (see Fig. 1). The mentioned belt is made up of six intrusive bodies, named El Fierro, Virorco, Escuela Las Pircas, Las Higueras, El Renegado, and Las Águilas. This contribution will be focused on specific rock samples obtained from outcrops and drill cores of El Fierro's intrusive body. The mentioned units are of economic interest for hosting base metal sulfide mineralizations (Cu–Ni–Co) and associated with them, accessory phases of platinum group minerals (MGP) (Mogessie et al., 2000; Ferracutti et al., 2005, 2013). According to Ferracutti et al. (2017) the mentioned units consist of intrusive bodies with primary magmatic stratification. This characteristic manifests itself as an alternation of levels of mafic and ultramafic rocks. The mineralizations present in the La Jovita–Las Águilas bodies are mainly restricted to levels of ultramafic rocks (Ferracutti et al., 2007a). Therefore, in the exploratory activity of a mining company, it is essential to be able to discriminate between mafic and ultramafic rocks during the description of drill core rock samples.

Mafic and ultramafic rocks belong to the large group of igneous rocks (volcanic or plutonic). It is worth mentioning that this group of rocks can be classified based on geochemical or mineralogical criteria. Regarding to geochemical criteria, igneous rocks classify in acid, intermediate, basic and ultrabasic rocks, according to their contents of SiO_2 (Cox et al., 1979; Bas et al., 1986; Wilson and Wilson, 1989). On the other hand, the mineralogical criteria is based on the modal percentages of certain minerals (such as olivine, pyroxenes, amphiboles, high Ca-plagioclase), which are identified by microscope or with the naked eye. Thus, this criterion classifies igneous rocks into felsic, intermediate, mafic and ultramafic rocks.

As previously said, the identification of the mentioned minerals is carried out from a petrographic analysis of rock samples, using a transmitted light microscope. To accomplish such studies, the samples must be processed in a Petrotomy Laboratory, where thin sections are made and used for the microscopic study. A mining company faces several challenges throughout this evaluation process. First of all, the time required for both the petrographic analysis and thin section preparation is incongruous with the timetables of mining companies. Secondly, a petrographic analysis is performed on a subset of rock samples that are not viable to do in drillings with hundreds of meters' worth of rock samples, as the ones studied in the present contribution.

As an alternative, the goal of lithologically characterizing drillings conducted on mafic–ultramafic bodies is suggested through the use of geochemical prospecting guides. This methodology consists of linking the concentrations of certain oxides and elements obtained from chemical analyses of rock powders with the presence of specific indicator minerals (mafic and plagioclase) used in the determination of mafic and ultramafic rocks (Pérez, 2021; Frank, 2022; Tanzola et al., 2024).

In recent decades, we have witnessed unprecedented growth in our ability to collect, store, and access massive volumes of data. This advancement has been possible thanks to the continuous increase in computational power (Washington et al., 2009), paving the way for the systematic processing and analysis of data, often using machine learning techniques and other statistical methods (Reichstein et al., 2019; Karpatne et al., 2018; Sun et al., 2022). In particular, machine learning is increasingly used to interpret geological data, overcoming the limitations of traditional geochemical classification diagrams, which are generally restricted to two or three variables. Machine learning algorithms allow the simultaneous use of multiple variables, demonstrating their effectiveness in the classification and prediction of geological phenomena.

Numerous studies have demonstrated the remarkable performance of machine learning algorithms such as random forests (RF), neural networks, and support vector machines (SVM) in identifying rock types, lithostratigraphic units, and tectonic environments. These techniques have been successfully applied across various geological contexts, offering improved accuracy and efficiency over traditional methods. For instance, Caté et al. (2018) implemented machine learning methodologies to classify lithological and alteration units in mineral deposits in Lalor, Snow Lake, Manitoba, Canada. Saporetti et al. (2018) conducted a comparative analysis of various machine learning techniques for the classification of carbonate-siliciclastic rocks in the South Provence basin, highlighting that Decision Trees and RF achieved the best results. Santos et al. (2022) extended the application of these techniques to the classification of rock masses, achieving high accuracy even with a reduced set of geomechanical variables, which could simplify and make geotechnical fieldwork more efficient. Sarantsatsral et al. (2021) demonstrated the effectiveness of the RF algorithm in predicting rock types at the Erdenet copper mine in Mongolia, using exploratory drilling data and spatial coordinates.

In the field of lithological mapping, Xie et al. (2018) conducted a comprehensive comparative analysis of various machine learning methods for lithological identification from well log data, concluding that ensemble methods such as RF and Gradient Tree Boosting offer the best performance. Kuhn et al. (2019) evaluated the applicability of RF and clustering techniques in the Central African Copperbelt, showing how these tools can complement and optimize conventional geological mapping methods at different stages of mineral exploration.

More recently, Saha et al. (2021) introduced an innovative methodology using the chemical composition of biotite to discriminate tectonic settings of igneous rocks using XGBoost and LightGBM algorithms. Alférez et al. (2021) achieved high accuracy in the classification of granitoids by applying convolutional neural networks to petrographic images, and Han et al. (2023) implemented the deep forest method for the lithological identification of igneous rocks, achieving high accuracy through the analysis of conventional log data. Zhao et al. (2023) developed a RF-based model for predicting formation environments of Cr-spinel in mafic–ultramafic rocks, overcoming the limitations of traditional empirical diagrams. Xing et al. (2023) proposed an innovative approach to classify rock types using machine learning with core and well log data, combining the flow zone index (FZI) method with supervised learning techniques and Shapley Additive Explanations (SHAP) to create an interpretable and accurate model applicable to hydrocarbon reservoirs. For their part, Tanzola et al. (2024) applied RF to validate a geochemical method for distinguishing between mafic and ultramafic rocks in the Virorco intrusive body, Argentina, achieving a 91% success rate in classifying these rock types based on geochemical parameters.

While these machine learning models offer high accuracy and efficiency, they often function as a “black box”, providing limited insight into their decision-making processes. Formalisms highlighting the importance of features (both global and local) for supervised learning of labeled data are lacking. SHAP are a recent development that allows for the quantitative estimation of model interpretability (Lundberg and Lee, 2017).

