International Journal of Safety and Security Engineering
Vol. 16, No. 7, July, 2026, pp. 1601-1616
Journal homepage: http://iieta.org/journals/ijsse

Preventive Model for Slope Failure Forecasting at the Quellaveco Mine Early Detection Approach Using MLP–PyTorch, Random Forest, and Logistic Regression with SMOTE

Eduardo Manuel Noriega-Vidal*, Alexander Sebastian Jimenez-Calvanapon, Christian Manuel Aranda-Vega, Manuel Oswaldo Castillo-Orbegoso

Department of Mining Engineering, Faculty of Engineering, Universidad Privada del Norte, Trujillo 13001, Peru

Corresponding Author Email: eduardo.noriega@upn.edu.pe

https://doi.org/10.18280/ijsse.160716

Received: 19 June 2026
Revised: 21 July 2026
Accepted: 27 July 2026
Available online: 31 July 2026

Keywords:
mining slopes, machine learning, MLP neural network, PyTorch, Random Forest, SMOTE, Quellaveco, SENAMHI

ABSTRACT

Neural Network implemented with PyTorch (21–20–10–5–1 architecture, Batch Normalization, 20% Dropout), Random Forest (200 trees), and Logistic Regression, on 1,500 historical records (2022–2024) with 21 geotechnical, environmental, and operational variables. Environmental variables were calibrated with data from four SENAMHI stations (Calacoa, Moquegua, Quinistaquillas, and Ubinas) using inverse distance weighting (IDW) interpolation. Facing class imbalance (91.33% stable / 8.67% failure), SMOTE was applied exclusively to the training set, generating 2,192 balanced samples. On the held-out test set (n = 300, original distribution) the Multi-Layer Perceptron (MLP) reached Accuracy = 0.870, Recall = 0.500, and AUC-ROC = 0.750; Random Forest obtained Accuracy = 0.867 but Recall = 0.423; Logistic Regression achieved the highest Recall (0.731) and AUC-ROC (0.785). These performance values are consistent with the variable-importance results presented below. The Gini importance ranking is led by the Factor of Safety (0.1513), followed by GSI (0.1184), RQD (0.0731), and Internal Friction Angle (0.0729). Because FS < 1.3 is also one of the two criteria used to label the target variable, the leading position of the Factor of Safety is reported as an internal consistency check of the model and not as an independent predictive finding; the ranking of the remaining variables, which do not intervene in the labelling rule, is unaffected by this constraint.

1. INTRODUCTION

Mining is one of the most important economic activities worldwide because of its contribution to the supply of strategic mineral resources, job creation, and industrial growth. However, open-pit operations permanently face risks associated with the geotechnical stability of their slopes, which can compromise operational safety, interrupt extractive activities, and generate significant economic and environmental impacts [1]. In large-scale operations such as the Quellaveco mine, in the Moquegua region, slope stability is a critical aspect because of the geological complexity and geomechanical variability of the rock mass. Stability is usually assessed through traditional methods such as limit-equilibrium and deterministic analyses, which estimate safety factors under given geotechnical conditions. These approaches, however, are limited when large volumes of information, non-linear relationships, and dynamic ground conditions are involved, and their ability to anticipate instability events early is restricted under highly variable operational scenarios.

In this context, machine learning techniques have gained relevance in slope-stability analysis because of their capacity to model complex relationships between geotechnical parameters and to improve predictive accuracy relative to conventional methods [2]. They allow large volumes of data from monitoring systems and operational records to be processed, which facilitates the identification of patterns associated with stability and instability conditions.

Geotechnical failures in mining slopes are among the main risks in open-pit operations because of their consequences for the safety of personnel and equipment and for operational continuity. In deep mines, structural discontinuities, lithological variations, and changing hydrogeological conditions increase the complexity of rock-mass behavior and make the timely identification of unstable zones difficult [3]. Moreover, progressive deformations and ground-degradation processes can develop gradually before a failure, which demands monitoring and analysis systems able to detect early patterns of instability.

Modern mining operations generate large volumes of information daily from geotechnical instruments, monitoring sensors, operational records, and production-control systems. A large part of these data, however, is not used in an integrated way for failure prevention, owing to the limitations of conventional methods in processing multivariable relationships and non-linear behavior [4]. This hinders preventive, real-time decision-making and reduces the capacity to anticipate instability events, especially in mines with high geological variability such as Quellaveco.

Multiple international studies have shown that slope failures arise from the simultaneous interaction of geomechanical, environmental, and operational factors [5]. Among the most frequently used geotechnical parameters are the cohesion of the rock mass, the internal friction angle, the slope inclination, the factor of safety, and the unconfined compressive strength. Environmental variables usually include precipitation, water infiltration, and pore pressure, given their influence on saturation and on the loss of ground strength, whereas operational variables commonly consider blasting vibrations, excavation depth, and monitored displacements [5]. In this study, these parameters served as input variables for training the machine learning models, while the target variable corresponded to the slope-stability condition, classified as stable or potential failure.

Traditional slope-stability approaches are constrained by the anisotropy of the rock mass and by the difficulty of representing geomechanical interactions, whereas algorithms such as Random Forest, support vector machines, gradient boosting, and neural networks have produced favorable results in classifying and predicting stability conditions [6]. A teacher–student modeling framework has been used to estimate the dry factor of safety from variables such as cohesion, unit weight, and internal friction angle, with Random Forest attaining coefficients of determination above 0.96 [7]. Similarly, an optimized hybrid support vector regression model has improved factor-of-safety prediction relative to conventional techniques [8].

The value of machine learning as a support tool for risk management and safety engineering has also been highlighted: the expansion of machine learning and deep learning has broadened predictive applications in structural monitoring, risk analysis, and safety systems, particularly where operational variability and data volumes are high [9]. In the Peruvian context, a supervised model trained on more than 65,000 records reached accuracies above 92% using Random Forest for landslide- and flood-related vulnerability estimation, and was framed as a decision-support tool rather than a replacement for expert judgment [10].

For this study, Logistic Regression and Random Forest were selected among the available supervised classifiers because of their advantages for geotechnical-stability problems. Logistic Regression makes it possible to interpret the influence of each variable on the probability of failure, which facilitates statistical analysis and offers adequate performance on binary datasets of stable and unstable conditions. Random Forest, in turn, handles non-linear relationships, reduces overfitting, works efficiently with heterogeneous and highly variable geotechnical data, and estimates the relative importance of the predictor variables, which improves the interpretation of rock-mass behavior in complex mining scenarios [11].

Recent work has extended these ideas. A hybrid physics–data approach that couples numerical analysis, neural networks, and monitoring data has been proposed to evaluate slope stability under future rainfall scenarios, reducing factor-of-safety errors to a maximum of 4.5% and cutting computational time by more than 48% [12]. A physics-based, machine-learning-informed model that applies XGBoost to the outputs of a pseudo-three-dimensional geotechnical model has reduced erroneous predictions of earthquake-induced landslides by 45% [13].

For deeply excavated road slopes, models built on slope height, inclination, cohesion, unit weight, and internal friction angle achieved high precision, with a hybrid SVM–SMA model reaching a coefficient of determination of 0.947 [14].

Comparative frameworks have reached similar conclusions. Using a database of 4,208 limit-equilibrium cases, artificial neural networks proved the most reliable for classifying safe and unsafe slopes among multiple linear regression, Random Forest, and support vector machines [15]. Supervised classifiers that combine geotechnical, geomorphological, and hydrological information from GIS and remote sensing—among them Random Forest, Logistic Regression, and neural networks—have efficiently flagged vulnerable highway slopes before extreme events [16]. A hybrid Bootstrap–GRU–Kriging model that integrates recurrent networks, geostatistics, and uncertainty analysis has provided more conservative and robust displacement-based instability predictions than conventional early-warning indicators [17].

Ensemble methods and data-balancing strategies have been especially influential. A comparison of limit-equilibrium methods against seven machine learning algorithms under dynamic loading showed that ensemble approaches attain the highest predictive accuracy and confirmed the strong influence of cohesion, internal friction angle, and slope geometry [18]. A probabilistic landslide-susceptibility framework based on ensemble learning and SVM-SMOTE balancing reached a precision of 95.73% and a sensitivity of 97.08%, underscoring that imbalance treatment is decisive for predictive capacity [19]. Likewise, genetic-algorithm-optimized models (GA-GBDT and GA-XGBoost) exceeded areas under the curve of 0.92 and outperformed Logistic Regression, showing that hyperparameter optimization and variable selection raise predictive reliability [20].

Despite these advances, the integrated application of machine learning to real open-pit operations remain limited, particularly for the early prediction of slope failures, and there is little evidence on the comparative performance of algorithms such as Logistic Regression and Random Forest under the geological and operational conditions of Peruvian mines such as Quellaveco. This gap motivates predictive approaches that jointly integrate geotechnical, environmental, and operational parameters to strengthen the preventive management of geotechnical risk in surface mining.

Accordingly, this study aimed to determine to what extent machine learning techniques allow geotechnical failures in the open-pit slopes of the Quellaveco mine (Moquegua, 2022–2024) to be predicted. The specific objectives were: (1) to calibrate the environmental variables associated with slope stability, such as precipitation and other meteorological parameters, using historical SENAMHI records and inverse distance weighting (IDW) interpolation, in order to build a consistent training database; (2) to implement and compare the performance of the Multi-Layer Perceptron (MLP) (PyTorch), Random Forest, and Logistic Regression models with SMOTE, using geotechnical, environmental, and operational variables for the classification of stability and instability conditions; and (3) to determine the contribution of these models to the early detection of failures through the evaluation of performance metrics and predictive capacity under realistic operational scenarios.

The article is organized into five sections. Section 2 presents the methodology, including the hydrometeorological analysis of the SENAMHI stations, the IDW calibration of the climatic variables, the data processing, the class balancing, and the development of the predictive models. Section 3 presents the results and the comparative analysis of the algorithms. Section

4 discusses the findings in relation to previous research, and Section 5 presents the main conclusions of the study.

2. METHODOLOGY

2.1 Hydrometeorological characterization of the study area

Daily records from four SENAMHI stations in the Moquegua region (Figure 1) were processed and weighted with IDW to obtain representative values for the Quellaveco area. This hydrometeorological calibration is the baseline on which the environmental variables of the predictive model were subsequently built; its correct characterization was therefore indispensable before addressing the machine learning methodology itself.

Figure 2 shows the monthly climatology of the four stations and the resulting IDW precipitation. The seasonal pattern of the southern Andean altiplano is observed: a wet period from December to March (maximum IDW PP: 4.70 mm/day in February) and a dry period from May to October (minimum: 1.76 mm/day in October). Calacoa registers the highest values (8.49 mm/day in February) due to its higher altitude and exposure to Amazonian air masses. The weighted mean annual precipitation resulted in 32.8 mm, consistent with the semi-arid character of Moquegua [17]. This pattern has direct implications for the pore pressure of the rock mass, a variable among those with the highest predictive importance in the trained model. It is worth noting that, although the weighted mean annual precipitation is relatively low in absolute terms, its seasonal concentration during the austral summer months generates infiltration peaks that can suddenly raise pore pressure, punctually increasing the probability of instability even in slopes with favorable geomechanical conditions during the rest of the year.

