Assessment of Multiple Machine Learning Models for Critical Mineral Exploration Data Quality in the Brazilian Shield

Humberto Alves Barbosa *

Posted Date: 26 May 2026

doi: 10.20944/preprints202605.1748.v1

Keywords: machine learning; critical minerals; Brazilian shield; Tocantins province; data quality; model construction; computational geoscience

 doi:10.20944/preprints202605.1748.v1

Article

Assessment of Multiple Machine Learning Models for Critical Mineral Exploration Data Quality in the Brazilian Shield

Humberto Alves Barbosa

Laboratório de Análise e Processamento de Imagens de Satélites (LAPIS), Federal University of Alagoas, A. C. Simões Campus, Alagoas 57072-900, Brazil; humberto.barbosa@icat.ufal.br or barbosa33@gmail.com; Tel.: +55-82-99999-3043

Abstract

The growing integration of artificial intelligence into mineral exploration has created new opportunities for improving target selection and decision-making in geologically complex regions. This study presents an integrated multiple machine learning framework designed to address the assessment of exploration data quality. The analysis was conducted using an extensive geophysical and geochemical dataset comprising 221 exploration sites distributed across the Brazilian Shield. Six widely adopted algorithms were comparatively evaluated, including Random Forest, XGBoost, AdaBoost, Decision Trees, K-Nearest Neighbors, and Logistic Regression. The results demonstrate that Random Forest achieved the highest accuracy in data quality classification (accuracy = 0.82, AUC = 0.85). Cross-validation confirmed model robustness (5-fold CV R^2 = 0.80 ± 0.02; accuracy = 0.82 ± 0.02). Feature importance and explainability analyses revealed that magnetic anomaly intensity, copper concentration, and alteration-related indices are the most influential predictors, reinforcing both the geological plausibility and the computational reliability of the models. This proposed methodology offers practical support for mineral exploration strategies across the Brazilian Shield and provides a scalable framework for future applications involving critical mineral systems.

Keywords: machine learning; critical minerals; Brazilian Shield; Tocantins Province; data quality; model construction; computational geoscience

1. Introduction

The global transition toward low-carbon energy systems has substantially increased the demand for critical minerals such as nickel, cobalt, lithium, copper, and rare earth elements [1]. This scenario has intensified the need for more efficient mineral exploration strategies capable of reducing uncertainty and optimizing investment decisions [2]. In this context, artificial intelligence and machine learning have become increasingly relevant as advanced analytical tools for processing the large volumes of geological, geochemical, geophysical, and remotely sensed data generated by modern exploration campaigns [3–6].

Exploration datasets are inherently complex. They typically combine variables derived from different acquisition techniques, spatial resolutions, and temporal frameworks, resulting in highly heterogeneous and multidimensional data structures [7]. In addition, the relationships between predictor variables and mineralization processes are rarely linear, which limits the performance of conventional statistical methods [8]. Machine learning methods offer an important advantage in this setting because they can capture complex non-linear interactions and hidden patterns that are difficult to identify through traditional interpretation alone [9].

Despite the rapid expansion of machine learning applications in geoscience, two major limitations remain. First, most predictive studies focus exclusively on prospectivity mapping or economic scoring without explicitly incorporating the reliability of the input data [10]. Second,

 doi:10.20944/preprints202605.1748.v1

exploration datasets often include inconsistencies related to sampling density, analytical procedures, acquisition standards, and legacy databases. [11]. Therefore, predictive models may learn noise and methodological artifacts rather than genuine geological controls, which can lead to biased results and overconfident predictions [12,13].

Previous studies have demonstrated the applicability of approaches such as logistic regression, support vector machines, random forests, neural networks, and evidence-based models for mineral prospectivity mapping [14]. More recently, multiple learning techniques have received considerable attention due to their ability to reduce overfitting, improve generalization, and model highly non-linear relationships [15]. However, these applications generally treat data quality as a preliminary filtering step rather than an explicit prediction target [16].

A more robust strategy is to incorporate data reliability directly into the learning framework. By simultaneously estimating data confidence and economic potential, the predictive workflow becomes more informative and better aligned with real exploration decision-making, where both opportunity and uncertainty must be considered. This dual-task perspective remains relatively underexplored in applied mineral exploration studies [18,19]. The Brazilian Shield provides an ideal natural laboratory for evaluating such an approach. As one of the largest Precambrian cratonic provinces in the world, it contains a wide diversity of lithological domains, tectonic histories, and mineralization styles [20]. Its geological complexity, combined with the availability of large public geoscientific databases, offers a robust environment for testing advanced machine learning frameworks [21].

In the exploration of complex datasets, use of multiple machine learning models may allow for reducing uncertainty and optimizing investment decisions across the Brazilian Shield. However, given the novelty of the method and the challenges in implementation, there are few studies on geological and geophysical predictors using explainable machine learning techniques. To conclude, this study selects multiple machine learning algorithms to develop an integrated predictive workflow to evaluate data quality, then compares the performance of multiple, and baseline models under identical validation conditions.

2. Geological Setting and Data-Driven Machine Learning Models

2.1. Geological Setting

The Brazilian Shield represents one of the most significant mineral provinces in the world [22]. It hosts major resources of critical minerals essential to contemporary industrial and energy systems. The Shield is predominantly composed of Archean and Proterozoic lithologies. Its long and complex tectonic evolution has resulted in a diverse range of mineral systems, including those hosting nickel, copper, cobalt, lithium, rare earth elements, and platinum group elements, which are essential for modern energy technologies. It is subdivided into several major geological provinces, each characterized by distinct tectonothermal histories and mineralization styles [23,24]. The regional geological framework of the Brazilian Shield and the spatial distribution of exploration sites considered in this study are shown in Figure 1.

