Mineralia Slovaca, Web ISSN 1338-3523, ISSN 0369-2086, https://doi.org/10.56623/ms.2024.56.2.4
56, 2 (2024), 165 – 180, © Authors 2024. CC BY 4.0

A comparative study of decision tree and support vector machine methods for gold prospectivity mapping

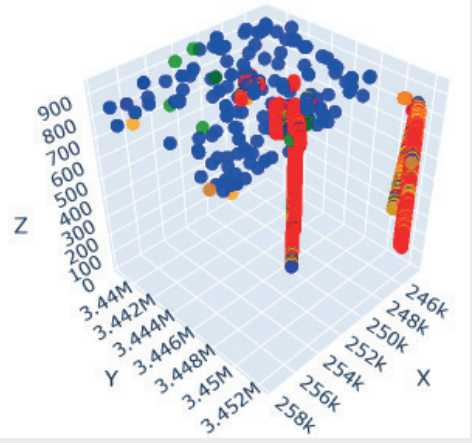
MOHAMMAD EBDALI[1] and ARDESHIR HEZARKHANI[2]*

[1, 2]Amirkabir University of Technology, Department of Mining Engineering, Tehran, Iran; ardehez@aut.ac.ir

Abstract: Elements geochemical anomalies mapping is one of the main goals of geochemical investigations. Since the field studies are time-consuming and costly, the data processing, applying machine learning methods could be applied, which is less expensive, faster, and more accurate. In this study, two machine learning methods, including decision tree algorithm and support vector machine, were applied for gold prospectivity mapping. To reach this aim, silver, copper, lead and zinc values were used as input variables. The comparison of the analytical results obtained from the two mentioned methods confirms that the DT (decision tree) model has sufficient accuracy and better performance than the SVM (support vector machine) model for preparing gold prospectivity mapping in the studied area.

Key words: decision tree, machine learning, mineral prospectivity mapping, support vector machine

Graphical abstract



Highlights

- Two machine learning methods, including decision tree algorithm and support vector machine, were applied for gold prospectivity mapping. To reach applicable results, the silver, copper, lead and zinc values were used as input variables.
- The comparison of the analytical results obtained from the two machine learning methods confirms that the DT (decision tree) model has sufficient accuracy and better performance than the SVM (support vector machine) method for preparing gold prospectivity mapping in the studied area.

1 Introduction

New techniques for estimating and evaluating variables have been created from research conducted over an extended period. New techniques like machine learning algorithms are used to address the issue of variable estimation (Dutta et al., 2010). These algorithms learn how data is related by using the examples provided to them. The attraction of these non-linear estimators is their capability to function like a black box. By providing sufficient data to these algorithms and subsequently educating them, they have the ability to grasp the connection between the input data, like sample coordinates, and the output data, such as the ore grade at specific points. With this approach, there is no need to take into account assumptions regarding linearity for components, coefficients, or relationships (Zhang et al., 2024). Machine learning algorithms are categorized into supervised learning methods and unsupervised learning methods (Chatterjee et al., 2010b; Jafrasteh et al., 2018). Both main categories have various techniques and algorithms which are utilized, based on the specific problem and the type and amount of data available. Decision tree is a commonly applied tool and technique for data mining proficiency. This method can be very beneficial in situations where the volume of data is extremely large. Data mining involves discovering valuable knowledge that is hidden and unknown within databases. There are two main categories of data mining methods: descriptive and predictive. Clustering is among the most well-known descriptive techniques, while classification is considered one of the most crucial predictive techniques. The clustering perspective is crucial in data mining for analysing large amounts of data and samples with different characteristics,

ORCID Nos. of authors: Mohammad Ebdali (http://orcid.org/0009-0002-4114-1611), Ardeshir Hezarkhani (http://orcid.org/0000-0002-1149-3440)

as it involves important methods and techniques (Anderberg, 1973). Data is clustered in the clustering method by maximizing similarity within groups and minimizing it between groups. The process of classification involves identifying data categories or concepts in order to create a model that can predict the categories of unknown objects. A classifier is a function that learns to assign a data item to a specific category from a set of predefined categories. Decision Tree, Bayes classifier, and Neural Network are a few popular classification techniques. A decision tree can generate easy-to-understand explanations of the connections within a set of data and is useful for tasks involving classification and prediction. This decision-making framework can be implemented through mathematical and computational methods that aid in categorizing and summarizing a dataset (Tsai & Yen-Jiun, 2009). Among the varied and effective methods in the artificial intelligence field for estimating and evaluating grades are artificial neural networks (Chatterjee et al., 2010a, b; Guo, 2010; Li et al., 2010; Mahmoudabadi et al., 2009; Nezamolhosseini et al., 2017; Samanta et al., 2005; Sayadi & Shahrabadi, 2008; Tahmasebi & Hezarkhani, 2010, 2012; Tsai & Yen-Jiun, 2009), neural fuzzy inference (Tahmasebi & Hezarkhani, 2010), random forest (Jafrasteh et al., 2018), and support vector machine (Maleki et al., 2014; Matías et al., 2004; Tenorio et al., 2015). In mining engineering, artificial neural networks are used not only for grade estimation, but also for tasks like blast network dimension estimation (Amnieh et al., 2012), processing factory design (Kotake et al., 2002; Singh et al., 2013), geological classification (Thuijsman, 1995), remote sensing data classification (Miller, 1995; Wang, 2005), and identifying failure models in Underground Mining (Lee & Sterling, 1992; Shahin et al., 2008). The decision tree algorithm, the foundation of the random forest algorithm, has been extensively utilized in geological and remote sensing applications as a successful classifier (Jhonnerie et al., 2015; Krishna et al., 2018; Masoumi et al., 2017).

Support vector machine, a supervised learning technique, is utilized for both classification and regression purposes. The SVM classifier operates by linearly classifying the data. It aims to choose the plain that has a higher confidence margin. Utilizing nonlinear programming techniques, the optimal line for the data is determined by solving the equation, as these methods are commonly used for solving constrained problems. Prior to the linear division, the data is transformed into a higher dimensional space using the kernel function to enable the machine to classify the data with high complexity. To address the issue of dealing with high dimensions, Lagrange's duality theorem is employed to convert the original minimization problem into a dual form. This new form involves a kernel function, a simpler version of the complex function phi that maps to the high-dimensional space by multiplying the phi function's vector. Various types of kernel functions, such as exponential, polynomial, and sigmoid kernels, can be utilized. Despite being newer than the artificial neural network method, the support vector machine method has quickly gained popularity due to its robust mathematical foundation (Kecman, 2001, 2004; Smola & Schölkopf, 2004). This technique is commonly employed in pattern recognition software, such as spatial data analysis (Dutta et al., 2010), creating natural radioactivity maps (Pozdnoukhov, 2005), estimating arsenic levels in bedrock from river sediments using support vector machine method (Twarakavi et al., 2006), identifying alteration zones linked to mineralization (Abbaszadeh et al., 2013, 2015; Soliman & Mahmoud, 2012), and classification of remote sensing images (Moorthi et al., 2011; Soliman & Mahmoud, 2012; Soliman et al., 2012). This approach has also been applied in determining the grade of slate deposits (Matías et al., 2004), glacial type platinum (Tenorio et al., 2015), gold (Chatterjee et al., 2010a), and iron (Maleki et al., 2014) in the grade estimation field. Hence, in this study, decision tree algorithm and support vector machine were applied to conduct gold prospectivity mapping.

2 Geological setting of the study area

The Janja area is located in Sistan and Paluchestan province. The region is situated in a flat sandy plain with sparse vegetation and few rock outcrops (Fig. 1), with altitudes ranging from 800 to 900 m a. s. l.

The research area is a section of the Zabul-Zahedan-Saravan subregion. This subregion is included in the flysch basin located in eastern Iran, commonly known as the eastern Iranian mountains. The region of the province contains a deep oceanic bedrock that is overlaid with a thick series of flysch deposits of Late Cretaceous-Oligocene age. The mentioned basin is believed to have originated after a collision within the continent between the Lot block to the west and the Afghan block to the east, following with formation of oceanic crust and ophiolitic complexes. While much of the crust has been lost in Neo-Alpine subduction zone, remnants can be seen along deep, longitudinal faults like the Nehbandan fault in the area. According to the geologic map of the area (Fig. 2), the Cretaceous Flysch (Kuf) is the oldest and most prominent solidified unit. This flysch is made up of alternating beds of grey to green greywackes and calcareous lichens. The majority of calcareous lichens displays a light green, weathered color, with occasional beds of fine-grained greywackes. In multiple areas of the region, thick beds of calcareous lichen (Kumr) units are visible. These limestone lichens are associated with the Cretaceous flysch and, in terms of lithology, are identical to the lower formations, however, they are consistently found below a layer of volcaniclastic sediments. The thickness of unit varies, reaching up to

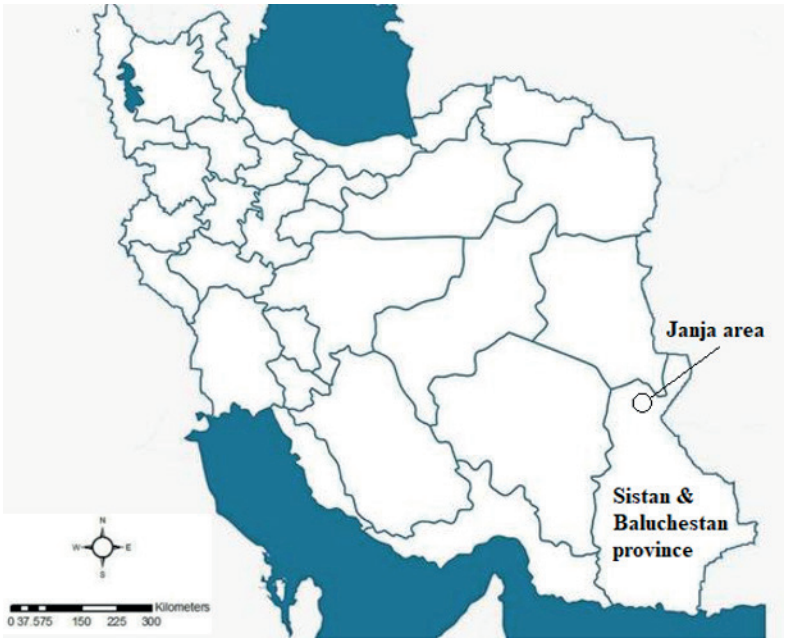


Fig. 1. Geographical location map of Janja area.