In this article, we present a study that offers a new perspective on the classification of mafic and ultramafic rocks by combining machine learning with advanced interpretability techniques. Using the RF algorithm as a basis, the research stands out for its exhaustive implementation of the SHAP method. Unlike previous studies that mainly focused on the application of various machine learning algorithms for lithological classification or that have superficially addressed SHAP, this work fully utilizes the potential of SHAP to interpret the RF model. We performed detailed analysis using SHAP, employing summary plots and force plots to achieve global and local interpretations. This facilitates a more complete understanding of the model and its predictions. Furthermore, we directly link the results obtained with SHAP with existing geological knowledge, explaining how the most important elements identified are related to the mineralogy of the rocks. The use

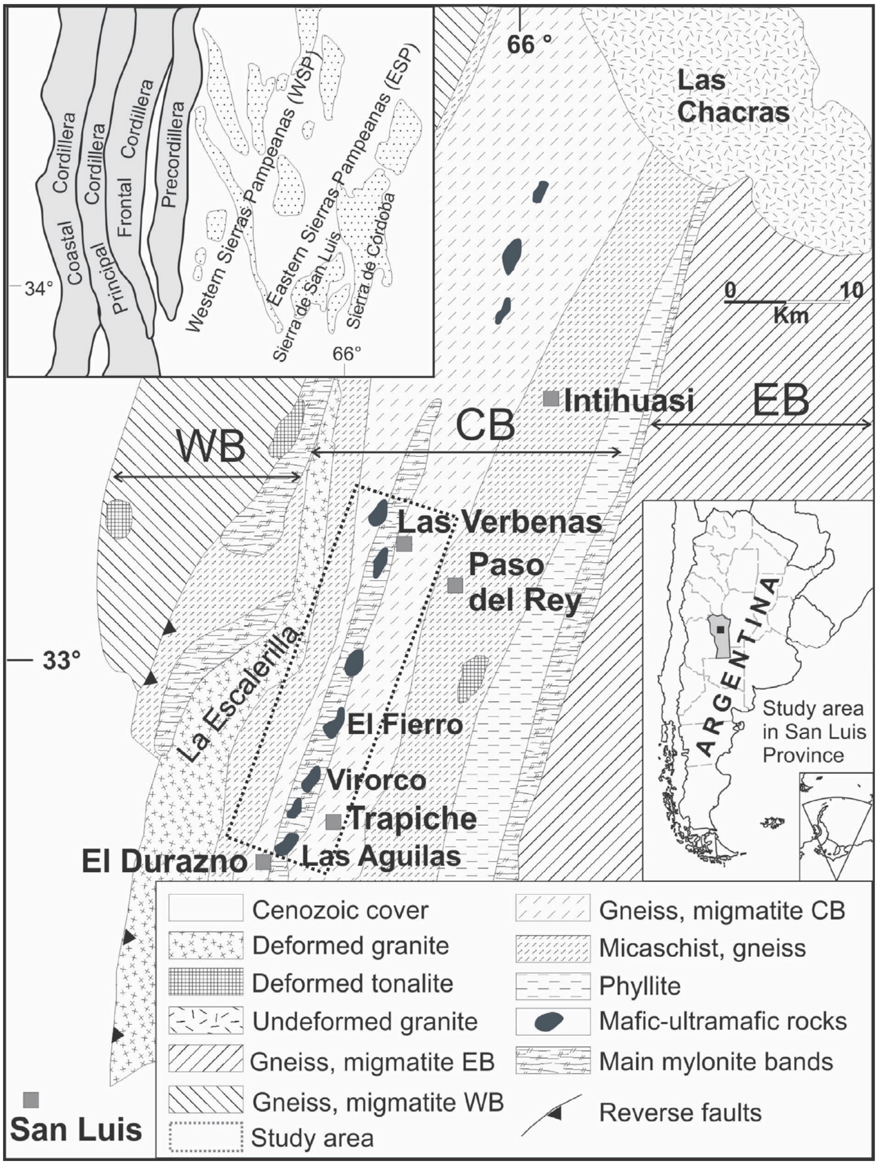


Fig. 1. Simplified geological map of Sierra de San Luis, Sierras Pampeanas, Argentina. The dotted rectangle indicates the area of study.
Source: Adapted from Delpino et al. (2007)

of SHAP allows understanding the relative importance of each geochemical feature in model decision making. We also present a detailed analysis of individual samples using SHAP force plots, providing a more granular view.

In summary, this study employs machine learning techniques, specifically the RF algorithm coupled with SHAP, to enhance the classification of mafic and ultramafic rocks within the La Jovita–Las Águilas belt. This approach not only aims to improve accuracy but also enhances interpretability by providing insights into the key factors influencing geological classifications.

2. Materials and methods

2.1. Data description

A total of 90 surface samples from the six intrusive bodies, mentioned above, were utilized as the training dataset. All data samples have a total of 19 attributes, which were used during the model training phase. These attributes include oxides (Al_2O_3, Fe_2O_3, MnO, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, and Cr_2O_3) and cations (Sr, V, Co, Ni, Cu, Zn, Ga, Sr, and Ba). These samples were previously labeled in terms of their mafic or ultramafic nature, based on microscopic petrographic analysis. The sampling process consisted of random points within each intrusive body, aiming to achieve the highest representativeness possible. Thus, the process was not influenced by any kind of bias. Firstly, rock fragments were obtained from the outcrops, from which microscopic sections and geochemical analyses were performed. Then, from the microscopic sections, it was determined whether each sample corresponded to a mafic or ultramafic rock. Consequently, each section classified in terms of mafic or ultramafic rock has associated geochemical data. In addition, the geochemical data of each sample were also available (Hauzenberger, 1997; Ferracutti, 2005; Ferracutti et al., 2007a,b, 2010, 2013, 2017; Cacace, 2019; Cacace et al., 2019; Tanzola et al., 2024).

Following the training, the RF methodology was applied to classify samples obtained from the CT-EF08_001 and CT-EF08_004 drill cores. As mentioned above, both drill cores were carried out on the El Fierro intrusive body by the Castillan Resources Corp. company during the year 2008. The geochemical data from these two drill cores consists of chemical analyses of major and trace elements, carried out by ALS