The São Francisco Craton includes the oldest rocks known in South America Shield. It is bounded by three composites, principally Neoproterozoic, orogenic belts, the Borborema Province to the north, the Tocantins Province to the west and southwest, and the Mantiqueira Province to the southeast. This geological heterogeneity exerts first-order control on the spatial distribution and geochemical signatures of critical mineral occurrences across the Shield [25]. Among these, the Serra Verde rare earth element deposit [26] in Tocantins Province, previously the Pela Ema tin mining centre, is located ~30 km West-Northwest (WNW) of the city of Minaçu, ~375 km North-Northeast (NNE) of Goiânia, the state capital, and 275 km North-Northwest (NNW) of Brasilia, in the north of Goiás State, Brazil (Location: 13° 29' 46"S, 48° 30' 42"W).

 doi:10.20944/preprints202605.1748.v1

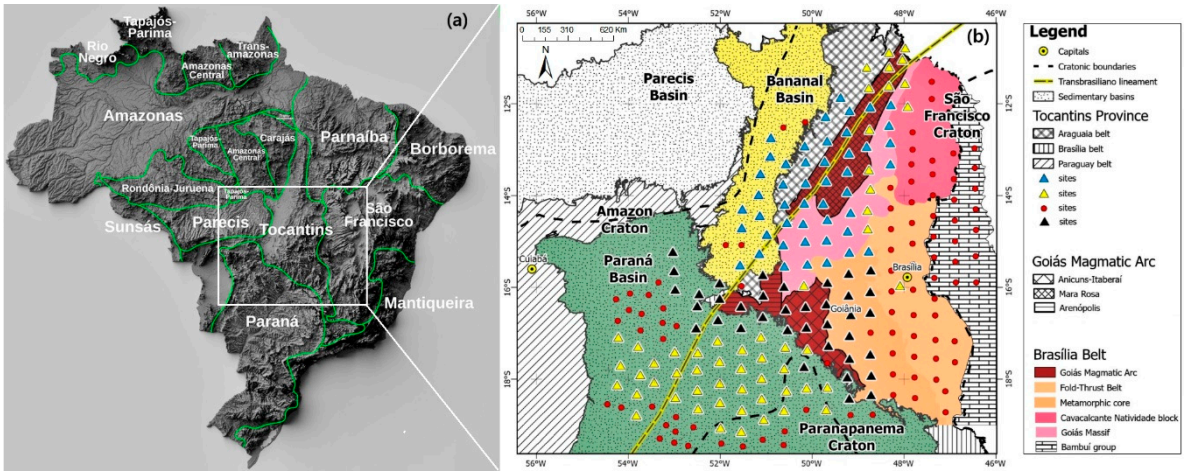

Figure 1. Geographical location of the study region. (a) Simplified geological map of the Brazilian Shield showing major provinces and its spatial distribution. Brazilian geological provinces: Transamazonas Province, Carajás Province, Amazonas Central Province, Tapajós-Parima Province, Rondônia-Juruena Province, Rio Negro Province, Sunsás Province, São Francisco Shield, Borborema Province, Tocantins Province, Mantiqueira Province, Amazonas Basin, Parnaíba Basin, Parecis Basin and Paraná Basin. (b) Spatial distribution of exploration sites (n=221).

2.2. Data-Driven Machine Learning Models

This study used multi-source dataset across the Brazilian Shield and integrates geological, geophysical, and geochemical attributes derived from publicly available surveys and exploration records (the Geological Survey of Brazil (GSB) at https://www.sgb.gov.br/) shown in Table 1. By encompassing a broad range of tectonic environments and mineral system types, the study region provides a robust testbed for evaluating machine learning-based predictive models in a real-world mineral exploration context [27]. This geological diversity is particularly well suited for data-driven approaches, as it enables models to learn complex, non-linear relationships between geological features and mineralization potential across varying spatial scales [28].

Table 1. Summary statistics of key geophysical, geochemical, and structural variables across the exploration dataset (n = 221), including mean, standard deviation, quartiles, and range.

Variable	Min	Mean	Max	Std Dev	25% Quartile	75% Quartile
Geophysical Variables						
Magnetic Anomaly (nT)	186762.2	52191.4	89467.7	12433.6	42577.8	61745.3
Gravity Anomaly (mGal)	-67.1	-23.1	22.7	15.5	-34.8	-11.4
EM Conductivity (S/m)	0.10	1.34	4.01	0.72	0.79	1.88
Cu	2.1	422.7	2144.7	317.3	186.5	586.1
Ni	1.2	84.2	455.7	70.3	33.8	122.5
Co	0.4	12.4	47.1	8.6	5.1	16.7
Li	0.6	16.1	87.1	13.7	6.7	23.4
Rare Earth Elements (REE)	3.0	45.2	185.3	31.8	22.7	63.7
Structural Variables						
Fault Distance (km)	0.1	4.1	22.5	3.6	1.4	6.2
Lineament Density (km/km²)	0.02	0.41	1.44	0.32	0.17	0.61
Fracture Intensity	0.01	0.35	1.25	0.23	0.14	0.55
Economic Potential Score	10.2	65.4	94.7	16.3	52.8	80.1

Given the heterogeneous nature of exploration data, a rigorous preprocessing workflow was applied prior to model training. Datasets compiled from heterogeneous sources commonly exhibit missing values, inconsistent scales, and mixed variable types, all of which can adversely affect

 doi:10.20944/preprints202605.1748.v1

machine learning model performance. To address these challenges, a structured preprocessing pipeline was implemented to ensure computational consistency and geological coherence across the dataset [29].

