minerals

MDPI

Article

Machine Learning Methods for Quantifying Uncertainty in Prospectivity Mapping of Magmatic-Hydrothermal Gold Deposits: A Case Study from Juruena Mineral Province, Northern Mato Grosso, Brazil

Victor Silva dos Santos [1,*], Erwan Gloaguen [1], Vinicius Hector Abud Louro [2] and Martin Blouin [1]

[1] Centre Terre Eau Environnement, Institut National de la Recherche Scientifique, 490 Couronne St, Quebec City, QC G1K 9A9, Canada; erwan.gloaguen@inrs.ca (E.G.); martin.blouin@inrs.ca (M.B.)
[2] Instituto de Geociências, Universidade de São Paulo-USP, Rua do Lago 562, São Paulo 05508-080, Brazil; vilouro@usp.br
* Correspondence: victor.santos@inrs.ca

Abstract: Mineral prospectivity mapping (MPM), like other geoscience fields, is subject to a variety of uncertainties. When data about unfavorable sites to find deposits (i.e., drill intersections to barren rocks) is lacking in MPM using machine learning (ML) methods, the synthetic generation of negative datasets is required. As a result, techniques for selecting point locations to represent negative examples must be employed. Several approaches have been proposed in the past; however, one can never be certain that the points chosen are true negatives or, at the very least, optimal for training. As a consequence, methodologies that account for the uncertainty of the generation of negative datasets in MPM are needed. In this paper, we compare two criteria for selecting negative examples and quantify the uncertainty associated with this process by generating 400 potential maps for each of the three ML methods utilized (200 maps for each criterion), which include random forest (RF), support vector machine (SVM), and k-nearest neighbors (KNC). The results showed that applying a geological constraint to the creation of negative datasets reduced prediction uncertainty and improved overall model performance but produced larger areas of very high probability (i.e., >0.9) when compared to using only the spatial distribution of known deposits and occurrences as a constraint. SHAP values were used to find approximations for the importance of features in nonlinear methods, and kernel density estimations were used to examine how they varied depending on the negative dataset used to train the ML models. Prospectivity models for magmatic-hydrothermal gold deposits were generated using data from the shuttle radar terrain mission, gamma-ray, magnetic lineaments, and proximity to dykes. The Juruena Mineral Province, situated in Northern Mato Grosso, Brazil, represented the case study for this work.

Keywords: mineral prospectivity mapping; machine learning; quantification of uncertainty; data integration; Juruena Mineral Province

check for updates

Citation: Silva dos Santos, V.; Gloaguen, E.; Hector Abud Louro, V.; Blouin, M. Machine Learning Methods for Quantifying Uncertainty in Prospectivity Mapping of Magmatic-Hydrothermal Gold Deposits: A Case Study from Juruena Mineral Province, Northern Mato Grosso, Brazil. *Minerals* **2022**, *12*, 941. https://doi.org/10.3390/min12080941

Academic Editor: Behnam Sadeghi

Received: 30 June 2022
Accepted: 22 July 2022
Published: 26 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The goal of mineral prospectivity mapping (MPM) is to define favorable areas for the discovery of new deposits. It is a multi-criteria process that can be guided by geological knowledge, i.e., understanding of the responsible mechanisms for ore-formation (knowledge-driven), geoscientific data that can serve as a signature or represent an ore-forming setting, capable of driving the search for new potential deposit areas (data-driven), or a hybrid approach (knowledge- and data-driven) [1]. Traditionally, the delineation of mineral exploration targets is conducted, for example, through the analysis of evidential layers in the Geographic Information System (GIS). Nevertheless, manual data processing is affected by the limited capacity of the human brain to process multiple variables at the

same time. Consequently, this leads to a lack of reproducibility, depending only on the assumption made by experts. Recently, Machine Learning (ML) algorithms have recently been used to overcome these limitations by automatically mapping relationships in n-dimensional datasets. When used to find the underlying models by learning relationships from evidential layers to classify probable mineralized locations, supervised learning methods may lead to more accurate prospective zone delimitation [2,3]. In MPM, supervised approaches consist of dichotomous problems in which models must be provided with two classes in order to classify whether a location is prospective or not: (i) a class of interest (e.g., known deposit locations); (ii) a class of contrast (e.g., non-deposit locations). Nonetheless, deposits are the result of rare events and non-random ore-forming processes, whereas non-deposits are the result of common and random geological processes, i.e., an undefined class with unknown locations [4].

Many authors have studied the distribution and behavior of spatial phenomena using techniques such as Fry analysis [5], fractal analysis [6], and point pattern analysis [7]. Deposits are point entities at regional or district scales, and their spatial distribution can be studied using such techniques, which contribute to a better understanding of the geological features that control their occurrence and geometry [4]. Following that, the spatial analysis of a class of interest (i.e., deposit locations or positive examples) may support the selection of contrasting examples (i.e., non-deposit locations or negative examples). However, there is still some uncertainty in this process because even when studying spatial patterns, it is difficult to map all scales at which mineralization controls operate. As a result, one can never guarantee that negative training datasets generated using random point locations constrained by geological or spatial distribution criteria represent true negatives (i.e., barren sites).

Previous work introduced a few methods for performing negative training sampling in MPM. One could, for example, use the spatial distribution of deposits to define areas where new deposits are unlikely to be discovered [8] and sample negative examples at 'unfavorable' lithologic units or drilling sites that appear barren [9]. Following that, ref. [10] investigated the effects of random negative sampling by comparing the potential maps generated by 50 negative training datasets produced from random point locations.

In this paper, we developed a machine learning-based framework for quantifying uncertainty in gold prospectivity mapping. Hundreds of potential maps were generated during the process using three widely used ML methods, including random forest (RF), support vector machine (SVM), and k-nearest neighbors classifier (KNC), to delimit prospective zones and quantify the uncertainty associated with negative random sampling. For each ML method, we generated 400 negative training datasets using random locations constrained by spatial, as well as spatial and geological, criteria. This methodology was applied to a real case study in Juruena Mineral Province, northern Mato Grosso State, Brazil.

2. Geological Setting of the Juruena Mineral Province

2.1. Regional Geology and Tectonics

The Juruena Mineral Province (JMP) is a gold-(polymetallic) region located in the south-central portion of the Amazonian Craton (AC). The AC is a tectonically stable area of approximately 4.5 million km^2 composed of two Precambrian shields (Guiana and Central Brazil), which are separated in its central area by the Phanerozoic covers of the Solimões and Amazon basins [11,12]. Early studies divided the region into six major geochronological provinces: Central Amazonian (>2.3 Ga), Maroni-Itacaiúnas (1.95–1.80 Ga), Ventuari-Tapajós (1.95–1.80 Ga), Rio Negro-Juruena (1.8–1.55 Ga), Rondonian-San Ignácio (1.55–1.3 Ga), and Sunsás (1.3–1.0 Ga) [13]. These provinces evolved due to a sequence of continental accretion events characterized by a succession of magmatic arcs. Lately, Refs. [13,14] associated the origin of Sunsás and some parts of Rondonian-San Ignácio and Maroni-Itacaiúnas provinces with continental collisional events.

Santos et al. [15,16], propose seven tectonic provinces, including Carajás (3.0–2.5 Ga), Amazônia Central (Archean), Transamazonas (2.26–2.01 Ga), Tapajós-Parima (2.03–1.88 Ga),

Rio Negro-Juruena (1.82–1.52 Ga), Rondônia-Juruena (1.82–1.4 Ga), and Sunsás-k'Mudku (1.45–1.10 Ga). Juliani et al. [17] divide the south portion of the AC into two Paleoproterozoic magmatic arcs along the E-W direction.

According to [18], the JMP has been tectonically juxtaposed to the Tapajós Mineral Province (1.69 to 1.63 Ga) through a roughly E-W-trending shear zone. The combination of new results and observations gathered by [19], together with pre-existing data from the literature, on the plutonic and volcanic rocks of the eastern portion of the JMP indicate long-lived magmatism in the southern part of the AC between 2037 and 1757 Ma. The region formed through tectonic switching is represented by alternating periods of crustal thickening during flat subduction and crustal extension during slab-rollback at a single active convergent margin of an accretionary orogen [19].