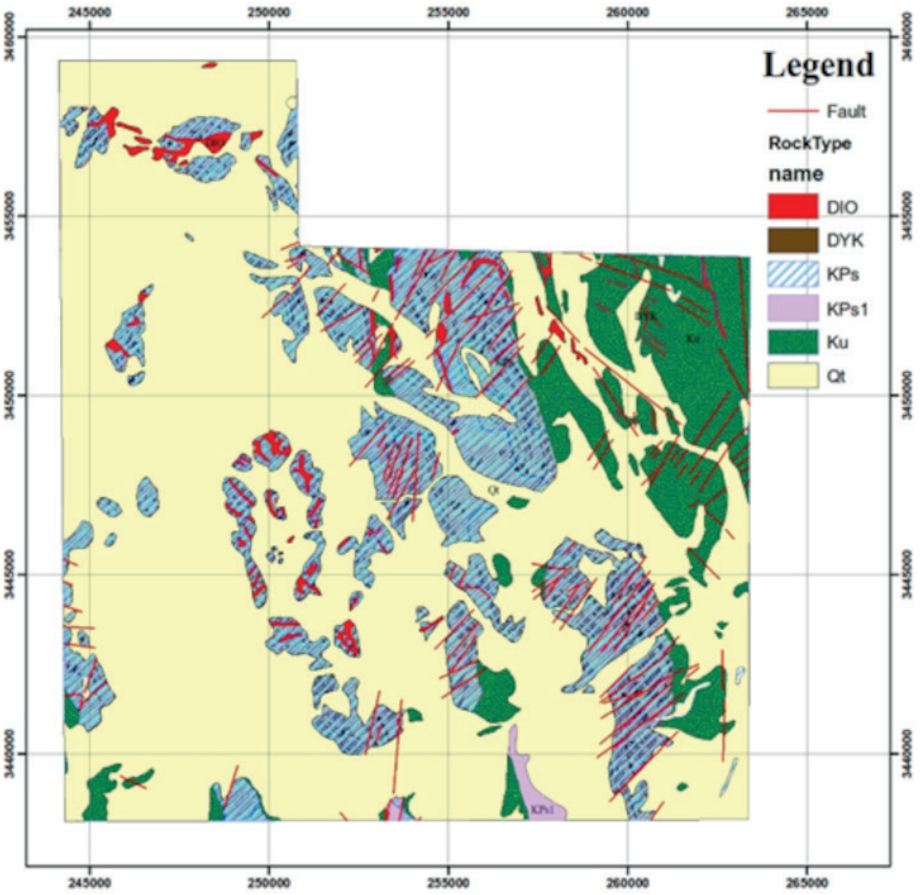
536 meters in the eastern part of the ophiolitic belt. The Sefidabe Formation (KPs) is one of the largest units found in the Janja area, particularly in the central and eastern parts of the zone. This structure is primarily made up of a series of volcaniclastic and pyroclastic deposits, which are divided into lower and upper sections in some areas. The bottom section (KPs1) is made up of thin to medium layers of volcaniclastic material, along with some volcanic flows, clastic sediments, and limestones. The upper section displays characteristics of stratified volcanic and pyroclastic rocks with pale green lichen layers alternated. In the north, these units are hundreds of meters thick, but their thickness decreases quickly towards the east and west. In certain areas, these units appear to be similar to Kuf sediments.

3 Data and methods

3.1 Sampling operation

Considering the extent of exploration and the dense waterway network in the area, a sampling system was developed for gathering geochemical and heavy mineral samples. This system consists of 285 geochemical samples and 59 heavy mineral samples from waterways (Fig. 3).

Alongside collecting samples from the sediment of waterways, 18 trenches were excavated on the surface where polymetallic veins are exposed, as well as in the area where intense operations were conducted. It was tried to design the trenches perpendicular to the trend of the veins based on the outcrops of the veins. Additionally, 16 boreholes



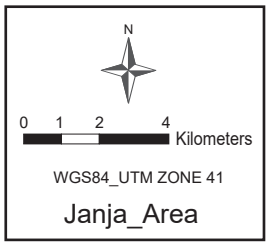


Fig. 2. Geological map of Janja area.

were drilled for polymetallic mineralization in the region. A combined total of 3 754 samples were gathered from trenches and boreholes excavated in the area.

3.2 Decision tree algorithm

The decision tree algorithm belongs among the most commonly used data mining algorithms. In data mining, the decision tree serves as a predictive model that is applicable for regression and classification models. If used for regression, it is termed a regression tree; if for classification, it is known as a classification tree. A set of binary decisions is used by the decision tree to carry out multi-stage classifications. Every choice separates a collection of pixels into two categories according to a set of logical conditions. A tree may contain any quantity of decision nodes. Information from various origins and categories can be utilized in constructing a decision tree (Zaremotlagh & Hezarkhani, 2017). Decision trees are effective methods for predicting and clarifying the connection between certain measurements and a target variable. The decision tree algorithm was utilized in this study to pinpoint potential mining locations. The decision tree picks a variable with a high information value to identify the promising range (equation [1]). The effectiveness of the chosen variable is assessed by utilizing equation [2] and the entropy reduction criterion (Zhang et al., 2019).

$$E(t) = \sum_{j}^{} p(j \mid t) Logp(j \mid t) p, \quad [1]$$

Where $E(t)$ is the entropy criterion – the relative frequency of class j in node t. Hence, the information gain for every split is calculated in the following manner:

$$GAIN_{Split} = E(p) - \sum_{i=1}^{k} \frac{ni}{n} E(j), \quad [2]$$

Where $E(p)$ is the overall entropy at parent node and is the sum of weighted entropy at each child node. The parent node is divided into k components and the number of records in the component is i. This maximizes the similarity and difference between the data sets in nodes (Shirali, 2016). In the decision tree framework, the forecast generated by the tree is described as a set of rules. Every route from the starting point to a last branch of the decision tree represents a rule, and ultimately, the leaf is marked with the category that contains the highest number of data entries. A feature selection criterion is a method to choose

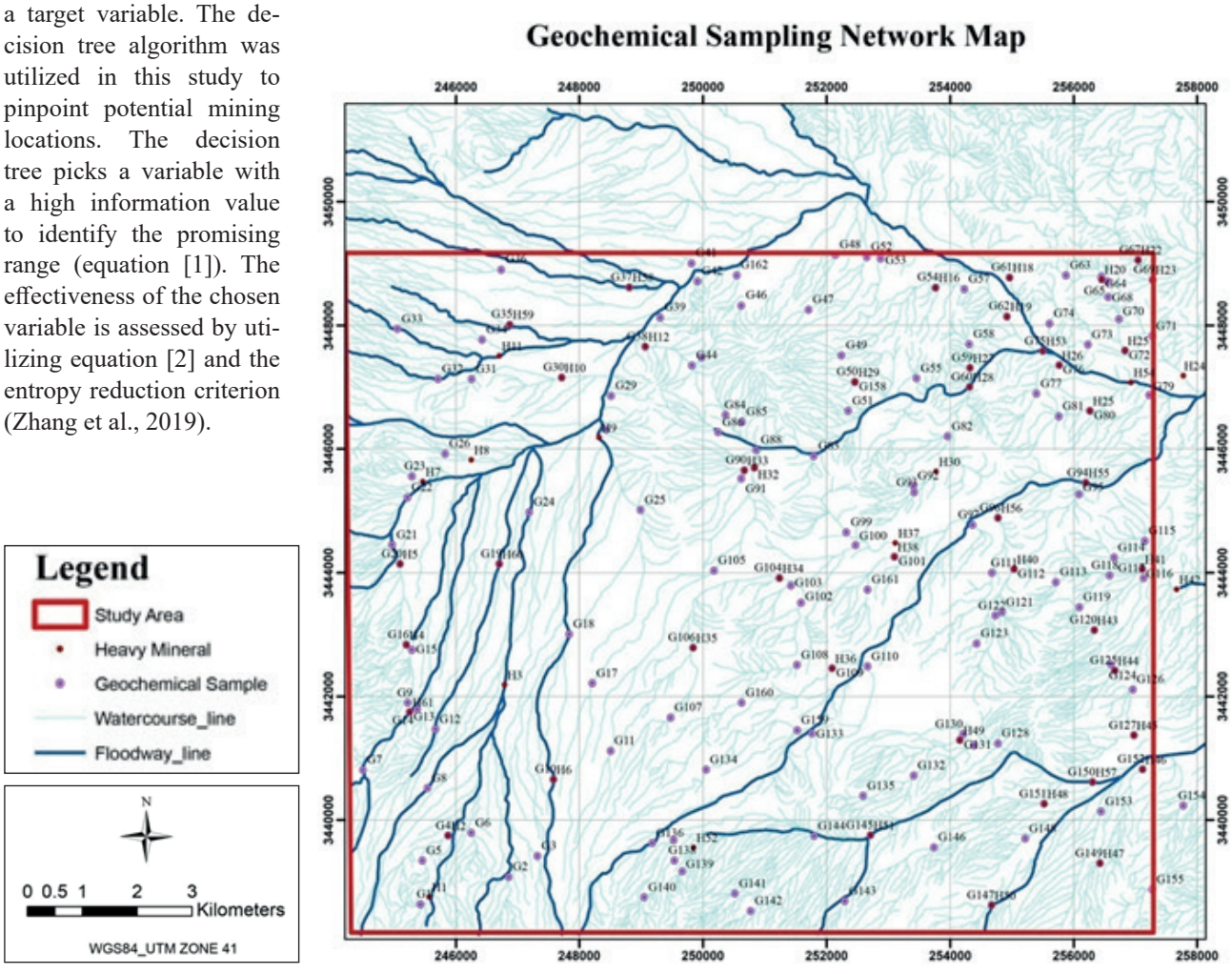


Fig. 3. Geochemical sampling network of waterway sediments in Janja area.