Missing values, corresponding to approximately 5% of the total dataset, were imputed using a K-nearest neighbors (KNN) imputation approach (k = 5), applied within each geological province to preserve regional geological coherence and maintains local geological context procedure reduces distortions that could arise from global imputation strategies [30]. Categorical variables, including lithology, geological formation, and tectonic province, were transformed using one-hot encoding, increasing the dimensionality of the feature space from 35 to 70 variables. Continuous predictors were standardized using z-score normalization, ensuring compatibility across algorithms, especially distance-sensitive models such as KNN [31].

Additional feature engineering was performed to improve the geological relevance of the predictors. Derived variables included: geochemical ratios (e.g., Cu/Ni and K/Th); alteration indices based on spectral proxies; distance to structural discontinuities; fault intersection density; and lithological contact proximity [32]. These variables were selected based on established mineral exploration criteria and geological controls commonly associated with critical mineral systems.

3. Methodology

3.1. Multiple Machine Learning Models, Training and Implementation Details

The method used in this study is described in a previous work [6], and is applied to six machine learning algorithms, which were selected to represent a spectrum of multiple and baseline modeling approaches commonly applied in computational geoscience [33]. The selection enables comparative evaluation of model complexity, predictive performance, and interpretability.

Random Forest (RF) models were implemented as bagging-based ensembles of decision trees with randomized feature selection, providing robustness against overfitting and strong performance on non-linear datasets [34]. XGBoost (XGB) was employed as a gradient boosting framework with regularization to enhance generalization in high-dimensional feature spaces. AdaBoost (AB) was included as an adaptive boosting method emphasizing sequential error correction using weak learners [35]. Decision Trees (DT) served as interpretable baseline models, while K-Nearest Neighbors (KNN) provided a non-parametric comparison sensitive to local data structure. Logistic Regression (LR) was applied as a linear baseline classifier to benchmark ensemble performance against traditional statistical methods [30,36]. Hyperparameter tuning was performed through grid search combined with five-fold cross-validation, ensuring a fair comparison among all models. Random Forest and XGBoost were chosen as the primary ensemble models due to their strong ability to capture complex non-linear relationships and handle high-dimensional datasets [37].

The proposed modeling framework was structured to address two interdependent objectives within a unified workflow: (i) the classification of exploration data quality into discrete ordinal categories (Excellent, Good, Fair, and Poor), and (ii) the prediction of economic mineral potential as a continuous variable. To ensure representativeness across different geological domains, the dataset was divided into training (80%) and testing (20%) subsets using stratified sampling based on geological provinces. This approach minimizes spatial bias and improves model generalization [37].

Model training and hyperparameter optimization were conducted using five-fold cross-validation on the training set. Regression performance was evaluated using R^2, adjusted R^2, root mean square error (RMSE), and mean absolute error (MAE), while classification performance was assessed using overall accuracy, macro-averaged precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Regression-based estimation of economic potential scores on a continuous scale from 0 to 100. By structuring the workflow in this manner, the framework enables simultaneous evaluation of predictive performance and data reliability, thereby embedding uncertainty quantification directly into the modeling process [38].

 doi:10.20944/preprints202605.1748.v1

3.2. Model Interpretability

To ensure that model output remains consistent with geological reasoning, multiple interpretability techniques were applied. Feature importance was extracted from tree-based models using impurity-based metrics, while coefficient-based interpretation was used for linear models. In addition, partial dependence analysis was employed to explore non-linear relationships between key predictors and model outputs. SHAP (SHapley Additive exPlanations) values were also calculated to quantify the contribution of individual features at both global and local levels. This approach allows for a more transparent interpretation of model behavior and facilitates the comparison between data-driven results and established geological knowledge [8].

All analyses were implemented in Python using widely adopted scientific computing libraries, including scikit-learn, XGBoost, pandas, and NumPy. Visualization and diagnostic procedures were conducted using matplotlib and seaborn [35]. To ensure reproducibility, all preprocessing steps, model configurations, and training procedures were systematically documented and version controlled. The computational workflow was designed to be easily transferable to other exploration datasets.

4. Results

The comparative analysis shown in Table 2 reveals clear differences in performance among the evaluated machine learning models across both regression and classification tasks. Ensemble-based approaches consistently outperform simpler baseline models, confirming their effectiveness in handling high-dimensional and non-linear geoscientific data. Among all tested methods, XGBoost and Random Forest demonstrate the most robust and stable behavior. Forest demonstrate the strongest and most stable performance, while simpler models such as Decision Trees, K-Nearest Neighbors, and Logistic Regression provide useful but comparatively limited predictive capability. According to the analysis results, the evaluated machine learning models exhibit distinct performance profiles across the dual tasks of economic potential prediction and exploration data quality assessment. This performance hierarchy remains consistent across both cross-validation and independent test datasets, suggesting that the observed results reflect intrinsic algorithmic capabilities rather than artifacts of data partitioning.

Table 2. Comparative performance metrics for all machine learning models across regression (economic potential prediction) and classification (data quality assessment) tasks on the test dataset.