2.2. The Juruena Mineral Province and Gold Mineralization

The JMP comprises a belt approximately 500 km long by 100 km wide constituted of Peixoto de Azevedo Domain (2.8–1.86 Ga), São Marcelo-Cabeça Group (1.85 Ga), Juruena Supersuite (1.81–1.75 Ga), Nova Monte Verde Complex (1.81–1.76 Ga), Colíder Group (1.81–1.76 Ga), Teles Pires suite (1.79–1.75 Ga), Roosevelt Group (1.76–1.74 Ga), Caiabis Group (1.98–1.37 Ga), and Alto Tapajós basin (0.30 Ga). The Peixoto de Azevedo Domain represents a remaining cratonic block of the southern part of the Ventuari-Tapajós Province (VTP), a continuation of the VTP to the south of the Alto Tapajós basin [20].

The geological framework of the eastern JMP is divided into five major lithostratigraphic units and has been simplified from [21] according to new proposals by [20], into (Figure 1): (i) 2.86–1.78 Ga Peixoto de Azevedo Domain; (ii) 1.82–1.78 Ga Colíder Group; (iii) 1.78–1.75 Ga Teles Pires suite; (iv) unconstrained rocks; and (v) 1.98–1.37 Ga Caibis Group, and Phanerozoic sedimentary cover sequences. The rocks in the province are mostly represented by Paleoproterozoic oxidized calc-alkaline, medium-to high-K, peraluminous to metaluminous I-type granites, volcanic and volcano-sedimentary sequences, with the presence of subordinate later A-type granites and volcanic rocks. Spatial analysis indicates a close relationship between the primary gold occurrences and I-type granites (calc-alkaline and oxidized), volcanic rocks, and volcano-sedimentary sequences that originated in continental magmatic arcs. Studies investigating hydrothermal alteration, styles, ore paragenesis, fluid inclusion, and isotopic data suggest the formation of the deposits in Paleoproterozoic magmatic-hydrothermal systems distributed in different crustal levels and positions [22].

Gold deposits in the JMP are mostly concentrated along with a set of WNW-ESE shear zones (so-called Peru-Trairão) and correspond to Cenozoic placers, eluvial deposits, and hydrothermal occurrences [23,24]. From 1980 to 1999, the region produced over 160 t of gold, mostly related to secondary ore sources. The region is currently focusing on primary types as part of a new exploration cycle. The primary gold mineralization was summarized based on the mineralization style and paragenesis of ore bodies as [25]: (i) veins, stockworks, and disseminated in shallow-emplaced granites; (ii) veins hosted in faults that crosscut granitoids; and (iii) epithermal gold-(polymetallic) veins in porphyries, with a more limited occurrence in volcanic and volcaniclastic rocks. The majority of gold occurrences in the JMP are associated with aplite and intermediate to mafic volcanic dykes [22].

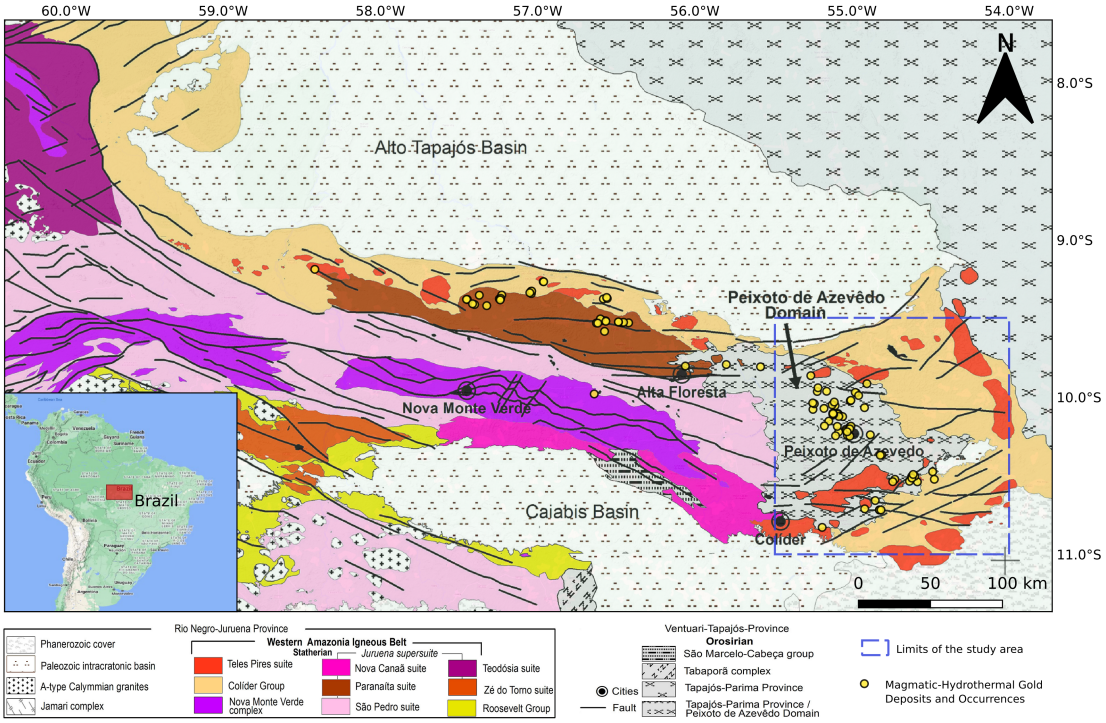

Figure 1. Simplified regional geological map modified from [20]. The limits of the study area are represented by the blue dashed lines, and the magmatic-hydrothermal gold deposits and occurrences in the JMP are represented by the yellow circles.

3. Materials and Methods

3.1. Feature Selection and Feature Engineering

We used airborne and geological data, freely available in the database of the Geological Survey of Brazil (GeoSGB-CPRM), and shuttle radar terrain mission (SRTM) data from the Earth Explorer platform of United States Geological Survey (USGS). The geological understanding of the magmatic-hydrothermal gold systems present in the JMP guided feature selection and engineering. The data were used to derive representations of source, transport, and alteration, which are key factors for ore-forming processes in the province.

The airborne data, i.e., gamma-ray, belong to Project 1121—Norte do Mato Grosso, coordinated by the Geological Survey of Brazil (CPRM) from 2012 to 2014. The survey consists of north-south lines spaced at 500 m, tie-lines separated by 10,000 m, terrain clearance of 100 m, and data collected each 8 (magnetic field) and 80 m (gamma-ray spectrometry), approximately. The gamma-ray data were used to compute the ratios between eTh/K and eU/K. This procedure allows highlighting possible potassic alteration or sericitized sites, commonly associated with the gold mineralization in the JMP [26,27]. Thus, the ML algorithms are expected to search for low ratio values due to the differential mobility between these three elements (K > Th and K > U) [28]. We created three maps to represent potential transport paths and sources for hydrothermal fluids in the region by buffering magnetic lineaments and dykes. Since the deposits are primarily found in lowland zones within the study area, the digital elevation model was left unchanged.

All of the features were re-interpolated to an identical regular grid with cell sizes of 125 m (Figure 2). To oversample and undersample the gamma-ray and elevation data, the linear and nearest neighbors methods were used, respectively. All the analyses were run using QGIS and Python Language.