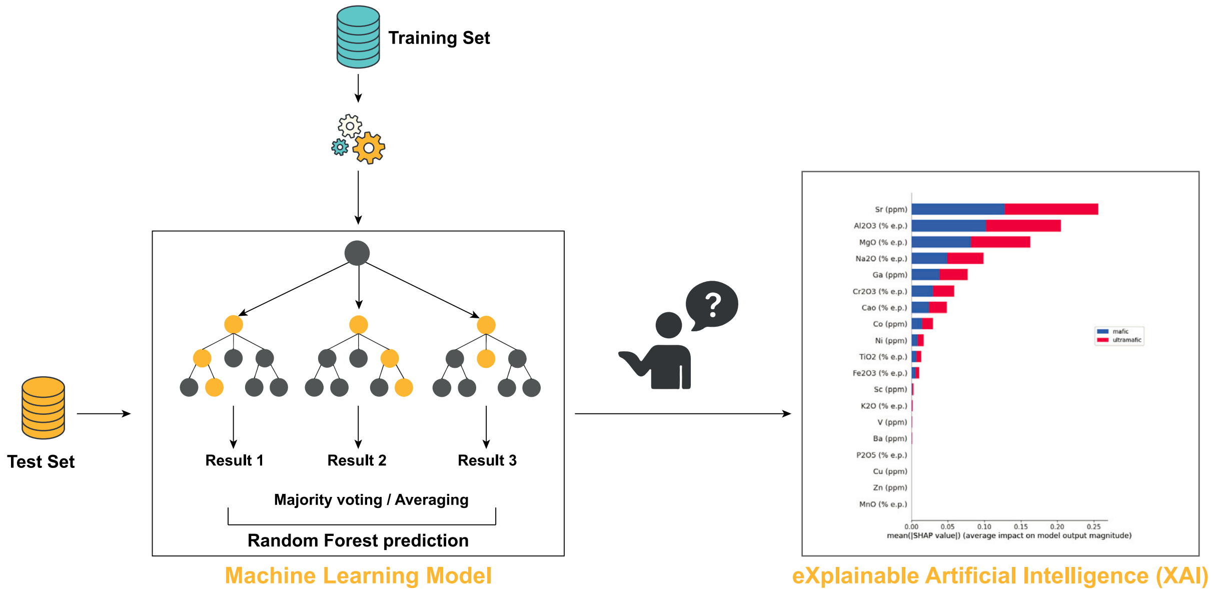


Fig. 2. The architecture of the proposed model.

Laboratory Group S.L. using the ME-ICP61 method, on pulverized rock samples. These major and trace elements analyzed in the two drill cores are the same as in the 90 rock surface samples. In reference to Tanzola et al. (2024), SiO_2 could not be used as a geochemical parameter to differentiate between mafic and ultramafic rocks. Due to an overlap of these oxide values in the collected 90 samples of both types of lithologies, it is not possible to establish a threshold value between mafic and ultramafic rocks. It should be noted that 76 out of a total of 90 samples are comprised within the mentioned overlap.

2.2. Random forest classifications

In this study, one of the most commonly used supervised machine learning models, known as random forest, was employed for the purpose of classifying mafic and ultramafic rocks intercepted by drill cores. Decision trees and RF are tools that are applied in multiple domains, such as classification, regression, and feature selection. Decision trees, for example, are generally used for classification and regression tasks. They imitate human reasoning in decision-making by building a tree-shaped structure. Each node in the tree represents a decision based on a specific characteristic of the dataset. Decision trees offer transparency, making them suitable for explaining the rationale behind a decision.

Random forest (Breiman, 2001) takes decision trees one step further. RF uses a collection of decision trees and aggregates their predictions (see Fig. 2) to make the final classification decision. This ensemble approach improves the accuracy and increases the robustness of the model. Each decision tree in a RF is trained with a subset of data and a random selection of features considered in each split. This randomness introduces diversity among the trees.

There are some key hyperparameters to adjust when training the model to improve its performance. The ‘*criterion*’ parameter determines the function used to measure the quality of a split at a node in the tree, with common options being ‘*gini*’ and ‘*entropy*’. The ‘*max_depth*’ parameter controls the maximum depth of individual trees in the forest, which influences the complexity of the model and its ability to capture complex patterns. Additionally, the ‘*n_estimators*’ parameter defines the number of trees in the forest, which affects the overall predictive power of the ensemble; a larger number often improves performance but comes with higher computational costs. Tuning these hyperparameters effectively in training is crucial to optimizing the RF model’s balance between accuracy and computational efficiency.

RF are particularly effective when dealing with high-dimensional datasets.

Table 1
Tuning parameters considered in GridSearchCV.

Tuning hyperparameters	List of values
max_depth	$\{8, 10, 12, 14\}$
n_estimators	$\{50, 100, 125, 225, 250\}$
criterion	$\{gini, entropy\}$
max_features	$sqrt$

2.3. Hyperparameter optimization

Hyperparameter tuning is a critical step in optimizing machine learning models for superior performance. In most machine learning applications, a common practice involves training multiple models on a given dataset and subsequently selecting the one that shows the best performance. When our model is trained too specifically on our training dataset, we risk overfitting, and larger errors will occur when applied to other test sets. Underfitting, on the other hand, can occur due to insufficient training of the model.

GridSearchCV (Pedregosa et al., 2011) is a method that allows the determination of the optimal hyperparameter values for a model. The selection of these hyperparameters can significantly influence the overall performance of the model. This method is widely used to exhaustively explore all possible combinations of hyperparameters. Initially, a range of values is defined for each hyperparameter, and various combinations of these are assessed in each iteration. Finally, the most successful combination is selected and integrated into the model training process. Table 1 shows the tuning parameters considered in GridSearchCV.

Hyperparameter tuning involves a trade-off between model accuracy and complexity. As we explore various combinations of hyperparameters, our goal is to find the right balance that results in high accuracy without overfitting the training data. As a result, we determined that the best model configuration is *criterion: “gini”, max_depth: 8, n_estimators: 50*.

2.4. Model performance evaluation

A *confusion matrix* is a table that allows the visualization of the performance of a model. It provides a visual representation of the proportion of correct and incorrect predictions. The rows of the table represent the True labels, and the columns the Predicted labels. The

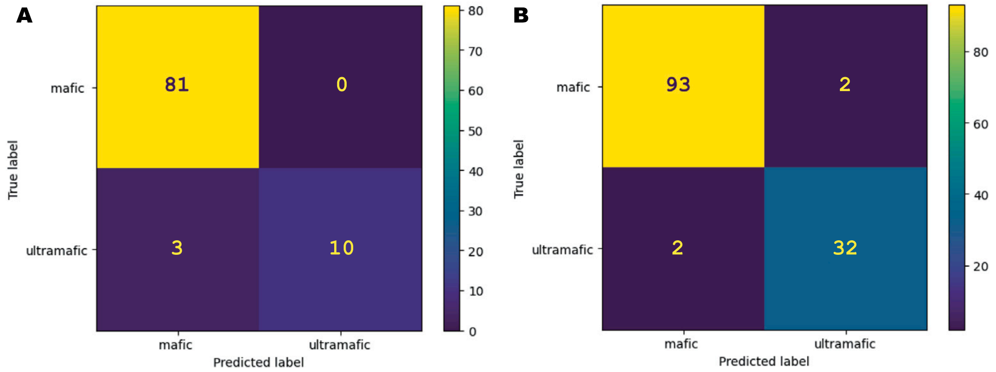


Fig. 3. Confusion matrix for the proposed model for the test set (a) CT-EF08_001 and (b) CT-EF08_004, respectively.

diagonal indicates the number of times the True and Predicted labels match. The values in the other cells represent cases where the classifier mislabeled a sample.