Model	Task	R^2/Accuracy	Adj R^2/Precision	RMSE/Recall	MAE/F1	CV Mean	CV Std	AUC-ROC
Linear Regression	Regression	0.21	0.08	8.80	6.91	0.27	0.03	-
KNN	Regression	0.09	-0.04	9.40	7.44	0.22	0.04	-
Decision Tree	Regression	0.42	0.39	7.20	5.62	0.48	0.04	-
AdaBoost	Regression	0.58	0.56	6.81	5.31	0.64	0.04	-
Random Forest	Regression	0.64	0.63	6.10	4.77	0.72	0.04	-
XGBoost	Regression	0.76	0.77	5.50	4.30	0.82	0.03	-
KNN	Classification	0.66	0.66	0.66	0.66	0.665	0.02	0.72
Logistic Regression	Classification	0.72	0.70	0.70	0.71	0.70	0.03	0.75
Decision Tree	Classification	0.74	0.73	0.73	0.73	0.73	0.03	0.77
AdaBoost	Classification	0.77	0.77	0.76	0.76	0.76	0.02	0.81
XGBoost	Classification	0.80	0.80	0.80	0.78	0.78	0.02	0.85
Random Forest	Classification	0.82	0.82	0.81	0.80	0.80	0.02	0.85

According to the analysis results shown in Figures 2 and 3 for regression-based estimation of economic potential scores, XGBoost achieves the highest predictive accuracy, with an R^2 value of 0.76 and RMSE of 5.50. Random Forest produced comparable results, although with slightly lower accuracy. These findings indicate that gradient-boosted and bagged tree are particularly effective at capturing the complex relationships among geological, geophysical, and geochemical variables that control mineralization potential. In contrast, baseline models, including single decision trees and

 doi:10.20944/preprints202605.1748.v1

linear regression, exhibit reduced predictive capacity, highlighting the importance of using more sophisticated approaches in complex geological settings. Model performance remained relatively stable across the Brazilina Shield, although slight reductions in accuracy were observed in areas characterized by lower data density or increased geological complexity. This suggests that the models generalize well within the bounds of the available feature space, while still reflecting underlying data limitations [39]. Magnetic anomalies play a key role in identifying structural and lithological controls, while copper concentrations and alteration signatures are well-known indicators of mineralization processes [40].

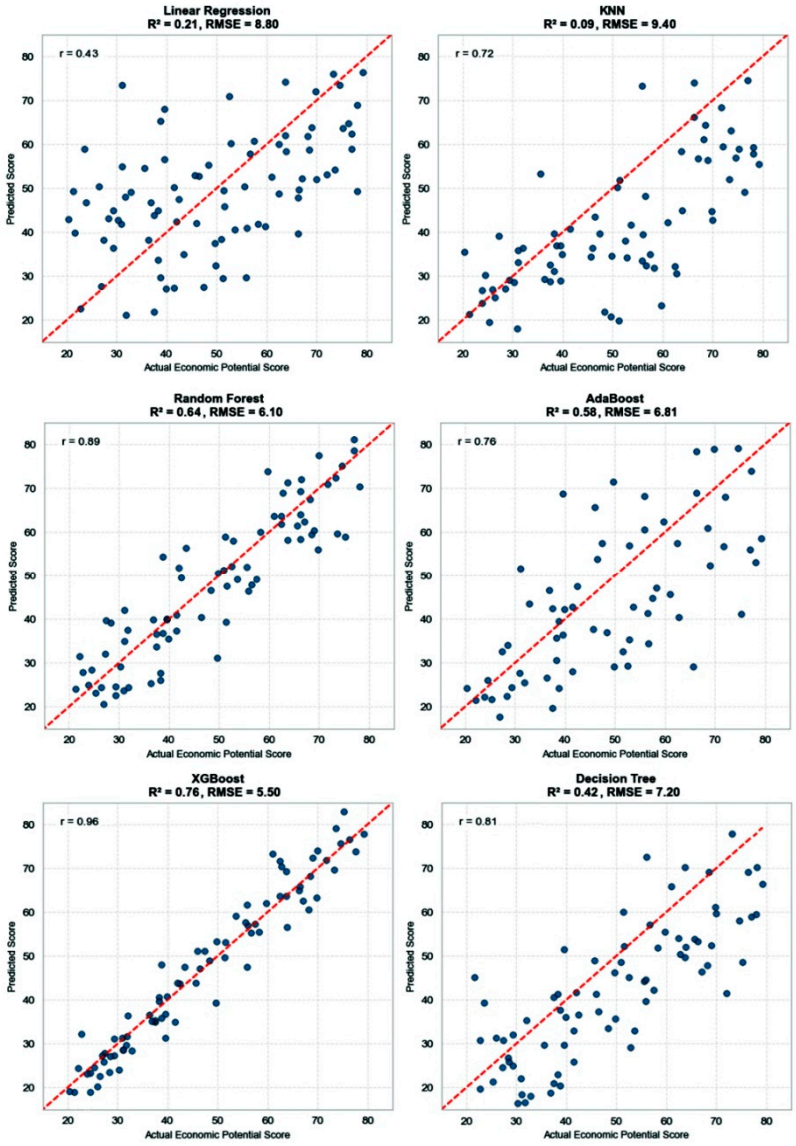

Figure 2. Observed versus predicted economic potential scores for the best-performing regression models. The dashed line represents perfect prediction.

 doi:10.20944/preprints202605.1748.v1

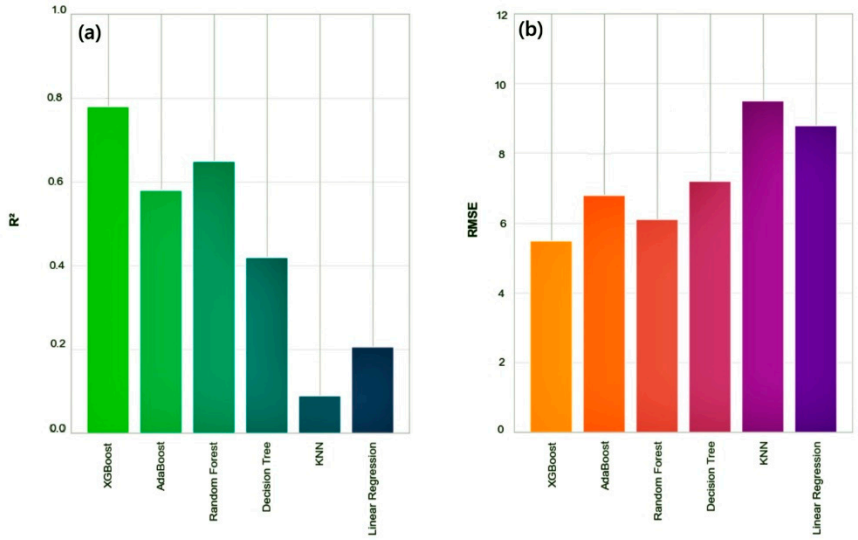

Figure 3. Comparison of six model accuracy. (a) R^2; (b) RMSE.