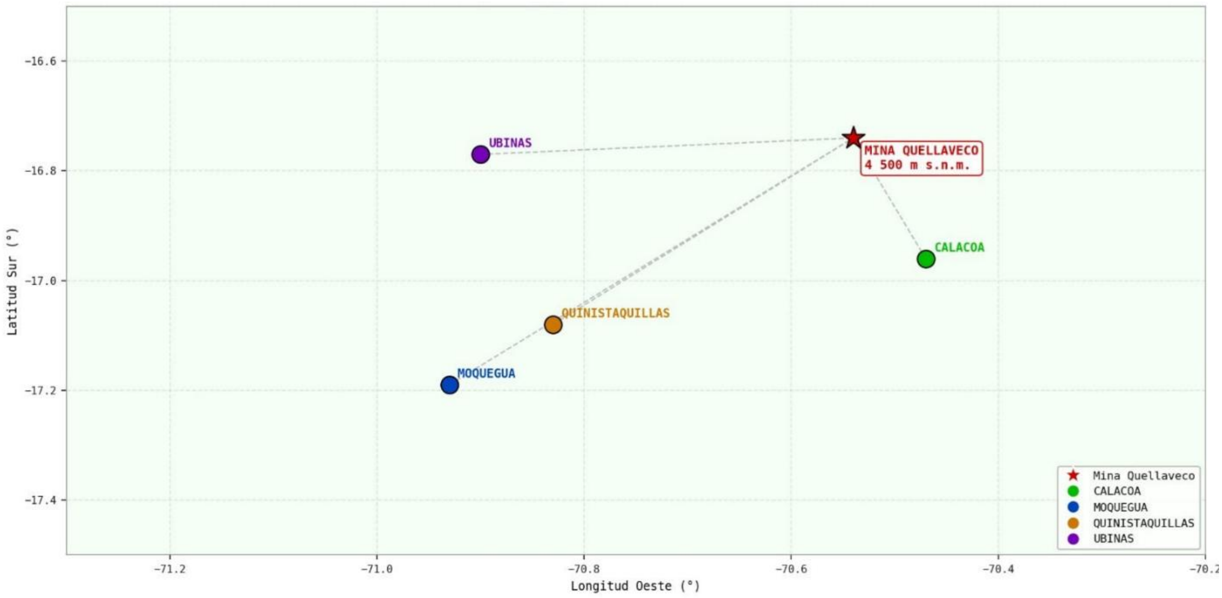

Figure 1. Location of the Quellaveco mine and the four SENAMHI stations
Note: Cartographic baseline: IGN / ANA, Peru.

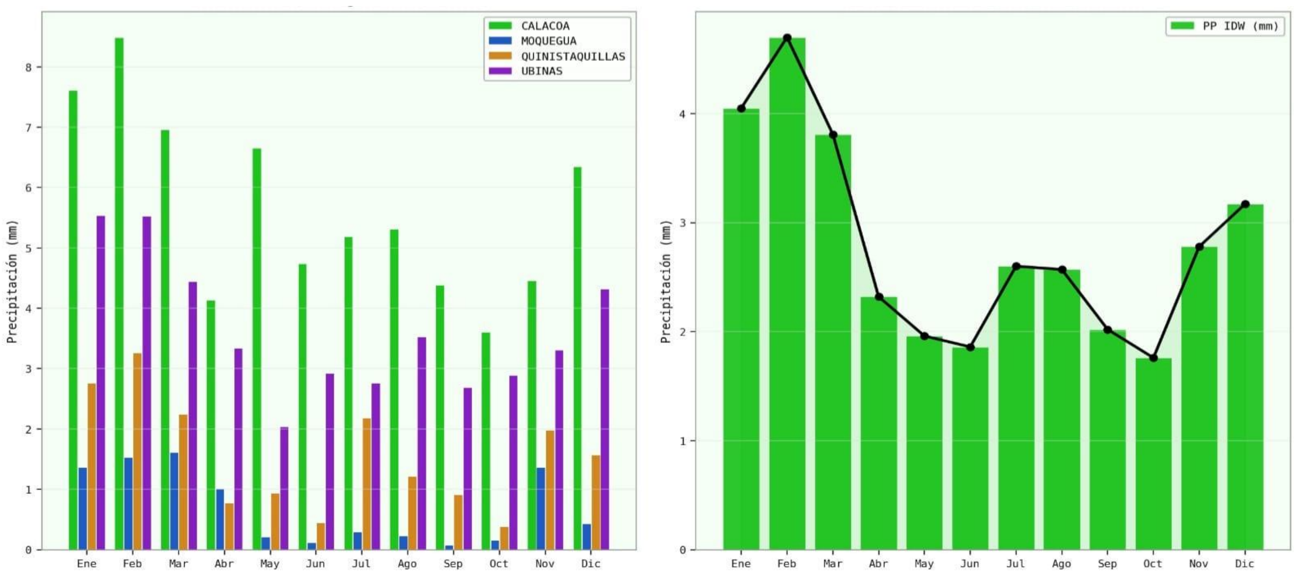

Figure 2. Monthly precipitation climatology by SENAMHI station (left) and mean inverse distance weighting (IDW) precipitation for the Quellaveco area (right)
Note: Period: 1964–2014.

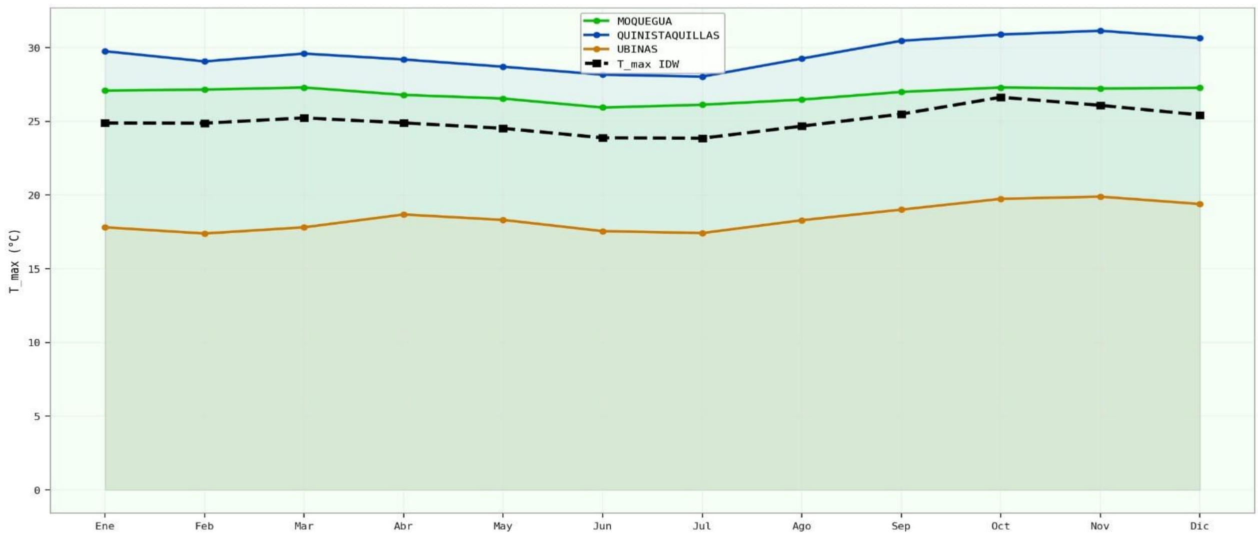

Figure 3. Time series of monthly T_max by SENAMHI station and inverse distance weighting (IDW) curve
Note: Quinistaquillas records the highest temperatures (up to 31.1 °C in November).

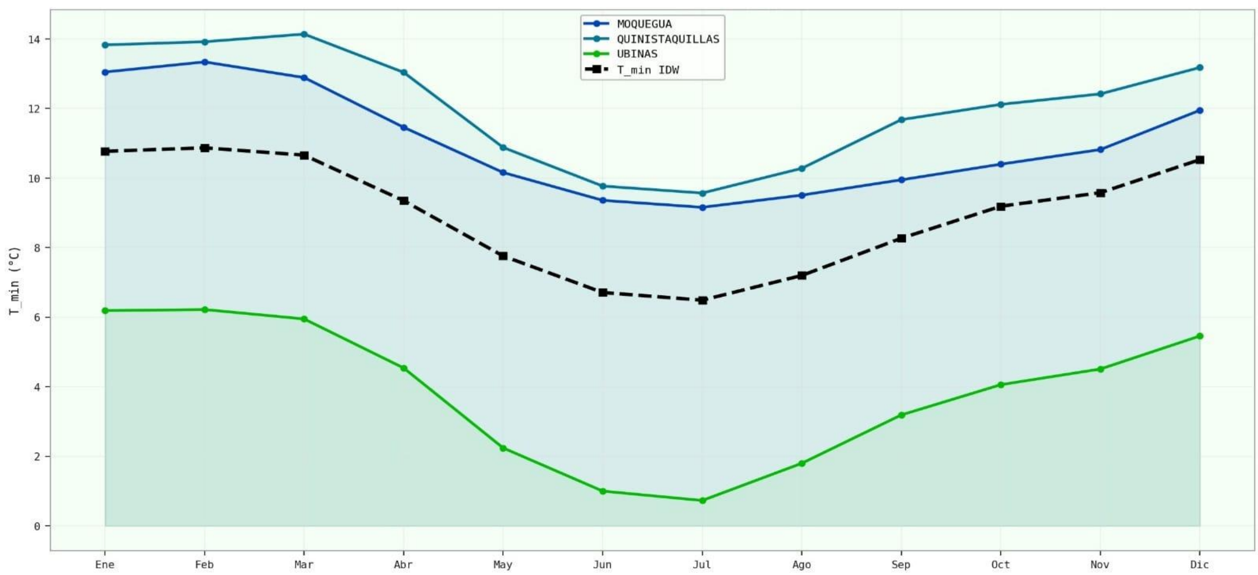

Figure 4. Time series of monthly T_min by SENAMHI station and weighted inverse distance weighting (IDW) curve
Note: Ubinas records the lowest minimum temperatures (down to 0.73 °C in July).