the best branch point criterion that separates the labeled classes in training data. Two popular methods for selecting features are the Gini coefficient and entropy index. The Gini index is a statistical metric employed to gauge the spread of data within a dataset. This index is determined by utilizing equation [3]:

$$Gini = 1 - \Sigma p^2, \quad [3]$$

Where p is the probability of occurrence of a particular class in the set. The Gini index is utilized in the decision tree algorithm to evaluate how similar the subsets formed by a feature are in terms of their composition. A lower Gini index makes a feature better for splitting data into more uniform subsets. The decision tree algorithm utilizes the Gini index to pick the feature that leads to the most significant decrease in the Gini index of the set. This indicates that the chosen characteristic results in the greatest level of uniformity in the subsets that are formed. Entropy plays a crucial role in machine learning, particularly in the development of decision trees. Entropy quantifies the level of instability or uncertainty in a dataset. This concept may aid in understanding how information is spread out in a set of data and how it can be used to properly split the data when constructing a decision tree. In order to determine entropy, we need first to determine the likelihood of each class appearing in the dataset. Next, these probabilities are used in determining the overall entropy, as outlined in equation [4]:

$$Entropy = -\Sigma(p(x)\ log_2\ p(x)), \quad [4]$$

Where $p(x)$ is the probability of occurrence of each class. The entropy value shows the level of instability or complexity present in the data. Entropy is utilized in decision trees to determine the best feature to split at each node. The objective is to choose the feature that results in the most significant decrease in entropy, ensuring that the partitioned data is more consistent and capable of making more precise forecasts. Pruning plays a crucial role in optimizing decision trees. Pruning involves removing sub-nodes from the decision tree, which is in contrast to splitting. When constructing a decision tree, several branches represent abnormalities in the training data resulting from outliers or noise. In some tree creation algorithms, pruning is is considered a part of the algorithm. While in other cases, pruning is solely employed to address the issue of overfitting. Several methods use statistical criteria to remove less reliable branches. Pruned trees are often smaller and less complicated, making them simpler to comprehend. Pruned trees generally have a quicker and more accurate performance in classifying test data compared to unpruned trees. Two popular methods are used for cutting trees. In the initial stages of constructing a tree, pre-pruning involves frequently stopping to prune the tree. As soon as a stop is created, the node transforms into a leaf. The post pruning method is the second case. It removes subtrees from a full-grown tree. Pruning a subtree involves removing branches at a node and replacing them with a leaf node. The size of the decision tree is another crucial parameter. A simpler decision tree is more communicative and transparent. Hence, the accuracy of the tree is significantly impacted by its level of complexity. Typically, the complexity of a tree is assessed by one of the following criteria: the total number of nodes, total number of leaves, tree depth, and number of features used.

3.3 Support vector machine

Support vector machines represent type of generalized linear models that rely on a linear combination of features to make decisions regarding classification and regression. Support vector machines, like artificial neural networks, have the ability to estimate multivariate functions with any desired level of precision and are suitable for representing various nonlinear and intricate procedures (Wu et al., 2018). Support vector regression, a form of support vector machine, is used for estimation across different issues. This approach involves educating algorithms and relies on the support vector machine classifier method, which is more comprehensive than the aforementioned method (Smola & Schölkopf, 2004). The method of support vector regression, rooted in statistical learning theory and focused on minimizing structural risk, was originally presented by Vapnik in the 1990s (Matías et al., 2004). In a regression model, it is essential to determine the relationship between the dependent variable y and a group of independent variables x (Smola & Schölkopf, 2004). This technique is applicable for predicting and classification problems involving two or more categories (Martínez-Ramón & Christodoulou, 2022). The aim of these issues is to create a classification standard that is effective for samples that have not been harvested and also has strong generalizability. The optimal hyperplane is defined as the linear separating plain with the largest margin and equal distance from the nearest points, with the goal of extending the boundary to cover all potential ranges. Typically, in linear classification tasks, a weight like W needs to be taken into account for a vector like X in order for this weight to effectively categorize the vectors into their correct classes. The optimal separator plate is chosen based on equation [5] (Abe, 2005; Tran et al., 2005).

$$\mathbf{w}^T x + b = 0, \quad [5]$$

Where $\mathbf{w}$ is the weight matrix; $\mathbf{T}$ is the output of the weighting matrix; x is the vector and b is the bias constant. The inner multiplication is used to express the relationship between vector x and weight $\mathbf{w}$. In the context

of the Support Vector Machine algorithm, the equation [5] typically represents a decision boundary. In other words, 0 (the numeral zero) indicates the decision boundary itself, which separates different classes in the feature space. The points that lie on this boundary satisfy the equation [5]:

- Positive Class: For instances of the positive class (e.g., class +1), the expression $\mathbf{w}^T x + b = 0 > 0$.
- Negative Class: For instances of the negative class (e.g., class –1), the expression $\mathbf{w}^T x + b = 0 < 0$.

A set of points is optimally separated by a plain based on the conditions that they are accurately classified in their respective classes. The maximum distance between the nearest points of each data class to the separating plain is shown in Fig. 4 (Sánchez, 2003; Tran et al., 2005). Therefore, it is necessary to calculate the parameters **w** and in order to meet the two specified conditions. To address this issue and regulate the precision of the data, equation [6] is formulated for the margin (Huang et al., 2006).

$$\mathbf{w}^T x + b = 0 = \begin{cases} x \le 1 & for \quad y_i = -1 \\ x \ge 1 & for \quad y_i = 1 \end{cases}, \qquad [6]$$

Where the y_i represents the class label for the i^{th} training example. In this case, it can take on values of −1 or 1, which correspond to two classes. Typically, $y_i = 1$ might represent the positive class and = −1 the negative class.

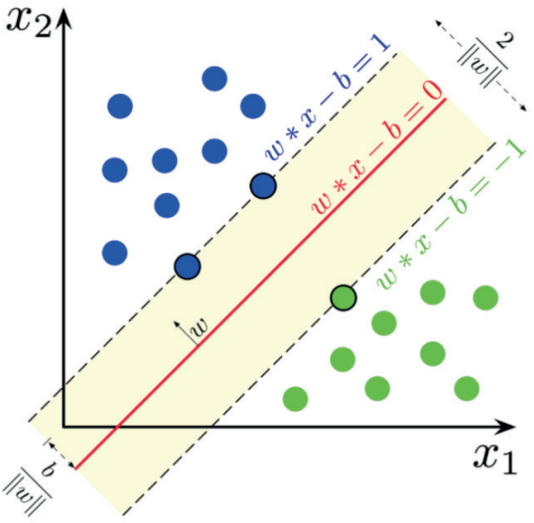


Fig. 4. Optimal separator plate and margins.

To find the ideal separation boundary with the right margin, the goal is to maximize the distance between both margins. Formula [7] illustrates the method for determining and optimizing the separation distance of these two thresholds (Martínez-Ramón & Christodoulou, 2022).

$$d(w, b; x) = \frac{|(\mathbf{w}^T x + b - 1) - (\mathbf{w}^T x + b + 1)|}{\|w\|} = \frac{2}{\|w\|}, [7]$$

Where $d(w, b; x)$ represents the distance from a point (given by the feature vector) to the decision boundary defined by $\mathbf{w}^T x + b = 0$. The constants 1 and −1 represent specific thresholds that define the margins for the separated classes. Typically, the decision boundary is set to equal 0 (the numeral zero), while the margins are defined as 1 and −1. The $\|w\|$ is called a soft function. According to the results obtained from equation [7], the goal is to maximize the desired margin (Abe, 2005; Huang et al., 2006; Martínez-Ramón & Christodoulou, 2022; Merler & Jurman, 2006; Sánchez, 2003). Occasionally in the linear system, there are situations where multiple data points are missing (Fig. 5). In this situation, the error function is necessary in order to attain an excellent separating plane. Equation [8] (Bishop & Nasrabadi, 2006; Wang, 2005) displays this function.

$$F(\xi) = \sum_{i=1}^{N} \xi i, \qquad [8]$$

Where the function $F(\xi)$ measures the total error resulting from the slack variables , which correspond to misclassified instances or instances that fall outside the margin. Put differently, ξi represents the classification error value. The data placed in the margin is the significant feature in Fig. 5. This data is used by the vector machine to accurately classify the data (Van Der Heijden et al., 2005; Wang, 2005).