Fig. 3(a) corresponds to the confusion matrix for the CT-EF08_001 test dataset. The dataset has 94 samples (81 mafic and 13 ultramafic). Only three samples were incorrectly predicted (ultramafic samples classified as mafic). Fig. 3(b) corresponds to the confusion matrix for the second test dataset. Of the 129 samples in the CT-EF08_004 set, 4 samples were miscategorized by the model.

However, the confusion matrix alone may be insufficient for evaluating our model's performance. A commonly used metric to evaluate the performance of a model is *f1_score*. This metric combines the model's recall and precision:

$$f1\_score = \frac{2 * (precision * recall)}{precision + recall}$$

The percentage of real positive cases that the model is able to identify is measured by recall. Precision, conversely, assesses the precision of the model's positive predictions. It indicates how many of the predicted positive cases are correct. Our model returned an f1_score of 0.8696 for the first test dataset and 0.9412 for the second. A high f1-score value indicates a good balance between precision and recall in a classification model.

Another metric used to evaluate the performance of our classification model is the *accuracy*. Expressed as a percentage, the accuracy score calculates the proportion of correctly predicted instances to the total number of instances in the data set. Our model reaches accuracy values close to 0.96, which indicates that the model has successfully classified most of the cases.

2.5. Model interpretation with the SHAP method

In the last decade, Artificial Intelligence (AI) appears to be in many different areas in human lives. Many times those AI models are based on complex algorithms, also called black boxes. The lack of transparency of black boxes creates questions about trust in those models and applications in real life. In that context, Explainable Artificial Intelligence (XAI) offers insights into understanding the outputs of these models. In recent years, tools have emerged with the objective of explaining the operation of black boxes. Moreover, studies have shown that these tools can be used as a feature selection tool, which can improve the accuracy of the models and reduce the computational costs of model training.

One of the most powerful tools in this field is SHAP (SHapley Additive exPlanations) method (Lundberg and Lee, 2017). This method provide a clear and intuitive way to understand the impact of different features on the model's predictions. SHAP uses cooperative game theory to interpret the predictions made by a model (Molnar, 2020). In this method, Shapley values (Shapley, 1953) are calculated, in which the independent variables are interpreted as players who collaborate to receive the payout, which in this case is the specific prediction made by the model minus the average value of all predictions. The players "share" the payment based on their contribution, and this distribution is calculated using the Shapley values. This mathematical basis behind SHAP makes it a solid interpretability theory. Furthermore, unlike other methodologies, SHAP allows for both local and global explanations of the model's results, i.e., the explanation of the influence of each variable on the model's observations and the importance of each variable in the overall model results. A disadvantage of the method is that it requires retraining the model on all possible subsets $S \subseteq F$, with F being the set of all independent variables, assigning each variable an importance value that represents the effect of said variable on the model prediction. However, SHAP has a fast implementation for tree-based models.

In the context of an RF model, SHAP values can help to understand why the model makes a particular prediction or what characteristics influence the decision (see Fig. 2). By quantifying the importance of each feature, SHAP values allow the identification of the features that drive the model's predictions, aiding both model understanding and debugging. These values can be especially valuable in explaining complex ensemble-based models like RF. SHAP is a post-hoc technique, that is, it is performed after the training phase of a model. From the output vector of the trained model it is possible to use the SHAP explainers, that is, an interface that allows the calculation of the shap values in an optimized way. From these SHAP values, we can interpret the data with different graphs.

2.5.1. SHAP summary plot

The SHAP summary plot is a visual representation that shows the contribution of each feature to the model predictions. This type of graph provides an overview of how different characteristics affect the model's predictions. Features are listed on the vertical axis according to their level of importance, and horizontal bars represent the contribution of each feature to the prediction (see Fig. 4). According to the model results, the most important chemical elements for classification are Al_2O_3, MgO, and Sr, while the elements K_2O, V, Zn, and Ba barely influence the decision.

The summary plot (see Fig. 5) combines feature importance with feature effects. The position on the y-axis is determined by the characteristic and on the x-axis by the Shapley value. Each point in the summary graph is a Shapley value for one feature and one instance. The overlapping points move in the y-axis direction, so we have an idea of the distribution of Shapley values per feature.

Fig. 5(a) shows that the Sr and Al_2O_3 variables have a wide range of impact on the model and that is why they are the most important variables. Additionally, we see that when the variable values are high, its influence helps to increase the prediction or the likelihood that the sample would be classified as mafic. However, in the next most important variable, MgO, we see the opposite effect. When the values are high, the probability of classifying the sample as mafic decreases.

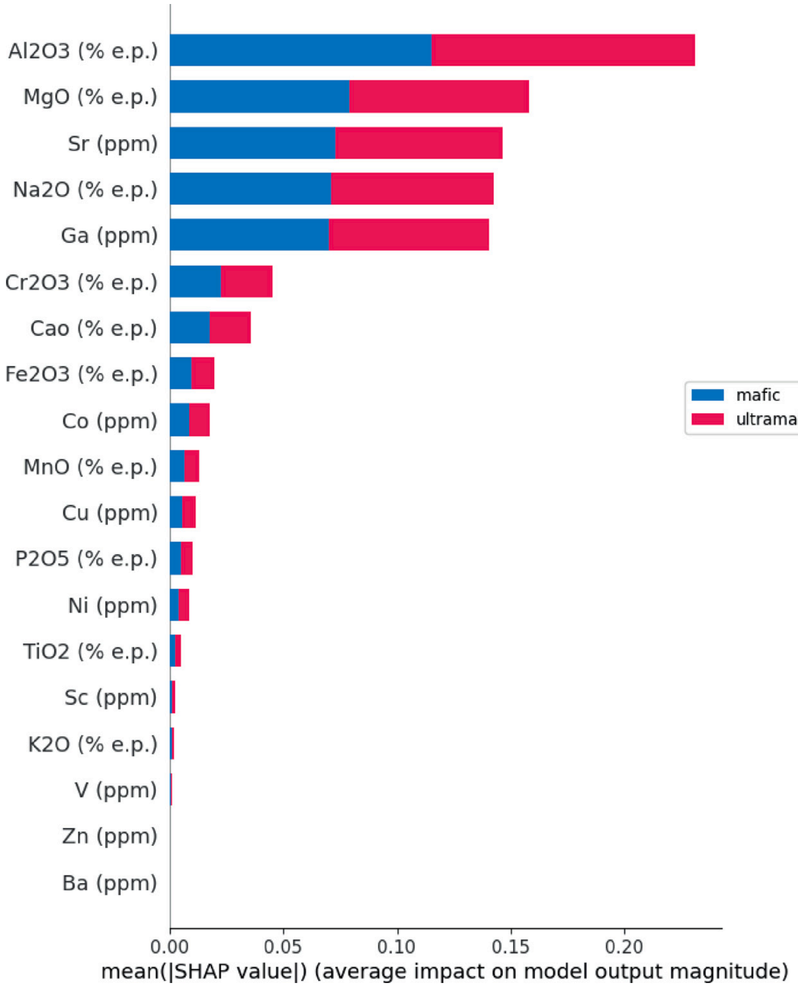


Fig. 4. SHAP Summary Plot: The plot shows the top 19 important features evaluated by the SHAP method and the effects of each feature on the outcome. Al_2O_3 was extracted as the most important feature.

