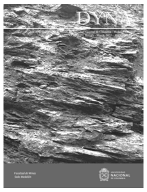
DYNA
http://dyna.medellin.unal.edu.co/



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Estimation of emerald mineralization probability using machine learning algorithms

Daniela Neva-Rodríguez & Luis Hernán Ochoa-Gutierrez

Universidad Nacional de Colombia, sede Bogotá, Facultad de Ciencias, Departamento de Geociencias, Bogotá, Colombia. dnevar@unal.edu.co, lhochoag@unal.edu.co

Received: January 16th, 2024. Received in revised form: August 9th, 2024. Accepted: August 23th, 2024.

Abstract
This research proposes a machine learning (ML) model that estimates the probability of emerald mineralization in rocks of the Western Emerald Belt (CEOC). Element concentrations, lithologies and coordinates were used as input variables and productivity as the target variable (176 samples). The variables were transformed to be integrated into the model. (1) Variable selection was performed using the Boruta method and backward elimination. (2) A logistic regression, a neural network, and a support vector machine were trained. (3) Calibration was achieved with the Platt method. (4) Calibration assessment was conducted by using the Brier score and calibration curves. The model selected was a calibrated support vector machine ($C = 0.19$ and $\lambda = 0.1$) that included 17 geochemical variables and the coordinates. The results were presented in a 3D plot. Assigning a probability value to each sample allows the mining targets to be ranked.

Keywords: gems; calibration; drillhole; mineral target.

Estimación de la probabilidad de mineralización de esmeraldas usando algoritmos de aprendizaje automático

Resumen
La investigación propone un modelo de aprendizaje automático para estimar la probabilidad de mineralización de esmeraldas en el Cinturón Esmeraldífero Occidental (CEOC). Se emplearon concentraciones elementales, litología y coordenadas como variables de entrada y la productividad como variable objetivo (176 muestras). Las variables fueron transformadas para ser integradas al modelo. (1) Se recurrió a los métodos Boruta y *backward elimination* para seleccionar las variables. (2) Una regresión logística (LR), una red neuronal de retropropagación (BPNN) y una máquina de vectores de soporte (SVM) fueron entrenadas. (3) Se usó la calibración de Platt y (4) se evaluó su desempeño usando la puntuación Brier y curvas de calibración. El modelo elegido fue una máquina de vectores de soporte calibrada ($C = 0.19$ y $\lambda = 0.1$) que incluyó 17 variables geoquímicas y las coordenadas. Asignar valores de probabilidad permitió jerarquizar los objetivos mineros.

Palabras clave: gemas; calibración; perforación; objetivo minero.

1 Introduction

The global demand for colored gemstones implies higher prices and industry growth [1]. However, Colombian emerald production has declined significantly, from 9.8 million carats in 2004 to approximately 964,000 carats in 2021 [2]. This negative trend indicates the need to establish new mining goals.

The mineral potential of economically viable areas was determined through geometric and geostatistical methods that require significant manual processing [3]. Geometric methods primarily use mean values or weighted averages for estimation in defined blocks, while geostatistical methods estimate mineral resources from the variability and spatial correlation characteristics of the original data [4]. These methods are efficient in early stages of a project, but some authors have pointed out that they perform poorly on very heterogeneous data sets [1].

How to cite: Neva-Rodríguez, D., and Ochoa-Gutierrez, L.H., Estimation of emerald mineralization probability using Machine Learning Algorithms. DYNA, 92(235), pp. 19-27, January - March, 2025.

Universidad Nacional de Colombia.
Revista DYNA, 92(235), pp. 19-27, January - March, 2025, ISSN 0012-7353
DOI: https://doi.org/10.15446/dyna.v92n235.112504

Mining recommendations are limited to determining if a rock is an emerald producer. Identifying potential mining sites can be challenging due to variations in mineral association within different rock formations. These discrepancies arise from changes in fluid composition over time and the interaction with the rock's composition [5].

The purpose of this research is to find a ML model applicable to drill data that estimates the probability of emerald mineralization in rocks of the Western Emerald Belt (CEOC). Lithological, geochemical data and coordinates were integrated and processed simultaneously to identify and hierarchize mining targets.

Determining and quantifying the mining potential of an emeraldiferous prospect is a challenging task. Significant amounts of data are available from emerald mines in Colombia due to the long history of exploration. However, no literature was identified regarding the useful of these datasets as training data in ML models for selecting new mining targets.

The absence of previous research of this nature may be attributed to experiments conducted by private companies that remain unpublished. Alternatively, stakeholders in the emerald industry may not be familiar with the use of ML algorithms for gemstone exploration, or there may be uncertainty about the significance of the data and methods.