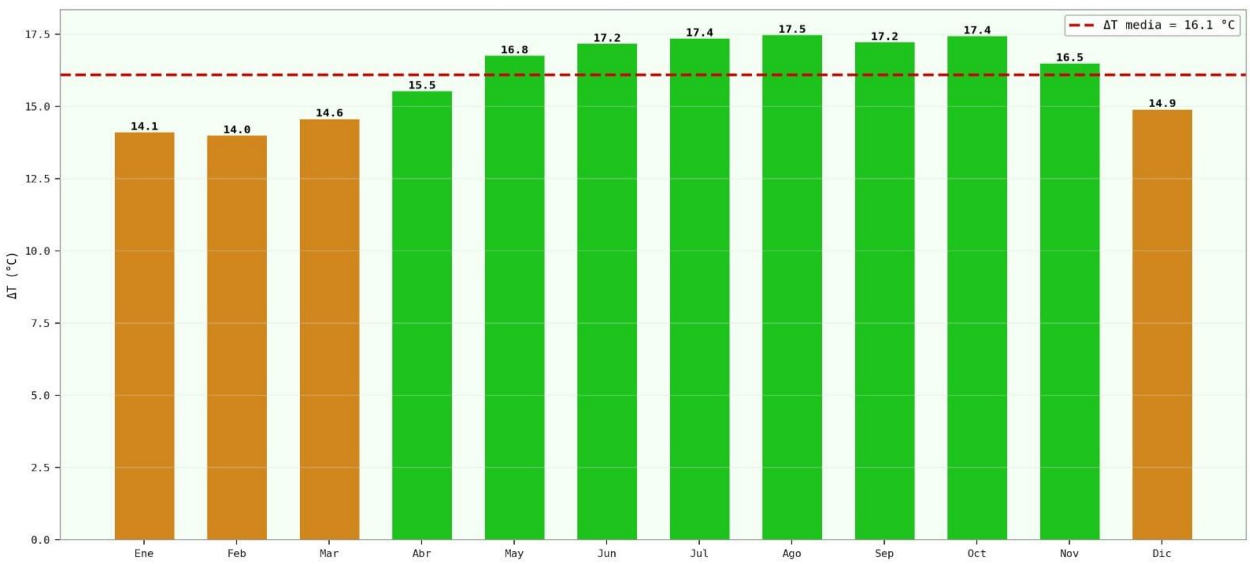

Figure 5. Inverse distance weighting (IDW)-weighted thermal amplitude ΔT = T_max − T_min
Note: Red line: annual mean ΔT = 16.1 °C. Green: ΔT ≥ 15 °C; amber: 13 ≤ ΔT < 15 °C.

Figure 6. Monthly inverse distance weighting (IDW) climatology: precipitation (bars), T_max, and T_min (lines)

Note: The figure reflects the semi-arid regime of the Quellaveco area with marked seasonality.

Amplitude Figures 3–5 show the monthly Tmax, Tmin, and the resulting IDW thermal amplitude ΔT. The IDW Tmax oscillates between 23.85 ℃ (July) and 26.63 ℃ (October), with an annual mean of 25.1 ℃. The mean amplitude is 16.1 ℃, with maximums in September–October (17.4 ℃). These thermomechanical cycles favor the propagation of structural discontinuities, affecting the GSI and the discontinuity index, two of the variables with the highest importance. The daily thermal amplitude, by inducing repeated cycles of expansion and contraction within the rock matrix, acts as a mechanical fatigue mechanism that, sustained over several years of operation, contributes to the progressive weakening of preexisting discontinuities and, therefore, to the gradual reduction of the shear strength of the rock mass.

Figure 6 synthesizes the three climatic components analyzed in this section — precipitation, maximum temperature, and minimum temperature — showing the semi-arid regime with marked seasonality that characterizes the study area: the months with the highest precipitation (December–March) coincide with relatively stable temperatures, whereas the dry period (May–August) exhibits the greatest daily thermal amplitude, a condition that, as noted, favors the mechanical fatigue of the rock mass.

2.2 Study design, area, and variables

This study followed a quantitative approach, because it analyzes numerical data on the geotechnical behavior of open-pit slopes. Geotechnical, environmental, and operational variables — such as cohesion, internal friction angle, precipitation, pore pressure, and blasting vibration — were statistically processed to implement predictive models for classifying stability and instability conditions [21]. Cumulative ground displacement was measured throughout the monitoring campaign, but it was not used as a predictor: it intervenes exclusively in the definition of the target variable (alert-threshold criterion, Section 2.1) and was therefore deliberately excluded from the 21 input variables listed in Table 1, in order to avoid any circularity between the label and the predictors. Consequently, no displacement variable appears in Table 1, and the description of the processed variables in this section is now consistent with that table.

A predictive, comparative, and evaluative design was adopted, since the research implemented, trained, and compared the performance of the Logistic Regression, Random Forest, and MLP algorithms in order to identify the model with the highest capacity for the early detection of slope failures at the Quellaveco mine. Accuracy, precision, recall, and F1-score were used to analyze model performance objectively [22].

The database comprises 1,500 historical records (January 2022–December 2024) obtained from geotechnical sensors, vibrating-wire piezometers, triaxial blasting accelerometers, and the SENAMHI stations described in Section 2.1. The class distribution is markedly imbalanced: 1,370 stable records (91.33%) and 130 potential-failure records (8.67%), a situation characteristic of geotechnical monitoring in real operations, where failure events are statistically infrequent but of high impact. Figure 7 shows this distribution. Imbalance of this kind, common in the detection of rare but critical events such as structural failures, financial fraud, or low-prevalence medical diagnoses, requires specific methodological tdreatment during training, because a classifier trained without adjustment tends to favor the majority class and to miss the failure cases that are precisely of greatest operational interest.

Table 1. Variables employed in the predictive models

Variable	Type	Unit	Range
Cohesion	Geotech.	kPa	30–350
Friction angle	Geotech.	°	15–55
RQD	Geotech.	%	5–100
GSI	Geotech.	–	15–85
UCS	Geotech.	MPa	5–120
Factor of safety	Geotech.	–	1.0–4.0
Slope angle	Geotech.	°	25–75
Slope height	Geotech.	m	50–400
Discontinuity index	Geotech.	/m	0.5–12
σ_1	Geotech.	MPa	1–25
σ_3	Geotech.	MPa	0.5–15
Precipitation *	Environ.	mm	0–60.6
Pore pressure	Environ.	kPa	5–250
Rock-mass moisture	Environ.	%	10–95
Groundwater flow	Environ.	m^3/d	0–35
Temperature *	Environ.	℃	6.5–31.1
Blasting vibration	Operat.	mm/s	0–28
Excavation depth	Operat.	m	45–380
Explosive charge	Operat.	kg	5–50
Number of drill holes	Operat.	–	8–80
Excavation velocity	Operat.	m/d	2–45
Stability condition	Target	0/1	Binary

Note: * Calibrated with SENAMHI (IDW, four stations).

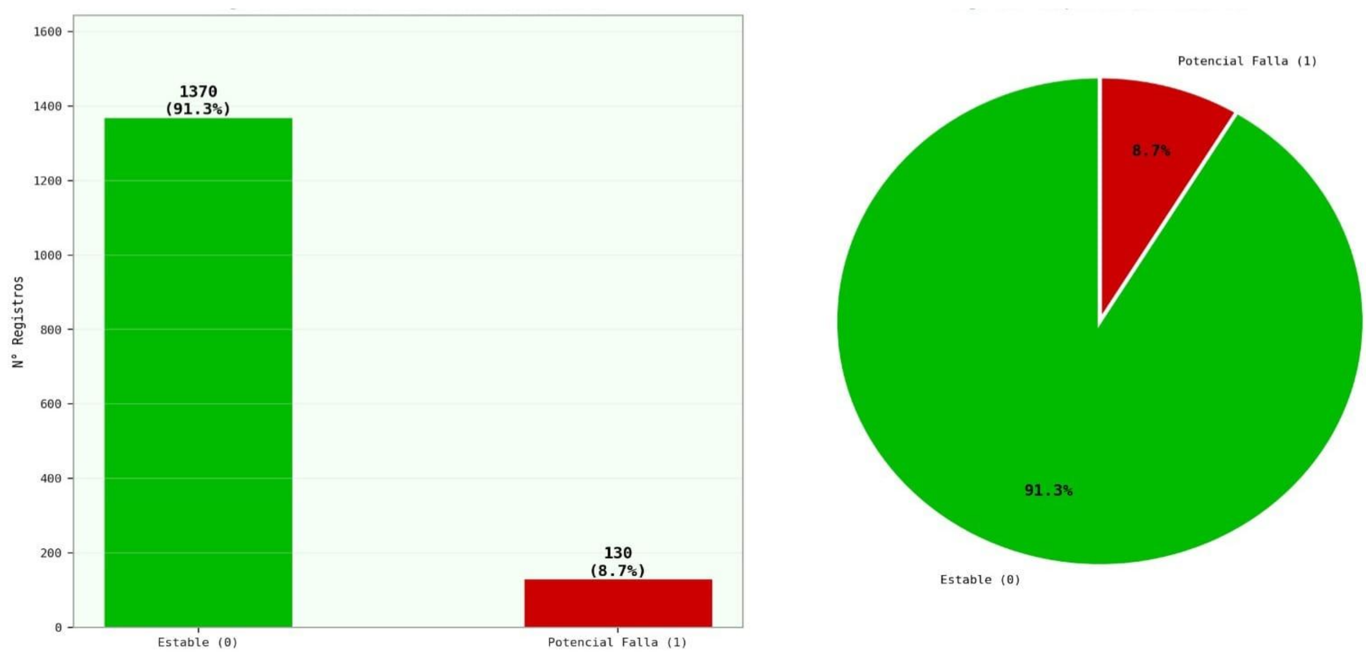

Figure 7. Class distribution in the database: Absolute frequency (left) and percentage proportion (right)
Note: The 91.33% / 8.67% imbalance motivates the use of SMOTE exclusively on the training partition.

