Geologia USP
Série Científica

https://doi.org/10.11606/issn.2316-9095.v25-222751
Revista do Instituto de Geociências - USP
Geol. USP, Sér. cient., São Paulo, v. 25, n. 4, p. 99-108, December, 2025

# Supervised learning for categorization of geological domain

*Aprendizagem supervisionda para categorização de domínios geologicos*

**Marcela Martins[1], Marcelo Monteiro da Rocha[1], Camila Duelis Viana[1],**

[1] Universidade de São Paulo, Instituto de Geociências, Rua do Lago, 562, Cidade Universitária, CEP 05508-080. São Paulo, SP, BR.
(marcela2.martins@usp.br; mmrocha@usp.br; camila.viana@usp.br)

Received on March 12, 2024; accepted on December 15, 2025.

## ABSTRACT

The geological model needs to be updated when new drill holes are added. This requires time and knowledge of the mineral deposit, as the new samples need to be categorized according to the relevant geological properties. This study employed six supervised machine learning algorithms (naive Bayes, k-Nearest Neighbors (kNN), Support Vector Machines (SVM), decision trees, random forest, and neural networks) to perform geological interpretation of a gold deposit of the metasedimentary type, located in the east-central region of the state of Bahia, with the objective of evaluating the ability of these algorithms to accurately classify the geological domains of new samples in a database. The results indicated that the random forest and neural networks algorithms successfully reproduced the geological interpretation of the mineral deposit, as the models generated were similar to those manually created by a geologist. This great performance is supported by the accuracies and precision obtained, which were 0.87 and 0.89, respectively. Therefore, it is recommended to use these algorithms for optimizing the geological model update process. However, the naive Bayes, kNN, SVM, and decision tree algorithms failed to categorize the geological domains of the samples correctly. As a result, some regions in the models exhibited distorted geological layers and a lack of statigraphic order due to interpretation erros. This is evidenced by the low accuracies and precision obtained-0.48, 0.73, and 0.75. Therefore, these algorithms are unsuitable for this task.

**Keywords:** Machine Learning; Supervised Learning; Geological Model; Gold Deposit.

## RESUMO

Atualização de modelo geológico quando novos furos são inseridos no modelo requer tempo e conhecimento do depósito mineral, pois as novas amostras precisam ser categorizadas de acordo com os domínios geológicos pertencentes. Este trabalho utilizou seis algoritmos de aprendizagem de máquina supervisionado (*naive bayes*, k-vizinhos mais próximos (kNN), *support vector machine* (SVM), árvore de decisão, *random forest* e rede neural) para realizar a interpretação geológica de um depósito de ouro do tipo metassedimentar localizado na região centro-leste do estado da Bahia, tendo com objetivo verificar a capacidade desses algoritmos em classificar corretamente os domínios geológicos das novas amostras de um banco de dados. Os resultados mostraram que os algoritmos de *random forest* e rede neural conseguiram reproduzir satisfatoriamente a interpretação geológica do depósito mineral, pois os modelos gerados foram semelhantes ao realizado manualmente por um geólogo. Esse bom desempenho é confirmado pelas acurácias e precisão obtidas, de 0,87 e 0,89, respectivamente. Portanto estes algoritmos são indicados para otimizar o processo de atualização de modelos geológicos. Porém, os algoritmos de *naive bayes,* kNN, *support vector machine* e árvore de decisão, não conseguiram categorizar corretamente os

domínios geológicos das amostras, com isso, algumas regiões nos modelos apresentaram camadas geológicas distorcidas e a ordem estratigráfica não foi respeitada devido aos erros na interpretação, fato evidenciado pelas baixas acurácias e precisão obtidas, sendo 0,48, 0,73 e 0,75. Portanto, estes algoritmos não são indicados para realizar esta tarefa.

**Palavras-chave:** Apredizado de máquina; Aprendizado Supervisionado; Modelo Geólogico; Depósito de Ouro.

## INTRODUCTION

The geological model is of great importance for the study of the economic viability of a mining venture, since it is through this model that mineral resources, choice of mining method, assessment of operating costs, profit forecast and others are estimated. Therefore, success is directly related to the accuracy and precision of the geological model (Abzalov, 2016).

The geological model consists of the graphical representation of shape, volume and extent of the different lithologies of a mineral deposit. This model is created based on information obtained from boreholes and/or channels, geological mapping and geophysical data. It is the result of the geological interpretation of the mineral deposit. It is worth noting that channel sampling refers to the continuous collection of samples along a defined line on the exposed surface of the mine working face, with the purpose of representing the lithology and grade of mineralization.

In geological modeling, resource domains are defined based on similar characteristics, such as lithological, mineralogical, structural, and others (Rossi e Deutsch, 2014). Each geological domain presents a unique statistical population, that is, it presents a distribution model and specific semivariogram. This uniqueness allows the estimation and/or simulation of contents to be made differently for each modeled geological domain (Rasera, 2014). In summary, it assumes resource domains are second-order stationary, that is, it admits that the mean and covariance are constant within a domain (Armstrong, 1998).

Currently, two types of methods are used for geological modelling, explicit and implicit (Mao et al., 2020). The first method is conventional and consists of the following steps: first, the lithological boundaries are manually interpreted through the construction of polygons in sections, following the connection of the polygons created with the help of a guideline. Finally, an interpolation by triangulation is performed to generate 3D solids that correspond to the geological domains (Cowan et al., 2003).