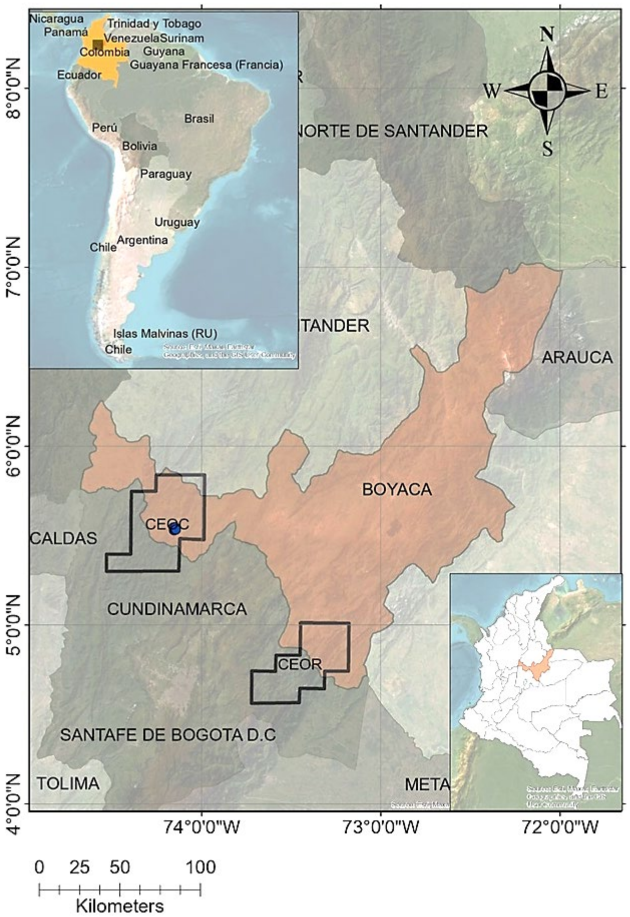


Figure 1. Emerald belts in Colombia - Western belt (CEOC) and Eastern belt (CEOR)
Source: Own elaboration and adapted from [6,7].

Colombian emerald deposits are found in two belts located in the Eastern Cordillera, Boyacá Department. The Western Emerald belt (CEOC) includes the mining districts of Muzo, Coscuez, La Pita, Peñas Blancas, and La Glorieta-Yacopí [7], The Eastern Emerald belt (CEOR) includes Chivor, Gachalá, and Macanal [8], (Fig. 1Figure 1).

Emerald deposits data from projects in the exploitation stage have been collected through various techniques over several decades, resulting in abundant and diverse information that includes lithological, mineralogical, and stratigraphic [9]; hydrothermal alteration, and geochemical [5,10,11] and geophysical data [12].

Giuliani et al., (1999) [10] made an important breakthrough by measuring the number of emerald crystals formed from the amount of beryl leached in five phases of hydrothermal alteration. Starting from the axiom that beryllium (Be) can be mobilized under hydrothermal conditions, they estimated the total gem reserves in the Chivor mines using a methodology developed, evaluated, and validated by Bourlès (1992) [17], to estimate the autigenic potential ^{10}Be/^{9}Be. The results indicated that 470 kg of beryllium was leached at the Chivor mine, and therefore, it can be concluded that if 100% of the leached beryllium was mobilized, there would be 720,000 emeralds, interpreted as crystals 1 cm in diameter by 5 cm in height.

Niño and Sheng-Rong (2017) [5] analyzed 17 samples belonging to seven mines in the Western Emerald Belt. They used the stable isotopes Oxygen-18 (^{18}O) / Oxygen-16 (^{16}O), and Carbon-12 (^{12}O) / Carbon-13 (^{13}O) as a measure of groundwater-mineral interactions to identify differences between calcite in productive and non-productive veins in the study area. They revealed that although the mineralogy of productive and non-productive veins is similar, it is possible to find variables that separate them using geochemical data; unfortunately, isotopic experiments on emeralds are rare.

1.1. Machine learning in mineral exploration

Machine Learning (ML) for the exploration of geological resources has primarily been restricted to massive and stratiform deposits [13]. Extending these techniques to gemstone deposits poses a challenge due to the discrete nature of gems, each varying in quality, making the estimation of quantity during exploration particularly challenging. The absence of prior scientific publications on using ML in gem exploration emphasizes the significance of this research. However, this does not imply that ML algorithms haven't been utilized in mineral exploration previously.

Zhang et al. (2018) [14] employed Support Vector Machines (SVM) and Artificial Neural Networks (ANN) to assess the prospectivity of a gold deposit in the Hatu region of Xinjiang, northwest China. The researchers compared the predictive results of the ANN and SVM models, demonstrating the superior predictive competence of SVM in this case.

Mohammadi and Hezarkhani (2018) [15] investigated a copper porphyry in the Urumieh-Dokhtar magmatic arc using drill hole data. The 3814 samples collected from 44 drillholes included lithology, alteration type, copper concentration, coordinates, and collar as input variables. This research focuses on a multi-class classification SVM in which the data were separated into four groups - zones (oxidation, leaching, trans, and hypogene zones). The model was trained on 80% of the data and evaluated on 20%; 25% of the area was classified as favorable, containing 78.6% of the known deposits.

Taboada et al., (2007) [16] analyzed data from a shale deposit in an area of low regional metamorphism in Spain using multiclass, ordinal, and regression SVMs with a Gaussian Kernel. The shales in this deposit were divided into three groups: first quality with no aesthetic defects; secondary quality with aesthetic defects, and a third group with no economic value. The input variables included intercalations with sheets or layers of sands, presence of closed or quartz-filled microfractures, planes of schistosity, sulfides affected by oxidation, degree of surface alteration, crenulation, and kink bands.

2 Materials and methods

Data is the core of this research and the most important part of ML. Fig.2, presents the samples from producing projects that were used as the data train for this research. Since the rock samples used as training data come from the Western Emerald Belt (CEOC), the application of the generated model is restricted to this belt.