For multiclass classification of exploration data quality, Random Forest achieved the best overall performance, reaching an accuracy of 0.82 and an AUC of 0.85 (Figure 4). XGBoost produced similar results, reinforcing the reliability of blended approaches for categorical prediction problems. Confusion matrix analysis indicates that the models are highly effective at distinguishing between high-quality and low-quality datasets. Misclassifications tend to occur primarily between adjacent classes, which is expected given the gradual and continuous nature of data quality. This behavior is particularly important from an operational perspective, as it minimizes the risk of incorrectly assigning high confidence to unreliable data. Intermediate classes exhibit higher classification uncertainty, reflecting the inherently continuous nature of data quality. However, multiple models successfully capture broad reliability trends, enabling effective pre-assessment of data confidence prior to detailed interpretation or additional data acquisition.

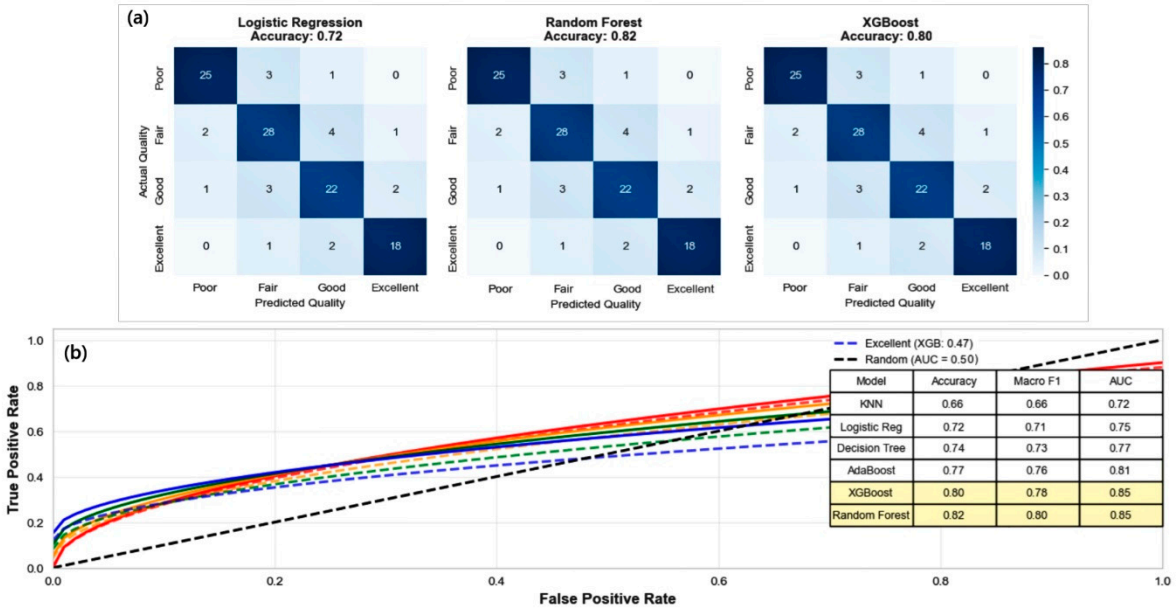

Figure 4. (a) confusion matrix and (b) receiver operating characteristic (ROC) curves for data quality classification using the Random Forest model.