Therefore, the result is based on the interpretation and geological knowledge of the geomodeler, and, as a result, several models can be created using the same database due to differences in interpretation. This type of modelling is more and time-consuming than implicit modelling (Cowan et al., 2003).

In implicit models, the generation of geological boundaries occurs automatically through mathematical interpolation functions. The process typically involves categorizing samples by domain, selecting the appropriate algorithm and parameters, and generating the geological surfaces by interpolating the sample data. Common interpolation techniques used in implicit modeling include radial basis functions (RBF), distance functions and signed distance functions. The geological knowledge of the deposit is fundamental in this type of modelling, as the geomodeler defines the model parameters, such as the interpolation method, influence range, and surface resolution, with the objective of adequately representing the geological boundaries and characteristics of the deposit.

The advantages of implicit modeling are reproducibility, speed and objectivity, as the algorithm automatically performs the interpolation of geological contacts through mathematical functions (Cowan et al., 2003). Regardless, it is of utmost importance that the geomodeler has a geological knowledge of the mineral deposit, as improper use of the tools of this method and erroneous categorization of the samples will result in a model that lacks generalization.

The process of updating geological models when new samples are entered into the database is still a process that requires technical knowledge and commitment of the geologist in implicit modeling. A possible way to streamline this process is the use of machine learning algorithms to determine the geological domains that the new samples belong to, since these algorithms may be able to identify the characteristics that define the domains.

Machine learning refers to the area of computer science where computers recognize patterns in data to make predictions about them (Samuel, 1959). In scientific literature, there are examples which prove machine learning efficiency in geological modelling, such as Zhang et al. (2023) who employed supervised machine learning algorithms, including support vector machines (SVM), k-nearest neighbors (kNN), random forest, AdaBoost, naïve Bayes, and artificial neural networks, to automatically delineate geo-

domain boundaries from a gold deposit in the Witwatersrand Basin, South Africa. The results indicated that, although several algorithms achieved good statistical performance, the SVM with a radial basis function (RBF) kernel was the most suitable for this application, as it produced simpler and more geologically realistic domain boundaries. Consequently, the authors demonstrated that the use of machine learning techniques for geodomain delineation is feasible, replicable and capable of reducing the subjectivity associated with traditional approaches.

Supervised learning is an area of machine learning that uses observed data (input data) with their respective classes (labels), that is, uses the observed data together with the classes to train the algorithm to build a model that represents the relationship between the input data and the output data. Thus, it is possible to predict the classes of new inputs. (Awad and Khanna, 2015).

This work aims to use supervised machine learning algorithms to assist in updating the geological model of a gold deposit; more precisely, to determine what geological domains are to which new samples entered the databases belong.

## CASE STUDY

The data for this study come from a gold mine. The studied deposit is in the east-central region of state of Bahia, within the Jacobina Basin, with comprises a Paleoproterozoic-age metasedimentary package that crops out in a series of N-S trending ridges, approximately 200 km in length and 8 to 10 km in width (Garayp e Frimmel, 2023). More detailed information regarding the exact location of the mine and specific aspects of the geological context has been omitted for confidentiality reasons. Therefore, all data related to geographic coordinates has been intentionally modified to preserve the confidentiality of the location.

The database consists of 673 boreholes and 4816 channels, containing 206304 samples: 176078 are diamond drill holes core, 30224 are channel samples (collected systematically from underground mining faces using continuous channel sampling, which involves cutting a uniform, linear groove along the rock face to obtain a representative sample of the exposed rock). The data extracted from this database for the development of this work included geographical coordinates, gold content (expressed in ppm), lithology, and the corresponding geological domain for each sample.

The lithology information corresponds to the rock nomenclature. The conglomerate is named based on the size and predominance order of pebbles found within the rock. Each pebble size is represented by an acronym, with the conglomerate nomenclature being composed of these acronyms, followed by 'PC,' meaning ‘pebble conglomerate’. The terminologies adopted for the pebbles are as follows, table 1.

To demonstrate how the nomenclature is made, follows an example: if the rock consists of 50% 5 mm pebble, 30% 35 mm pebble and 20% 70 mm pebble, then the rock nomenclature will be ‘SLVLPC’. Quartzite is designated as ‘qto’ in its nomenclature, and when it has pebbles, the size of the pebble is added to the end of the name. For example, quartzite with pebbles smaller than 4 mm will be designated ‘qto-s’.

**Table 1.** Pebble size classes used for conglomerate nomenclature.

| Size class | Acronym | Pebble size (mm) |
|---|---|---|
| Very Small | VS | <4 |
| Small | S | 4-16 |
| Medium | M | 16-32 |
| Large | L | 32-64 |
| Very Large | VL | >64 |

As this work aims to use supervised learning algorithms for the geological interpretation of new samples, it is required to have information about the classes that, in this study, represent the geological domains. This data are obtained through the geological interpretation of the mineral deposit, that is, the geological model. Then, it was necessary to validate the geological model of the mineral deposit to verify if any area needs to be reinterpreted. This step aimed to confirm that the model stratigraphy was correct, ensuring that geological domains were positioned in the proper stratigraphic order. The validation also involved checking whether the geometry of the geological bodies was coherent with the geological context and whether the dip and strike directions of the domains agreed with the geological mapping data. This validation, as well as the database validation, is extremely important for a good performance of the algorithms because the results must reflect the existing patterns in the input data, therefore wrong instances and classes can erroneously distort the output model. Thus, it is essential to ensure the accuracy of the instances and classes to obtain good results. This included verifying the spatial coordinates of all samples and boreholes, confirming whether the drill holes were correctly positioned on the surface topography or at valid underground drilling locations, and ensuring that lithological records were consistent with the expected geology.