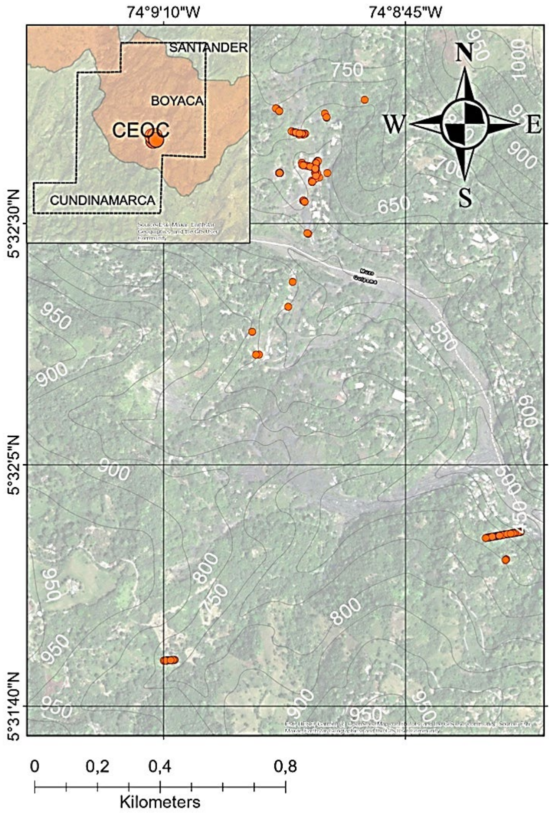


Figure 2 Sample's location
Source: Own elaboration

Table 1. Train samples.

Train samples	Input variables	Output variables	Quantity
Productive	49	1	88
Sterile	49		88
Total			176

Source: The authors

Each of the 176 samples utilized as training examples originates from projects in the production stage, including 49 associated variables. These samples are categorized into either productive (88 samples) or sterile (88 samples) classes. (Table 1). The rock samples were analyzed by inductively coupled plasma mass spectrometry (ICP-MS), which quantifies most of the elements in the periodic table at trace levels; 170 of the 176 samples were prepared by 4-acid digestion and the remainder were prepared by aqua regia.

2.1 Variables

Table 2 presents the input and output variables associated with each example. There are 49 input variables.

The numerical values of latitude, longitude, and elevation describe the location of the samples and thus the relations between geologic bodies. Lithology is a categorical variable that can take three values: mudstone, vein, and breccia.

The elemental concentrations of major elements are determined as percentages and minor elements in parts per million (ppm). The lithology variable includes values such as vain, breccia, and mudstone. One-hot coding and binary codification were employed to convert lithology and productivity into numerical form, given that both represent categorical data.

Table 2.
Input variables description.

Description	Variable	Format	Transformation method
Location (3)	Longitude Latitude Altitude	Numeric (m)	Normalization (Cube 1x1x1)
Type of Rock (1)	Lithology	Categoric (Veins, mudstone, Breccia)	One-hot codification (1-0)
Trace elements (36)	Ag, As, Ba, Be, Bi, Cd, Ce, Co, Cr, Cs, Cu, Ga, Hf, In, La, Li, Mn,Mo, Nb, Ni, Pb, Rb, Sb, Sc, Se, Sr, Ta, Te, Th, Tl, U, V, W,Y, Zn, Zr.	Numeric (ppm)	Z-score
Major elements (9)	Al, Ca, Fe, K, Mg, Na, P, S, Ti.	Numeric (%)	Z-score
Total – Input variables: 49			

Source: The authors

Since ML algorithms are sensitive to the range and distribution of attributes, and outliers can corrupt the training process, all outliers were considered as real observable values and removed by the flooring and capping process. All values below the 1st percentile and above the 99th percentile were replaced by the value of the 1st percentile and the 99th percentile. After the removal of outliers, the data underwent standardization using the z-score. This step was crucial due to the varied reporting of elemental concentrations, with some values expressed in percent and others in parts per million.

According to Colombian Emerald Good Practices Guide [17], the key to calculating resources and reserves is the rigorous delineation of geological bodies. This information was integrated into the model through the coordinates of each sample. Coordinates were transformed into vectors by situating each point within a unit cube with dimensions of 1x1x1. The use of the new unit avoided including parts per million and millions of meters in the same model and ensures the maintenance of spatial relationships. The scale was calculated by dividing 1.0 by the range in which the initial coordinates are found separately for longitude, latitude, and altitude.

2.2 *Proposal*

The research was divided into two phases: modeling and application (Fig.3). Each step in the process impacts the following one. data selection and classification are essential to calibrate the models.

Since the training values have as output a categorical value "mineralization of emeralds" or "absence of mineralization" and the desired result of this research is a probability, the problem was approached from two angles.

A classification was performed using logistic regression, SVM and ANN to divide the samples into prospective and non-prospective, then the probability of emerald mineralization was determined using the Platt calibration method [18]. This method allows a comparison between the classification into prospective samples and the probability of emerald mineralization, translated as confidence.

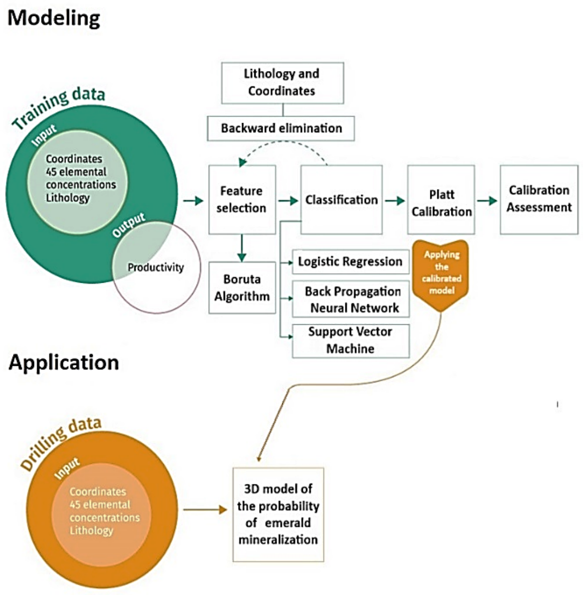


Figure 3 Methodology
Source: Own elaboration

Table 3.
Backward elimination - data sets.