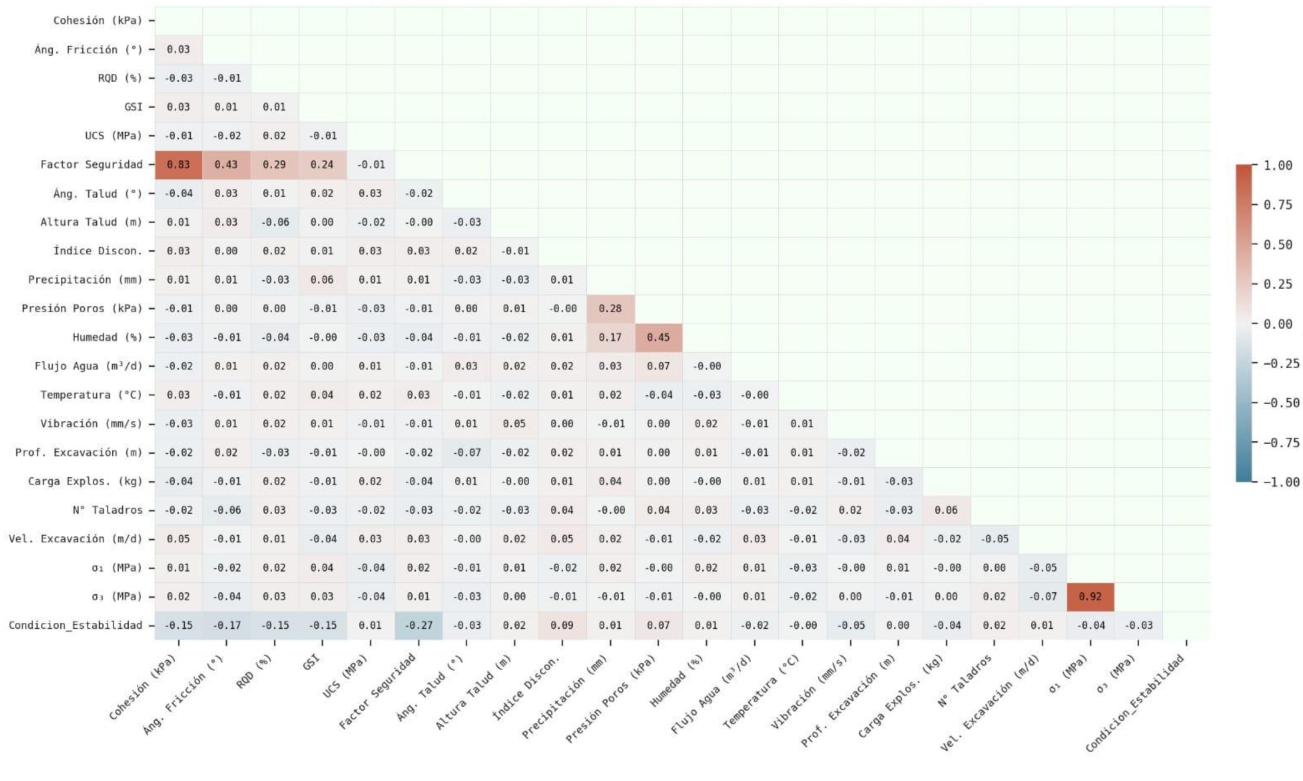

Figure 8. Pearson correlation matrix between the 21 predictor variables and the target variable (stability condition)
Note: Red tones: positive correlation; blue tones: negative correlation. The high collinearity between factor of safety–cohesion (r = 0.83) and σ_1–σ_3 (r = 0.92), and the moderate association between pore pressure and rock-mass moisture (r = 0.45), stand out.

Each record was labeled according to the stability condition observed in the corresponding monitoring window, using measurable, auditable field criteria rather than a subjective assessment. A record was assigned to the potential-failure class (1) when the monitored sector met at least one of the following criteria, defined by the mine's geotechnical protocol and by the applicable Peruvian regulation (MEM-Peru, RCD N°023-2016-EM): (i) a back-calculated factor of safety below the admissible regulatory threshold (FS < 1.3) during the window, and/or (ii) cumulative measured displacement exceeding the alert threshold established for the monitored bench (based on the displacement-rate / inverse-velocity criterion derived from prism and slope-stability-radar monitoring). Records that met none of these criteria throughout the window were labeled as stable (0). This operational definition links the target variable directly to quantitative field conditions and makes the labeling reproducible.

Limitation of the labelling scheme. The factor of safety participates in the definition of Class 1 (FS < 1.3) and is, at the same time, one of the 21 predictors listed in Table 1. For the subset of positive records generated by that criterion, the classifiers can therefore recover part of the labelling rule instead of learning an independent relationship, and any importance measure obtained for FS is partly determined by construction. Cumulative displacement, the second labelling criterion, was excluded from the predictor set precisely for this reason. The factor of safety was retained because it is a continuous quantity that is back-calculated in real time in the operation and carries geomechanical information beyond the 1.3 threshold, and because removing it would have made the input set inconsistent with the variables actually available to the geotechnical team. Nevertheless, no claim of discovery is attached to its importance: Sections 3.2 and 5 state explicitly that the high Gini value of FS reflects the labelling rule and that an ablation experiment in which FS is removed from the input set is required before that ranking can be read as independent evidence. This ablation is identified as immediate future work.

Twenty-one predictor variables grouped into three categories (Table 1) and one binary target variable were selected. The geotechnical variables capture the intrinsic strength of the rock mass; the environmental variables integrate the climatic and hydrogeological effects calibrated with the SENAMHI stations; and the operational variables record the impact of extractive activities on slope stability. This tripartite grouping also allows the relative contribution of each dimension — intrinsic, environmental, and anthropic — to be evaluated independently, which facilitates the physical interpretation of the variable-importance analysis.

Before training, collinearity among the 21 predictor variables and the target variable was assessed with a Pearson correlation matrix (Figure 8). This exploratory analysis identifies strong linear relationships between variable pairs that could introduce redundancy in tree-based models or instability in the Logistic Regression coefficients. As expected, a high correlation appears between the factor of safety and cohesion ($r = 0.83$) and between the principal stresses σ_1 and σ_3 ($r = 0.92$), consistent with their intrinsic geomechanical relationship. In contrast, the linear correlation of each individual variable with the stability condition is comparatively low (maximum $|r| = 0.27$ for slope angle), which confirms that geotechnical failure responds to non-linear interactions among multiple factors rather than to a single dominant predictor and justifies the use of machine learning models able to capture such interactions.

2.3 Preprocessing and SMOTE

Before the predictive models were implemented, the database underwent a preprocessing stage to guarantee the quality, consistency, and reliability of the information used for training, reducing errors associated with incomplete records, outliers, and scale differences among variables [23]. Because the geotechnical, environmental, and operational variables have different ranges and units, they were standardized with StandardScaler to mitigate the influence of heterogeneous scales, optimize convergence, and improve the stability of the predictions [24]. All processing was carried out with Python-based data-analysis and machine learning tools.

To avoid optimistic bias, SMOTE was never applied to the complete dataset before partitioning. A stratified 80% / 20% split first isolated a held-out test set (n = 300) that preserved the original class proportions. Within the remaining 80% used for training, SMOTE (k = 5 nearest neighbors) was applied independently inside the training portion of each of the ten cross-validation folds, so that the synthetic minority samples generated for a given fold never informed the fold reserved for validation.

The final models were then fitted on the fully balanced training set (2,192 samples, 50% / 50%) and evaluated once on the untouched test set. Because the test set contained no synthetic data and never participated in the resampling, no information leaked between subsets, and the reported test metrics reflect the expected performance on realistic operational data. Figure 9 illustrates the class distribution of the training partition before and after applying SMOTE.

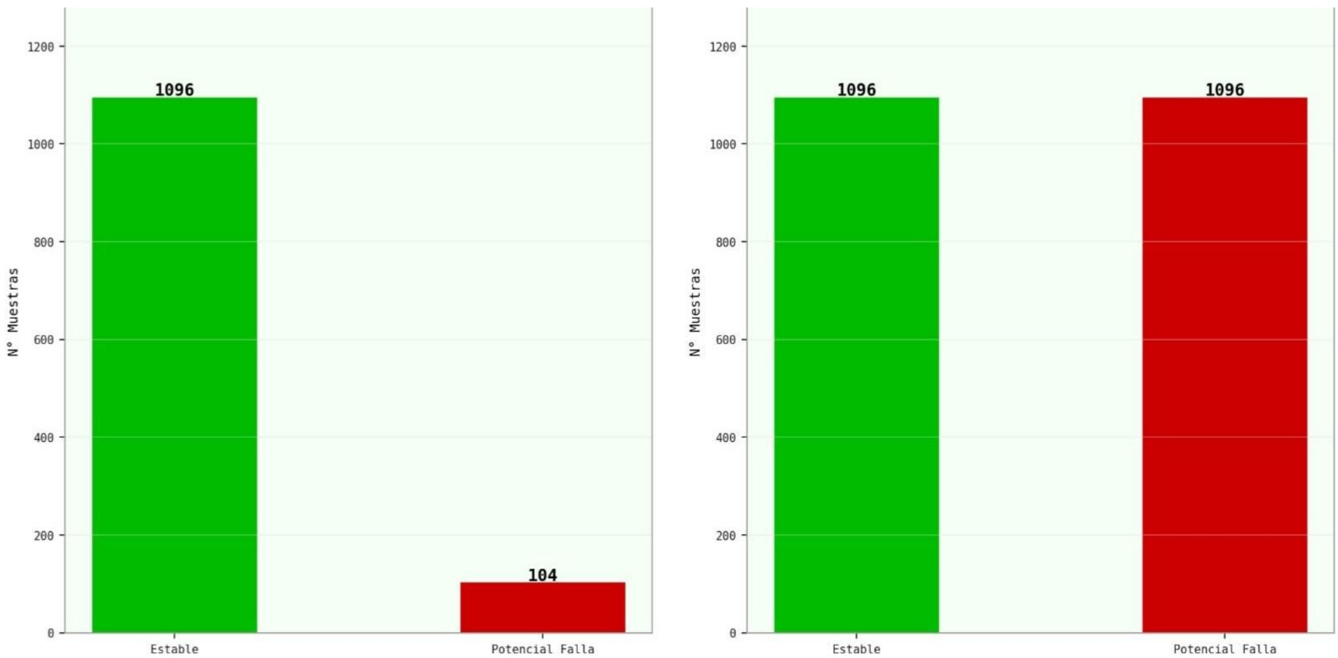

Figure 9. Effect of SMOTE on the training set: before (left) 1,096 stable / 104 failure (imbalance 91.33% / 8.67%); after (right) 2,192 perfectly balanced samples (50% / 50%)

Note: The test set was not modified, so that evaluation reflects real conditions.

For illustrative purposes, Figure 10 presents a representative schematic cross-section of a typical slope at the Quellaveco mine, showing the physical elements that the variables in Table 1 characterize numerically: the geometry of the benches, the potential failure surface, the position of the monitoring piezometers, and the water table. The range of the factor of safety shown (1.25–2.48) illustrates the real geomechanical variability that the predictive model must capture from the 21 input variables.

2.4 Predictive models: Multi-Layer Perceptron, Random Forest, and Logistic Regression

The implemented neural network (Figure 11) follows a 21–20–10–5–1 architecture with double regularization in each hidden layer: batch normalization to stabilize the activation distribution between layers, and dropout (20%) to reduce overfitting by randomly deactivating neurons during training. The binary cross-entropy (BCE) loss function (Eq. (1)) is the standard choice for binary classification:

$$\zeta_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i ln\hat{y}_{i_i} + (1-y_i)\ln(1-\hat{y}_i)\right] \quad (1)$$

where, $y_i \in \{0, 1\}$ is the true label and $\hat{y}_i = \sigma(f(x_i)) \in (0, 1)$ is the probability predicted by the network.