### Class Imbalance

The geological model of the studied deposit comprises 11 geological domains, which are defined according to the lithological characteristics, such as rock type, pebble size, packaging, thickness and others. In the stratigraphy of the deposit, there is intercalation of quartzite domains with conglomerate domains, and mineralization occurs in all conglomerate domains and two quartzite domains. The definition of the geological domains of this deposit is a very complex task since the deposit is composed exclusively of conglomerates and quartzites and presents a large number of faults that displace the layers. Thus, the geological knowledge of the mineral deposit is of great importance for modeling the ore bodies.

The geological domains were numerically coded from 0 to 13, with the categorization being done in ascending order from top to bottom, that is, value 0 represents the top domain, value 1 is at the base of the class 0, and so on. Figure 1 represents a section in the geological model containing all classes which are present. To verify the distribution of the samples in the geological domains, a bar chart was created with quantities of instances by class, represented in Figure 2. Only in the neural network algorithm, the classes were coded through the one-hot encoding method; this method creates a tensor, where the columns represent the classes: when the column value is 1, it indicates the presence of the specific class, and when it is zero, it means the absence of this class, i.e., a Bolean classification.

As shown in Figure 2A, the data set is imbalanced, that is, when classes are formed by different amounts of instances (Chawla et al., 2004). Unbalanced data can impact the performance of the algorithm, as the model generated by unbalanced data may not be able to distinguish the class that has the least number of instances (Chawla et al., 2004). That is, it tends to predict data from a minority class as if it belongs to a majority class. Imbalanced data is inherent to geological databases due to the nature of mining operations, which involve drilling holes and creating channels in areas with a higher likelihood of ore presence. Another factor contributing to data imbalance is the variation in thicknesses of geological domains. Greater thicknesses often yield more samples compared to thinner domains. This can be verified through statistical analysis, as shown in Table 2: while Class 9 exhibits the highest mean value, Class 1 has the largest sample count due to its more significant thickness as compared to Class 9. A statistical analysis was performed for each domain class, as presented in Table 2.

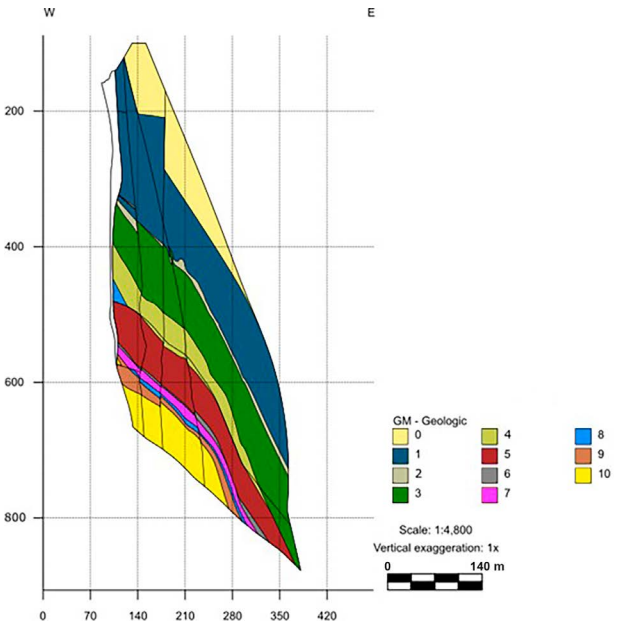


**Figure 1.** Geological section with all geological domain.

It is worth noting that, in this case study, no technique was applied to address class imbalance, as the objective was to preserve the original data distribution and avoid discarding samples from majority classes, which could lead to the loss of important information for the model. Furthermore, the focus of this work is on analyzing the model's behavior in a realistic scenario where imbalanced data is inherent to geological datasets. However, in the training set, an effort was made to maintain the original distribution of the dataset. This can be observed in Figure 2B, which shows the class distribution in the training set, clearly reflecting the same imbalance present in the dataset.

## Supervised training and testing of learning algorithms

Six machine learning algorithms were used: decision tree, kNN, naive Bayes, SVM, neural network and random forest. The input variables are: coordinates, as the sample positions influence geological domain configurations; lithology, as changes in lithologies can alter geological domains; and ore content, which indicates whether the domain is mineralized or not. The variable that represents the class corresponds to the geological domains obtained from the implicit geological model; it is the information that the model given by the algorithm needs to predict. It is worth mentioning that the instances used to train the algorithms are the same information that geologists use to do the modeling in this deposit, those are information the relevant for geological interpretation since the use of non-relevant information can negatively influence the performance of the algorithms, as this information hinders learning (Singh e Singh, 2020) and increases the computational cost (Dougherty, 2013).