Variables	LR	ANN	SVM	CANN	CSVM
Elemental concentrations latitude, longitude, altitude, lithology	LR1	ANN1	SVM1	CANN1	CSVM1
Elemental concentrations latitude, longitude, altitude	LR2	ANN2	SVM2	CANN2	CSVM2
Elemental concentrations lithology	LR3	ANN3	SVM3	CANN3	CSVM3
Elemental concentrations	LR4	ANN4	SVM4	CANN4	CSVM4

Source: The authors

2.3 *Feature selection*

Variable selection reduces the model's complexity and eliminates noise in the data; this, in turn, reduces the computational cost of training. Boruta and backward elimination methods were employed to assess the relevance of the variables based on their nature.

2.3.1 Selection of geochemical variables (Boruta Method)

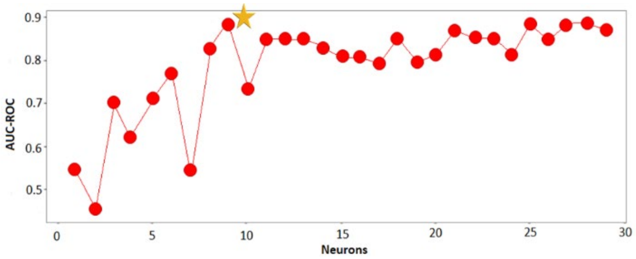
Synthetic variables were created by rearranging the values of the original variables. Subsequently, the significance of each original variable was evaluated by comparing it with its synthetic counterparts in the random forest algorithm. A total of 1000 iterations were performed, leading to two possible outcomes: either the variable exceeded the threshold or it did not. The right tail of the binomial distribution was used to identify variables that consistently met the threshold, categorizing them as significant [19].

2.3.2. Location and lithological variables selection (Backward elimination method)

The influence of the variables in each model (LR, ANN, and SVM) was evaluated through tests in which the input variables were varied to determine which had the greatest influence on classification [20]. The test groups were formed by subtracting coordinates and lithology (Table 3)

2.1 *Classification*

This problem is limited to two labels or classes (binary classification) because each sample belongs only to the productive or non-productive group. The models used to perform the classification task were logistic regression, back-



propagation neural network and support vector machine.

Figure 4. Hyperparameters selection - Elbow method for ANN.
Source: Own elaboration

The neural network and support vector machine hyperparameters were chosen through grid search with 5-fold cross-validation and evaluated using AUC ROC. Unlike neural networks and support vector machines, it is not necessary to select hyperparameters for logistic regression (LR).

2.4.1. Backpropagation neural network (ANN).

Nine neurons were selected using the elbow method because the model's performance did not significantly improve beyond this point (Fig.4).

The number of layers was limited to one due to the small amount of data; adding more layers would increase the model's complexity. The logistic function was selected as the activation function for its range of 0 to 1, enabling interpretation as probability. An adaptive learning rate was utilized with the gradient descent method.

2.4.2. Support vector machine (SVM).

The values C = 0.19, λ = 0.1 and a RBF kernel was selected. This hyperparameter combination optimized the model's performance according to the area under the ROC curve (Fig. 5).

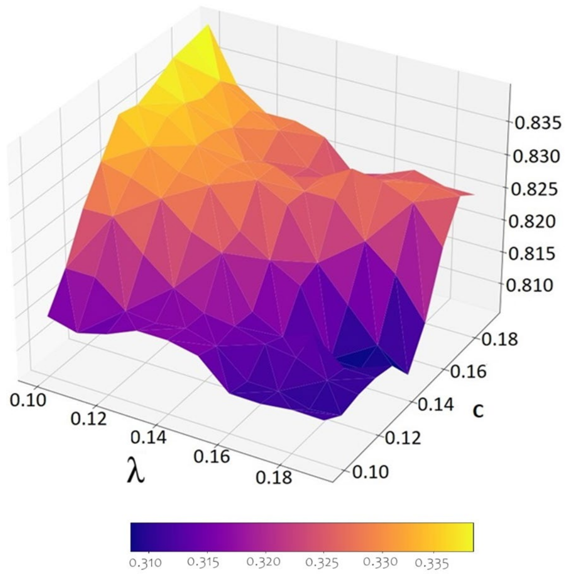


Figure 5. Hyperparameters selection - Gridsearch for SMV.
Source: Own elaboration

2.2 Calibration

The Platt's calibration can be understood as a dual training process. A calibrated algorithm generates true probabilities [21]. Since logistic regression inherently produces well-calibrated probabilities, it is not necessary to calibrate these models.

Neural networks provide probabilities, but because the development of these algorithms has focused on increasing capacity in classification tasks, modern ANNs are poorly calibrated, unlike those of a decade ago [22]. Support vector machines do not provide probabilities natively, so they require a calibration process to provide reliable probabilities [23].

Neural networks and support vector machines were trained, and then, using the estimated probabilities, the values of A and B are calculated to obtain calibrated probabilities [18].