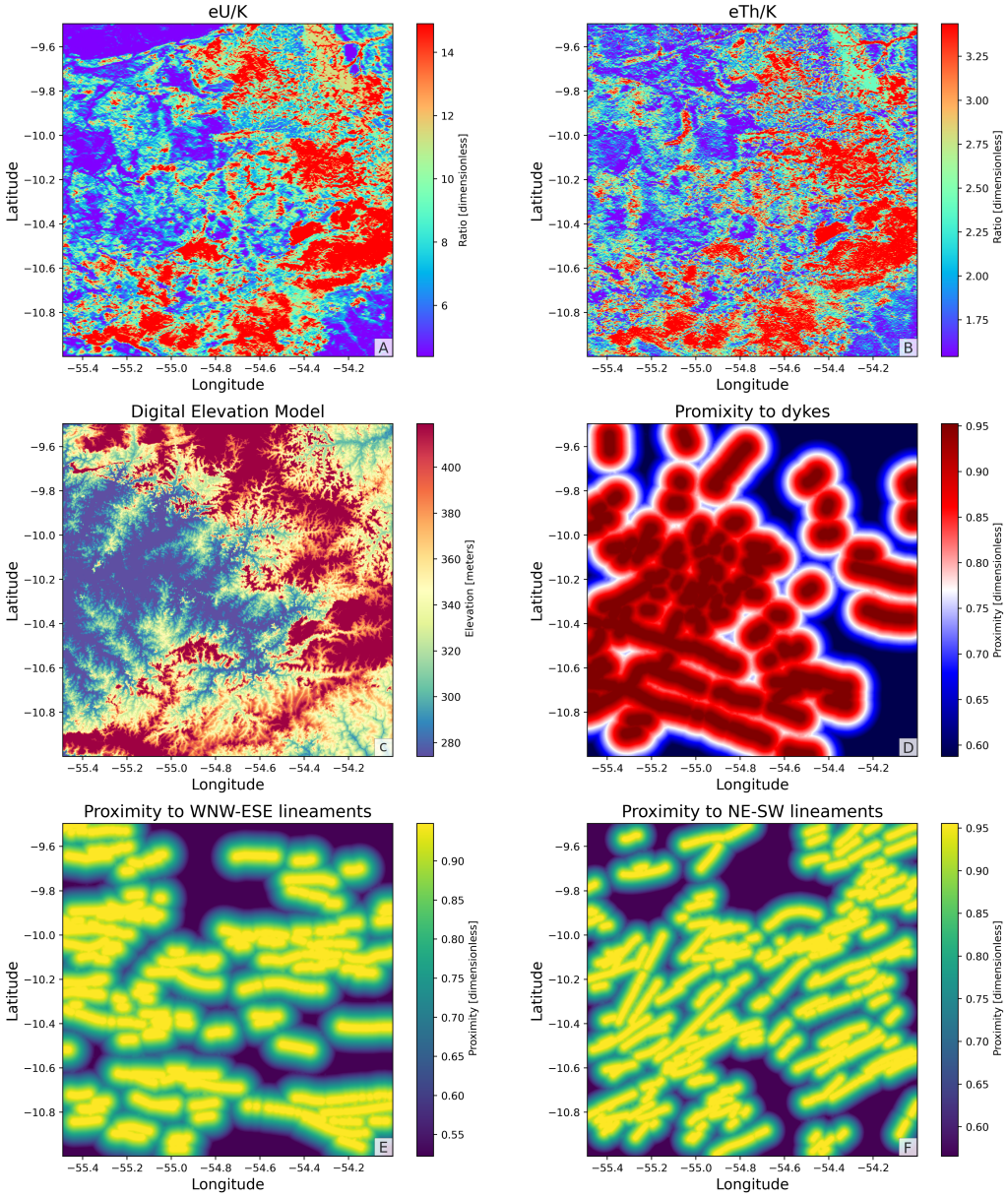

Figure 2. Features used to map prospective areas for magmatic-hydrothermal gold deposits. (**A**) Ratio between equivalent uranium and percentage of potassium; (**B**) ratio between equivalent thorium and percentage of potassium; (**C**) elevation from SRTM data; (**D**) proximity to mapped dykes; (**E**) proximity to WNW-ESE magnetic lineaments; (**F**) proximity to NE-SW magnetic lineaments.

3.2. Negative Sampling Methodology

In geoscientific problems, model validation presents issues due to spatial autocorrelation. That happens because the data violate the assumption of independence of most standard statistical procedures once natural phenomena create spatial continuity of physical/chemical properties [29,30]. Given this, to avoid over-optimistic results, it is preferable to split the data into training and test sets according to geographic regions. In this paper, the training and test sets are composed only of samples within the north and south portions of the study area, respectively.

Two sampling constraints were used to ensure that negative datasets were created using random point locations known as 'unfavorables' for deposits of the type sought: (i) point pattern analysis (PPA) was used to determine how close the deposits are to one another in space. This allows one to determine how far away from known deposits the likelihood of discovering a new one is sufficiently low and thus set the minimum distance

between positive and negative examples (MDPN). In our case, an MDPN of 9.92 km was set according to the major breaking point before the sill present in the graph in Figure 3, which corresponds to a likelihood of 3.45% of finding a new deposit; (ii) the lithologic unit favorability is a metric that represents the strength of the spatial association (SA) between rock units and known deposits. The SA is calculated as the proportion of a unit's area covered by known deposits divided by the proportion of the study area occupied by the same unit [1] (Figure 4). It represents how likely a lithologic unit is to host magmatic-hydrothermal gold deposits in this context and could be used as another evidential feature. For this study, however, we used this metric as the second constraint for creating negative datasets. As a result, negative examples were chosen only in locations where the SA is zero.

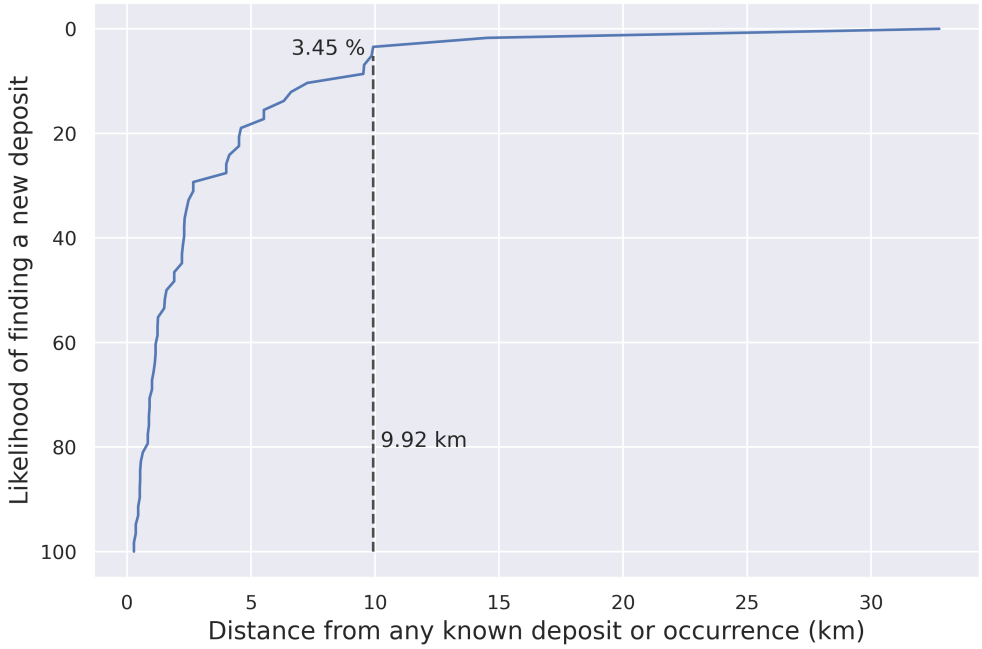

Figure 3. Likelihood of finding new deposits vs. distance from any known deposit or occurrence.

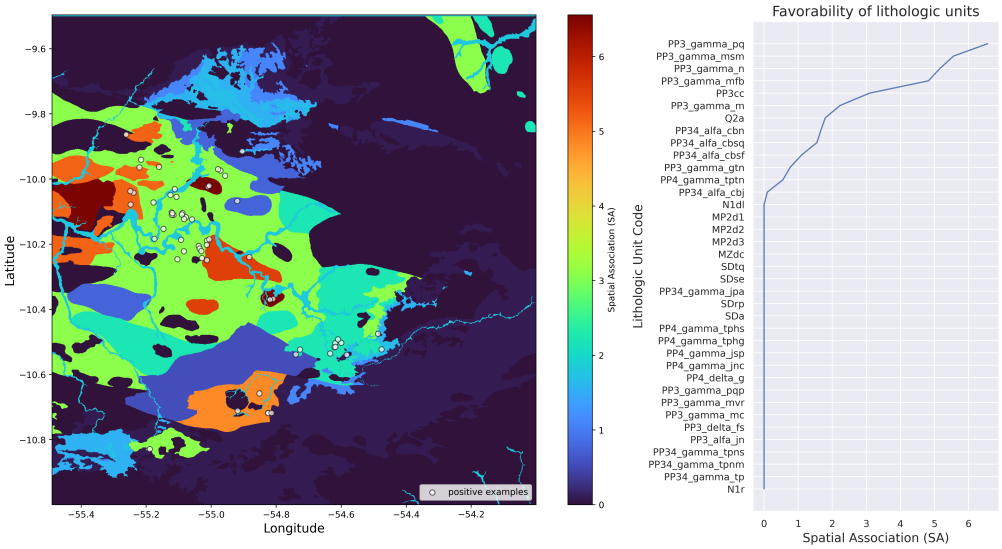

Figure 4. Spatial association between lithologic units and gold deposits in the study area. On the left, the map displays the lithologic units colored according to their spatial association to gold deposits. On the right, the graph shows the unit codes vs. their spatial association to deposits, in increasing order.