The input numerical variables, coordinates and gold content were normalized for the use of neural networks and SVM algorithms. Normalization consists of transforming the raw values into a specific interval, that is, variables will provide a uniform contribution to the algorithm, thus avoiding larger values will mask smaller values (Singh e Singh, 2020). Data normalization also reduces the effects of outliers (Singh e Singh, 2020). It is worth mentioning that data normalization does not always improve the performance of algorithms; thus, it is of great importance to analyze data to decide on its execution (Sing e Singh, 2020). In this work, the data were normalized according to Equation 1 and applied to the neural network and SVM algorithms. Therefore, the data average equals zero, and the variance equals one.

$$z = \frac{(x - \mu)}{\sigma} \quad (1)$$

Where:
$z$: is the Z-Score Normalization;
$x$: is the feature;
$\mu$: is the mean value of that variable and
$\sigma$: is the standardized value of variable.

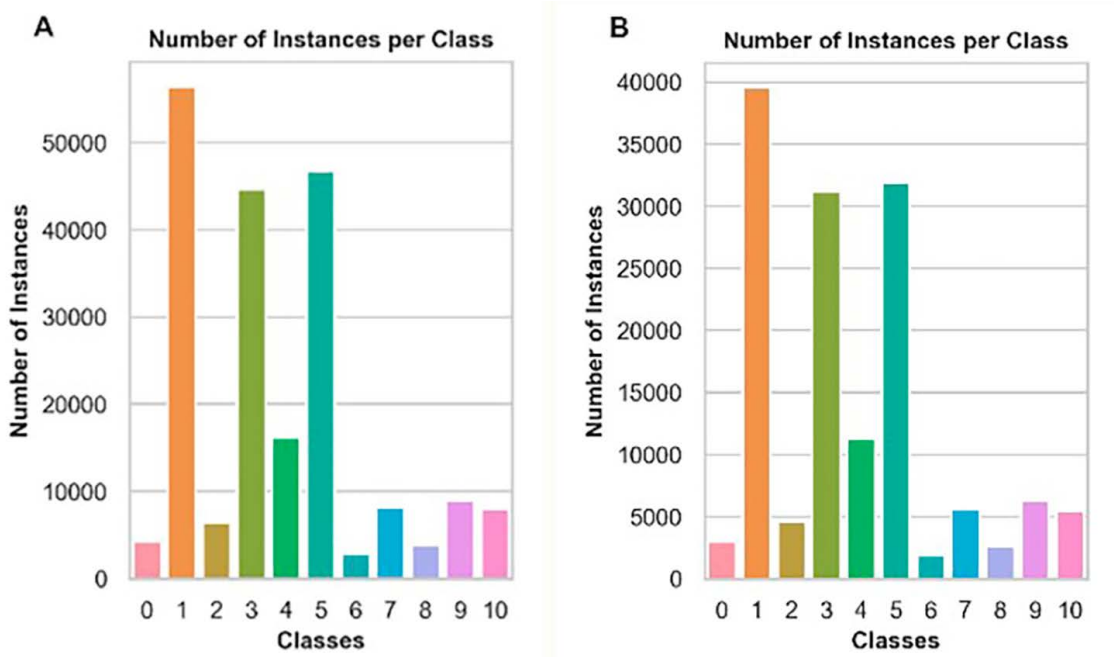


**Figure 2.** Bar chart with sample distribution by class in the dataset (A) and in the training set (B).

When the training and test sets were divided, the proportions of the classes were considered, that is, the classes had similar proportions in the test and training set as result of the class imbalance found in the database. To split the datasets, it was also considered to keep samples of each drill hole in the same set, that is, there cannot be samples of one hole in the test and training set because when geological models are updated, holes are inserted, and not just one core sample. There is a spatially homogeneous distribution of the training and test data. The training set makes up 70% of the database and the test set constitutes 30% of it.

It is important to note that, in this case study, a separate validation set was not used. Instead, the k-fold cross-validation technique was applied, as it is more suitable for small datasets, such as the one used in this study. K-fold cross-validation allows the model to use all samples for both training and validation across different iterations, resulting in a more robust evaluation of model performance. On the other hand, splitting the dataset into training, validation and test sets can reduce the representativeness of each subset, thereby compromising the model´s generalization capability.

## RESULTS

Several tests were performed to define the best parameters to be used in each algorithm. In this work, only the best results obtained for each algorithm will be presented. It is important to note that there was no need to define any parameters for the Naive Bayes algorithm, as it is based on statistical learning. In the case of the kNN algorithm, the best performance was achieved with k=13 using the Euclidean distance as the similarity metric. The best SVM model was obtained using the RBF kernel (Radial Basis Fuction) and a penalty parameter C =1, with one-vs-all strategy for multiclass classification, where a separate model is trained for each class against all others. For the decision tree algorithm, the best results were achieved with a maximum depth of 13, using entropy as the splitting criterion, a minimum of two samples required to split a node, and a minimum of five samples per leaf. The random forest algorithm performed best with a set of 250 decision trees, where each node is split based on at least two samples, and each leaf must contain a minimum of three samples. Finally, the best neural network, of the feedforward type, was composed of two hidden layers: the first with 256 neurons and ReLU activation function, and the second with 128 neurons and tanh activation, followed by an output layer with 11 neurons and softmax activation function, suitable for multiclass classification.