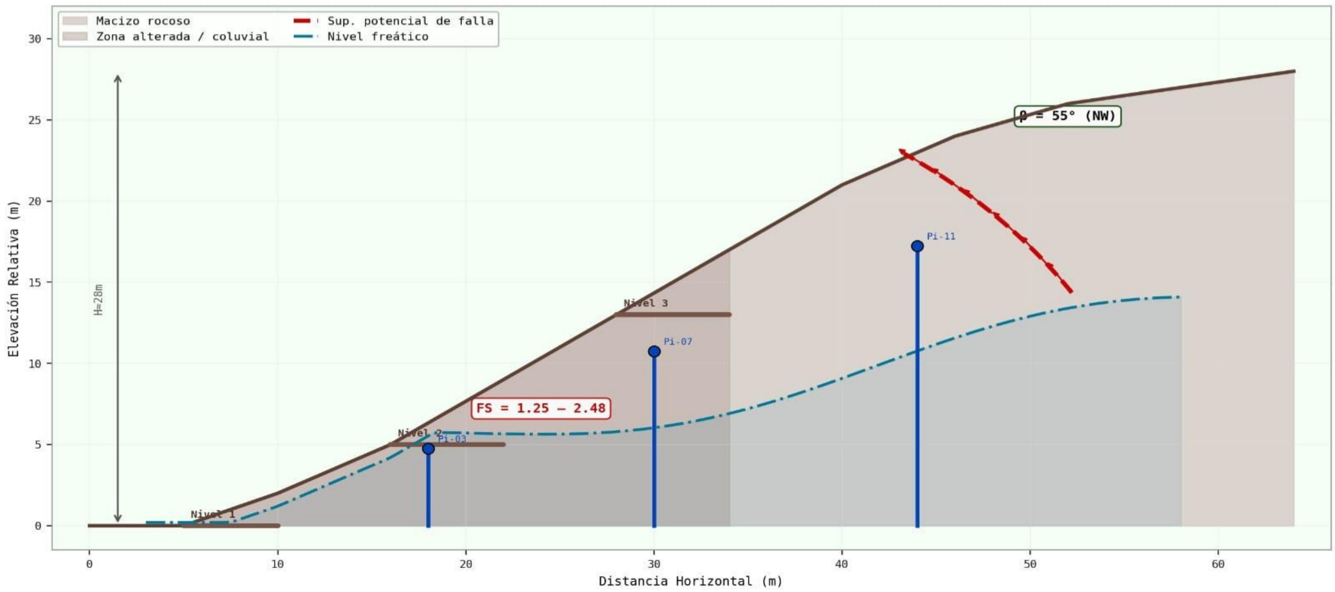

Figure 10. Schematic cross-section of the slope—Quellaveco mine
Note: Benches, potential failure surface, monitoring piezometers, and water table (ISRM 2007 / NTP 339.134 / MEM-Peru Regulation RCD N°023-2016-EM).

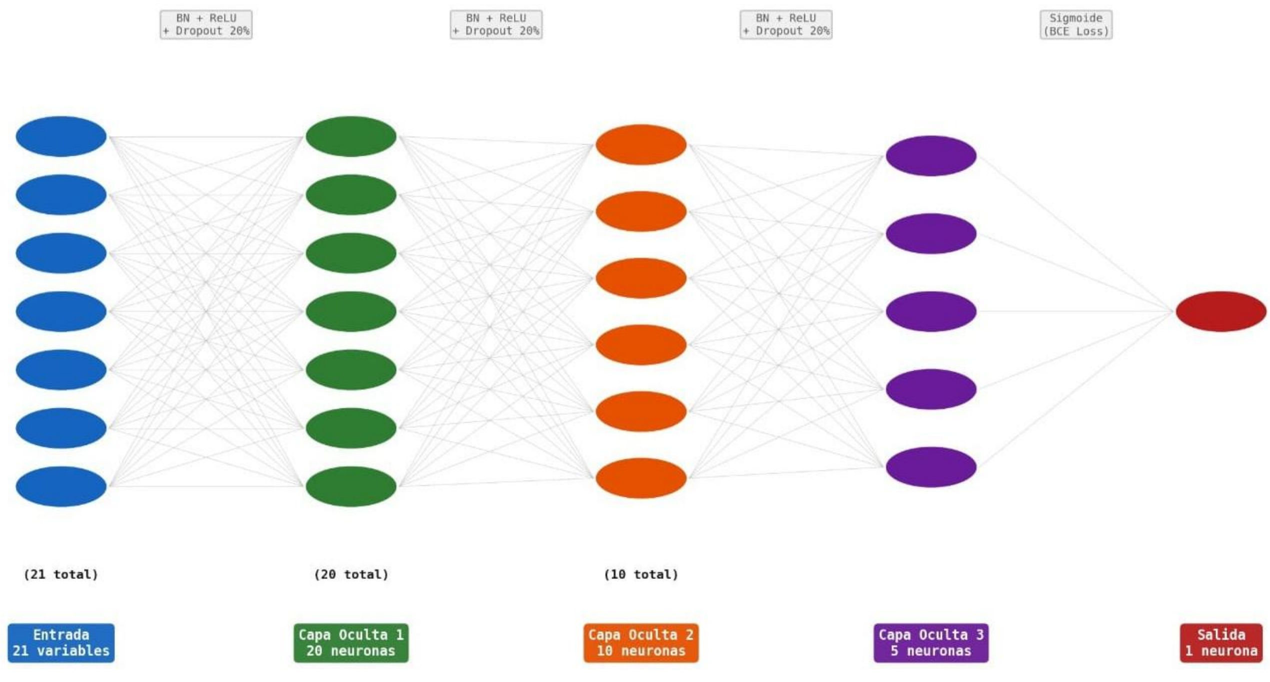

Figure 11. Architecture of the 21–20–10–5–1 Multi-Layer Perceptron (MLP) neural network
Note: Each hidden layer incorporates batch normalization, ReLU activation, and 20% dropout. The output layer applies a sigmoid function over the BCE loss. The pyramidal architecture progressively reduces dimensionality to extract hierarchical features from the input variables.

Training used 300 epochs, the Adam optimizer (η = 10^{-3}, weight decay = 10^{-4}), and a CosineAnnealingLR scheduler for the progressive reduction of the learning rate, which favors smooth convergence toward the minimum of the loss function. The classification threshold was set at 0.5 on the sigmoid output. The combination of batch normalization and dropout responds to a complementary regularization strategy: the former accelerates and stabilizes convergence by normalizing intermediate activations, while the latter reduces the network's dependence on individual neurons, lowering the risk of overfitting given a training set of moderate size.

Logistic Regression was used as a baseline binary classifier because it estimates the probability of a geotechnical event and allows the individual influence of the predictor variables to be interpreted, which facilitates the statistical reading of the results [25]. Random Forest was implemented because it handles non-linear relationships and highly variable geotechnical data. Built from multiple decision trees trained on random subsets of samples and variables, it reduces overfitting and improves robustness; it also manages heterogeneous variables, identifies complex patterns, and estimates the relative importance of each predictor, which strengthens the interpretation of slope behavior and the early detection of instability in open-pit mining [26].

Random Forest was configured with $B = 200$ decision trees, a maximum depth of 15, min_samples_split = 4, and class-weight adjustment (class_weight = "balanced"). The model averages the predictions of the individual trees, reducing variance through ensemble diversity:

$$\hat{y}RF(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{2}$$

Logistic Regression, with L2 regularization, max_iter = 1,000, and balanced weights, estimates the probability of failure through the sigmoid function:

$$P(Y = 1|(x)|) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \tag{3}$$

The simplicity of this post-normalization linear model allows the contribution of each variable to be interpreted directly through the β coefficients. Such interpretability is a practical advantage in mining, where geotechnical teams often need not only a prediction but also a physical justification of the variables driving it, so that the alerts generated by the model can be validated in the field by specialized personnel.

2.5 Evaluation and validation

The performance of the models was evaluated with supervised-classification metrics oriented toward accuracy and predictive capacity, which allowed the behavior of Logistic Regression, Random Forest, and the MLP in identifying stability and instability conditions to be compared. The models were evaluated with the standard binary-classification metrics of (Eqs. (4)-(6)).

$$Accuray = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{6}$$

where *TP, TN, FP,* and *FN* are the true positives, true negatives, false positives, and false negatives, respectively. In addition, the area under the ROC curve (AUC-ROC) was computed as a threshold-independent measure of discriminability.

2.6 Tools and study population

Data processing and model implementation were carried out with Python, given its wide use in machine learning and predictive-analysis research [27]. Figure 12 summarizes the overall workflow followed in this study, from data acquisition to the early-detection system.

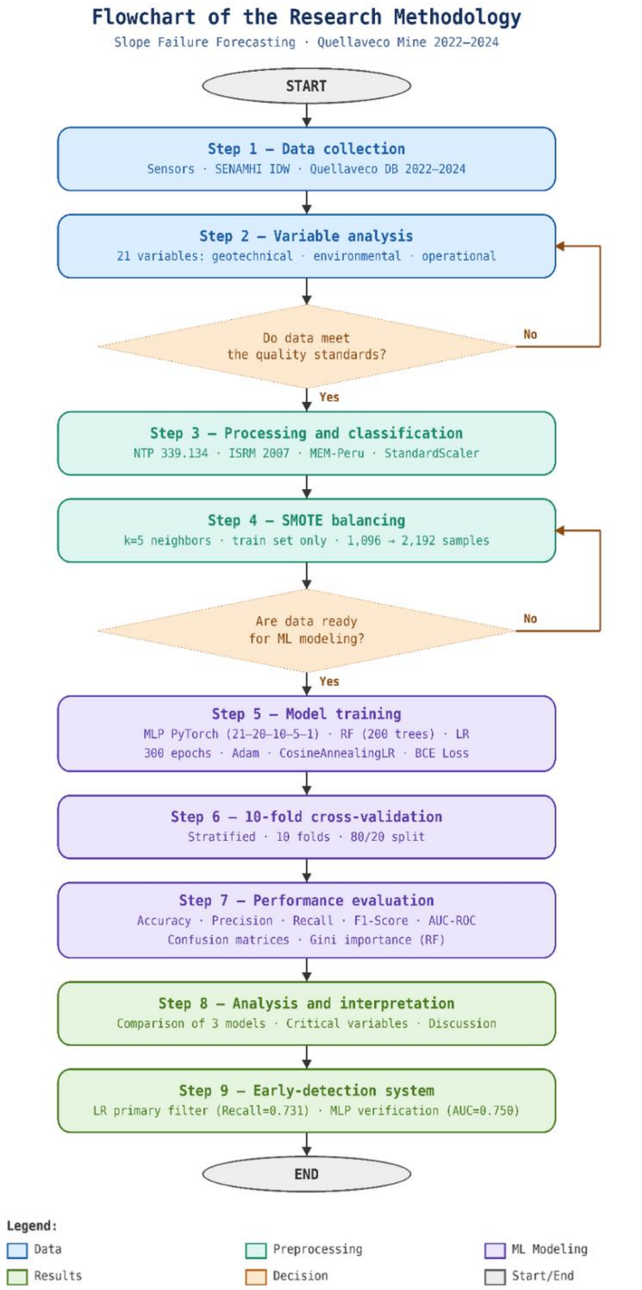

Figure 12. Flowchart of the research methodology
Note: Blue: data collection; teal: preprocessing; purple: ML modeling; green: analysis and results; amber: decision points with feedback to the previous step when data do not meet the quality criteria.