 doi:10.20944/preprints202605.1748.v1

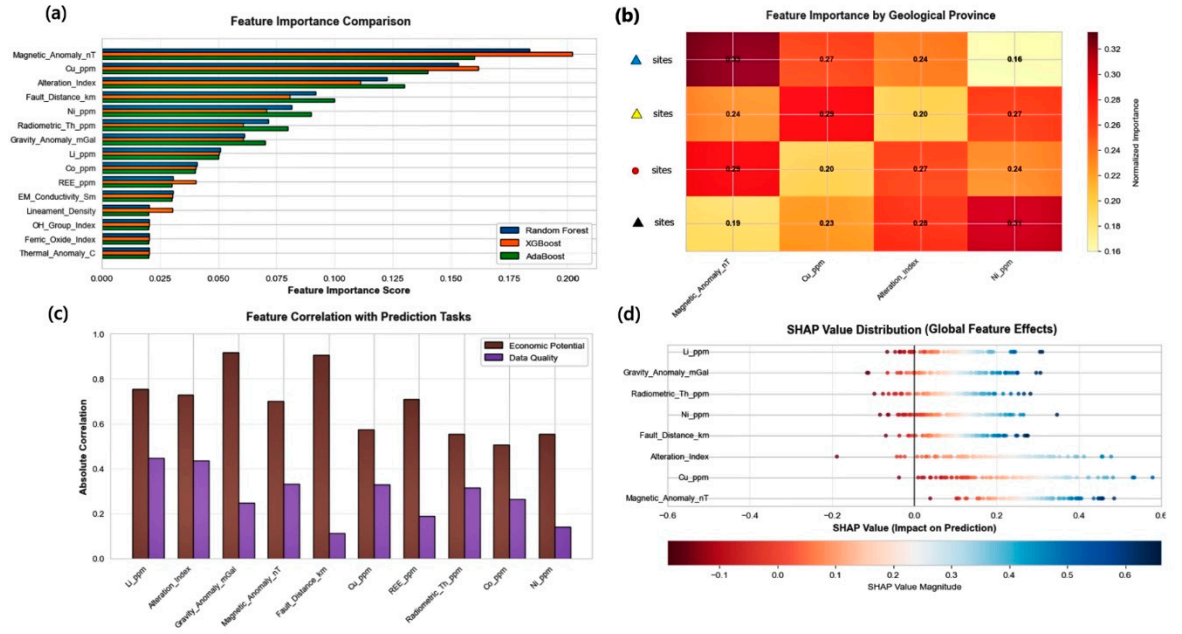

Figure 5. (d) global SHAP feature importance summary showing the relative contribution of key (a) geophysical, (b) geological, and (c) geochemical variables to model predictions.

Feature importance analysis across different models reveals a consistent hierarchy of predictive variables. Magnetic anomaly intensity, copper concentration, alteration indices, and structural proximity variables emerge as the most influential features in both economic potential prediction and data quality classification (prediction tasks). Secondary predictors include lithologically conditioned geochemical ratios and radiometric parameters, reflecting the multivariate nature of mineral systems (Figure 5).

SHAP-based explainability analyses confirm these findings and provide additional insight into non-linear feature interactions (Figure 5). High magnetic anomalies and elevated copper concentrations are positively associated with increased economic potential, while inconsistencies in geochemical and geophysical signals are linked to lower data quality classifications. Importantly, feature importance rankings remain broadly stable across geological provinces, suggesting transferable predictive patterns within Precambrian shield environments.

5. Discussion

The superior performance of multiple methods observed in this study can be attributed to their ability to represent complex, non-linear interactions characteristic of mineral exploration datasets and reduce overfitting through aggregation mechanisms. XGBoost benefits from its gradient boosting architecture, which iteratively minimizes prediction errors [34,35]. This makes it especially effective for continuous prediction tasks such as economic potential estimation. Random Forest, on the other hand, demonstrates greater robustness to data variability and noise, which explains its superior performance in data quality classification [7,12]. These differences highlight the importance of selecting algorithms according to the specific objectives of the analysis. From a computational perspective, these properties are essential for geoscientific applications where predictor variables are often correlated, noisy, and unevenly distributed across geological domains.

The feature of importance hierarchies identified across multiple models exhibit strong consistency with established geological understanding of mineral systems in Precambrian shield environments [23,27]. Importantly, the models identify these relationships independently of explicit geological rules, providing quantitative validation of empirically derived exploration criteria. This convergence between data-driven inference and geological reasoning enhances confidence in the

 doi:10.20944/preprints202605.1748.v1

interpretability and operational relevance of machine learning outputs, addressing a common critique of black-box modeling approaches in geoscience [29].

Variations in model performance across geological provinces underscore the dependency of machine learning algorithms on training data representativeness [32]. Regions with dense historical exploration and well-constrained geological frameworks yield higher predictive accuracy, while structurally complex or underexplored terrains present greater challenges. This behavior is consistent with the fundamental principle that machine learning models extrapolate from observed patterns rather than infer new geological processes [3].

Machine learning models are inherently dependent on observed patterns and may struggle in regions where data are sparse or poorly constrained [32]. Nevertheless, the relatively stable performance observed across Brazilian Shield suggests that multiple models are capable of providing useful insights even under imperfect data conditions. However, the persistence of operationally useful performance in data-sparse settings suggests that multiple models can support early-stage targeting and data quality screening even under imperfect information. Iterative refinement through targeted data acquisition, informed by initial model outputs, offers a practical pathway for progressively improving model reliability.

The findings of this study extend beyond the specific case of the Tocantins Province to broader applications in computational geoscience. The dual-task framework provides a template for integrating predictive modeling and uncertainty characterization in other subsurface exploration contexts, including energy resources, groundwater assessment, and environmental monitoring. From an exploration strategy perspective, the ability to jointly evaluate potential and reliability supports more sophisticated portfolio management, enabling prioritization based on both expected reward and associated risk. Critically, the framework emphasizes the complementary relationship between machine learning and geological expertise. Rather than replacing traditional interpretation, ensemble models function as scalable analytical tools that synthesize multivariate data and highlight patterns warranting further geological investigation.

Despite its advantages, model performance remains constrained by the quality, diversity, and representativeness of available training data [32]. Unobserved variables and previously unknown mineralization processes cannot be captured by the models. Future work may address these limitations through the integration of additional data types, including unstructured geological information and temporal datasets. The incorporation of adaptive learning strategies and hybrid modeling approaches represents a promising direction for further research.

5. Conclusions

This study demonstrates that multiple machine learning provides an effective and interpretable computational framework for addressing two interrelated challenges in mineral exploration: economic potential prediction and exploration data quality assessment. By integrating economic potential prediction and data quality assessment within a single framework, the proposed methodology advances beyond conventional workflows that treat data reliability as a secondary concern. The main conclusions are as follows:

(1) Comparative evaluation of six machine learning algorithms shows that XGBoost achieves the highest predictive accuracy for continuous economic potential estimation, while Random Forest provides superior performance and stability for multiclass data quality classification. These results highlight the importance of algorithm selection based on task-specific objectives and data characteristics rather than reliance on a single modeling approach. Importantly, multiple methods consistently outperform baseline models, confirming their suitability for high-dimensional, heterogeneous geoscientific datasets.