The first performance evaluation of the algorithms was performed through stratified k-fold cross-validation in the training set with 5 folds. The metrics obtained in each algorithm are shown in Tables 3, 4, 5, 6,7 and 8. It is important to note that the metrics presented here were calculated considering the weight of the classes because it guarantees a fairer evaluation, since this work uses an unbalanced data set.

Evaluate the algorithm performance was using the algorithms in the set of tests. Table 9 shows to the mean of the global metrics of each algorithm used in the for each cross-validation fold, computed as the average across all classes.

The performance metrics show that the worst performing algorithm was the naive bayes, and the best performing was the neural network. In the cross-validation, it is

**Table 2.** Statistic analyzes of geological domains.

| | Samples | Mean | Variance | Standard Deviantion | C.V.* |
|---|---|---|---|---|---|
| Class 0 | 4265 | 0.06 | 2.332 | 1.53 | 24.01 |
| Class 1 | 56432 | 0.52 | 6.372 | 2.52 | 4.82 |
| Class 2 | 6421 | 0.21 | 2.449 | 1.57 | 7.414 |
| Class 3 | 44636 | 0.25 | 1.354 | 1.16 | 4.716 |
| Class 4 | 16153 | 0.17 | 5.927 | 2.43 | 14.62 |
| Class 5 | 46738 | 1.03 | 23.95 | 4.89 | 4.735 |
| Class 6 | 2804 | 0.54 | 2.52 | 2.52 | 4.691 |
| Class 7 | 8165 | 1.58 | 4.76 | 4.76 | 3.008 |
| Class 8 | 3773 | 0.75 | 3.18 | 3.18 | 4.255 |
| Class 9 | 8992 | 5.61 | 200.8 | 14.17 | 2.525 |
| Class 10 | 7925 | 0.17 | 3.727 | 1.93 | 11.66 |

C.V.*: is coefficient of Variantion

**Table 3.** Mean class-averaged metrics for each fold in cross-validation for naïve bayes.

| | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|
| Fold 1 | 0.48 | 0.6 | 0.53 | 0.49 |
| Fold 2 | 0.47 | 0.57 | 0.5 | 0.47 |
| Fold 3 | 0.48 | 0.57 | 0.51 | 0.48 |
| Fold 4 | 0.51 | 0.67 | 0.56 | 0.51 |
| Fold 5 | 0.49 | 0.66 | 0.55 | 0.49 |

**Table 4.** Mean class-averaged metrics for each fold in cross-validation for kNN.

| | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|
| Fold 1 | 0.75 | 0.76 | 0.75 | 0.75 |
| Fold 2 | 0.79 | 0.78 | 0.78 | 0.79 |
| Fold 3 | 0.76 | 0.76 | 0.76 | 0.76 |
| Fold 4 | 0.77 | 0.77 | 0.77 | 0.77 |
| Fold 5 | 0.72 | 0.71 | 0.71 | 0.71 |

**Table 5.** Mean class-average metrics for each fold in cross-validation for SVM.

| | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|
| Fold 1 | 0.74 | 0.69 | 0.7 | 0.74 |
| Fold 2 | 0.76 | 0.7 | 0.71 | 0.76 |
| Fold 3 | 0.74 | 0.7 | 0.7 | 0.74 |
| Fold 4 | 0.74 | 0.69 | 0.7 | 0.74 |
| Fold 5 | 0.71 | 0.68 | 0.67 | 0.71 |

**Table 6.** Mean class-average metrics for each fold in cross-validation for decision tree.

| | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|
| Fold 1 | 0.77 | 0.77 | 0.77 | 0.77 |
| Fold 2 | 0.8 | 0.8 | 0.8 | 0.8 |
| Fold 3 | 0.81 | 0.81 | 0.81 | 0.81 |
| Fold 4 | 0.81 | 0.79 | 0.79 | 0.79 |
| Fold 5 | 0.74 | 0.74 | 0.74 | 0.74 |

**Table 7.** Mean class-average metrics for each fold in cross-validation for random forest.

| | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|
| Fold 1 | 0.85 | 0.85 | 0.85 | 0.85 |
| Fold 2 | 0.8 | 0.8 | 0.8 | 0.8 |
| Fold 3 | 0.87 | 0.87 | 0.87 | 0.87 |
| Fold 4 | 0.86 | 0.86 | 0.86 | 0.86 |
| Fold 5 | 0.8 | 0.8 | 0.8 | 0.7 |

**Table 8.** Mean class-average metrics for each fold in cross-validation for neural network.

| | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|
| Fold 1 | 0.87 | 0.87 | 0.87 | 0.86 |
| Fold 2 | 0.88 | 0.89 | 0.88 | 0.88 |
| Fold 3 | 0.89 | 0.89 | 0.88 | 0.88 |
| Fold 4 | 0.89 | 0.89 | 0.88 | 0.88 |
| Fold 5 | 0.85 | 0.86 | 0.85 | 0.84 |

possible to notice that all models presented a good generalization, since the metrics did not vary more than 10%, thus suggesting that the utilized models are robust and are not sensitive to small variations in the data, so when these models are applied to new data, they perform similar results to those previously found. The random forest algorithm also presented good results and SVM and decision tree algorithms presented satisfactory results.