The study population consisted of the geotechnical, environmental, and operational records associated with the behavior of the open-pit slopes at the Quellaveco mine, obtained through monitoring systems, operational reports, and geotechnical controls. The sample consisted of the records selected for training and validating the predictive models, integrating geotechnical, environmental, and operational variables corresponding to stable and unstable conditions. Given the predictive focus of the research, all available records that met the quality and consistency criteria established during data cleaning were used, which improved the learning capacity of the algorithms and adequately.

3. RESULTS

3.1 Descriptive statistics

Table 2 summarizes the main statistics. The factor of safety had a mean of 2.42 (SD = 0.31), and the GSI showed high dispersion (mean = 55.3, SD = 14.8), indicating the geomechanical heterogeneity of the Quellaveco rock mass. This relatively high dispersion of the GSI suggests the coexistence, within the same pit, of sectors of good structural quality alongside significantly more fractured zones, which reinforces the need for sector-differentiated geotechnical monitoring rather than uniform stability criteria for the whole operation.

3.2 Variable importance

Figure 13 shows the Gini importance obtained with Random Forest. The factor of safety heads the ranking (Gini = 0.1513), but this position must be read with caution and is not interpreted here as a predictive discovery: FS < 1.3 is one of the two criteria used to assign Class 1 (Section 2.1), so a substantial part of its importance reflects the labelling rule itself rather than an independently learned relationship. Its high Gini value therefore confirms the internal consistency of the model rather than the predictive value of the variable, and a confirmatory ablation without FS remains pending. The GSI (0.1184) reflects the structural quality of the rock mass, followed by the RQD (0.0731) and the internal friction angle (0.0729); none of these three variables intervenes in the definition of the target variable, so their ranking is free of the circularity affecting FS and constitutes the informative part of the importance analysis. The joint predominance of these geotechnical variables over the environmental and operational ones suggests that, although climate and extractive activity modulate the probability of failure, the baseline condition of the rock mass remains the first-order determining factor, consistent with the classic fundamentals of rock mechanics applied to slope design.

3.3 Multi-Layer Perceptron learning curves

Figure 14 presents the BCE loss and accuracy curves over 300 epochs. The training loss decreases from about 0.68 to about 0.28, while the validation curve converges to about 0.38 from epoch 100 onward, with a stable gap that indicates the absence of significant overfitting. The validation accuracy stabilizes around 0.87 from epoch 120. The stability of the gap between the training and validation curves, without an increasing divergence as the epochs progress, indicates that the adopted regularization scheme — batch normalization plus dropout — fulfilled its role and allowed the network to generalize reasonably beyond the training data.

Table 2. Descriptive statistics—Main variables (n = 1,500)

Variable	Mean	SD	Min.	Max.
Factor of safety	2.42	0.31	1.00	4.00
GSI	55.3	14.8	15	85
Cohesion (kPa)	150.4	48.2	30	350
Friction angle (°)	35.2	9.1	15	55
RQD (%)	58.3	18.6	5	100
UCS (MPa)	52.8	21.4	5	120
Precipitation (mm) *	22.1	18.3	0	60.6
Pore pressure (kPa)	78.4	52.1	5	250
Vibration (mm/s)	7.3	6.8	0	28

Note: * Calibrated with SENAMHI (Calacoa).

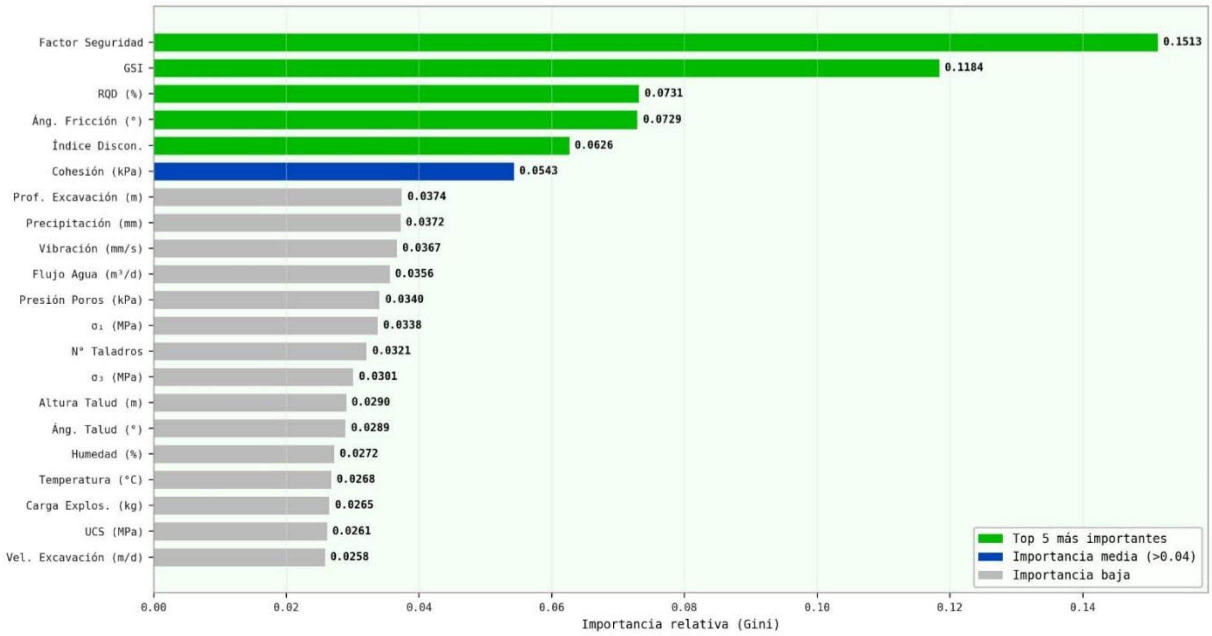

Figure 13. Gini importance—Random Forest

Note: Green bars: top five; blue bar: medium importance (Gini > 0.04). The leading position of the factor of safety is conditioned by its participation in the labelling rule (FS < 1.3) and is not interpreted as an independent finding; GSI, RQD, and internal friction angle do not intervene in that rule.

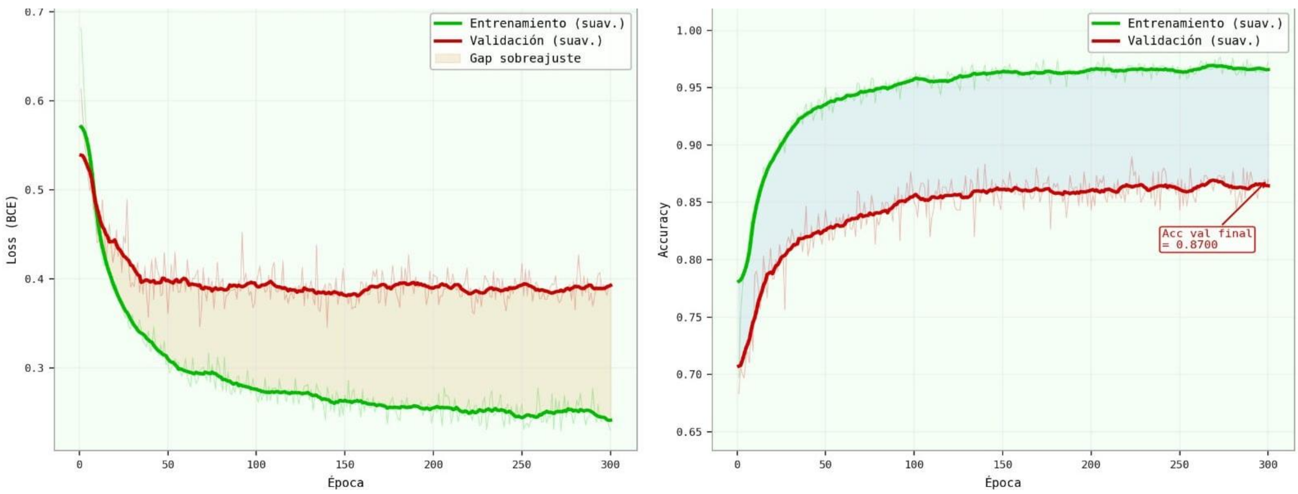

Figure 14. Multi-Layer Perceptron (MLP) learning curves (PyTorch)
Note: Green: training; red: validation. Amber zone: overfitting gap. Stable convergence from epoch 120 (final validation accuracy = 0.870).

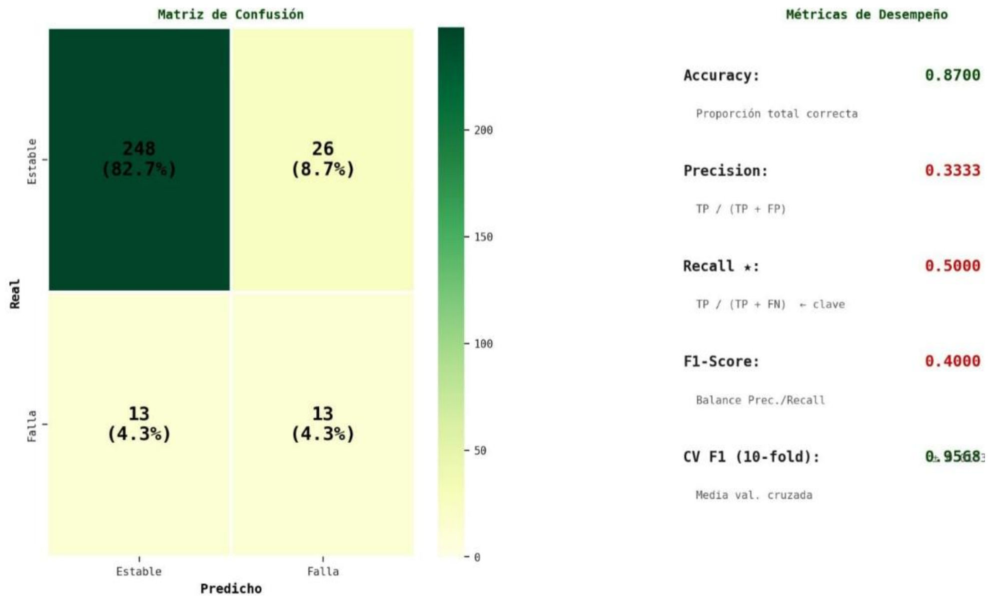

Figure 15. Confusion matrix—Multi-Layer Perceptron (MLP) (PyTorch)
Note: TP = 13 (50.0% of failures detected), FP = 26, FN = 13, TN = 248.