An ML-based framework was used to evaluate how these criteria influence the uncertainty of the predictions. The following steps comprise this framework: (1) create the negative dataset using the constraint(s); (2) merge it with the positive dataset created using the positive examples available in the study area (i.e., deposits and/or occurrences); (3) divide the new dataset into training and testing sets; (4) use the training set for training the model(s); (5) generate the potential map of the test area; (6) repeat the process n-times, storing the predictions and metrics for each realization; and (7) calculate the mean and standard deviation of the potential maps. This procedure was used to generate the negative datasets (Figure 5). Each model type resulted in the generation of 400 potential maps.

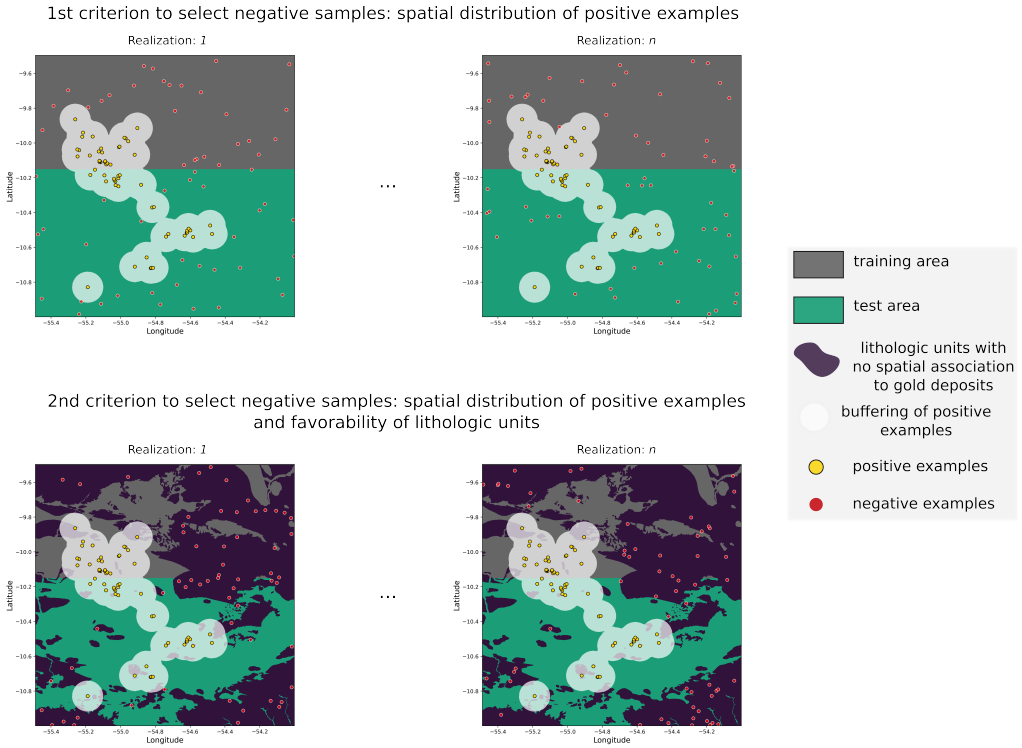

Figure 5. Scheme representing the process of generation of negative datasets using the two adopted criteria. All figures display the study area divided between training and testing. The region in white represents a buffer in which the maximum distance from positive examples is 9.92 km.

3.3. Model Training

The models were trained using the known magmatic-hydrothermal gold deposits and occurrences found in the records of the GeoSGB-CPRM. In the study area, there are 11 deposits, of which 6 belong to the training area and 5 to the test area. Of all mineral occurrences, 21 are in the training area, and 27 are in the test area. The negative datasets contain the same number of samples as the sum of positive examples in the correspondent areas, i.e., the ML models were trained with 32 positive and 32 negative examples.

A random search algorithm was used to define the optimal hyperparameters [31]. This algorithm performs a random search for the best hyperparameters in a parameter space defined by the user. It also performs cross-validation on batches of the training set. We used 100 iterations and 5 folds to determine the best hyperparameters for the three model types (Table 1).

Table 1. Optimal hyperparameters defined for the two adopted criteria using random search.

Model	Hyperparameters			
RF	n-estimators	bootstrap	criterion	max-depth
	300	False	gini	5
SVM	kernel	gamma	C	
	RBF	0.5	0.25	
KNC	n-neighbors	weights	algorithm	p
	7	distance	auto	2

3.4. Algorithms

3.4.1. Random Forest

Random forest (RF) consists of a combination of tree predictors that depend on values of a random vector sampled independently and with the same distribution for all trees in the forest [32]. It is an ensemble machine learning method that combines multiple decision trees, realizing repeated predictions of the same phenomenon represented by the training dataset [1].

In ensemble methods, for the kth tree, a random vector Θ_k is generated, independently of the past random vectors $\Theta_1, \ldots, \Theta_{k-1}$, but presenting the same distribution. A tree grows using the training set and Θ_k, resulting in a classifier $h(x, \Theta_k)$ where x is an input vector. In dichotomous problems, as the classifications in MPM, the model iteratively splits the dependent variable (root mode) into binary pieces (leaf nodes) in every decision tree. The trees then search along all the splits in order to find the best one that maximizes the purity of the resulting trees [33].

There are many approximations for measuring impurity. The most frequently used ones are entropy, gain-ratio, Gini index, and chi-square. The entropy, for example, attempts to maximize the mutual information, constructing an equal probability node in the decision (Equation (1)):

$$Entropy = -\sum_j p_j \ \log_2 \ p_j; \tag{1}$$

While the classification error can be measured as (Equation (2)):

$$ClassificationError = 1 - max \ p_j; \tag{2}$$

where p_j is the probability of class j.

3.4.2. Support Vector Machine

Support vector machines (SVMs) are supervised learning methods used to model datasets in classification and regression problems. The algorithm initially seeks to construct a linear classifier to separate different classes with the widest decision boundaries. In more complex cases, in which the origin feature spaces are not linearly separable, the input data have to be transformed into a higher n-dimensional feature space through various kernel functions, where a classification problem can be performed by an (n − 1) dimensional plane referred to as a hyperplane. A soft margin kind is utilized to reduce the misclassification [34].

For linear problems, in a hyperplane decision function $f(x) = sgn((w \cdot x) + b)$, the Vapnik–Chervonenkis (VC) dimension is controlled by the norm of the weight vector w, and given a training set $(x_1, \ y_1), \ldots, (x_l, \ y_l)$, with $x_i \in \mathbb{R}^n$, and $\ y_i \in \pm 1$, a border/margin that better separates classes can be found by minimizing $\frac{1}{2}\|w\|^2$ subject to $y_i \cdot ((x_i \cdot w) + b) \geq 1$ for $i = 1, \ldots, l$, [35].

In nonlinear problems, for a training dataset defined as $\{x_i\}_{i=1}^n$ with n feature vectors, a labeled target dataset $\{y_i\}_{i=1}^n$ is associated with each vector x_i. The labels in this study,

are $y = 1$, and $y = 0$, indicating, respectively, deposits and non-deposits. As the input data cannot be linearly separated in the original feature space, they are mapped into a higher-dimensional space H by a mapping function $\Phi : \mathbb{R}^n \rightarrow H$ [36].

There are four popular kernel functions used in SVM algorithms, which include linear, radial basis, polynomial, and sigmoid functions. For geoscientific problems, the radial basis function (RBF) is considered the most suitable due to its simplicity and adaptability [37]. The RBF can be described as (Equation (3)):

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2),\ \gamma > 0, \tag{3}$$

in which γ determines the width of the RBF.

3.4.3. K-Nearest Neighbors

K-nearest neighbors (KNN) is one of the most fundamental and simple classification methods. It aims to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. The KNN classifier is commonly based on the Euclidean distance, seeking to segment and find patterns by learning from a training set $(x_{i1}, x_{i2}, \ldots, x_{ip})$, with n being the total number of input samples $(i = 1, 2, \ldots, n)$, and p the total number of features $(j = 1, 2, \ldots, p)$ [38]. The Euclidean distance between sample x_i and x_l $(l = 1, 2, \ldots, n)$ is defined as in Equation (4).

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + \cdots + (x_{ip} + x_{lp})^2}. \tag{4}$$

Classification typically involves partitioning samples into training and testing categories. Consider x_i and x as the training and test sample, respectively, and let ω be the true class of a training set and $\hat{\omega}$ be the predicted class for a test set $(\omega, \hat{\omega} = 1, 2, \ldots, \Omega)$, where Ω is the total number of classes. During the training process, only the true class ω is used. When testing, the model predicts $\hat{\omega}$ of each sample. For $K = 1$, the predicted class of test sample x is set equal to the class ω of its nearest neighbor, where m_i is a nearest neighbor to x if Equation (5) is true.

$$d(m_i, x) = min_j\{d(m_j, x)\}. \tag{5}$$

Generalizing $\forall\ K \in \mathbb{Z}_+^*$, the predicted class of test sample x is set equal to the most frequent true class among K nearest training samples. This forms the decision rule $D: x \rightarrow \hat{\omega}$ [38].