The confusion matrix and metrics for each of the classes were calculated to evaluate how the imbalanced dataset influenced the performance of the algorithms. Although confusion matrices were generated for all models, only the confusion matrix of the neural networks are presented (Figure 3), as it achieved the best overall performance. The metrics are in Table 10.

Implicit models were generated using the interpretations of the test set algorithms to visually evaluate the performance of the algorithms. Information from the training set was also used to develop the models, as the objective of the work is to evaluate whether supervised learning algorithms can interpret new drills, so the information from the older drills is also part of the model, that is, the training set

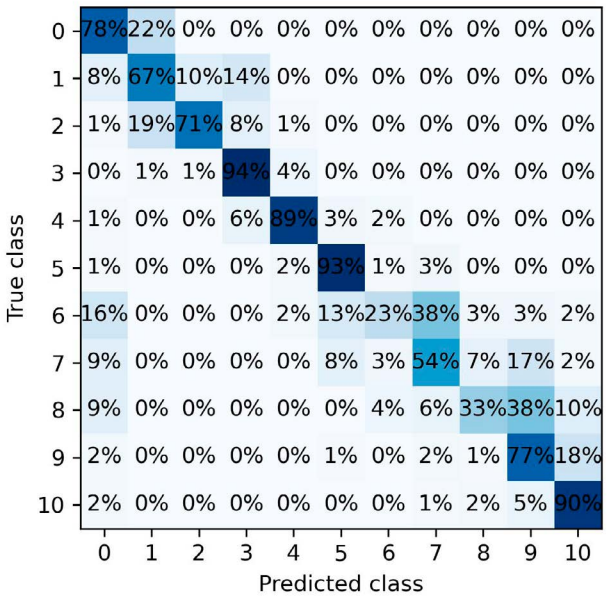


**Figure 3.** Confusion matrix of neural network.

**Table 9.** Mean class-average metrics for validating with test set.

| | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|
| Naïve Bayes | 0.48 | 0.59 | 0.52 | 0.48 |
| kNN | 0.73 | 0.73 | 0.73 | 0.73 |
| SVM | 0.75 | 0.7 | 0.71 | 0.75 |
| Decision Tree | 0.8 | 0.8 | 0.8 | 0.8 |
| Random Forest | 0.87 | 0.86 | 0.86 | 0.87 |
| Neural Network | 0.89 | 0.89 | 0.89 | 0.88 |

**Table 10.** Global metrics to each class using neural network on the test set.

| | Accuracy | Precision | F1 | Recall |
|---|---|---|---|---|
| Class 0 | 0.78 | 0.49 | 0.6 | 0.78 |
| Class 1 | 0.95 | 0.96 | 0.95 | 0..95 |
| Class 2 | 0.71 | 0.77 | 0.74 | 0.71 |
| Class 3 | 0.94 | 0.94 | 0.94 | 0.94 |
| Class 4 | 0.89 | 0.85 | 0.87 | 0.86 |
| Class 5 | 0.93 | 0.96 | 0.95 | 0.93 |
| Class 6 | 0.23 | 0.4 | 0.29 | 0.23 |
| Class 7 | 0.54 | 0.62 | 0.58 | 0.54 |
| Class 8 | 0.33 | 0.56 | 0.42 | 0.33 |
| Class 9 | 0.77 | 0.67 | 0.72 | 0.77 |
| Class 10 | 0.9 | 0.76 | 0.83 | 0.9 |

represents the old drills and the test set represents the new holes. Figure 4 shows the sections of the implicit models using the algorithms.

The geological model of the naive bayes and kNN shows bodies with a non-geological shape. It is also noteworthy that the interpretation of the geological domains did not respect the stratigraphy, that is, the domain of the top of the stratigraphy running at the base of the model. Thus, these models are not representative of the deposit. Although the decision tree and SVM algorithms outperform the naive bayes when generating geological models, they still exhibit areas of interpretation errors within the geological domains. Furthermore, these models fail to accurately represent the true geology of the study area. On the other hand, the random forest and neural network algorithms show models that are similar to those generated through the interpretations of performed manually sampling; few samples were interpreted wrongly, so regions with errors occur locally that are considered inexpressive, because they do not distort the shapes of the geological layers.

## DISCUSSION

The focus of this study was not to perform the interpretation in a deposit with little information and geological knowledge about it, as it is likely that the algorithms do not perform well with little data, the an increase in relevant information to the algorithm tends to improve its performance, emphasizing that, often, the necessary information is restricted due to the high cost of drilling boreholes and channels, so balance is necessary: sufficient data is required for the algorithm to recognize patterns effectively, while ensuring the cost remains feasible for acquiring the necessary information.