(2) Model interpretability analyses reveal that key predictors-including magnetic anomaly intensity, copper concentration, alteration indices, and structural proximity metrics-dominate both prediction and classification tasks. The convergence of these data-driven insights with established geological understanding reinforces the credibility of machine learning outputs and

 doi:10.20944/preprints202605.1748.v1

supports their integration into exploration decision-making processes. Rather than functioning as black-box predictors, the models provide quantitative validation of known exploration indicators while enabling scalable analysis across large datasets. The results show that XGBoost is the most effective model for continuous prediction tasks, while Random Forest provides superior performance in classification problems. Multiple methods consistently outperform simpler approaches, confirming their suitability for complex geoscientific applications.

(3) The application of the framework to the Brazilian Shield illustrates its effectiveness within a geologically complex Precambrian environment characterized by diverse mineralization styles and variable data quality. Feature importance analyses highlight the relevance of magnetic, geochemical, and structural variables, reinforcing the consistency between data-driven results and geological interpretation.

Overall, this work contributes to the advancement of computational geoscience by demonstrating how machine learning can be used not only to predict mineral potential, but also to quantify the reliability of the data that support those predictions.

Author Contributions: Experiment design; experimentation; data analysis; writing, H.A.B has read and agreed to the published version of the manuscript.

Funding: This study received partial funding from Capes through Notice no.28/2022—PDPG Social Vulnerability & Human Rights [Grant Number 88881.705050/2022-01].

Data Availability Statement: The datasets analyzed in this study are derived from publicly available geological, geophysical, and mineral exploration databases provided by Brazilian federal and provincial agencies. The repository with the data sets used in this study is available from the corresponding author upon reasonable request. The machine learning code (Python scripts and Jupyter notebooks) developed for this study is openly available upon reasonable request.

Computer Code Availability: The machine learning code (Python scripts and Jupyter notebooks) developed for this study is openly available upon reasonable request.

Acknowledgments: The author thanks the anonymous reviewers and editor for their valuable feedback and recommendations.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1
2. Batchelor, T., Grunsky, E.C. Robust multivariate analysis of geochemical data. *Computers & Geosciences* **2016**, 96, 111–123. https://doi.org/10.1016/j.cageo.2016.08.002
3. Cracknell, M.J., Reading, A.M. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences* **2014**, 63, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008
4. Carranza, E.J.M. Geochemical anomaly and mineral prospectivity mapping in GIS. *In: Handbook of Exploration and Environmental Geochemistry* **2009**, vol. 11. Elsevier, Amsterdam.
5. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. IEEE *Geoscience and Remote Sensing Magazine* **2016**, 4, 22-40, doi:10.1109/MGRS.2016.2540798
6. Barbosa, H. A.; Buriti, C. O.; Kumar, T. L. Deep learning for flash drought detection: a case study in northeastern Brazil. *Atm sphere* **2024**, v. 15, n. 7, p. 761. DOI: 10.3390/atmos15070761.
7. Breiman, L., 2001. Random forests. *Machine Learning* **2001**, 45, 5–32. https://doi.org/10.1023/A:1010933404324

8. Lundberg, S.M., Lee, S.I. A unified approach to interpreting model predictions. A*dvances in Neural Information Processing Systems* **2017**, 30, 4765–4774.
9. Molnar, C. Interpretable Machine Learning. *2nd ed. CRC* **2022**, Boca Raton, FL.
10. Schodde, R. The discovery cost curve: Implications for exploration strategy. *Minerals Engineering* **2014**, 68, 70–80. https://doi.org/10.1016/j.mineng.2014.02.004
11. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M. Graph neural networks: A review of methods and applications. *AI Open 1* **2021**, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001
12. Rodríguez-Galiano, V., Sánchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M.. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews* **2015**, 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001
13. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1
14. Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences* **2014**, 63, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008
15. Prodhan, F.A.; Zhang, J.; Hasan, S.S.; Pangali Sharma, T.P.; Mohana, H.P. A review of machine learning methods for drought hazard monitoring and forecasting: Current research trends, challenges, and future research directions. *Environmental Modelling & Software* **2022**, 149, 105327, doi:https://doi.org/10.1016/j.envsoft.2022.105327.
16. Felsche, E.; Ludwig, R. Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations. *Natural Hazards and Earth System Sciences* **2021**, 21, 3679-3691, doi:10.5194/nhess- 529 21-3679-2021
17. Thurston, P.C., Ayer, J.A., Goutier, J., Hamilton, M.A., 2008. Depositional gaps in Abitibi greenstone belt stratigraphy: A key to exploration for syngenetic mineralization. *Economic Geology* **2008**, 103, 1097–1134. https://doi.org/10.2113/gsecongeo.103.6.1097.
18. Morellato, L. P. C., Camargo, M. G. G., Gressler, E. A Review of Plant Phenology in South and Central America. In: *Phenology: An Integrative Environmental Science* **2013**, 91-113. https://doi.org/10.1007/978-94-007-6925-0_6
19. Grunsky, E.C.. The interpretation of geochemical survey data. *Geochemistry: Exploration, Environment, Analysis* **2010**, 10, 27–74. https://doi.org/10.1144/1467-7873/09-219
20. Silva G.F., Silva A.D.R., Souza Gaia S.M. An overview of critical and strategic minerals of Brazil. *An overview of critical and strategic minerals potential of Brazil* **2024**. Brasília, DF, Geological Survey of Brazil, 35 p. Available online at: https://rigeo.sgb.gov.br/ handle/doc/24748 / (accessed on 18 February 2026).
21. Kenkmann, T., Vasconcelos, M. A. R., Crósta, A. P., Reimold, W. U, The complex impact structure Serra da Cangalha, Tocantins State, *Brazil: Meteoritics & Planetary Science* **2011**, 46 (6): 875 – 889. DOI: 10.1111/j.1945-5100.2011.01199.x
22. Almeida, F.F.M., Hasui, Y., Brito Neves, B.B., and Fuck, R.A, Brazilian Structural Provinces: An introduction. *Earth – Science Reviews* **1981**, 17: 1 – 29. DOI: 10.1016/0012-8252(81)90003-9.
23. Frasca, A.A.S., Lima H.A.F., Moraes, L.L., and Ribeiro, P.S.E. Geologia e recursos minerais da Folha Gurupi – SC.22 – Z – D. Estado do Tocantins. Scale 1:250.000, Final Report: CPRM: Goiânia, Brazil, **2010**, 180 pp. Available online at: https://rigeo.cprm.gov.br/handle/doc/10907 (accessed on 05 Abril 2026).
24. Martins-Ferreira, M.A.C., Campos, J.E.G., Von Huelsen, M.G., and Neri, B.L, Paleorift structure constrained by gravity and stratigraphic data: The Statherian Araí rift case: *Tectonophysics* **2018b**, 738 – 739: 64 - 82. DOI: 10.1016/j.tecto.2018.05.014.
25. Soares, J.E.P., Stephenson, R., Fuck, R.A., Lima, M.V.A.G., Araújo, V.C.M., Lima, F.T., Rocha, F.A.S., and Trindade, C.R. Structure of the crust and upper mantle beneath the Parnaíba Basin, Brazil, from wide-angle reflection data: In: Daly, M.C., Fuck, R.A., Julià, J., MacDonald, D.I.M., and Watts, A.B. (eds) Cratonic Basin