4. Results and Discussion

Two hundred potential maps were generated for each of the two aforementioned negative dataset criteria. Following that, we created six maps that represented the average predicted prospectivities for each model type and criterion (Figure 6). Additionally, the standard deviation was calculated for each of these scenarios in order to create maps that represented the inconsistency of the predictions in the test area locations (Figure 7).

The maps generated using the second criterion show optimistic results compared to those obtained using the first criterion. It is demonstrated by the increase in the area associated with very high probabilities (i.e., >0.9): when adopting the first criterion, the RF, SVM, and KNC models predicted very high probabilities in 15.2%, 8.2%, and 21.4% of the test area; whereas adopting the second criterion the RF, SVM, and KNC models predicted very high probabilities in 17.7%, 16.6%, and 26.4% of the test area, respectively. On the other hand, there was a reduction in the predictions' average standard deviation associated with deposit and occurrence locations (Figure 8).

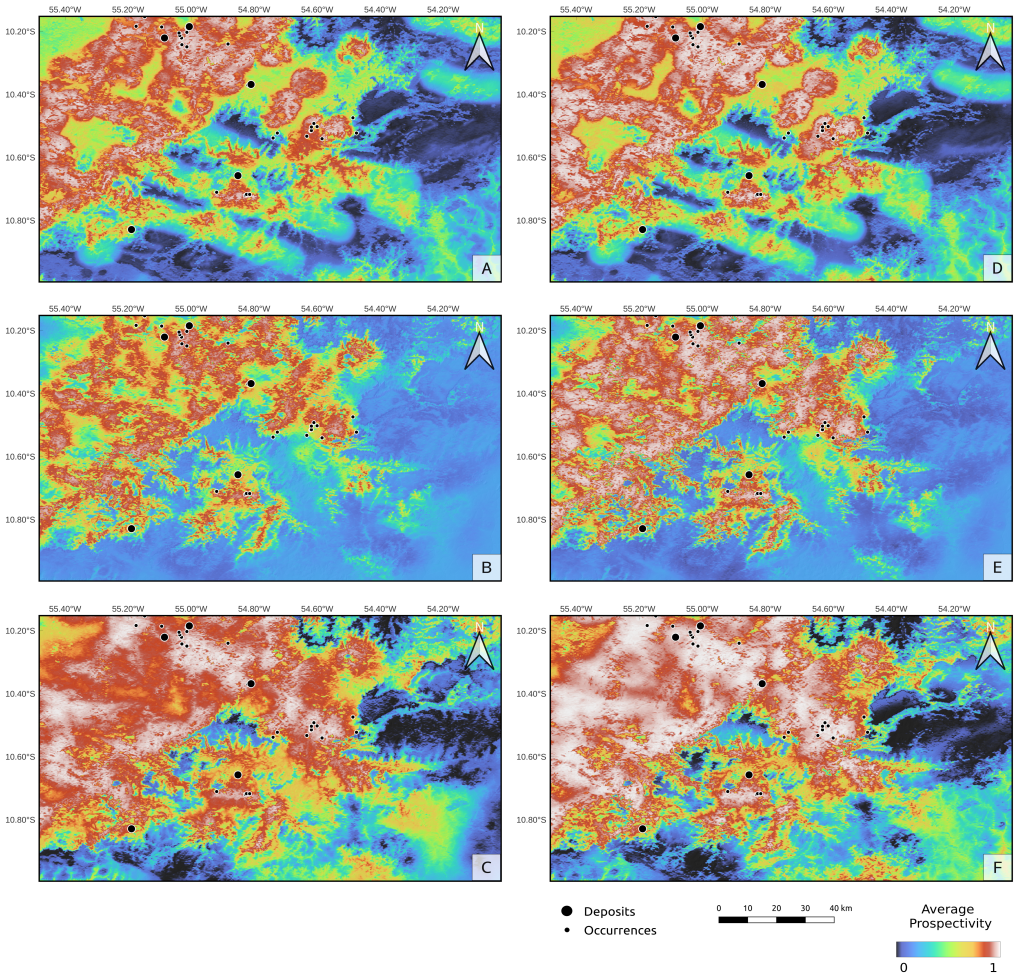

Figure 6. Potential maps. The left column shows the average prospectivity of the models when the first criterion was used to create negative datasets: (**A**) random forest; (**B**) support vector machine; (**C**) k-nearest neighbors. The right column shows the average prospectivity of the models when the second criterion was used to create negative datasets: (**D**) random forest; (**E**) support vector machine; (**F**) k-nearest neighbors.

The AUC scores in the three ML models were stored at each realization in order to analyze their performances in all scenarios. This way, graphs exhibiting the models' reliability could be generated (Figure 9). These graphs display the cumulative mean and standard deviation of the AUC scores along with the realizations. They also show the effect of the criteria used to generate the negative training datasets on the general model's performance. The standard deviations of the AUCs, for example, converge from 125 realizations in both cases; however, for models in which both the spatial distribution of deposits and favorability of lithologic units were used to constrain the selection of negative examples, the overall performances tend to be more stable, with standard deviations ranging between 0.037 and 0.052 as opposed to 0.044 and 0.063 in the first case. Furthermore, it appears that using geological knowledge to limit the space of possibilities for negative examples leads to a significant increase in performance. The cumulative AUCs in the first case ranged from 0.63 (for SVM models) to 0.77 (for KNC models), whereas in the second case, we achieved scores ranging from 0.77 (for RF models) to 0.86. (for KNC models). This is reflected in a reduction of positive example misclassification. The mode of the confusion matrices shows that when both criteria were used, most RF and KNC models correctly predicted all deposit locations (Figure 10).

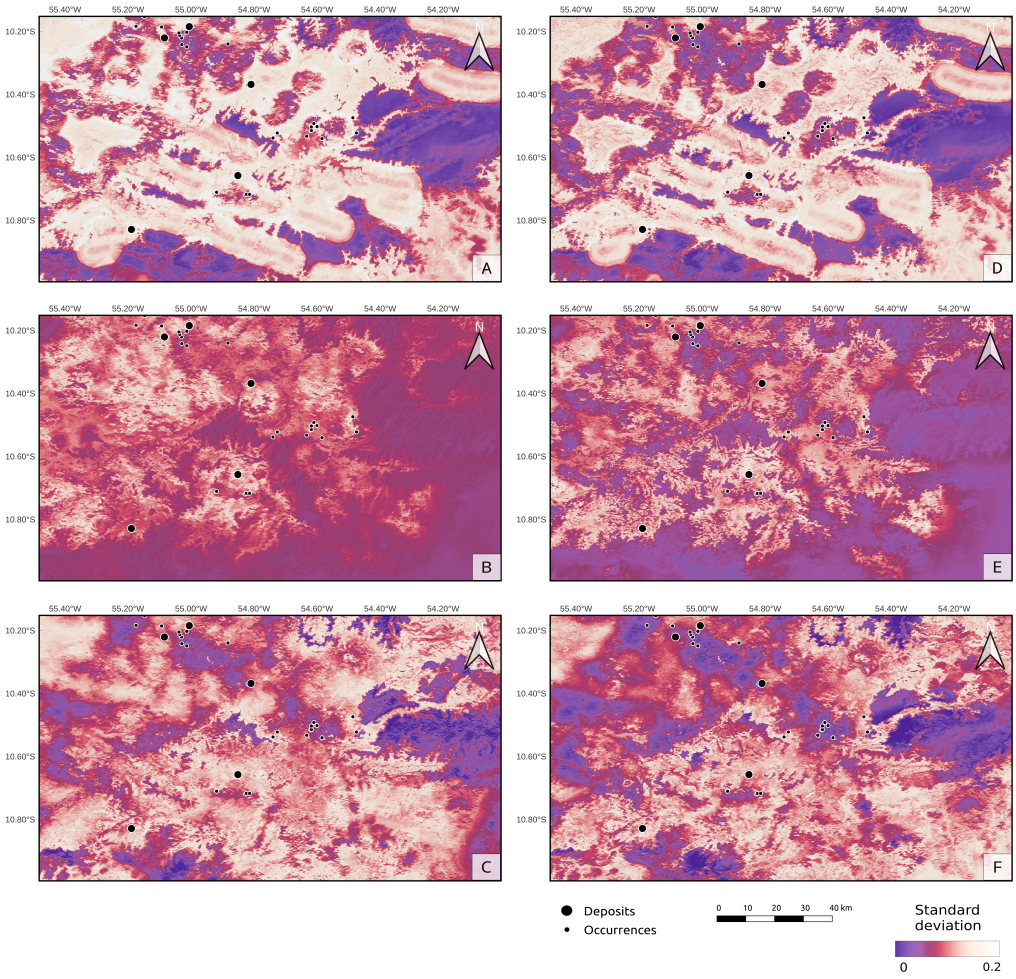

Figure 7. Standard deviation maps. The left column shows the standard deviation of the predictions when the first criterion was used to create negative datasets: (**A**) random forest; (**B**) support vector machine; (**C**) k-nearest neighbors. The right column shows the standard deviation of the predictions when the second criterion was used to create negative datasets: (**D**) random forest; (**E**) support vector machine; (**F**) k-nearest neighbors.