In eq. (1), $P(y = 1|x)$ is the probability of 1 or 0 depending on whether mineralization is present or not; $f(x)$ is the score provided by the support vector machine or neural network; and A and B are parameters learned by the algorithm and were found to obtain calibrated probabilities.

$$P(y = 1|x) = \frac{1}{1 + exp(A\,f\,(x) + B)} \qquad (1)$$

2.3 Calibration assessment

The Brier Score was originally proposed to measure the evaluation of probability forecasts in meteorology [24]; however, its applications have become widespread. The Brier Score for a sample of n binary predictions is given by eq. (*2*), where f_t is the predicted probability of the event occurring, and o_t is 1 or 0, depending on whether the event subsequently occurred or not; lower score values indicate better estimated probabilities [25].

$$BS = \frac{1}{n}\sum_{t=1}^{n} f_t - {o_t}^2 \qquad (2)$$

3 Results

The Boruta algorithm identified seventeen chemical elements in the green zone as variables that enable data division into two classes (productive and non-productive).

Twenty-eight chemical elements were placed in the orange zone, indicating that their characteristics did not meet the threshold to be considered relevant and were therefore disregarded (Fig.6).

The research seeks to obtain probability values using classification as an intermediary process. Nonetheless, an assessment of the capability to classify between productive and non-productive samples was carried out both before and after calibration.

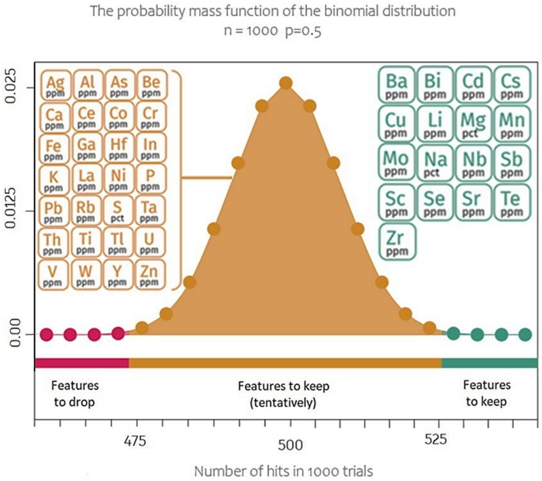


Figure 6. Geochemistry variable selection - Boruta Method.
Source: Own elaboration

The Fig. 7 displays the accuracy, precision, sensitivity, F1 score, specificity, and area under the ROC curve for the three evaluated algorithms and the subsets chosen according to the backward elimination method. The logistic regression algorithm exhibits superior classification performance, achieving better metrics in all scenarios. After calibration, the classification ability of the neural networks decreased, while the classification ability of the support vector machine increased.

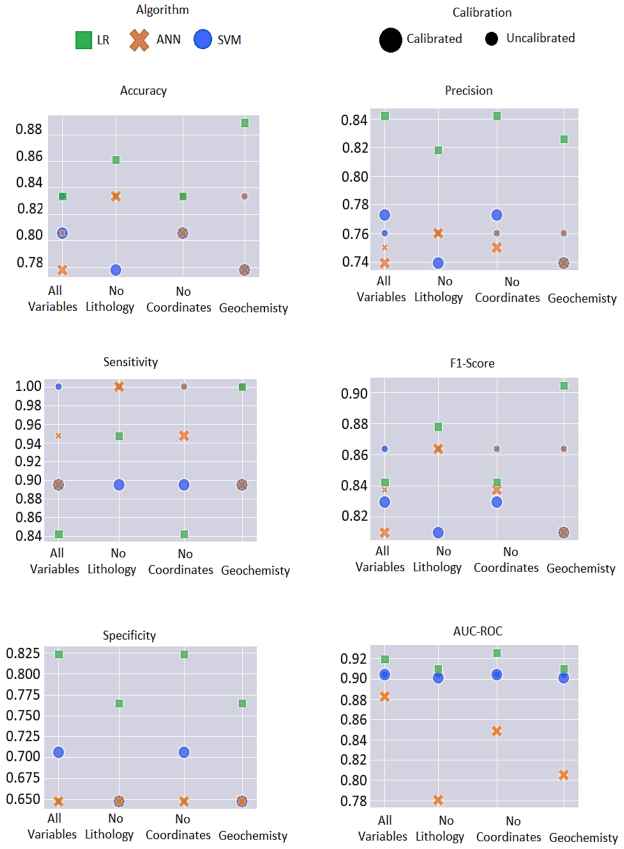


Figure 7. Classification metrics.
Source: Own elaboration

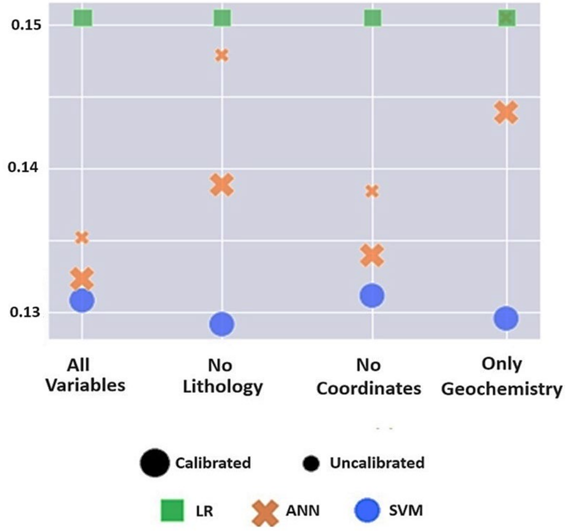


Figure 8. Brier score results.
Source: Own elaboration

The Brier score is negatively oriented, which means that smaller values indicate better forecasts [25]. Since support vector machines do not generate inherent probabilities, there are no blue points in Fig. 8 showing the pre-calibration probability score for this model.