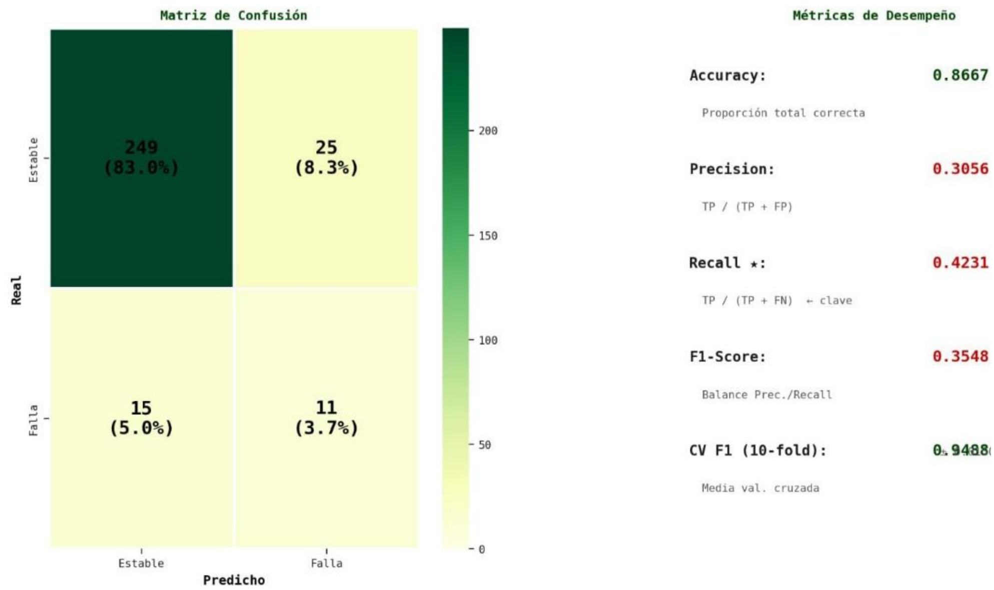

Figure 16. Confusion matrix—Random Forest
Note: Overall accuracy 86.7% but recall = 42.3%: only 11 of the 26 real failures are detected.

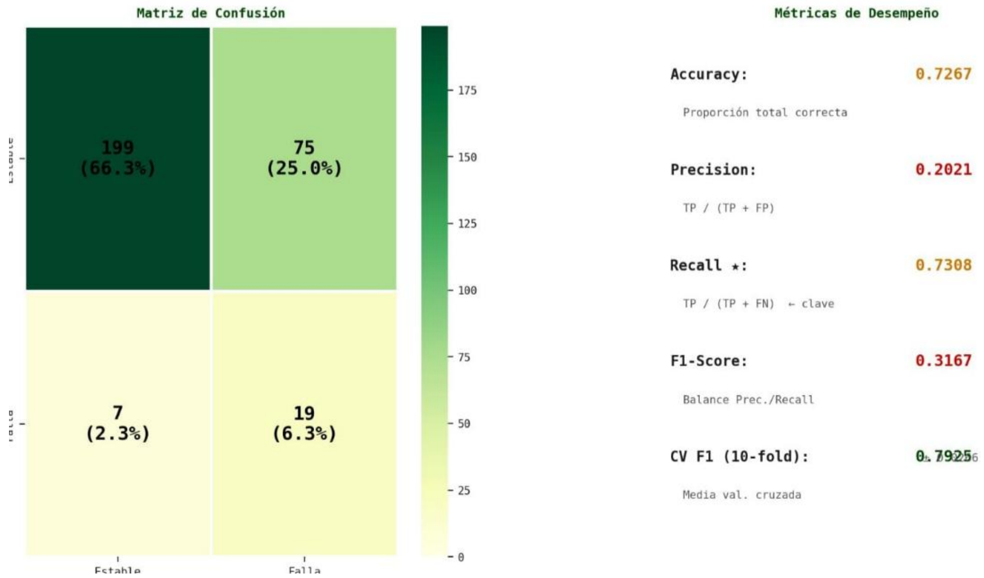

Figure 17. Confusion matrix—Logistic Regression

Note: Recall = 73.1%: 19 of the 26 real failures are detected, the best performance among the three models for early detection, at the cost of a higher false-alarm rate (FP = 75).

3.4 Model performance

The confusion matrices (Figures 15–17) reveal distinct patterns. The MLP detected 13 of the 26 real failures (TP = 13, FN = 13, FP = 26, TN = 248). The Random Forest, despite a comparable overall accuracy (0.867), detected only 11 of the 26 failures (TP = 11, FN = 15, FP = 25, TN = 249), revealing a bias toward the majority class. The Logistic Regression maximized recall (0.731) by detecting 19 of the 26 failures (TP = 19, FN = 7, TN = 199), at the cost of 75 false alarms (FP = 75). This contrast illustrates the classic trade-off between aggregate accuracy and sensitivity to the minority class: the models with the highest overall accuracy (the MLP and the Random Forest) detect fewer real failures than the Logistic Regression, which deliberately sacrifices overall accuracy to maximize the detection of the events that actually matter to the operator. These test-set results for the three models are summarized in Table 3.

Table 3. Summarizes the metrics on the held-out test set (n = 300, original unmodified distribution)

Model	Accuracy	Precision	Recall	F1	AUC
MLP (PyTorch)	0.870	0.333	0.500	0.400	0.750
Random Forest	0.867	0.306	0.423	0.355	0.754
Regression Logistic	0.727	0.202	**0.731**	0.317	**0.785**

Note: Recall = prioritized metric for early detection; bold = best value per column; MLP = Multi-Layer Perceptron (MLP).

3.5 ROC curves

Figure 18 confirms that the three models outperform the random classifier (AUC = 0.50). The Logistic Regression reached the highest AUC (0.785); the Random Forest (0.754) slightly exceeded the MLP (0.750). The steep ascent of the curves at FPR < 0.2 means that a substantial proportion of the potential-failure records is recovered while the false-alarm rate remains low. It does not imply anything about the severity of the events: the target variable is strictly binary (stable / potential failure) and contains no severity grading, so the ROC curve cannot distinguish more severe from milder failures. Any such reading has been removed from the manuscript. In practical terms, this shape of the ROC curve means that the three models correctly identify a significant proportion of real failures while incurring a comparatively low false-alarm rate, which is valuable for an early-warning system that must operate continuously without generating alarm fatigue among the monitoring personnel.

3.6 Metrics comparison

Figure 19 compares the four metrics. No model exceeds 90%, which is expected given the extreme imbalance. Only the Logistic Regression recall (0.731) surpasses the 70% threshold, meaning that roughly three of every four failures are detected in advance. This must be interpreted in the light of the asymmetric cost of errors: even a partial reduction of the undetected-failure risk represents a substantial improvement over purely reactive monitoring, in which intervention occurs only once the instability is already visible in the field.

3.7 Operational trade-off and evaluation of the two-stage detector

Because recall was prioritized, the precision of the Logistic Regression is low (0.202): of every 100 alerts it raises, only about 20 correspond to genuine potential-failure conditions, while the remaining 80 are false alarms. On the test set this amount to 75 false positives out of 274 truly stable records (a false-positive rate of 27.4%). In an operation where each false alarm can trigger an unnecessary verification or a temporary stoppage, this cost is not negligible and must be managed explicitly rather than merely acknowledged.

The two-stage Logistic Regression + MLP scheme could not be tested on a sample-by-sample basis in this study. The joint per-record predictions of the two classifiers were not exported, so the behaviour of the combination was only bounded analytically, by combining the marginal test-set results of the two models under an assumption of conditional independence that cannot be verified with the available outputs. Two canonical configurations were projected in this way (Table 4): a strict conjunction (an alert is confirmed only if both models flag the record) and a disjunction (an alert is raised if either model flags it). The resulting figures are analytical projections, not measured performance, and they do not constitute an evaluation of the cascade; they are reported only to delimit the order of magnitude of the trade-off. No deployment configuration is recommended on this basis.

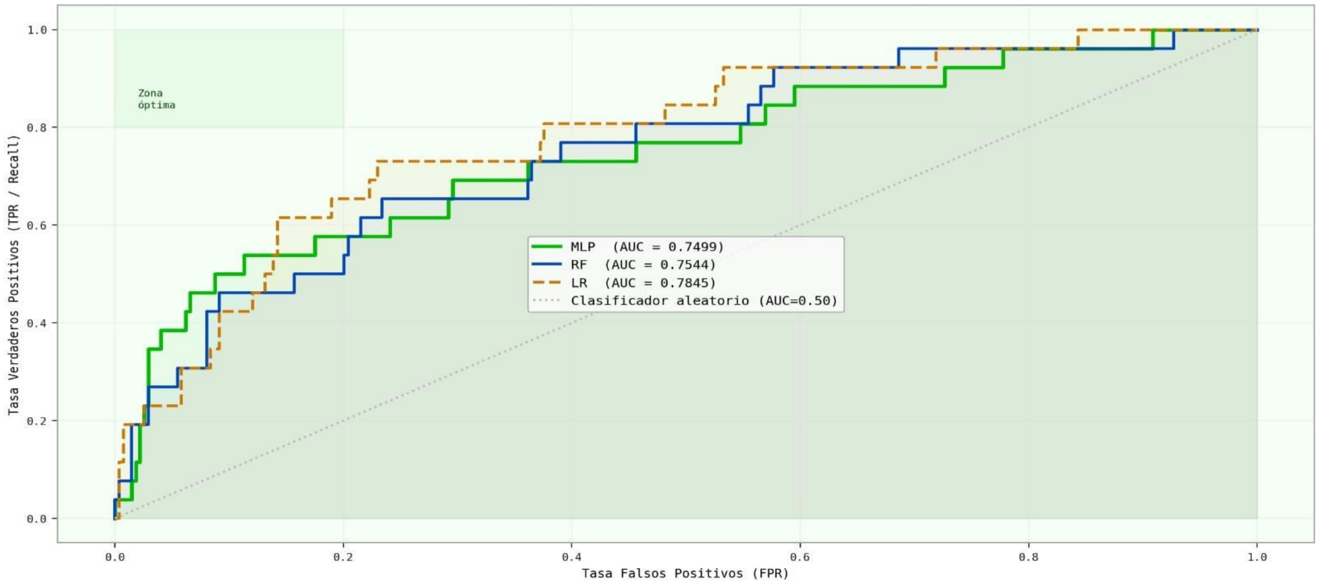

Figure 18. Comparative ROC curves

Note: Logistic Regression AUC = 0.785; Random Forest = 0.754; MLP = 0.750. Steep ascent at FPR < 0.2: a high proportion of failures is recovered at a low false-alarm rate. The target variable is binary and carries no severity grading, so no distinction between more and less severe events can be derived from this curve.