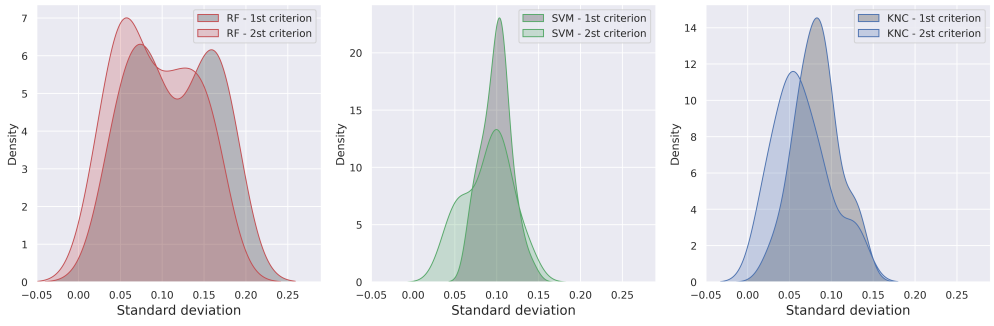

Figure 8. Kernel density estimation of the standard deviation of predictions related to deposit and occurrence locations.

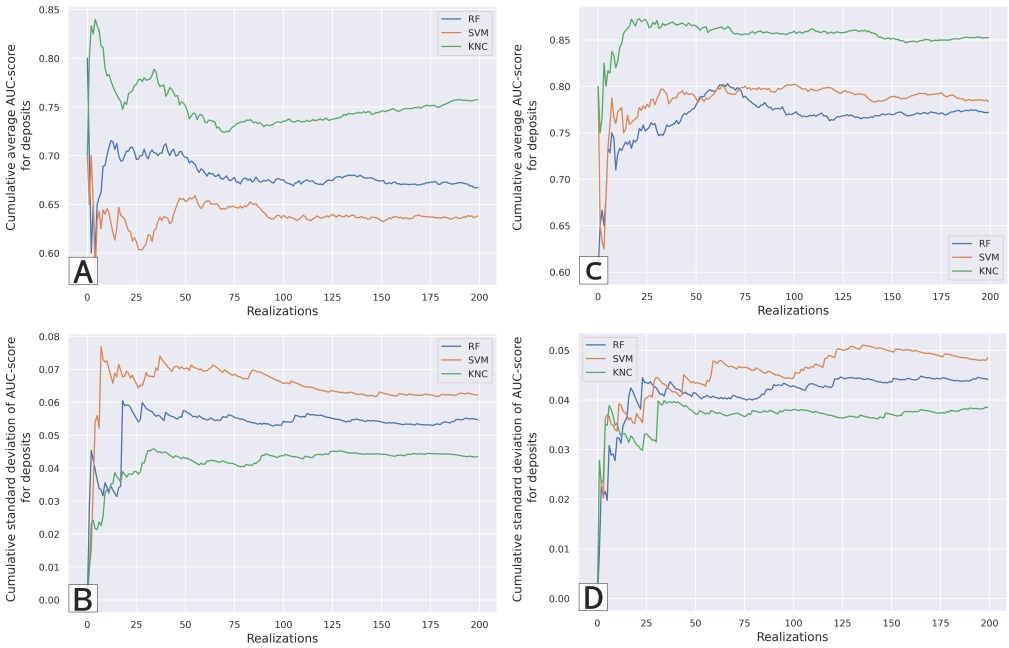

Figure 9. The area under the curve (AUC)-score of the deposits in the test area along with the realizations. The left column shows the cumulative statistics of the AUC when the first criterion was used to create negative datasets: (**A**) cumulative average AUC; (**B**) cumulative standard deviation of the AUC. The right column shows the cumulative statistics of the AUC when the second criterion was used to create negative datasets: (**C**) cumulative average AUC; (**D**) cumulative standard deviation of the AUC.

Effective prospectivity maps require models with reasonable input, i.e., proxies representing some or all of the mechanisms responsible for ore formation. Obtaining reliable and interpretable outputs, however, is just as important as providing appropriate input information to models. Although, due to the 'black-box' nature of machine learning modeling processes, it can be a difficult task. The feature importance is used to interpret the results of ML models. However, when using nonlinear methods, a simpler approximation is required in order to explain individual predictions from ML models. For this reason, we calculated SHAP values to approximate the feature importance of the SVM and KNC models [39]. The kernel density estimations (KDE) of the 'feature importance' of all models—trained using all negative training datasets generated—could thus be computed (Figure 11). The KDEs demonstrate that the three ML methods approximate linear solutions that are similar. Besides that, it is interesting to see how much the individual importance of the features varies depending on the negative dataset provided. This means that the models can map to the same solution in various ways. Thus, by running multiple models, one can obtain the average (sometimes converging) outcome and its associated uncertainty. In this case, we can see that the elevation and proximity to dykes are the two most important features, regardless of the criterion used. It demonstrates that, while the overall performance and stability of the models varied, the geological interpretation for the found solutions is the same for both negative example generation criteria.