In the results obtained, the influence of imbalanced data on algorithm performance was evident. As anticipated, algorithms tended to misclassify minority classes as majority classes, resulting in poor metrics for the minority classes. Sampling in minority classes would tend to solve

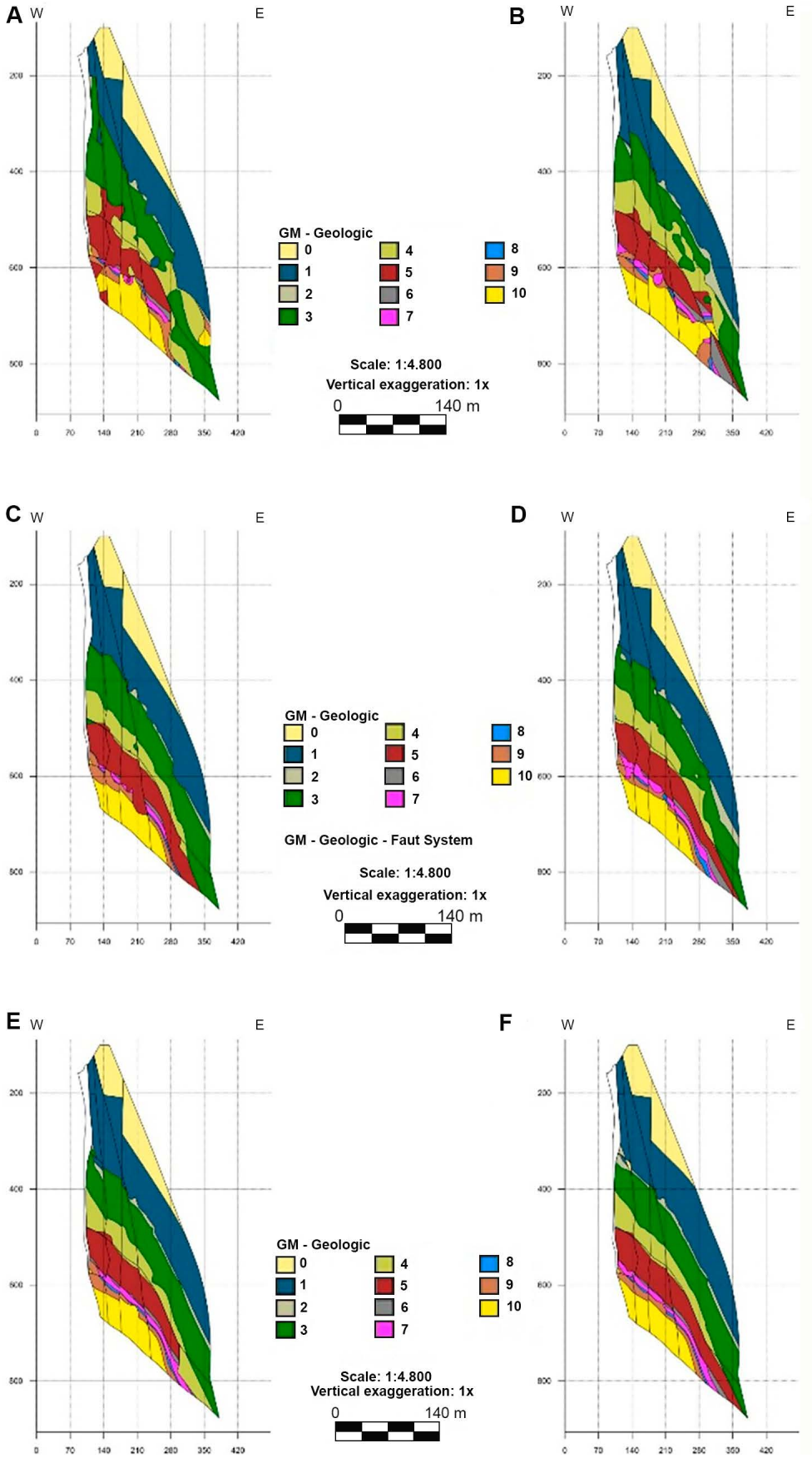


**Figure 4.** (A) Geological section with geological interpretations performed by naïve Bayes; (B) kNN; (C) SVM; (D) decision tree; (E) random forest and (F) neural network.

this problem of imbalanced data as the representativeness of minority classes increased, but this situation is not feasible in mining, as it is not possible to carry out sampling equally among geological domains, due to the different thicknesses of the layers and the tendency to carry out sampling in regions that show the possibility of having ore, since sampling in sterile places represents loss for the mining venture. A study using techniques for imbalanced data, such as oversampling and undersampling, is necessary to verify if there is a significant improvement in the performance of the algorithms that validate the need to use these techniques.

The naive Bayes algorithm showed worse results when compared to the other algorithms. The implicit models generated with the interpretation performed with naive Bayes in the set of tests did not show layers with geological formats due to many samples being miscategorized. The performance of the naive Bayes algorithm may be related to the fact that this algorithm is based on the theory of conditional independence that considers the samples being independent of each other, but in geostatistics its assumed that the regionalized variables have spatial continuity, that is, closer samples tend to be more similar in relation to distant ones. The algorithms random forest and neural network showed the best results among the algorithms used. The interpretations performed by these two algorithms were like those performed by a geologist, so these algorithms can help optimize the updating of geological models. It is important to emphasize that both algorithms were capable of accurately interpreting entire drill holes. This was ensured by structuring the train-test split so that each drill hole was assigned exclusively to either the training or test set. This strategy provided a more realistic assessment of the models' ability to generalize to unseen drill holes.