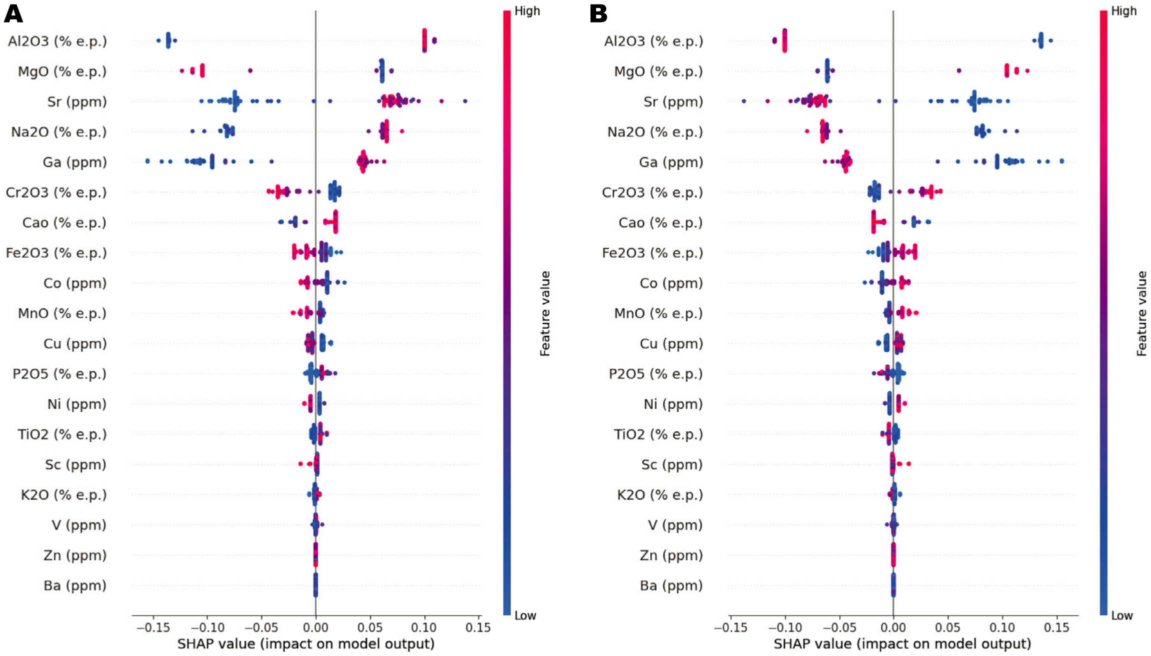
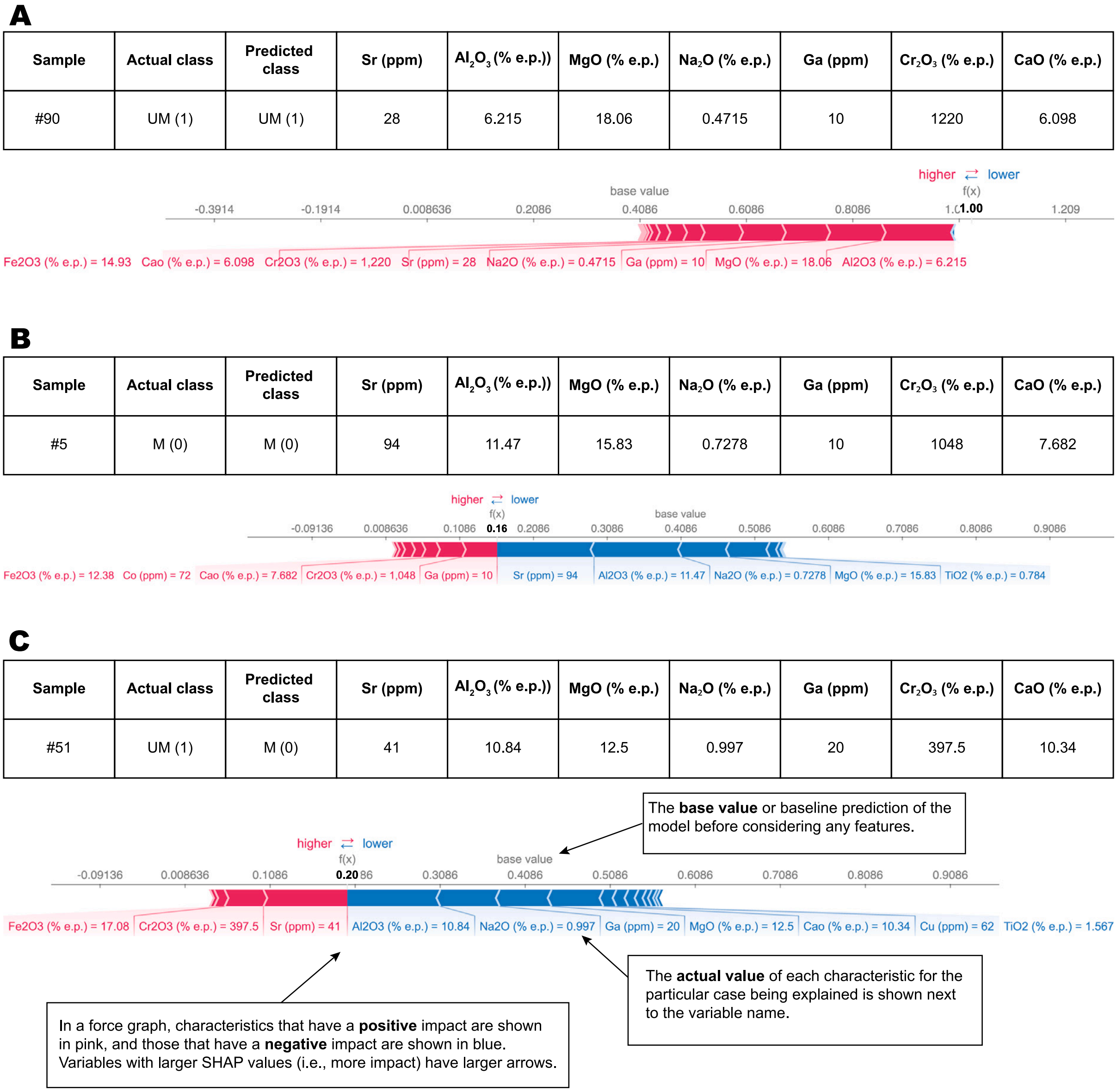


Fig. 5. SHAP summary plot for (a) mafic samples (b) ultramafic samples.