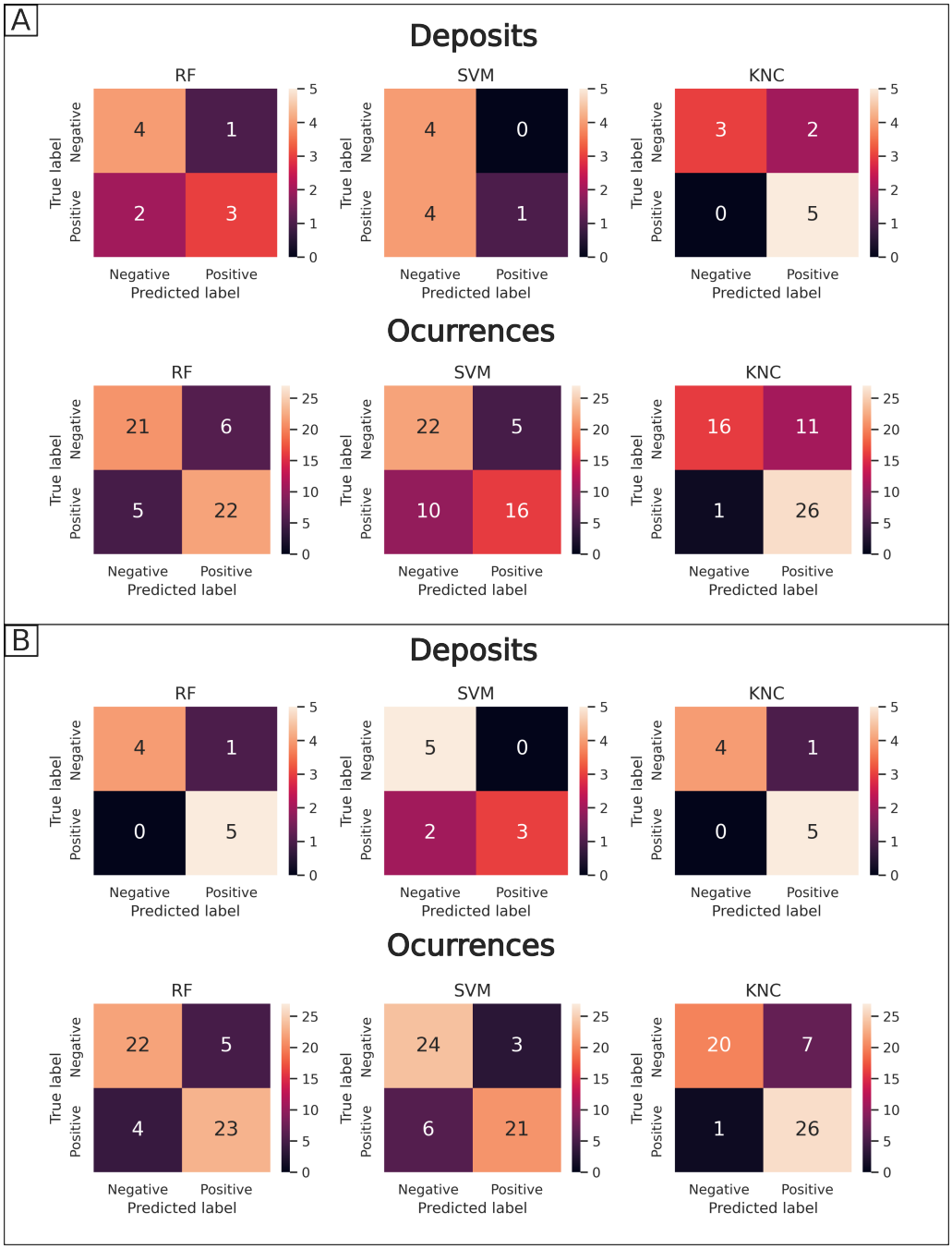

Figure 10. RF, SVM and KNC confusion matrices. (**A**) Ground truth vs. predictions for gold deposits and occurrences in which the first criterion to create negative datasets was adopted. (**B**) Ground truth vs. predictions for gold deposits and occurrences in which the second criterion to create negative datasets was adopted.

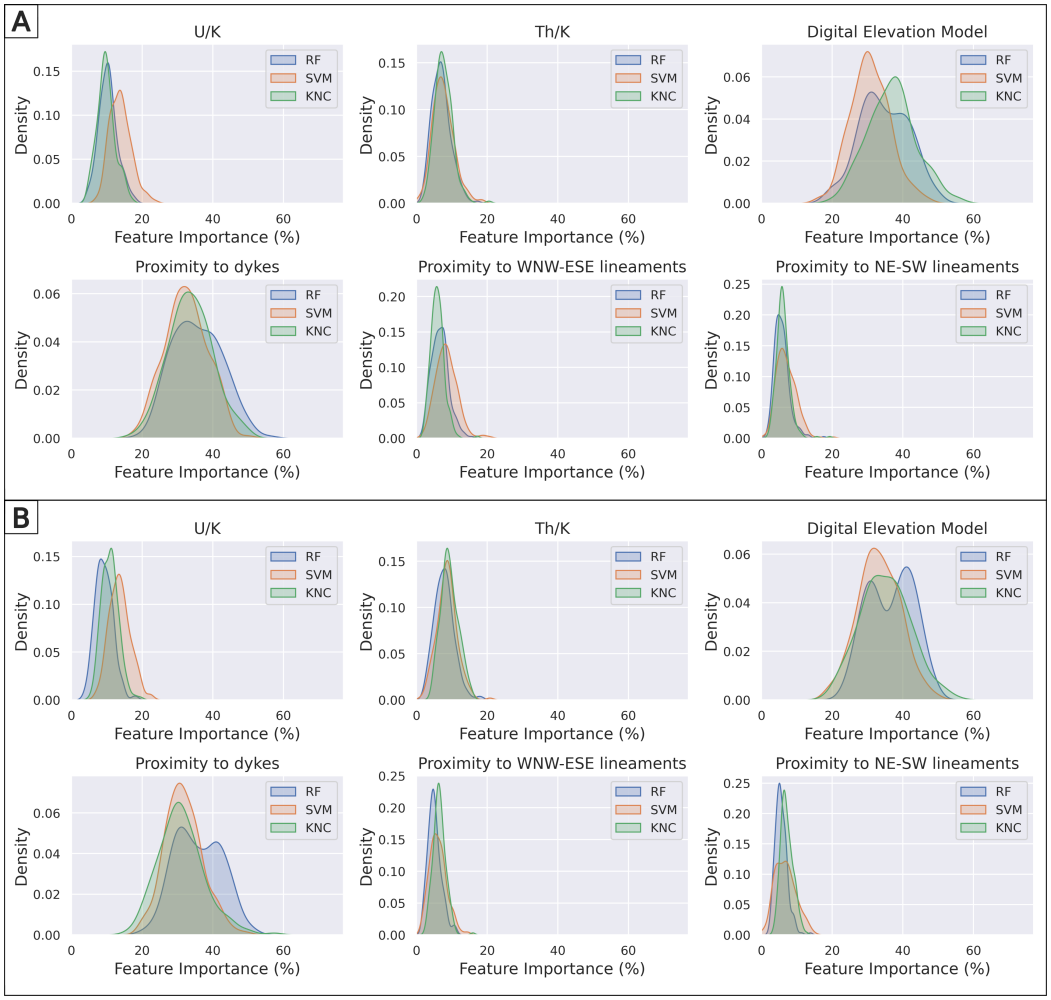

Figure 11. Kernel density estimation of the feature importance. (**A**) Estimated density of the feature importance of all realizations in which the first criterion to create negative datasets was adopted. (**B**) Estimated density of the feature importance of all realizations in which the second criterion to create negative datasets was adopted.

5. Conclusions

This study examined the effects of negative dataset generation in MPM when two different criteria were used to limit the selection of negative examples: (1) negative examples chosen outside the buffering of positive examples; (2) negative examples chosen in locations both outside the buffering of positive examples and without any spatial association to gold deposits. During the process, 200 potential maps for three ML methods were generated, including random forest, support vector machine, and k-nearest neighbors classifier. We were able to quantify the models' uncertainty by calculating the mean and standard deviation of the potential maps, which revealed where the models' predictions are more stable. Furthermore, we were able to see that, while using the second criterion produced more optimistic results, it reduced the standard deviation of predictions in deposit and occurrence locations. Besides that, when we added the favorability of lithologic units to the constraints, the number of false negatives in the RF and KNC predictions decreased to zero. The features had similar importance from one criterion to another, implying that the outcomes have similar interpretations or physical meaning. The elevation and proximity to dykes were the two most important features in this study. The Juruena Mineral Province in Northern Mato Grosso, Brazil, was used as a real case study for this research.

Author Contributions: Conceptualization, V.S.d.S., V.H.A.L., E.G. and M.B.; Methodology, V.S.d.S. and E.G.; software, V.S.d.S. and E.G.; validation, V.S.d.S., E.G. and V.H.A.L.; formal analysis, V.S.d.S. and E.G.; investigation, V.S.d.S., E.G., V.H.A.L. and M.B.; writing—original draft preparation, V.S.d.S., E.G. and V.H.A.L.; writing—review and editing, V.S.d.S., E.G., V.H.A.L. and M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the scripts and associated data are available on GitHub: https://github.com/victsnet/MPM---Juruena-Mineral-Province.git (accessed on 29 June 2022).

Acknowledgments: We would like to thank the Agency of Innovation of the University of Sao Paulo (AUSPIN) for the administrative support at the beginning of this project.

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

Abbreviations

The following abbreviations are used in this manuscript:

JMP	Juruena Mineral Province
AC	Amazonian Craton
RF	Random Forest
SVM	Support Vector Machine
KNC	K-Nearest neighbors Classifier
MDPN	Minimum Distance between Positive and Negative examples
AUC	Area Under the Curve
GeoSGB-CPRM	Database of the Geological Survey of Brazil