It is worth noting that when considering aspects such as ease of use, energy efficiency and computational complexity, random forest stands out as a more practical and accessible option for mining applications. Unlike neural networks, which require extensive preprocessing—such as input normalization, random forest operates efficiently even with raw data and fewer configuration steps. Its robustness to imbalanced and heterogeneous datasets, as well as its flexibility in handling different data types, makes random forest particularly well-suited for geological data, which often exhibit these characteristics.

In mining contexts, where geological models need to be constantly updated with new drilling data, solutions that require less preparation and processing are preferable. Although the results obtained by neural network and random forest were very similar, the latter strikes a balance between solid results, simplicity and low computational cost.

## CONCLUSION

The objective of this study was to use supervised machine learning algorithms for geological interpretation when new drill holes are added to the model of a gold deposit. The naive Bayes, kNN, decision tree, and SVM algorithms are not advisable for this activity, as they incorrectly interpreted the samples, as they presents lower predictive performance and higher misclassification rates. Naive Bayes showed the poorest performance with an accuracy of 0.48 Although kNN, SVM and decision tree achieved moderate accuracies, respectively, 0.73, 0.75 and 0.8 Consequently, it would take a considerable amount of time for a geologist to correct the distorted areas in the model.

On the other hand, the use of random forest and neural network algorithms can optimize the geological model update process. The categorization results of these algorithms were close to those performed by a geologist, with accuracies of 0.87 for the Random forest and 0.89 for the neural network, thus making it generate representative models of the mineral deposit.

In practice, random forest and neural network are best suited for deposit with dense sampling, where the volume and quality of data are sufficient to support model training. Their successful implementation requires the involvement of a qualifies professional with knowledge in both geology and machine learning. Moreover, to maintain model accuracy over time, it is essential to retrain the algorithms as new drilling data becomes available. Without regular updates, the models may lose reliability due to the natural variability of geological formations. Thus, despite their potential, the practical application of these techniques depends on careful planning, technical expertise, and continuous refinement.

## REFERENCES

Abzalov, M. (2016). Quantitative Geological Model. In: Abzalov, M (Eds). Applied mining Geology (v.1, 335-3350). Switzerland: Springer. https://doi.org/10.1007/978-3-319-39264-6

Armstrong, M. (1998). Regionalized Variables. In: Armstrong, M. Basic Linear Geostatistics (v.1, 20-34). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-58727-6

Awad, M., Khanna, R. (2015). Machine Learning. In: Awad, M. e Khanna, R. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Berkeley: Apress open, 1-18. https://doi.org/10.1007/978-1-4302-5990-9_1

Chawla, N. V., Japkowicz, N., Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter, 6, 1-6. https://doi.org/10.1145/1007730.1007733

Cowan, E. J., Beatson, R. S., Ross, H. J., Fright, W. R., McLennan, T. J., Evans, T. R., Carr, J. C., Lane, R. G., (2003). Practical Implicit Geological Modelling. In: Dominy, S. (ed), 5th International Mining Geology Conference, 8, 89-99. Australia: The Australasian Institute of

Mining and Metallurgy. Available at: https://www.researchgate.net/publication/281685127_Practical_Implicit_Geological_Modelling. Accessed on: Dec 16, 2025.

Dougherty, G. (2013). Classification. In: Dougherty, G (Ed). Patter Recognition and Classification- An Introduction, v.1, 9-26. New York: Springer. https://doi.org/:10.1007/978-1-4614-5323-9

Garayp, E., Frimmel, H. E. (2023). A modified paleoplacer model for the metaconglomerate-hosted gold deposits at Jacobina, Brazil. Mineralium Deposita. https://doi.org/10.1007/s00126-023-01220-9

Mao, X., Zhang, W., Liu, Z., Ren, J., Bayless, R. C., Deng, H. (2020). 3D Mineral Prospectivity Modeling for the Low-Sulfidation Epithermal Gold Deposit: A Case Study of the Axi Gold Deposit, Western Tianshan, NW China. Mineral, 10(3), 233. https://doi.org/10.3390/min10030233

Rasera, L. G. (2014). Geoestatística de múltiplos pontos aplicada à simulação de modelos geológicos em grids estratigráficos. Dissertação (Mestrado). Porto Alegre: Escola de Engenharia de Minas, Metalúrgica e de Materiais, Universidade de Rio Grande do Sul -UFRGS. Available at: https://lume.ufrgs.br/bitstream/10183/117142/2/000929816.pdf.txt5178f9de3f751d6f65932823152e5c1fMD52THUMBNAIL000929816.pdf.jpg000929816.pdf.jpgGenerated. Accessed on: Dec 16, 2025.

Rossi, M. E., Deutsch, C. V. (2014). Geological Controls and Block Modeling. In: Rossi, M.E. & Deutsch, C. V. Mineral Resource Estimation ( 29-50). New York: Springer. https://doi.org/10.1007/978-1-4020-5717-5_3

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, v. 3, 210-229. https://doi.org/10.1147/rd.33.0210

Singh, D., Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Zhang, S. E., Nwaila, G. T., Bourdeau, J. E., Ghorbani, Y., Carranza, E. J. M. (2023). Machine Learning-Based Delineation of Geodomain Boundaries: A Proof-of-Concept Study Using Data from the Witwatersrand Goldfields. Natural Resources Research, 32, 879-900. https://doi.org/10.1007/s11053-023-10159-7