A

Sample	Actual class	Predicted class	Sr (ppm)	Al_2O_3 (% e.p.))	MgO (% e.p.)	Na_2O (% e.p.)	Ga (ppm)	Cr_2O_3 (% e.p.)	CaO (% e.p.)
#90	UM (1)	UM (1)	28	6.215	18.06	0.4715	10	1220	6.098

B

Sample	Actual class	Predicted class	Sr (ppm)	Al_2O_3 (% e.p.))	MgO (% e.p.)	Na_2O (% e.p.)	Ga (ppm)	Cr_2O_3 (% e.p.)	CaO (% e.p.)
#5	M (0)	M (0)	94	11.47	15.83	0.7278	10	1048	7.682

C

Sample	Actual class	Predicted class	Sr (ppm)	Al_2O_3 (% e.p.))	MgO (% e.p.)	Na_2O (% e.p.)	Ga (ppm)	Cr_2O_3 (% e.p.)	CaO (% e.p.)
#51	UM (1)	M (0)	41	10.84	12.5	0.997	20	397.5	10.34

The **base value** or baseline prediction of the model before considering any features.

In a force graph, characteristics that have a **positive** impact are shown in pink, and those that have a **negative** impact are shown in blue. Variables with larger SHAP values (i.e., more impact) have larger arrows.

The **actual value** of each characteristic for the particular case being explained is shown next to the variable name.

Fig. 6. SHAP forces plots for local interpretability. Force plot for sample (a) #90 classified correctly as ultramafic, for sample (b) #5 classified correctly as mafic, and for sample (c) #51, which is classified as mafic, but belongs to ultramafic samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.5.2. Local interpretability

Local explanations allow analyzing the model classification for selected data points (Lundberg et al., 2020). These are illustrated with SHAP force plots. Following Molnar (2020), SHAP values are associated with different “forces” that increase or decrease the model prediction. Each prediction starts from the base value, which is given by the average of all probabilities for each sample present in the dataset if none of the input attributes is known (Lundberg and Lee, 2017).

Fig. 6 shows force plots visualizing the Shapley values of the features. Feature values in pink increase prediction. The size of the bar shows the magnitude of the effect of the characteristic. Feature values in blue cause the prediction to decrease. The sum of all SHAP values explains why the model’s prediction differed from the baseline.

The force graph for sample #90 is shown in Fig. 6(a). The model predicted 1.00 (absolutely certain that it is an ultramafic sample). The values of Al_2O_3, MgO, Ga, and Na_2O have significantly increased their chances of the sample being classified as ultramafic.

The graph in Fig. 6(b) shows the force graph for sample #5. The model predicted 0.16 (16% probability that it was an ultramafic sample). In this case, the values of Sr, Al_2O_3, Na_2O and MgO contribute negatively and significantly decrease the chances of being classified as ultramafic.

Finally, in Fig. 6(c) we study a false negative sample where its real class according to geologists is ultramafic, but it was classified as mafic. We observe that this sample is characterized by relatively high values of Sr and Al_2O_3, and low MgO values, which coincides with the general behavior expected by a mafic.

The utilization of SHAP values forces plots offers a vital tool for understanding and interpreting machine learning models. These plots provide a clear and intuitive insight into how individual features influence model predictions. By visualizing the contribution of each feature to a prediction, scientists can gain a deeper understanding of model behavior, identify critical drivers, and detect potential issues like biases or outliers. It is very convenient to use it for error analysis or a deep understanding of a particular case.

3. Results and discussion

Our model based on RF and interpreted using SHAP demonstrates great potential for the automated classification of mafic and ultramafic rocks. This approach not only provides a powerful tool for geological classification but also illustrates the value of combining machine learning techniques with traditional geological knowledge.

SHAP analysis applied to our RF model has provided valuable insights into the relative importance of chemical elements in the classification of mafic and ultramafic rocks. As shown in the SHAP summary plot, the most important chemical elements for classification are Al_2O_3, MgO, and Sr. This is consistent with some of the parameters proposed by the authors previously mentioned, used in the determination of mafic and ultramafic rocks from the bonding with the presence of mafic minerals and plagioclase. The prominence of Al_2O_3 in our model is consistent with classical mineralogical criteria. As stated in Section 1, according to Streckeisen (1973), the threshold value in mineral percentage in mafic rocks is $> 10\%$ of plagioclase and in ultramafic rocks is $< 10\%$. Since Al_2O_3 is a primary component of plagioclase, its high concentration in a sample strongly suggests the presence of a mafic rock. The importance of MgO in our model reflects its crucial role in the composition of mafic minerals. We observe an inverse relationship between MgO content and plagioclase abundance, consistent with the transition from ultramafic to mafic compositions. This relationship captured by the model underscores its ability to discern complex geochemical patterns. Concerning Sr, plagioclase exhibits a notably higher partition coefficient compared to the other minerals found in these rocks. Consequently, elevated Sr concentrations are associated with a concurrent rise in the modal proportion of plagioclase within the rock. These associations between geochemistry and the presence of the mentioned minerals are reflected in the SHAP values force plots.

Notably, we achieved high accuracy in classifying the CT-EF08_001 (96.80%) and CT-EF08_004 (96.90%) core datasets. These results indicate that our methodology is as effective as traditional methods based on established geological criteria (Tanzola et al., 2024). This alignment between a data-driven approach and traditional geological knowledge validates the robustness of our model.

Despite promising results, it is crucial to acknowledge the potential limitations of our approach. Our training dataset may not fully represent global geological diversity. The samples used could be biased towards certain regions or types of geological formations, limiting the model's applicability to significantly different geological contexts. Additionally, we have not explicitly considered analytical errors in geochemical measurements. These uncertainties could impact classification, especially for samples near the boundary between mafic and ultramafic compositions. These boundary cases could be sources of classification error.