References

1. Carranza, E.J.M.; Laborte, A.G. Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm. *Ore Geol. Rev.* **2015**, *71*, 777–787. [CrossRef]
2. Zuo, R.; Carranza, E.J.M. Support vector machine: A tool for mapping mineral prospectivity. *Comput. Geosci.* **2011**, *37*, 1967–1975. [CrossRef]
3. Rodriguez-Galiano, V.; Sanchez-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 Geol. Rev.* **2015**, *71*, 804–818. [CrossRef]
4. Carranza, E.J.M. Controls on mineral deposit occurrence inferred from analysis of their spatial pattern and spatial association with geological features. *Ore Geol. Rev.* **2009**, *35*, 383–400. [CrossRef]
5. Vearncombe, J.; Vearncombe, S. The spatial distribution of mineralization; applications of Fry analysis. *Econ. Geol.* **1999**, *94*, 475–486. [CrossRef]
6. Ford, A.; Blenkinsop, T.G. Combining fractal analysis of mineral deposit clustering with weights of evidence to evaluate patterns of mineralization: Application to copper deposits of the Mount Isa Inlier, NW Queensland, Australia. *Ore Geol. Rev.* **2008**, *33*, 435–450. [CrossRef]
7. Gatrell, A.C.; Bailey, T.C.; Diggle, P.J.; Rowlingson, B.S. Spatial point pattern analysis and its application in geographical epidemiology. *Trans. Inst. Br. Geogr.* **1996**, *21*, 256–274. [CrossRef]
8. Carranza, E.; Hale, M.; Faassen, C. Selection of coherent deposit-type locations and their application in data-driven mineral prospectivity mapping. *Ore Geol. Rev.* **2008**, *33*, 536–558. [CrossRef]
9. Nykänen, V.; Lahti, I.; Niiranen, T.; Korhonen, K. Receiver operating characteristics (ROC) as validation tool for prospectivity models—A magmatic Ni–Cu case study from the Central Lapland Greenstone Belt, Northern Finland. *Ore Geol. Rev.* **2015**, *71*, 853–860. [CrossRef]
10. Zuo, R.; Wang, Z. Effects of random negative training samples on mineral prospectivity mapping. *Nat. Resour. Res.* **2020**, *29*, 3443–3455. [CrossRef]
11. Almeida, F.F.M.; Hasui, Y.; Brito Neves, B.B.; Fuck, R.A. Brazilian structural provinces: An introduction. *Earth-Sci. Rev.* **1981**, *17*, 1–29. [CrossRef]
12. Santos, J.O.S.; Hartmann, L.A.; Gaudette, H.E.; Groves, D.I.; Mcnaughton, N.J.; Fletcher, I.R. A new understanding of the provinces of the Amazon Craton based on integration of field mapping and U-Pb and Sm-Nd geochronology. *Gondwana Res.* **2000**, *3*, 453–488. [CrossRef]

13. Tassinari, C.C.; Macambira, M.J. Geochronological provinces of the Amazonian Craton. *Episodes-Newsmag. Int. Union Geol. Sci.* **1999**, *22*, 174–182. [CrossRef]
14. Tassinari, C.C.G.; Macambira, M.J.B. A evolução tectônica do Craton Amazônico. In Proceedings of the Congresso Brasileiro de Geologia, SBG, Araxá. 2004. Available online: https://repositorio.usp.br/item/001407316 (accessed on 29 June 2022).
15. Santos, J.O.S. Geotectônica dos escudos das Guianas e Brasil-Central. In *Geologia, Tectônica e Recursos Minerais do Brasil*; CPRM: Brasília, Brazil, 2003; Volume 4, pp. 169–226.
16. Santos, J.O.S.; Hartmann, L.A.; Faria, M.d.; Riker, S.R.; Souza, M.D.; Almeida, M.E.; McNaughton, N.J. A compartimentação do Cráton Amazonas em províncias: avanços ocorridos no período 2000–2006. *Simpósio De Geol. Da Amaz.* **2006**, *9*, 2006.
17. Juliani, C.; Carneiro, C.d.C.; Carreiro-Araújo, S.A.; Fernandes, C.; Monteiro, L.; Crósta, A. Estruturação dos arcos magmáticos paleoproterozoicos na porção sul do Cráton Amazônico: Implicações geotectônicas e metalogenéticas. *Simpósio De Geol. Da Amaz.* **2013**, *13*, 157–160.
18. Scandolara, J.; Correa, R.; Fuck, R.; Souza, V.; Rodrigues, J.; Ribeiro, P.; Frasca, A.; Saboia, A.; Lacerda Filho, J. Paleo-Mesoproterozoic arc-accretion along the southwestern margin of the Amazonian craton: The Juruena accretionary orogen and possible implications for Columbia supercontinent. *J. S. Am. Earth Sci.* **2017**, *73*, 223–247. [CrossRef]
19. Trevisan, V.G.; Hagemann, S.G.; Loucks, R.R.; Xavier, R.P.; Motta, J.G.; Parra-Avila, L.A.; Petersson, A.; Gao, J.F.; Kemp, A.I.; Assis, R.R. Tectonic switches recorded in a Paleoproterozoic accretionary orogen in the Alta Floresta Mineral Province, southern Amazonian Craton. *Precambrian Res.* **2021**, *364*, 106324. [CrossRef]
20. Rizzotto, G.J.; Alves, C.L.; Rios, F.S.; Barros, M.A.d.S. The Western Amazonia Igneous Belt. *J. S. Am. Earth Sci.* **2019**, *96*, 102326. [CrossRef]
21. Assis, R.R.; Xavier, R.P.; Creaser, R.A. Linking the Timing of Disseminated Granite-Hosted Gold-Rich Deposits to Paleoproterozoic Felsic Magmatism at Alta Floresta Gold Province, Amazon Craton, Brazil: Insights from Pyrite and Molybdenite Re-Os Geochronology. *Econ. Geol.* **2017**, *112*, 1937–1957. [CrossRef]
22. Juliani, C.; Rodrigues de Assis, R.; Virgínia Soares Monteiro, L.; Marcello Dias Fernandes, C.; Eduardo Zimmermann da Silva Martins, J.; Ricardo Costa e Costa, J. Gold in Paleoproterozoic (2.1 to 1.77 Ga) Continental Magmatic Arcs at the Tapajós and Juruena Mineral Provinces (Amazonian Craton, Brazil): A New Frontier for the Exploration of Epithermal–Porphyry and Related Deposits. *Minerals* **2021**, *11*, 714. [CrossRef]
23. Paes de Barros, A.J. Granitos da Região de Peixoto de Azevedo: Novo Mundo e Mineralizações auríferas Relacionadas-Província Aurífera Alta Floresta (MT). 2007. Available online: http://repositorio.unicamp.br/jspui/handle/REPOSIP/287713 (accessed on 15 June 2022).
24. Miguel, E., Jr. Mineralizações Auríferas do Lineamento Peru-Trairão, Província Aurífera de Alta Floresta-MT: Controle Estrutural e Idade U-Pb Das Rochas Hospedeiras. Master's Thesis, Universidade Estadual de Campinas, Unicamp, Brazil, 2011.
25. Santos, J.O.S.; Groves, D.I.; Hartmann, L.A.; Moura, M.A.; McNaughton, N.J. Gold deposits of the Tapajós and Alta Floresta Domains, Tapajós–Parima orogenic belt, Amazon Craton, Brazil. *Miner. Depos.* **2001**, *36*, 278–299. [CrossRef]
26. Abreu, M. *Alteração Hidrotermal e Mineralização Aurífera do Depósito de Novo Mundo, Região de Teles Pires-Peixoto de Azevedo, Província de Alta Floresta (MT)*; Trabalho de Conclusão de Curso; Instituto de Geociências, Universidade Estadual de Campinas: Campinas, Brazil, 2004; 29p.
27. Assis, R.R. Depósitos Auríferos Associados ao Magmatismo Félsico da Provincia de Alta Floresta (MT), Craton Amazonico: Litogeoquímica, Idade Das Mineralizações e Fonte dos Fluidos. Ph.D. Thesis, Universidade Estadual de Campinas, Unicamp, Brazil, 2015.
28. Shives, R.B.; Charbonneau, B.; Ford, K.L. The detection of potassic alteration by gamma-ray spectrometry—Recognition of alteration related to mineralization. *Geophysics* **2000**, *65*, 2001–2011. [CrossRef]
29. Griffith, D.A. *Spatial Autocorrelation: A Primer*; Association of American Geographers: Washington, DC, USA, 1987.
30. Legendre, P. Spatial autocorrelation: trouble or new paradigm? *Ecology* **1993**, *74*, 1659–1673. [CrossRef]
31. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
32. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
33. Rodriguez-Galiano, V.; Chica-Olmo, M.; Chica-Rivas, M. Predictive modelling of gold potential with the integration of multisource information based on random forest: A case study on the Rodalquilar area, Southern Spain. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1336–1354. [CrossRef]
34. Suykens, J.A.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
35. Schölkopf, B.; Simard, P.; Smola, A.J.; Vapnik, V. Prior knowledge in support vector kernels. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 30 November–5 December 1998; pp. 640–646.
36. Burges, C.J. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167. [CrossRef]
37. Zuo, R.; Zhang, Z.; Zhang, D.; Carranza, E.J.M.; Wang, H. Evaluation of uncertainty in mineral prospectivity mapping due to missing evidence: A case study with skarn-type Fe deposits in Southwestern Fujian Province, China. *Ore Geol. Rev.* **2015**, *71*, 502–515. [CrossRef]
38. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
39. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3–7.