 doi:10.20944/preprints202605.1748.v1

Formation: A case study of the Parnaíba Basin of Brazil: *Geological Society, London, Special Publication*, **2018**, 472. DOI: 10.1144/SP472.9

26. Brasil Mineral, 2024a. Mineração Serra Verde inicia produção comercial em Minaçu. *Brasil Mineral*, 11 jan. **2024**. Available online at: https:// www.brasilmineral.com.br/noticias/mineracao-serra-verde-iniciaproducao-comercial-em-minacu / (accessed on 16 February 2026).
27. Ribeiro, P.S.E., Frasca, A.A.S., Carneiro, J.S.M., Hattingh K., Rezende, E.S., Martins, F.R. Mapa geológico e de recursos minerais do Estado do Tocantins, Escala 1:500.000: Maps and GIS files, *CPRM*, **2022** Goiânia, Brazil. Available online at: https://rigeo.sgb.gov.br/handle/doc/22530 (accessed on 26 December 2025).
28. Brown, W.M., Gedeon, T.D., Groves, D.I., Barnes, R.G. Artificial neural networks: A new method for mineral prospectivity mapping. *Australian Journal of Earth Sciences* **2000**, 47, 757–770. https://doi.org/10.1046/j.1440-0952.2000.00807.x
29. Lundberg, S.M., Lee, S.I. A unified approach to interpreting model predictions. A*dvances in Neural Information Processing Systems* **2017**, 30, 4765–4774.
30. Cover, T.M., Hart, P.E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **1967**, 13, 21–27. https://doi.org/10.1109/TIT.1967.1053964
31. Chen, Y., Wu, W. 2017. Application of one-class support vector machine to quickly identify multivariate anomalies from geochemical exploration data. *Geochemistry: Exploration, Environment, Analysis* **2017**, 17, 231–238. https://doi.org/10.1144/geochem2016-013
32. Dickson, B.L., Scott, K.M. Interpretation of aerial gamma-ray surveys adding the geochemical factors: *Journal of Australian Geology & Geophysics*, **1997**, 17 (2): 187 – 200.
33. Chen, Y., Wu, W.. Application of one-class support vector machine to quickly identify multivariate anomalies from geochemical exploration data. *Geochemistry: Exploration, Environment, Analysis* **2017,** 17, 231–238. https://doi.org/10.1144/geochem2016-013.
34. Breiman, L. Random forests. *Machine Learning* **2001**, 45, 5-32, doi:10.1023/a:1010933404324.
35. Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 785–794. https://doi.org/10.1145/2939672.2939785
36. Sittaro, F.; Hutengs, C.; Semella, S.; Vohland, M. A Machine Learning Framework for the Classification of Natura Habitat Types at Large Spatial Scales Using MODIS Surface Reflectance Data. *Remote Sensing* **2022**, 14, 823, 540 doi:10.3390/rs14040823
37. Freund, Y., Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. J*ournal of Computer and System Sciences* **1997**, 119–139. https://doi.org/10.1006/jcss.1997.1504
38. L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller. A unifying review of deep and shallow anomaly detection, *Proc. IEEE* **2021**, vol. 109, no. 5, pp. 756–795.
39. Cheng, M.; Zhong, L.; Ma, Y.; Wang, X.; Li, P.; Wang, Z.; Qi, Y. A New Drought Monitoring Index on the Tibetan Plateau Based on Multisource Data and Machine Learning Methods. *Remote Sensing* **2023**, 15, 512, doi:10.3390/rs15020512
40. Paixão, M. A. P., Nilson, A. A., and Dantas, E. L. The Neoproterozoic Quatipuru ophiolite and the Araguaia fold belt, central-northern Brazil, compared with correlatives in NW Africa: In: Pankhurst, R. J., Trouw, R. A. J., Brito Neves, B. B., and Wit, M. J.. (eds.). West Gondwana: Pre-Cenozoic Correlations Across the South Atlantic Region. Bath, UK: *The Geological Society Publishing House* **2008**, 294: 297-318. DOI: 10.1144/SP294.16.