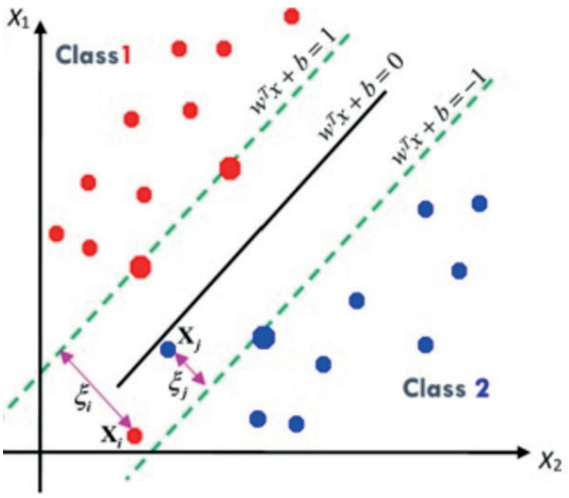


Fig. 5. A linear integral system with error rate ξi (Zhang et al., 2019).

Hence, equation [9] represents the optimization problem in non-separable linear systems. The vector **w** determines the hyper plain of the generalized optimal isolator through equation [9] (Bishop & Nasrabadi, 2006).

$$M in_{\mathbf{w},b} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi i, S.t \quad yi\,(\mathbf{w}^T x_i + b) \ge 1 - \xi i \qquad [9]$$

Where $M\ in_{w,b}$ denotes the objective function that is attempted to be minimized $\frac{1}{2}\mathbf{w}^{T}\mathbf{w}$ during the training of the model. It represents the total loss that needs to be minimized – represents the regularization component of the objective function. It is often associated with maximizing the margin, the factor of $\frac{1}{2}$ is used for mathematical convenience, particularly when taking derivatives, $\sum_{i=1}^{N}$ ξi represents the total slack variable penalties for all training samples, N is the total number of training samples in the dataset, and ξi is the slack variable corresponding to the i^{th} training example. The ξi measures the extent to which the i^{th} example falls within the margin or is misclassified. If ξi = 0, the example is correctly classified; if ξi > 0, it is either misclassified or lies within the margin. The y_i is the class label of the i^{th} training example, which takes values of 1 or −1. It indicates which class the example belongs to. The x_i is the feature vector of the i^{th} training sample. Each component of this vector corresponds to a feature of that sample. The constraint $y_i(\mathbf{w}^{T}x_i + b) \geq 1 - \xi i$ ensures that each training example is correctly classified with a margin of at least 1 minus the slack variable. For a correctly classified point ($y_i = +1$), the condition requires that the output of the decision function $\mathbf{w}^{T}x_i + b$ is at least 1. For misclassified points or those near the margin (when ξi > 0), it provides leeway by allowing a margin of less than 1. The parameter C is the interaction coefficient to maximize margins and minimize performance error. In cases such as this one, Lagrange multipliers are utilized, appearing as Lagrange multipliers in equation [10] in the latest correlation involving α, β (Wang, 2005).

$$Lp\ (\mathbf{w},\ b,\ \xi,\ \alpha,\ \beta) = \frac{1}{2}\ \mathbf{w}^{T}\ \mathbf{w} + C\ \sum_{(i=1)}^{N}\xi_i - \sum_{(i=1)}^{N}\alpha_i \{y_i\ [\mathbf{w}^{T}\ x_i + b] - 1 + \xi_i\ \} - \sum_{(i=1)}^{N}\ \beta_i\ \xi_i \qquad [10]$$

Where α_i is Lagrange multiplier associated with the constraints of the SVM optimization problem. Each corresponds to the i^{th} training example and measures how much the corresponding constraint contributes to the Lagrangian. Non-zero indicates that a data point is either a support vector or on the margin and β_i is Lagrange multipliers associated with the slack variables ξ_i. Each β_i represents the penalty associated with the i^{th} slack variable. A non-zero β_i indicates that there is some violation of the minimum margin requirement.

The fundamental issue of equation [10] can be converted into a dual problem by dual classical Lagrange. Equation [11] defines the dual aspect of this relationship.

$$Max\mathbf{W}(\alpha,\beta) = Max_{\alpha,\beta}(M\ in_{w,b,\xi}L(\mathbf{w},b,\xi,\alpha,\beta), \qquad [11]$$

Setting the derivative of equation [11] with respect to **W** and equal to zero yields the values of equations [12]:

$$\begin{cases} \dfrac{\partial L}{\partial \mathbf{w}} = o \quad then \quad \mathbf{w} = \sum_{i=1}^{N}\alpha_i\ \beta_i\ \mathrm{x}_i \\ \dfrac{\partial L}{\partial b} = o \quad then \quad \mathbf{w} = \sum_{i=1}^{N}\alpha_i\ \mathrm{y}_i = o\ , \\ \dfrac{\partial L}{\partial \xi} = o \quad then \quad \alpha_i\ + \beta_i = C \end{cases} \qquad [12]$$

Equation [11] can be used to express the relationships in the form of equation [13] to derive the basic equation of the vector machine in the linear integral state (Wang, 2005).

$$MaxL_d\ (\alpha) = \sum_{i=1}^{N}\ \alpha_i - \frac{1}{2}\ \sum_{i=1}^{N}\ y_i\ y_i\ \alpha_i\ \alpha_i x_i^{T}\ x_i, \qquad [13]$$

$$S.t = \begin{cases} o \leq \alpha_i \leq C \\ \sum_{i=1}^{N}\ \alpha_i\ y_i = o \end{cases}$$

It is evident that the linearly separable system functions in the same way as the linearly separable system. The only thing that sets them apart is the adjustment of the boundaries of the Lagrange coefficients. The parameter C must be established in these systems in order to determine the classifier's extra capacity. Typically, a vector such as x in a higher-dimensional space is represented as a linear vector machine in a higher space (also known as a feature space) depicted in Fig. 6, with the input space still being nonlinear.

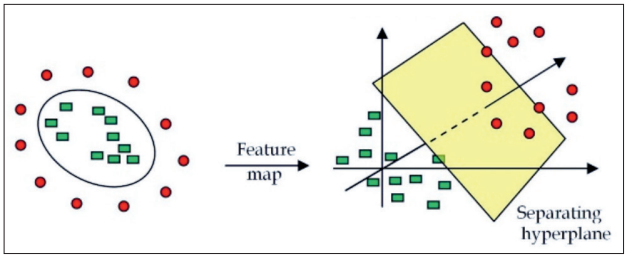


Fig. 6. Data classification in higher space.

It is important to mention that Table 1 displays the most frequently utilized kernel functions for linear inseparable problems.

Tab. 1

The prevailing kernel functions employed in linearly inseparable issues

Type	Kernel function
Linear	K(x_i,x_j)=$\gamma x_i\ x_j$
Polynomial	$K(x_i\ x_j) = (\gamma x_i\ x_j + r)^{d}, \gamma > o^{\circ}$
RBF	$K(x_i\ x_j) = exp\{-\gamma\|\|x_i - x_j\|\|^{2}\}, \gamma > o^{\circ}$
Sigmoid	$K(x_i\ x_j) = \tanh(\gamma x_i\ x_j + r), \gamma > o^{\circ}$

4 Results and analysis

4.1 Gold grade estimation using decision tree algorithm

Regarding this matter, the data extracted from geochemical samples and the analysis results from exploratory boreholes were evaluated in relation to outlier values, number of communities, and normality. Next, this information was sent to Python 3.11 software to create a model of potential gold mineralization areas, using factors such as rock type, alteration zones, and silver, copper, lead, and zinc values as independent variables. The model building process involves two phases: training the model and assessing its accuracy, which will be detailed further. To prepare the model for training, the dataset was split randomly into two sets in a 70 : 30 ratio. Training involved 70 % of the data while the remaining 30 % was used for validation during model evaluation. In order to prevent overfitting during modeling, the pruning technique was applied before the tree was allowed to grow, aiming to decrease complexity. In order to achieve this goal, the confidence factor, which ranges from 0 to 1, was utilized. This study utilized a confidence factor of 0.95 to prune the decision tree. The scikit-learn library's classifier decision tree was employed to develop and utilize a decision tree for data categorization purposes. This class constructs a decision tree using the CART algorithm. The decision tree is constructed using the CART algorithm through recursive and alternating divisions. Once the model is in place, it might not function effectively, therefore tweaking the corresponding hyperparameters is necessary to achieve the intended results. Some of the decision tree hyperparameters are as follows:

1. The minimum number of samples in each node to create branching and tree growth.
2. The least number of leaves that show the results of decisions.
3. Maximum depth of tree (number of rows).
4. The largest number of components that are randomly considered in the posterior direction of the tree.
5. Tree node evaluation indices to analyse the information obtained from different nodes for the purpose of branching, for example Gini coefficient and entropy.

Fig. 7. Correlation diagram between the variables of Janja dataset.