To address these limitations and further improve our model, we propose in future research to incorporate samples from a broader range of geological contexts to improve the robustness and applicability of the model. Furthermore, we believe that including additional elemental relationships between chemical elements could provide complementary information for classification. We also consider it appropriate to carry out a detailed analysis of the sensitivity of the model to analytical uncertainties and evaluate their impact on the classification; investigate misclassified samples or those close to the classification limit to refine the model and improve accuracy in these challenging cases.

4. Conclusions

Our study successfully applied SHAP to interpret a RF classifier trained to differentiate between mafic and ultramafic rock samples. This approach not only achieved high classification accuracy but also provided valuable insights into the geochemical factors driving the classification. Our proposed model achieved an impressive accuracy of around 96% in classifying mafic and ultramafic samples. This high accuracy demonstrates the potential of machine learning approaches in geological classification tasks.

The SHAP analysis revealed that Al_2O_3, MgO, and Sr were the most important chemical attributes in the model's predictions. This aligns well with established geological knowledge, as these elements are closely associated with the mineralogical differences between mafic and ultramafic rocks. The use of SHAP allowed us to move beyond the "black box" nature of machine learning models, providing both global and local interpretations of the model's decisions. This interpretability is crucial for building trust in the model and for gaining new geological insights.

Looking forward, we foresee several avenues for further enhancement and expansion of this research.

- *Expanding the dataset:* To improve the model's robustness and generalizability, future work should focus on incorporating samples from a broader range of geological contexts.
- *Incorporating additional features:* Exploring additional elemental relationships and geochemical ratios could provide complementary information for classification.
- *Web application development:* To make this tool accessible to the wider geological community, we propose developing a web application that integrates these machine learning and interpretability techniques.
- *Oversampling techniques:* To address potential class imbalance and further improve model robustness, future work could explore oversampling techniques or collect additional samples, particularly of underrepresented classes.
- *Exploring alternative machine learning models:* Although our RF model has demonstrated excellent performance, future research could investigate and compare other machine learning algorithms to potentially uncover more efficient or accurate approaches for geological classification.

In conclusion, our study demonstrates that the combination of RF classification and SHAP interpretation offers a powerful and insightful approach to geological classification tasks. This methodology not only provides accurate classifications but also offers interpretable results that can enhance our understanding of the geochemical factors distinguishing mafic and ultramafic rocks. As we continue to refine and expand this approach, it has the potential to become a valuable tool in Geosciences, complementing traditional methods and potentially uncovering new insights in rock classification and characterization.

Funding

This work was partially supported by PGI 24/N048 and PGI 24/ZN38 research grants from the Secretaría General de Ciencia y Tecnología, Universidad Nacional del Sur (Argentina), and by 28720210100824CO (PIBAA) granted by National Council for Scientific and Technical Research (CONICET).

CRediT authorship contribution statement

Antonella S. Antonini: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Juan Tanzola:** Writing – original draft, Validation, Resources, Data curation. **Lucía Asiain:** Writing – review & editing, Resources, Investigation, Data curation, Supervision. **Gabriela R. Ferracutti:** Writing – review & editing, Resources, Investigation, Data curation, Supervision. **Silvia M. Castro:** Writing – review & editing, Supervision, Funding acquisition. **Ernesto A. Bjerg:** Writing – review & editing, Resources, Investigation, Data curation, Supervision. **María Luján Ganuza:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data is available in the articles cited in the paper.

References

Alférez, G.H., Vázquez, E.L., Ardila, A.M.M., Clausen, B.L., 2021. Automatic classification of plutonic rocks with deep learning. Appl. Comput. Geosci. 10, 100061.

Bas, M.J.L., Maitre, R.W.L., Streckeisen, A., Zanettin, B., IUGS Subcommission on the Systematics of Igneous Rocks, 1986. A chemical classification of volcanic rocks based on the total alkali-silica diagram. J. Petrol. 27, 745–750.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Cacace, F.E., 2019. Petrología y Geoquímica de los Cuerpos Máficos-Ultramáficos Estratificados de la Sierra Grande de San Luis (Ph.D. thesis). Universidad Nacional del Sur.

Cacace, F.E., Ferracutti, G.R., Ntaflos, T., Asiain, L.M., Bjerg, E.A., 2019. Petrografía, geoquímica y estratigrafía ígnea del cuerpo intrusivo escuela las pircas, sierra grande de san luis, Argentina. Rev. Asoc. Geol. Argent. 76, 229–253.

Caté, A., Schetselaar, E., Mercier-Langevin, P., Ross, P.-S., 2018. Classification of lithostratigraphic and alteration units from drillhole lithogeochemical data using machine learning: A case study from the lalor volcanogenic massive sulphide deposit, Snow Lake, Manitoba, Canada. J. Geochem. Explor. 188, 216–228.

Cox, K.G.K.G., Bell, J.D.J.D., Pankhurst, R.J., 1979. The interpretation of igneous rocks / K. G. Cox, J. D. Bell, and R. J. Pankhurst. The Interpretation of Igneous Rocks. G. Allen & Unwin, London.

Delpino, S.H., Bjerg, E.A., Ferracutti, G.R., Mogessie, A., 2007. Counterclockwise tectonometamorphic evolution of the pringles metamorphic complex, Sierras Pampeanas of San Luis (Argentina). J. South Am. Earth Sci. 23, 147–175. http://dx.doi.org/10.1016/j.jsames.2006.09.019, URL: https://www.sciencedirect.com/science/article/pii/S089598110600099X.

Ferracutti, G., 2005. Geología y Mineralizaciones Asociadas a la Faja de Rocas Máficas–Ultramáficas de las Sierras Pampeanas de San Luis (Ph.D. thesis). Tesis doctoral. Universidad Nacional del Sur, Bahía Blanca, Argentina.

Ferracutti, G., Bjerg, E., Hauzenberger, C., Mogessie, A., Cacace, F., Asiain, L., 2017. Meso to neoproterozoic layered mafic-ultramafic rocks from the virorco back-arc intrusion, Argentina. J. South Am. Earth Sci. 79, 489–506. http://dx.doi.org/10.1016/j.jsames.2017.09.016.

Ferracutti, G., Bjerg, E., Mogessie, A., 2007a. Metales base y preciosos en las águilas, sierra de san luis: Mineralogía, génesis y evolución. Rev. Asoc. Geol. Argent. 62, 434–446.

Ferracutti, G., Bjerg, E., Mogessie, A., 2007b. Petrología y geoquímica de las rocas máficas-ultramáficas del área río las águilas-arroyo de los manantiales, provincia de san luis. Rev. Asoc. Geol. Argent. 62, 405–416.