Since it is unnecessary to calibrate the LR model [26], there are no variations observed in the green points pre and post-calibration (Fig. 8) or green calibration curves (Figure 9Figs. 9, 10). The Brier score suggests that the support vector machine without lithology provided the best calibrated estimated probabilities.

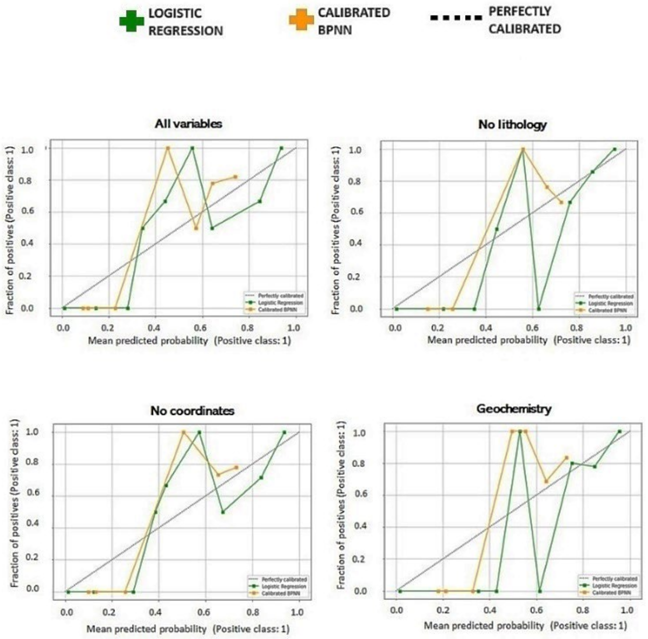


Figure 9. Calibration curves- before Platt calibration
Source: Own elaboration

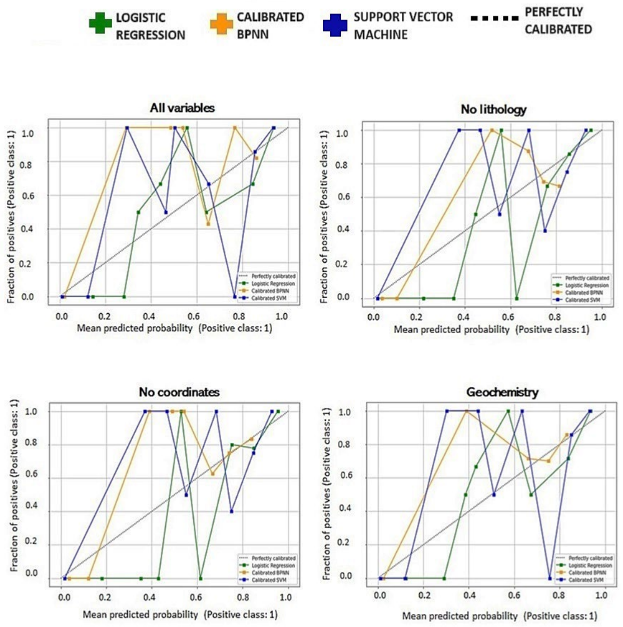


Figure 10. Calibration curves - after Platt Calibration
Source: Own elaboration

4 Discussion

This research used 176 samples; however, 6 were prepared through the aqua regia method instead of the four-acid method, introducing bias to the data that should be avoided. Future experimental designs need to consider the rigor required to apply machine learning models. The collected information must be highly accurate, or the data analysis and application will be unreliable [27].

The wealth of data collected on emeralds over the years must be reinterpreted using ML techniques to provide mining recommendations. It is important to recognize the bias inherent in any data to ensure data quality; this is no small matter, as poor-quality data input leads to unreliable data output.

Studying the geological processes associated with emeralds is essential to identify the data that most efficiently describes the deposits. The use of elemental concentrations may not be the most appropriate descriptor to define such deposits because it does not efficiently incorporate mineralogical information into the models. Partial analysis and isotope data, such as those used by Niño and Sheng-Rong (2017) [5], should be integrated.

This methodology proposes integrating data from various sources, including spatial, geochemical, and lithological variables, and analyzing them simultaneously. Geological body boundaries are important in gemstone exploration; therefore, not only must more training data be integrated, but the spatial density of the data must be high enough to allow the models to integrate the geometry of the geological bodies.

Metrics of ML models should be assessed based on their use and stakeholders' interests. If the classification process is more important, it is essential to evaluate whether the business need is to minimize false positives or maximize true positives. The primary aim of these models is to estimate the probability of emerald mineralization for ranking mining targets. Hence, the selection of the best model was based on the calibration performance.