By applying various hyperparameters, the model was eventually fine-tuned to its optimal state. In this framework, the model requires a minimum of 5 samples in each node for tree development, a minimum of 3 leaves, a maximum tree depth of 4, and uses the entropy index as the appropriate criterion. Furthermore, because the 3rd and 4th categories of gold hold greater significance in terms of discovery, a weighting method was implemented for these classes. This description assigned weights of 0.2, 0.2, 0.3, and 0.3 to classes 1,

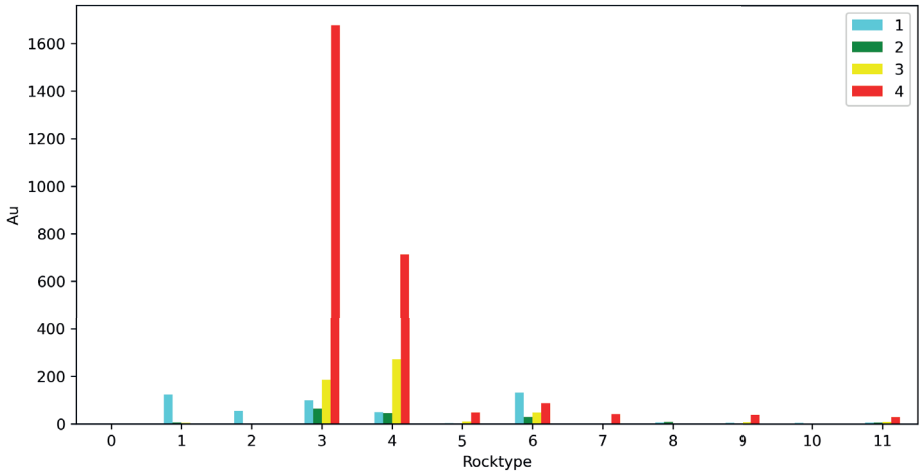


Fig. 8. The quantity of gold found in various rock types in the Janja area (3 – diorite, 4 – hornfels, 6 – schist, 7 – clay).

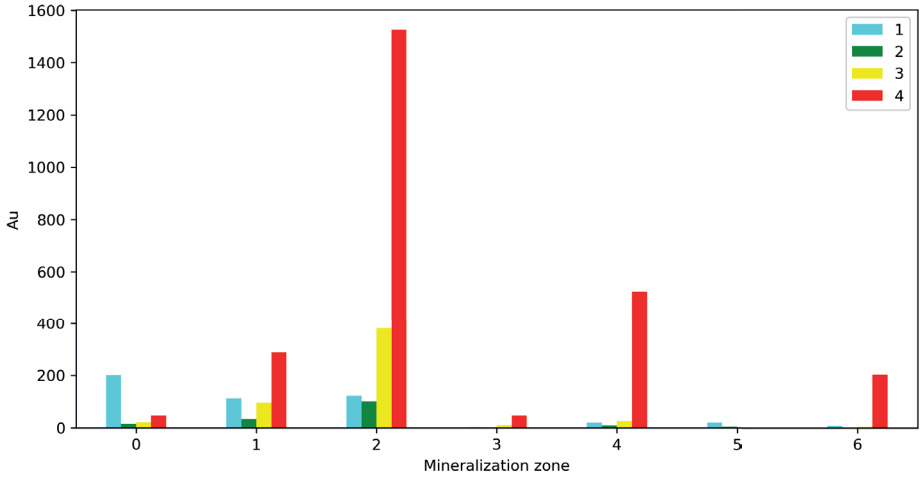


Fig. 9. The quantity of gold found in various alteration zones in the Janja area, specifically in zones 12 (propylitic) and 17 (potassic).

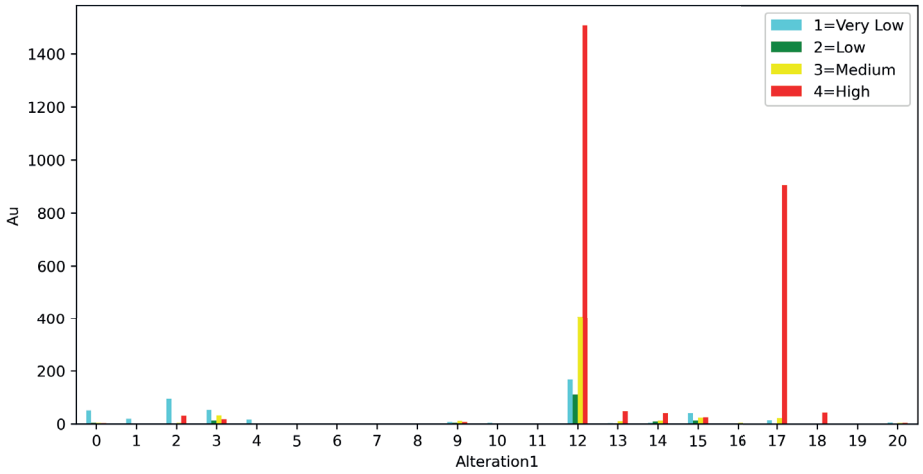


Fig. 10. The quantity of gold in different mineralization zones in the Janja area, including zones 2 (hypogene), 4 (oxyhypogene), and 1 and 6 (supergene and oxidation zones).

2, 3, and 4, respectively. According to the results of this algorithm depicted in Fig. 7, the correlation between gold and type 1 alteration stands at 0.597, with type 2 alteration at 0.575, with mineralization zone at 0.393, with the type of mineralization host rock at 0.049, and with silver, copper, lead and zinc show no relation. The types of variables such as alteration type 1 and type 2, mineralization zone, and rock type have been discussed in a broad manner, with each having distinct variations. The correlation of each type with mineralization will vary accordingly.

It is worth noting that diorite, hornfels, amphibole-rich schist, and clay are the most effective in gold particle deposition, as shown in Fig. 8.

Furthermore, Fig. 9 illustrates the presence of gold particles increasing in areas with propylitic, potassic, clay, siliceous, iron oxide III, and chlorite alterations, in descending order.

Another factor that influences the buildup of gold particles, specifically the mineralization zone, was also examined, with the outcomes displayed in Fig. 10. According to the results of the decision tree model, the hypogene mineralization zone followed by the oxyhypogene zone will be given higher priority for gold particle accumulation. In the following stage, the oxidation and oxysupergen areas provide ideal conditions for the gold particles to be deposited.

Ultimately, the decision tree algorithm is employed to showcase the three-dimensional model of the area under study, as depicted in Fig. 11. Indeed, the identification of potential mineral-rich zones can be achieved through the utilization of this algorithm.

Evaluating the accuracy of a classifier is crucial once it has been constructed. Using test data is more beneficial for evaluating the precision of a classifier. After constructing the model with the training data, it is important to assess its accuracy in predicting the class label of the samples using the test data. The classifier's accuracy on test samples is determined by the number of samples correctly classified by the model. Certainly, the error rate can be computed in place of accuracy. Another crucial aspect revolves around the cross validation of data, which is essential for assessing the efficacy of machine learning models and relies on where the model gathers the necessary training information from within the dataset. In the specific model, the sensitivity of the model to this issue was also assessed by assigning values of 5 and 10 for this criterion. Ultimately, the model that was achieved underwent evaluation. During the model assessment, the following metrics were calculated: overall accuracy (91.7 %), prediction accuracy for positive cases (91.9 %), coverage of positive samples (91.7 %), and F1 score (91.8 %).

4.2 Gold grade estimation using support vector machine method

The radial basis function (RBF) is chosen as the kernel function to assess the grade of Janja polymetallic deposit through the support vector machine classification method. The RBF kernel transforms the samples in a nonlinear manner to a space with higher dimensions. In contrast to other kernel types, this kernel encounters fewer computational issues and can effectively handle data sets with high dimensions. Additionally, this kernel function has a lower number of parameters when compared to functions like polynomial functions, ultimately decreasing the model's complexity (Hsu et al., 2003; Lin et al., 2008). The effectiveness of a model using the support vector machine technique heavily relies on the parameters chosen for the model. In order to achieve a model with strong generalization abilities, it is important to carefully decide on the model parameters. Selecting the best parameters for the model can impact its performance quality; hence, a model with incorrect parameters might yield undesirable outcomes (Kecman, 2004; Oliver & Webster, 2014). In order to find the best parameters for the model – such as C, γ, and kernel type, a network search approach was employed, utilizing a 10-point cross-validation. This approach involves setting up a uniform grid in

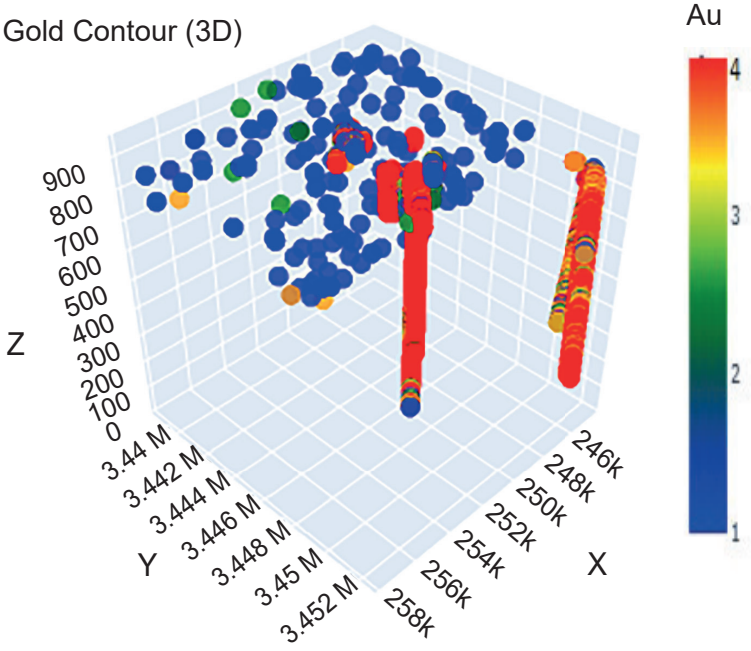


Fig. 11. Three-dimensional diagram of promising mineral areas of the studied region.

the parameter space that needs to be explored, followed by assessing all grid points in order to identify the best overall point. Ultimately, the grid search will identify the best possible point among all points within the grid for the specified parameter. The network search method begins by creating a large network in the parameter space and gradually defines smaller networks as it approaches the optimal point, ultimately converging towards the overall optimal point in the parameter space (Hsu et al., 2003). Table 2 displays the range of values to search for each model parameter. The primary concept of this approach is to identify the best parameters of the model in order to minimize the model error. As stated, the cross-validation method is utilized in conjunction with the network search method for this purpose. During k-fold validation, the training set is initially partitioned into k equal-sized subsets. Each subset is tested in consecutive order by the model trained based on the remaining k-1 subsets. Thus, each individual sample in the training set is calculated only one time. Therefore, the accuracy of cross-validation will be the proportion of correctly predicted data. Indeed, the RMSE error value is achieved for the specified parameters in every test subset through the implementation of cross-validation technique. The mean RMSE for each of the k test subsets is shown as a way to evaluate the model's performance. This introduces the optimal parameters of the model as those that result in the lowest RMSE. Another benefit of using cross-validation is its ability to address the issue of overfitting in the model (Che & Hu, 2008; Hsu et al., 2003).