Ferracutti, G.R., Bjerg, E.A., Mogessie, A., 2010. Caracterización de la mineralización de las águilas basada en indicadores litogeoquímicos y relaciones entre Cu-Ni y EGP. Rev. Asoc. Geol. Argent. 67, 205–215.

Ferracutti, G., Bjerg, E., Mogessie, A., 2013. Petrology, geochemistry and mineralization of the Las Águilas and Virorco mafic–ultramafic bodies, San Luis Province, Argentina. Int. J. Earth Sci. 102, 701–720. http://dx.doi.org/10.1007/s00531-012-0834-8.

Ferracutti, G.R., Mogessie, A., Bjerg, E.A., 2005. Chemical and mineralogical profile of the Las Águilas mafic-ultramafic drill core, San Luis Province, Argentina. Mitt. Österreichischen Mineral. Ges..

Frank, G., 2022. Análisis Litogeoquímico y Geoestadístico de las Perforaciones SL_10_04 Y SL_14_04, Cuerpo Intrusivo Máfico–Ultramáfico las Águilas, Provincia de San Luis (Ph.D. thesis). Trabajo Final de Licenciatura. Universidad Nacional del Sur, Bahía Blanca, Argentina.

Han, R., Wang, Z., Wang, W., Xu, F., Qi, X., Cui, Y., Zhang, Z., 2023. Igneous rocks lithology identification with deep forest: Case study from eastern sag, Liaohe basin. J. Appl. Geophys. 208, 104892.

Hauzenberger, C.A., 1997. The Sierras de San Luis, Central-Argentina: Metamorphic, Metallogenic, and Geochemical Investigations. na.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V., 2018. Machine learning for the geosciences: Challenges and opportunities. IEEE Trans. Knowl. Data Eng. 31, 1544–1554. http://dx.doi.org/10.1109/TKDE.2018.2861006.

Kilmurray, J., Villar, L., 1981. El basamento de la sierra de san luis y su petrología. In: Geología y Recursos Minerales de la Provincia de San Luis, Relatorio Del VIII Congreso Geológico Argentino. pp. 33–54.

Kuhn, S., Cracknell, M.J., Reading, A.M., 2019. Lithological mapping in the central African copper belt using random forests and clustering: Strategies for optimised results. Ore Geol. Rev. 112, 103015.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. http://dx.doi.org/10.1038/s42256-019-0138-9.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17, Curran Associates Inc., Red Hook, NY, USA, pp. 4768–4777. http://dx.doi.org/10.48550/arXiv.1705.07874.

Mogessie, A., Hauzenberger, C., Hoinkes, G., Felfernig, A., Stumpfl, E., Bjerg, E.A., Kostadinoff, J., 2000. Genesis of platinum-group minerals in the las aguilas mafic-ultramafic rocks, San Luis Province, Argentina: Textural, chemical and mineralogical evidence. Mineral. Petrol. 68, 85–114. http://dx.doi.org/10.1007/s007100050005.

Molnar, C., 2020. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, second ed. URL: https://christophm.github.io/interpretable-ml-book.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. http://dx.doi.org/10.48550/arXiv.1201.0490.

Pérez, S.R., 2021. Análisis Litogeoquímico Vinculado a la Presencia de Minerales Del Grupo Del Espinelo En la Perforación SL_01_04, Cuerpo ígneo Máfico–Ultramáfico Las Águilas (Ph.D. thesis). Trabajo Final de Licenciatura. Universidad Nacional del Sur, Bahía Blanca, Argentina.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, f., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. http://dx.doi.org/10.1038/s41586-019-0912-1.

Saha, R., Upadhyay, D., Mishra, B., 2021. Discriminating tectonic setting of igneous rocks using biotite major element chemistry — A machine learning approach. Geochem. Geophys. Geosyst. 22, e2021GC010053.

Santos, A.E.M., Lana, M.S., Pereira, T.M., 2022. Evaluation of machine learning methods for rock mass classification. Neural Comput. Appl. 34, 4633–4642.

Saporetti, C.M., da Fonseca, L.G., Pereira, E., de Oliveira, L.C., 2018. Machine learning approaches for petrographic classification of carbonate-siliciclastic rocks using well logs and textural information. J. Appl. Geophys. 155, 217–225.

Sarantsatsral, N., Ganguli, R., Pothina, R., Tumen-Ayush, B., 2021. A case study of rock type prediction using random forests: Erdenet copper mine, Mongolia. Minerals 11, 1059.

Shapley, L.S., 1953. A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games (AM-28), vol. II. Princeton University Press, Princeton, pp. 307–318. http://dx.doi.org/10.1515/9781400881970-018.

Streckeisen, A., 1973. Plutonic rock : Classification and nomenclature recommended by the iugs subcommission on the systematics of igneous rocks. Geotimes 18, 26–30, URL: https://api.semanticscholar.org/CorpusID:232549475.

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S.M., Wang, J., Lin, C., Cristea, N., Tong, D., Carande, W.H., Ma, X., et al., 2022. A review of earth artificial intelligence. Comput. Geosci. 159, 105034. http://dx.doi.org/10.1016/j.cageo.2022.105034.

Tanzola, J., Ferracutti, G., Asiain, L., Antonini, A., Luján Ganuza, M., 2024. Applied geochemistry for the discrimination between mafic and ultramafic rocks in Cu-Ni-pge-bearing layered complexes: A case study at the La Jovita–Las Águilas belt, Sierra Grande de San Luis, Argentina. J. South Am. Earth Sci. 104755. http://dx.doi.org/10.1016/j.jsames.2023.104755.

Washington, W.M., Buja, L., Craig, A., 2009. The computational future for climate and Earth system models: On the path to petaflop and beyond. Phil. Trans. R. Soc. A 367, 833–846. http://dx.doi.org/10.1098/rsta.2008.0219.

Wilson, M., Wilson, B.G., 1989. Igneous petrogenesis a global tectonic approach. URL: https://api.semanticscholar.org/CorpusID:133626460.

Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., Tu, M., 2018. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. J. Pet. Sci. Eng. 160, 182–193.

Xing, Y., Yang, H., Yu, W., 2023. An approach for the classification of rock types using machine learning of core and log data. Sustainability 15, 8868.

Zhao, J., Xue, S., Li, Y., Niu, Y., Wang, X., Zhang, X., Wang, L., Xin, Y., Zhang, R., Wang, X., 2023. Machine learning prediction of mafic–ultramafic rock-related Cr-spinel formation environments and its application to the tectonic settings of magmatic sulfide deposits. Ore Geol. Rev. 105841.