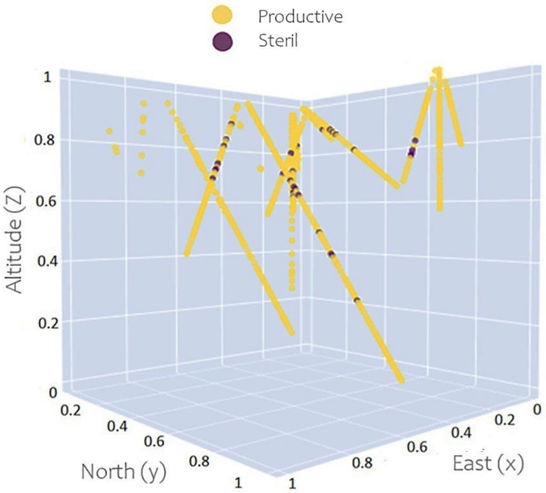


Figure 11. 3D Classification model.
Source: Own elaboration

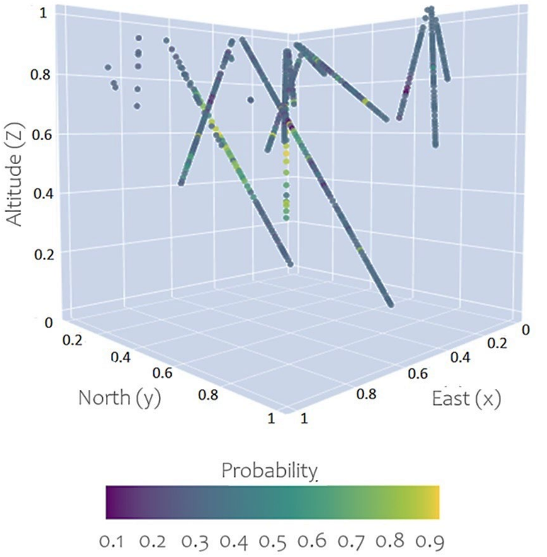


Figure 12. 3D Calibration model.
Source: Own elaboration

The model CSVM2 achieved the lowest Brier score among the models evaluated. Two intervals have been identified as having the highest probability for emerald mineralization. This model should be used with caution because it overestimates the highest probabilities, suggesting that more data should be integrated.

Emeralds, like other gemstones such as rubies and diamonds, occur as discrete particles in parts per billion. The concentration of emeralds cannot be accurately measured in drill samples, and their value varies from deposit to deposit based on size and quality [28].

The advantages of using calibrated models over classification models (Fig. 11Figure 11) for this application are obvious. Calibrated models hierarchize mining targets by assigning a numerical value, allowing for informed decision-making, without restricting it to the mere occurrence or absence of mineralization. Additionally, utilize free software to present the outcomes in 3D. The Fig. 12 shows two continuous segments with estimated probabilities above 85%.

5 Conclusions

The data from existing emerald exploration and exploitation campaigns should be processed using machine learning techniques. It is imperative to study the characteristics of this ore to determine which are the variables that more accurately describe emerald deposits and to identify the biases that exist in the data.

The model SVM; C = 0.19 and λ = 0.1 calibrated by Platt calibration method (CSVM2) was selected because generate the best calibrated probabilities, as the numerical value associated with the classification is more important in this application than the classification itself; however, the selected model overestimates the highest probabilities, so it should be used with caution.

Presenting the estimated probabilities on a 3D model is extremely useful for mining companies because it allows them to quickly identify the best segments (higher probability) and consider other factors such as access roads and previous mining works.

Geologic models or data with higher spatial density should be integrated to describe geologic unit boundary information more accurately and improve the models' calibration.

Although there are examples of machine learning (ML) applications in mineral exploration, its use in gemstone exploration remains limited. This presents a significant challenge due to the lack of established or measured precedents in this field.

According to Abu-Mostafa et al. (2012) [29], for a problem to be addressed using machine learning methods, it must meet three conditions: (1) there must be a pattern in the data, (2) the problem cannot be explained through a simple mathematical model, and (3) adequate data must be available.

The problem addressed in this research meets these conditions: there is a relationship between emerald mineralization and changes in measured chemical elements. This relationship cannot be explained by a polynomial model, thus requiring data to identify and understand these relationships, and there is sufficient data available to determine the necessary parameters.

Given this, the mining industry must ensure that decision-makers understand the model and its limitations. Additionally, it is crucial to commit to the continuous search for high-quality training data, as the more comprehensive the training data, the better the model will be. Human supervision is also important for interpreting and validating the results.

This research sets out a methodology for gem exploration, however, each of the steps performed from data exploration to calibration is a discipline. Efficient application of these techniques to the gem exploration industry involves ongoing research that evaluates the spectrum of possibilities available.

Acknowledgements

This research is part of a master's thesis submitted by the first author in 2023. The authors like to express their gratitude Esmeraldas Mining Services for providing the data for this study.

References

[1] Cartier, L.E., Gemstones and sustainable development: perspectives and trends in mining, processing and trade of precious stones. Extr Ind Soc. 6(4), pp. 1013-1016, 2019. DOI: https://doi.org/10.1016/j.exis.2019.09.005

[2] Agencia Nacional Minera - Unidad de planeación minero-energética UPME. Producción histórica de esmeraldas en Colombia. [online]. 2021.

[3] Dominy, S.C., Noppé, M.A., and Annels, A.E., Errors and uncertainty in mineral resource and ore reserve estimation: the importance of getting it right. [online]. 2004. Available at: https://bit.ly/3os0e4M. DOI: https://doi.org/10.2113/11.1-4.77

[4] Quintín, J., Estudios de la estimación y simulación geoestadística para la caracterización de parámetros geólogo-industriales en el yacimiento laterítico Punta Gorda. Minería y Geología. 21, 2005.

[5] Niño, G., and Sheng-Rong, S., Geological and geochemical analyses for emerald exploration in the Muzo Formation along the Western Emerald Belt, Colombia. National Taiwan University, 2017.

[6] Melo, R.T., Notas sobre el contexto tectonoestratigráfico de formación de las esmeraldas colombianas. Boletín Geológico. 45(45), pp. 37-48, 2019. DOI: https://doi.org/10.32685/0120-1425/boletingeo.45.2019.486

[7] Moreno, G., Terraza, R., Montoya, D., Geología del cinturón esmeraldífero oriental. Boletín de Geología. [online]. 31, pp. 51-67, 2009. Available at: https://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-02832009000200004&nrm=iso

[8] Pignatelli, I., Guiliani, G., Ohnenstetter, D., et al., Colombian Trapiche Emeralds: recent advances in understanding their formation. Gems & Gemology. 51, pp. 222-259, 2015. DOI: https://doi.org/10.5741/gems.51.3.222

[9] Groat, L.A., Giuliani, G., Marshall, D.D., Turner, D., Emerald deposits and occurrences: a review. Ore Geol Rev. 34(1-2), pp. 87-112, 2008. DOI: https://doi.org/10.1016/j.oregeorev.2007.09.003

[10] Mendoza, J., Anotaciones geoquímicas para exploración de esmeraldas en la región Muzo-Coscuez con base en la relación Na/K y elementos traza. Geología Colombiana. 21, pp. 89-98, 1996.