Tab. 2

The search interval for each of the model parameters in estimating gold grade

Support vector machine model parameter	Search range
C	{0.01, 0.1, 1, 10, 100}
γ	{0.01, 0.1, 0.2, 0.3}
Kernel	{Linear, Polynomial, RBF}

Table 3 displays the best C and γ parameters, as well as the kernel type, determined through a grid search method using 10-fold cross-validation. Once the model's best parameters were chosen and the lowest error from cross-validation was achieved, the entire training dataset (90 % of the total data) was used to train the optimal model. The process of training the model was carried out using Python 3.11 software. Following the training phase using the best parameters, the support vector machine model's effectiveness and efficiency were assessed using the test data (10 % of the overall dataset).

Tab. 3

Optimal values of C, γ and Kernel parameters applying the gird search method based on cross-validations that is divided into 10 folds

Optimal parameters of the model	The optimal value of the parameter
C	1
γ	0.1
Kernel	RBF

SVM methods have been employed in order to improve the accuracy of gold grade predictions. To achieve this goal, the dataset needs to be separated into training and testing sets. The input data consisted of the silver, copper, lead, and zinc values with gold value being the output data. This network underwent training using various combinations of training and testing data, including ratios of 90/10, 80/20, 70/30, 60/40, and 50/50. The results that were obtained can be found in table 4. It was found that the best scenario involved using 90 % of the data for training and 10 % for testing.

Tab. 4

Impact of varying selection ratios of training and testing data on support vector machine algorithm

Training / testing [%]	Accuracy [%]	RMSE
90/10	82	0.917
80/20	80	0.964
70/30	76.8	0.949
60/40	76.5	0.923
50/50	76.3	0.911

Following the calculation of the training and testing data ratio, visual patterns were generated using three-dimensional diagrams to show the correlation between the actual and predicted values of gold. Furthermore, within these illustrations, gold anomaly regions were depicted in pairs based on the concentration of the input elements. Since gold has a stronger correlation with copper compared to silver, lead, and zinc, pairs of input elements are first examined before drawing diagrams to represent anomalies of gold, as shown in Fig. 12.

5 Conclusions

Today, the implementation of variable estimation is a recent approach that has facilitated the decision-making process in various fields of study. Estimating the

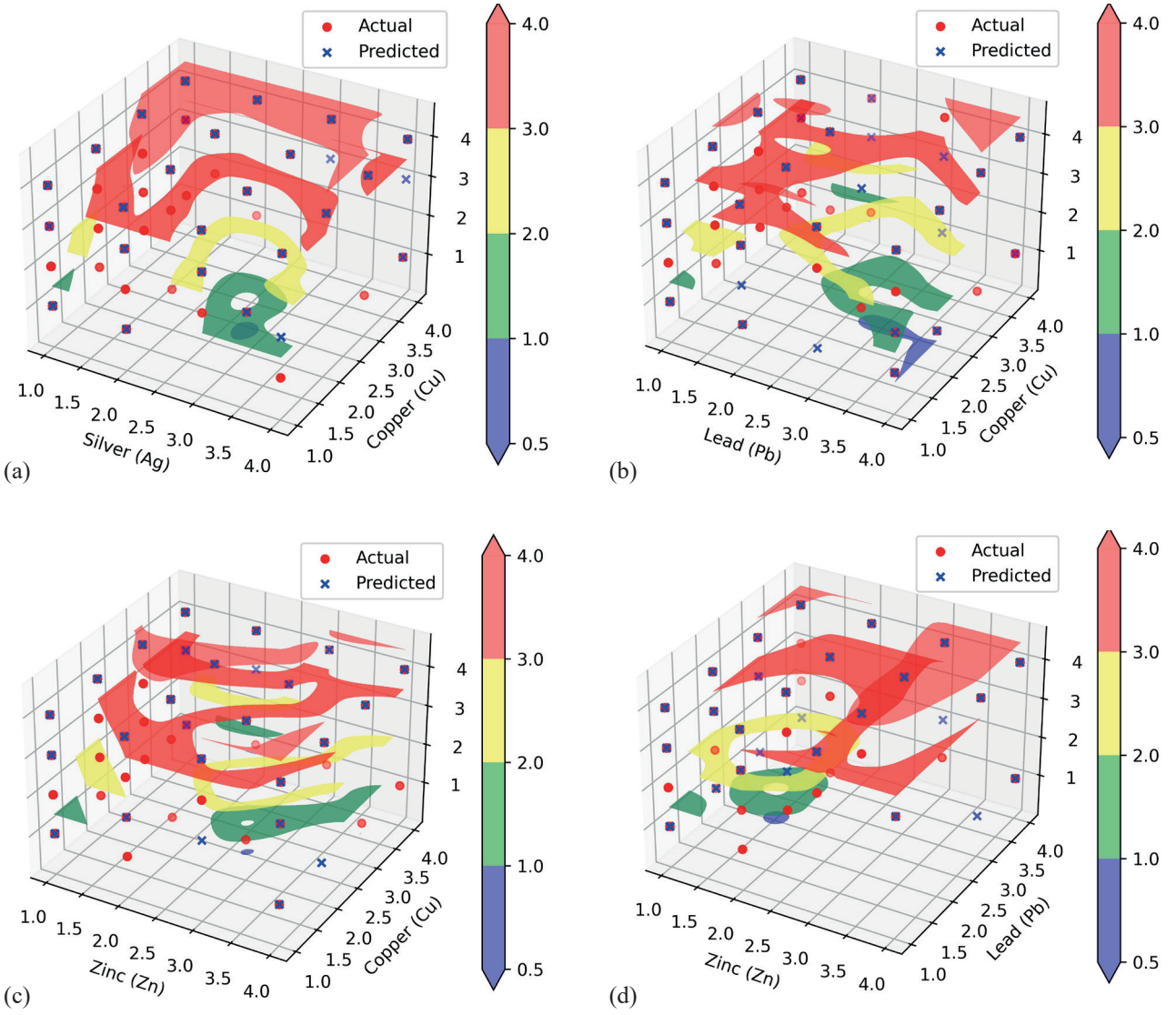


Fig. 12. 3D diagram of gold anomaly zones based on grade of: a) silver and copper, b) lead and copper, c) zinc and copper, and d) zinc and lead.

grade of mineral deposits is a crucial aspect of evaluating them in geosciences. There exist several techniques to determine this crucial factor. Years of thorough research have resulted in the creation of innovative techniques for estimating variable amount in the deposit. One of the methods available is the decision tree algorithm, which is a simple and powerful machine learning algorithm suitable for classification and regression tasks. It can be easily understood and is capable of processing many different types of data. Nonetheless, it is susceptible to overfitting and may be unreliable. Although they have limitations, decision trees are utilized in many fields, making them a useful tool for analysing data and making predictions. If the data includes logical conditions or is divided into various categories, the decision tree algorithm is the appropriate selection. Using different classification algorithms is recommended when dealing with an excessive amount of numerical variables in the data. By utilizing the decision tree algorithm, it is evident that diorite, hornfels, amphibole-rich schist, and clay show the highest potential for gold particle deposition. Diorite plays a significant role in rock formations containing gold deposits and is considered the source rock. Indeed, the thermal engine is situated within the diorite rock, causing the hydrothermal solutions to move due to the heat produced. Hornfels does not play a significant role in gold mineralization due to its low porosity, which results in poor permeability to hydrothermal fluids, preventing mineralization in this rock. When pyrite crystals are found in schist or clay, it indicates the presence of either pyrite or

its phantom form representing gold. This shows that gold has been displaced within the sulfur phase due to shifts in oxidation-reduction conditions. The gold particles are typically found in reducing environments like clay minerals and schists, where they can be released from complexes and precipitated due to hydrothermal activity, making these rocks favorable hosts for gold. The producer may consist of diorite while the host rock can be composed of clay and schist rocks. Hence, it is recommended to obtain a thin sample from the schists and use electric charge to determine if gold particles are present. Additionally, it is important to determine the phase of any gold particles found in the schist sections: Is it in the crystallization phase or in the mineralogical phase like pyrite or chalcopyrite? Furthermore, there has been a build-up of gold particles in the alteration zones of propylenic, potassic, clay, siliceous, iron III oxide and chlorite in descending order. It is important to point out that the propylitic and argillic alteration zones possess geochemical attributes required for the accumulation of gold. Mineralization is favored in the propylitic zone of hydrothermal systems due to its high oxidation levels and acidic environment. Atmospheric waters also have a part to play. This area is optimal for disrupting gold compounds. In the potassic zone, temperatures exceeding 450 degrees Celsius prevent gold complexes from breaking unless they are already enclosed in pyrite and chalcopyrite. However, typically, the area high in potassium does not have a significant amount of gold deposits, whereas the surrounding propylitic and argillic zones are rich in gold because gold complexes are the final phase to deteriorate and, if they bond, they stay firm even in colder temperatures. In other words, gold complexes either do not enter the game or come until the end of the game. The accumulation of gold particles in the mineralization zones was also investigated. According to the decision tree model's results, the hypogene mineralization zone followed by the oxyhypogene will be given higher priority for the aggregation of gold particles. In the following stage, the oxidation and oxysupergen regions are conducive zones for the accumulation of gold particles. Indeed, the hypogene mineralization area corresponds to potassic and phyllic modification. The propylitic zone has a similar level of oxidation as the hypogene zone, but actually displays more argillic characteristics. The supergene zone includes clay minerals, iron oxide, and chlorite, with siliceous alteration also being part of the argillic zone. Indeed, there is a high concentration of gold in the hot section of the mine, which goes against the low gold mineralization found in Sarcheshme and Songun mines in Iran. In those mines, gold is mostly deposited in the cold areas near the mass and in veins of types 4 and 5, which contain pyrite, anhydrite, and calcite. In conclusion, the decision tree algorithm can accurately pinpoint mineral-rich areas with a high success rate of 92 %, as demonstrated in the relevant section where the diagram was displayed.

Another effective and powerful machine learning technique is the support vector machine method, employed for predicting the gold quality in Janja region. Firstly, the support vector machine method was used to determine the best values for the model parameters, followed by the implementation of the final model with these optimal parameters. The assessment of the model indicates that it will be effective in predicting the gold grade, which is a crucial factor in assessing the quality of the deposit. The primary objective of this study is to develop a method for predicting the gold grade based on its distinct characteristics and properties. Due to the limited amount of data in numerous instances, analysing it can be challenging and expensive. However, usually during the initial phases of research, analysis, and forming conclusions and decisions should be made according to the limited data available. Appropriate estimation and forecasting methods are necessary for this task. This research aimed to forecast the gold grade quantity and accompanying anomaly ranges in Janja polymetallic deposit by utilizing statistical techniques and support vector machine algorithm. It can be observed that the support vector machine technique is able to accurately predict the distribution of gold grades within a certain range. Hence, it is possible to conclude that employing support vector machine techniques in geochemical research on low grade and dispersed elements like gold not only saves time and money but also identifies patterns by interpolating between inputs and outputs while reducing the error between predicted and actual values. The results of deposit modeling indicate that it has been successful in predicting the gold grade. Moreover, the economic analysis and mine design can be carried out using the resulting grade model.

Acknowledgement

The completion of this research paper would not have been possible without the support and guidance by Professor Ardeshir Hezarkhani, Amirkabir University of Technology. His dedication and overwhelming attitude towards helping his students is solely responsible for completing this research paper. The encouragement and insightful feedback were instrumental in accomplishing this task.

Authors would also like to express a gratitude to experts from the State Geological Institute of Dionýz Štúr (SGUDS, Slovak Geological Survey) for their constructive criticism and valuable feedback, enhancing the quality of this paper. We also consider it our duty to express our gratitude to the esteemed reviewers, Professor Ladislav Vizi (SGUDS) and an anonymous reviewer for their valuable comments, improving the scientific level of this paper.

References

Abbaszadeh, M., Hezarkhani, A. & Soltani-Mohammadi, S., 2013: An SVM-based machine learning method for the separation of alteration zones in Sungun porphyry copper deposit. *Geochemistry*, 73, 545–554.

Abbaszadeh, M., Hezarkhani, A. & Soltani-Mohammadi, S., 2015: Classification of alteration zones based on whole-rock geochemical data using support vector machine. *Journal of the Geological Society of India*, 85, 500–508.

Abe, S., 2005: Support vector machines for pattern classification. *London, Springer.*

Amnieh, H. B., Siamaki, A. & Soltani, S., 2012: Design of blasting pattern in proportion to the peak particle velocity (PPV): Artificial neural networks approach. *Safety Science*, 50, 1913–1916.

Anderberg, M. R., 1973: The broad view of cluster analysis. *Cluster analysis for applications*, 1, 1–9.

Bishop, C. M. & Nasrabadi, N. M., 2006: Pattern recognition and machine learning. *New York, Springer.*

Chatterjee, S., Bandopadhyay, S. & Machuca, D., 2010a: Ore grade prediction using a genetic algorithm and clustering based ensemble neural network model. *Mathematical Geosciences*, 42, 309–326.

Chatterjee, S., Bandopadhyay, S. & Machuca, D., 2010b: Ore Grade Prediction Using a Genetic Algorithm and Clustering Based Ensemble Neural Network Model. *Mathematical Geosciences*, 42, 309–326.

Che, X. L. & Hu, L., 2008: Grid resource prediction approach based on Nu-Support Vector Regression. In: *International Conference on Machine Learning and Cybernetics*, IEEE, 778–783.

Dutta, S., Bandopadhyay, S., Ganguli, R. & Misra, D., 2010: Machine Learning Algorithms and Their Application to Ore Reserve Estimation of Sparse and Imprecise Data. *Journal of Intelligent Learning Systems & Applications*, 2, 86–96.

Guo, W. W., 2010: A novel application of neural networks for instant iron-ore grade estimation. *Expert Systems with Applications*, 37, 8729–8735.

Hsu, C. W., Chang, C. C. & Lin, C. J., 2003: A practical guide to support vector classification. *Department of Computer Science, National Taiwan University*, 1396–1400.

Huang, T. M., Kecman, V. & Kopriva, I., 2006: Kernel based algorithms for mining huge data sets. *Heidelberg, Springer.*

Jafrasteh, B., Fathianpour, N. & Suárez, A., 2018: Comparison of machine learning methods for copper ore grade estimation. *Computational Geosciences*, 22, 1371–1388.

Jhonnerie, R., Siregar, V. P., Nababan, B., Prasetyo, L. B. & Wouthuyzen, S., 2015: Random forest classification for mangrove land cover mapping using Landsat 5 TM and ALOS PALSAR imageries. *Procedia Environmental Sciences*, 24, 215–221.

Kecman, V., 2001: Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. *MIT press.*

Kecman, V., 2004: Support vector machines basics. School of Engineering, *University of Auckland.*

Kotake, N., Suzuki, K., Asahi, S. & Kanda, Y., 2002: Experimental study on the grinding rate constant of solid materials in a ball mill. *Powder Technology*, 122, 101–108.

Krishna, G., Sahoo, R. N., Pradhan, S., Ahmad, T. & Sahoo, P. M., 2018: Hyperspectral satellite data analysis for pure pixels extraction and evaluation of advanced classifier algorithms for LULC classification. *Earth Science Informatics*, 11, 159–170.

Lee, C. & Sterling, R. L., 1992: Identifying probable failure modes for underground openings using a neural network. *International journal of rock mechanics and mining sciences & geomechanics abstracts*, 29, 1.

Li, X. L., Xie, Y. L., Guo, Q. J. & Li, L. H., 2010: Adaptive ore grade estimation method for the mineral deposit evaluation. *Mathematical and Computer Modelling*, 52, 1947–1956.

Lin, S. W., Lee, Z. J., Chen, S. C. & Seng, T. Y., 2008: Parameter determination of support vector machine and feature selection using simulated annealing approach. *Applied soft computing*, 8, 1505–1512.

Mahmoudabadi, H., Izadi, M. & Bagher Menhaj, M., 2009: A hybrid method for grade estimation using genetic algorithm and neural networks. *Computational Geosciences*, 13, 91–101.

Maleki, S., Ramazi, H. R. & Moradi, S., 2014: Estimation of Iron concentration by using a support vector machineand an artificial neural network-the case study of the Choghart deposit southeast of Yazd, Yazd, Iran. *Geopersia*, 4, 201–212.

Martínez-Ramón, M. & Christodoulou, C., 2022: Support vector machines for antenna array processing and electromagnetics. *Springer Nature.*

Masoumi, F., Eslamkish, T., Abkar, A. A., Honarmand, M. & Harris, J. R., 2017: Integration of spectral, thermal, and textural features of ASTER data using Random Forests classification for lithological mapping. *Journal of African Earth Sciences*, 129, 445–457.

Matías, J. M., Vaamonde, A., Taboada, J. & Gonzalez-Manteiga, W., 2004: Support vector machines and gradient boosting for graphical estimation of a slate deposit. *Stochastic Environmental Research and Risk Assessment*, 18, 309–323.

Merler, S. & Jurman, G., 2006: Terminated ramp-support vector machines: a nonparametric data dependent kernel. *Neural Networks*, 19, 1597–1611.

Moorthi, S. M., Misra, I., Kaur, R., Darji, N. P. & Ramakrishnan, R., 2011: Kernel based learning approach for satellite image classification using support vector machine. *IEEE recent advances in intelligent computational systems,* 107–110.

Nezamolhosseini, S. A., Mojtahedzadeh, S. H. & Gholamnejad, J., 2017: The application of artificial neural networks to ore reserve estimation at choghart iron ore deposit. *Analytical and Numerical Methods in Mining Engineering*, 73–83.

Oliver, M. A. & Webster, R., 2014: A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena,* 113, 56–69.