[11] Ottaway, T.L., The geochemistry of the Muzo emerald deposit, Colombia. University of Toronto, Toronto, Canada, 1991, 216 P.

[12] Ochoa, L.H., Evaluación magnetometría, radiométrica y geoeléctrica de depósitos esmeraldíferos. Geofísica Colombiana. 7, pp. 13-18, 2003.

[13] Dumakor-Dupey, N.K., and Arya, S., Machine learning—a review of applications in mineral resource estimation. Energies (Basel). 14(14), art. 4079, 2021. DOI: https://doi.org/103390/en14144079

[14] Zhang, N., Zhou, K.,and Li, D., Back-propagation neural network and support vector machines for gold mineral prospectivity mapping in the Hatu region, Xinjiang, China. Earth Sci Inform. 11(4), pp. 553-566, 2018. DOI: https://doi.org/10.1007/s12145-018-0346-6

[15] Mahvash-Mohammadi, N., and Hezarkhani, A., Application of support vector machine for the separation of mineralised zones in the Takht-e-Gonbad porphyry deposit, SE Iran. Journal of African Earth Sciences. 143, pp. 301-308, 2018. DOI: https://doi.org/10.1016/j.jafrearsci.2018.02.005

[16] Taboada, J., Matías, J.M., Ordóñez, C., and García, P.J., Creating a quality map of a slate deposit using support vector machines. J

Comput Appl Math. 204(1), pp. 84-94, 2007. DOI: https://doi.org/10.1016/j.cam.2006.04.030

[17] Bonilla, G., Castaño, A., Nieto, M.A., and Parra, S., Guía de buenas prácticas de la esmeralda colombiana. Published online 2020.

[18] Platt, J.C., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers. 10(3), pp. 61-74, 1999.

[19] Kursa, M., Jankowski, A., and Rudnicki, W., Boruta - A system for feature selection. Fundam Inform. 101, pp. 271-285, 2010. DOI: https://doi.org/10.3233/FI-2010-288

[20] Nguyen, H.B., Xue, B., Liu, I., and Zhang, M., Filter based backward elimination in wrapper based PSO for feature selection in classification. In: 2014 IEEE Congress on Evolutionary Computation (CEC). 2014, pp. 3111-3118. DOI: https://doi.org/10.1109/CEC.2014.6900657

[21] Kaplan, D., Bayesian statistics for the social sciences. The Guilford Press. 2014, 318 P.

[22] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q., On Calibration of modern neural Networks. Proceedings of the 34th International Conference on Machine Learning. Vol. 70, 2017. DOI: https://doi.org/10.48550/ARXIV.1706.04599

[23] Niculescu-Mizil, A., and Caruana, R., Obtaining Calibrated probabilities from Boosting. In: UAI. Vol 5, 2005, pp. 413-420.

[24] Brier, G.W., Weather review verification of forecasts expressed in terms of probability. Monthy Weather Review. [online]. 78(1), pp. 1-3, 1950. [Accessed November 13th, 2023]. Available at: https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml. DOI: https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[25] Roulston, M.S., Performance targets and the Brier score. Meteorological Applications. 14(2), pp. 185-194, 2007. DOI: https://doi.org/10.1002/met.21

[26] Niu, L., A review of the application of logistic regression in educational research: common issues, implications, and suggestions. Educ Rev (Birm). 72(1), pp. 41-67, 2020. DOI: https://doi.org/10.1080/00131911.2018.1483892

[27] Kilkenny, M.F., and Robinson, K.M., Data quality: "Garbage in – garbage out." Health Information Management Journal. 47(3), pp. 103-105, 2018. DOI: https://doi.org/10.1177/1833358318774357

[28] Burgess, J., Buxton, N., Dyck, D., and Oosterveld, M., Thurston M., CIM Estimation best practice committee & mineral reserves best practices guidelines. Guidelines Specific to Particular Commodities Rock Hosted Diamonds. CIM Estimation Best Practice Committee. Published online, 2008.

[29] Abu-Mostafa, Y.S., Magdon-Ismail, M., and Lin, H.-T., Learning from data, AMLBook.com., vol. 1. AMLBook.com, 2012.

D. Neva-Rodríguez, is BSc. in Geology with a MSc. in Geomatics from the Universidad Nacional de Colombia, has dedicated her career to exploring metallic minerals such as gold and copper, along with precious gemstones like emeralds, focusing on spatial data analysis for mineral exploration.
ORCID: 0009-0004-3254-4667

L.H. Ochoa-Gutierrez, is a BSc. Eng. in Civil Engineer in 1988, obtained a PhD. in Systems Engineering in 2017, a MSc. in Geophysics in 2003 and a MSc. in Geomatics, 2007. With over two decades as an Associate Professor at the Universidad Nacional de Colombia, Ochoa has played a valuable role in advancing research at the intersection of machine learning and Geosciences, leaving an enduring mark on academia.
ORCID: 0000-0002-3607-7339