Pozdnoukhov, A., 2005: Support vector regression for automated robust spatial mapping of natural radioactivity. *Automatic mapping algorithms,* 57.

Samanta, B., Ganguli, R. & Bandopadhyay, S., 2005: Comparing the predictive performance of neural networks with ordinary kriging in a bauxite deposit. *Mining Technology*, 114, 129–139.

Sánchez, A. V. D., 2003: Advanced support vector machines and kernel methods. *Neurocomputing*, 55, 5–20.

Sayadi, A. R. M. M. & Shahrabadi, H., 2008: Reserve Evaluation of Esfordi Phosphate Mine using Geostatistical and Artificial Neural Network, *Scientific Quarterly Journal of Geosciences*, 18, 102–109.

Shahin, M. A., Jaksa, M. B. & Maier, H. R., 2008: State of the art of artificial neural networks in geotechnical engineering. *Electronic Journal of Geotechnical Engineering*, 8, 1–26.

Shirali, R., 2016: Classification trees and rule-based modeling using the C5. 0 algorithm for self-image across sex and race in St. Louis. *St. Louis, Washington University.*

Singh, V., Banerjee, P. K., Tripathy, S. K., Saxena, V. & Venugopal, R., 2013: Artificial neural network modeling of ball mill grinding process. *J. Powder Metall. Min.*, 2, 106.

Smola, A. J. & Schölkopf, B., 2004: A tutorial on support vector regression. *Statistics and computing*, 14, 199–222.

Soliman, O. S. & Mahmoud, A. S., 2012: A classification system for remote sensing satellite images using support vector machine with non-linear kernel functions. In: *8th International Conference on Informatics and Systems (INFOS)*, BIO-181, IEEE.

Soliman, O. S., Mahmoud, A. S. & Hassan, S. M., 2012: Remote sensing satellite images classification using support vector machine and particle swarm optimization. In: *Third International Conference on Innovations in Bio-Inspired Computing and Applications*, IEEE, 280– 285.

Tahmasebi, P. & Hezarkhani, A., 2010: Application of adaptive neuro-fuzzy inference system for grade estimation; case study, Sarcheshmeh porphyry copper deposit, Kerman, Iran. *Australian Journal of Basic and Applied Sciences*, 4, 408–420.

Tahmasebi, P. & Hezarkhani, A., 2012: A hybrid neural networks-fuzzy logic-genetic algorithm for grade estimation. *Computers & geosciences*, 42, 18–27.

Tenorio, V. O., Bandopadhyay, S., Misra, D., Naidu, S. & Kelley, J., 2015: Support vector machines applied for resource estimation of underwater glacier-type platinum deposits. *Application Of Computers an d Operations Research in the Mineral Industry*, 18, 309–323.

Thuijsman, F., 1995: Artificial neural networks: an introduction to ANN theory and practice. *Springer Science & Business Media.*

Tran, Q. A., Li, X. & Duan, H., 2005: Efficient performance estimate for one-class support vector machine. 26, no. 8. *Pattern Recognition Letters*, 26, 1174–1182.

Tsai, C. F. & Yen-Jiun, C., 2009: Earnings management prediction: a pilot study of combining neural networks and decision trees. *Expert Systems with Applications*, 36, 7183–7191.

Twarakavi, N. K., Misra, D. & Bandopadhyay, S., 2006: Prediction of arsenic in bedrock derived stream sediments at a gold mine site under conditions of sparse data. *Natural Resources Research*, 15, 15–26.

Van Der Heijden, F., Duin, R. P., De Ridder, D. & Tax, D. M., 2005: Classification, parameter estimation and state estimation: an engineering approach using MATLAB. *John Wiley*.

Wang, L. (ed.), 2005: Support vector machines: theory and applications. *Springer Science & Business Media.*

Wu, W., Li, A. D., He, X. H., Ma, R., Liu, H. B. & Lv, J. K., 2018: A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China. *Computers and Electronics in Agriculture*, 144, 86–93.

Zaremotlagh, S. & Hezarkhani, A., 2017: The use of decision tree induction and artificial neural networks for recognizing the geochemical distribution patterns of LREE in the Choghart deposit, Central Iran. *Journal of African Earth Sciences*, 128, 37–46.

Zhang, S., Carranza, E. J. M., Fu, C., Zhang, W. & Xiang, Q., 2024: Interpretable Machine Learning for Geochemical Anomaly Delineation in the Yuanbo Nang District, Gansu Province, China. *Minerals,* 14, 5.

Zhang, S., Xiao, K., Carranza, E. J. M. & Yang, F., 2019: Maximum entropy and random forest modeling of mineral potential: Analysis of gold prospectivity in the Hezuo-Meiwu district, west Qinling Orogen, China. *Natural Resources Research*, 28, 645–664.

Porovnávacia štúdia počítačových metodík rozhodovacieho stromu a podporného vektora na mapovanie perspektív zlatonosnosti

Mapovanie geochemických anomálií je jedným z hlavných cieľov geochemického výskumu. Keďže terénne štúdie sú časovo náročné a nákladné, veľkou pomocou pri hodnotení získaných údajov je ich spracovanie metodikami počítačového spracovania, hlavne algoritmom rozhodovacieho stromu (*Decision Tree*; DT). Je to jednoduchý a výkonný nástroj pri strojovom učení, vhodný na klasifikačné a regresné úlohy. Poskytuje lacnú, rýchlu a pomerne presnú alternatívu spracovania získaných údajov. Okrem algoritmu DT článok prezentuje aj podporný vektorový nástroj (*Support Vector Machine*; SVM). Obe metodiky boli aplikované pri mapovaní perspektívy zlata v oblasti Janja vo východnom Iráne. Porovnanie analytických výsledkov získaných z týchto dvoch metód potvrdzuje, že model DT má dostatočnú presnosť a viac správnych výsledkov ako model SVM. Má to vplyv na proces prípravy a projektovania ďalších fáz prieskumu a mapovania perspektívy zlata študovanej oblasti.

Skúmaná oblasť Janja v provinciách Sistan a Balúčistan (východný Irán) sa vyznačuje charakteristickou piesočnatou rovinou v nadmorskej výške 800 – 9 000 m, riedkou vegetáciou, ale aj malou odkrytosťou skalného podložia (obr. 1). Skúmaná oblasť zahŕňa subregión Zábul – Zahedan – Saravan. Flyšové sekvencie majú vrchnokriedový až oligocénny vek. Vznik tejto zóny sa interpretuje zrážkou litosférických mikroplatní, ktorá bola spojená s exhumáciou ofiolitových komplexov. V centrálnej a východnej časti zóny vystupuje formácia Sefidabe (označenie KPs na obr. 2). Je to jedna z najrozsiahlejších jednotiek v oblasti Janja. Pozostáva z vulkanických a pyroklastických sekvencií, ktorých prítomnosť zohrala úlohu pri vzniku miestnej zlatorudnej mineralizácie. Vzhľadom na rozsah prieskumu a hustú sieť vodných tokov sa v oblasti realizoval systematický odber 285 geochemických vzoriek a 59 vzoriek ťažkých minerálov z vodných tokov (obr. 3). Popri odbere vzoriek z vodných tokov sa vzorkovanie robilo aj na povrchu v miestach odkrytosti polymetalických žíl.

Aplikácia algoritmu DT dokladá, že diorit a s ním susediace horniny metamorfované teplom, bridlice a zvetraninový plášť (hlina) vykazujú najvyšší potenciál na prítomnosť zlatých častíc. Diorit má tradične významnú úlohu v horninových sekvenciách obsahujúcich ložiská zlata a považuje sa za zdrojovú horninu mineralizácie. Teplo dioritovej magmy napomáha hydrotermálnym procesom. Naopak, kontaktné rohovce v susedstve dioritových telies zohrávajú negatívnu úlohu pri vzniku zlatorudnej mineralizácie pre ich nízku pórovitosť, sťažujúcu prestup hydroteriem a kryštalizačné procesy mineralizácie. Kryštály pyritu nachádzajúce sa v bridlici alebo hline indikujú vyzrážanie zlata zo sulfátovej fázy v dôsledku zmeny oxidačno-redukčných podmienok. Častice zlata sa zvyčajne nachádzajú v redukčnom prostredí, ako sú ílovité minerály a bridlice. Tam sa môžu vyzrážať v dôsledku hydrotermálnej aktivity. Vďaka tomu sú tieto horniny perspektívne na zlatorudnú mineralizáciu. Zdrojovou horninou v prípade výskytov v oblasti Janja je teda diorit a hostiteľským prostredím sú bridlice a zvetraninový plášť. Dôležité je aj zistenie formy výskytu zlata v bridliciach (napr. v pyrite alebo chalkopyrite) a skúmanie propylitizovaných a argilitových alteračných zón. Mineralizácia sa prednostne vyskytuje v propylitickej zóne hydrotermálneho systému v dôsledku vysokej úrovne oxidácie a kyslého prostredia. Úlohu zohráva aj atmosférická voda. Oblasti s vysokým obsahom draslíka však zvyčajne nedisponujú väčším množstvom zlatorudných ložísk. Naopak, susediace okolité propylitické a argilitové zóny sú bohaté na zlatorudnú mineralizáciu. Záver článku prináša hodnotenie, že algoritmus rozhodovacieho stromu (DT) môže určiť oblasti bohaté na vyhľadávanú mineralizáciu s vysokou úspešnosťou.

Doručené / Received: 22. 10. 2024
Prijaté na publikovanie / Accepted: 17. 12. 2024