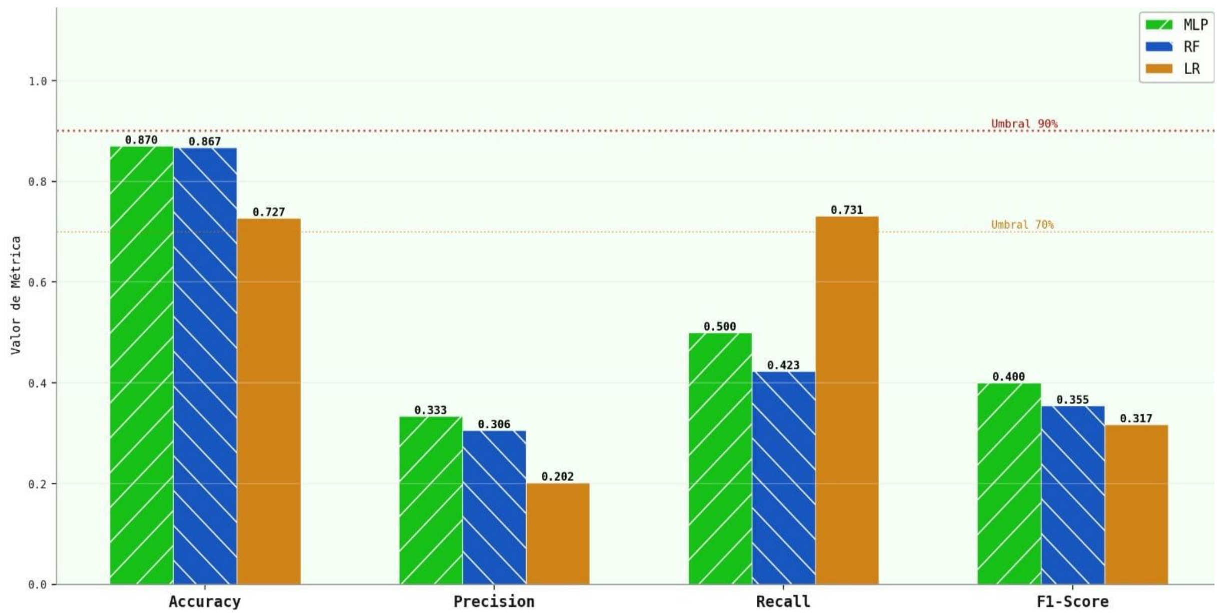

Figure 19. Metrics comparison

Note: Red line: 90% threshold; amber line: 70% threshold. Only the Logistic Regression recall (0.731) exceeds 70%.

Table 4. Analytical projection of two-stage Logistic Regression + Multi-Layer Perceptron (MLP) detection

Configuration	Recall	Precision	TP	FP
Logistic Regression alone	0.731	0.202	19	75
Logistic Regression AND MLP (confirm)	~0.37	~0.57	~9	~7
Logistic Regression OR MLP (either)	~0.87	~0.19	~22	~94

Note: Not an empirical evaluation; derived from the marginal test-set results of the two models under an unverified conditional-independence assumption (reference base: 26 failures and 274 stable records). Values for the combined rows are analytical estimates and are rounded.

Under this projection the strict conjunction would reduce false alarms (from 75 to about seven) and raise precision to roughly 0.57, but would lower recall to about 0.37, which would be unacceptable for a safety-critical detector whose purpose is precisely not to miss failures; the disjunction would push recall to about 0.87 while worsening the false-alarm burden. Neither symmetric rule therefore appears promising. A configuration that merits empirical testing—and that is offered here as a hypothesis rather than as a validated recommendation—is a role-differentiated cascade in which the Logistic Regression acts as a high-sensitivity primary screen that raises every alert (recall = 0.731) and the MLP, which has higher precision (0.333), a comparable AUC, and a probability output, is used as a secondary triage layer that orders the alerts by predicted probability. Because the target variable is binary, this ordering ranks alerts by model confidence and not by failure severity. Until the joint predictions are exported and the scheme is measured directly on the test set, and subsequently in the field, this architecture must be regarded as an untested proposal.

4. DISCUSSION

The results show that machine learning is a viable alternative for the early prediction of geotechnical failures in mining slopes. Although the MLP achieved the highest overall accuracy (0.870) and the Random Forest a comparable value (0.867), the Logistic Regression provided the best performance for early detection, with a recall of 0.731 and an AUC-ROC of 0.785, correctly identifying 19 of the 26 failures in the test set. This is consistent with evidence that machine learning algorithms can model complex relationships between geotechnical and environmental variables and overcome several limitations of traditional slope-stability approaches [6]. The results also confirm that integrating geomechanical, hydrogeological, and climatic variables improve the identification of patterns associated with potentially unstable conditions.

The superiority of the Logistic Regression in sensitivity partially contrasts with results that identified Random Forest as the best performer for factor-of-safety estimation, with coefficients of determination above 0.96 [7]. That difference is attributable to the nature of the present problem: whereas that work addressed the continuous prediction of the factor of safety under controlled conditions, this study addressed a binary classification with strong class imbalance, where failures represent only 8.67% of the records. In this setting, the Logistic Regression identified critical events more effectively, while the Random Forest tended to favor the majority class, raising overall accuracy but reducing the detection of real failures.

The findings also differ from those reporting coefficients of determination above 0.92 with an optimized hybrid support vector regression model [8]. Although such advanced algorithms are effective for estimating the factor of safety, the present results show that statistically simpler models can offer operational advantages when the primary objective is the early detection of risk conditions.

In real mining, where the cost of a false negative can imply significant economic losses, equipment damage, or harm to personnel, the ability to identify potentially unstable events in time is more relevant than maximizing overall accuracy alone.

This is precisely why the low precision of the Logistic Regression, analyzed in Section 3.7, must be managed through the proposed screen-and-triage architecture rather than by shifting to a higher-accuracy but lower-recall model.

The results are also consistent with work reporting accuracies above 92% with Random Forest for natural-hazard vulnerability estimation, whose authors frame predictive models as decision-support tools rather than substitutes for professional judgment [10]. In agreement with that stance, the model developed for Quellaveco does not aim to replace specialized geotechnical evaluation but to complement existing monitoring with early warnings based on historical data and real SENAMHI environmental variables, which increases the representativeness of the operational conditions and strengthens preventive capacity.

Finally, the predictive importance of the factor of safety, the GSI, and the internal friction angle confirms the relevance of parameters traditionally used in the geomechanical evaluation of slopes. The leading position of the factor of safety is conditioned by its role in the labelling rule and is not read as evidence of predictive value; the informative part of the ranking is the weight of the GSI, the RQD, and the internal friction angle, none of which intervenes in the definition of the target variable. Their predominance within the Random Forest model shows that slope stability remains strongly conditioned by the shear-strength properties of the rock mass; at the same time, the simultaneous incorporation of climatic and hydrogeological variables captured operational conditions that conventional approaches rarely integrate jointly. The results therefore support the use of machine learning models as complementary tools for strengthening the preventive management of geotechnical risk in large-scale operations such as the Quellaveco mine.

5. CONCLUSION

The three machine learning models demonstrated discriminative capacity for the stability condition at the Quellaveco mine, with AUC-ROC values between 0.750 and 0.785, all clearly above the random classifier. Calibrating the environmental variables through IDW interpolation of four SENAMHI stations placed precipitation and temperature within climatically plausible ranges for the Moquegua region and provided a physically meaningful environmental context, while applying SMOTE inside each cross-validation fold mitigated the extreme class imbalance without introducing data leakage. On the held-out test set, the Logistic Regression reached the highest recall (0.731), detecting three of every four failures, whereas the MLP offered the best overall balance (accuracy = 0.870, AUC = 0.750) and the Random Forest, despite a comparable accuracy, detected the fewest failures.

Among the predictors, the GSI (0.1184), the RQD (0.0731), and the internal friction angle (0.0729) emerged as the most influential variables that are independent of the labelling rule, consistent with the fundamentals of rock mechanics and with the ISRM 2007, NTP 339.134, and MEM-Peru (RCD N°023-2016-EM) frameworks. The factor of safety obtained the highest Gini value (0.1513), but since FS < 1.3 is one of the two criteria used to define the positive class, that value is reported as an internal consistency check and not as an independent predictive finding.

Given the asymmetric cost of errors, a role-differentiated two-stage detector—with the Logistic Regression as a high-sensitivity primary screen and the MLP, exploiting its higher precision and probability output, as a triage layer that orders the resulting alerts for field verification—is proposed as a hypothesis for future testing rather than as a validated recommendation. The projected figures (false alarms falling from 75 to about seven under a strict conjunction, at a recall of roughly 0.37, against a recall near 0.87 under a disjunction) are analytical projections obtained from the marginal results of the two models under an unverified conditional-independence assumption; the cascade was not evaluated sample by sample and no deployment configuration is recommended on that basis. The main limitations are the moderate size and single-site scope of the database, the fact that the two-stage scheme was projected analytically rather than measured or field-tested, and the partial circularity introduced by retaining the factor of safety as a predictor while FS < 1.3 also participates in the definition of the positive class. Future work should therefore repeat the training with the factor of safety removed from the input set in order to establish the importance ranking free of that constraint, export the joint model predictions to measure the cascade empirically, expand and diversify the historical database, explore alternative class-balancing strategies, and field-test the detector before integrating it into the operational geotechnical monitoring system of the Quellaveco mine.

REFERENCES

[1] Santillana Santos, M.E. (2006). La importancia de la actividad minera en la economía y sociedad peruana. Economic Commission for Latin America and the Caribbean. CEPAL. https://www.cepal.org/es/publicaciones/6312-la-importancia-la-actividad-minera-la-economia-sociedad-peruana.

[2] Mahmoodzadeh, A., Alanazi, A., Mohammed, A.H., et al. (2024). Comprehensive analysis of multiple machine learning techniques for rock slope failure prediction. Journal of Rock Mechanics and Geotechnical Engineering, 16(11): 4386-4398. https://doi.org/10.1016/j.jrmge.2023.08.023

[3] Gladious, J., Paul, P.S., Mukhopadhyay, M. (2025). Machine learning based prediction of geotechnical parameters affecting slope stability in open-pit iron ore mines in high precipitation zone. Scientific Reports, 15: 21868. https://doi.org/10.1038/s41598-025-99026-4

[4] Jiang, S., Li, J., Zhang, S., Gu, Q., Lu, C., Liu, H. (2022). Landslide risk prediction by using GBRT algorithm: Application of artificial intelligence in disaster prevention of energy mining. Process Safety and Environmental Protection, 166: 384-392. https://doi.org/10.1016/j.psep.2022.08.043

[5] McQuillan, A., Mitelman, A., Elmo, D. (2023). Back-analysis of structurally controlled failure in an open-pit mine with machine learning tools. Geotechnics, 3(4): 1207-1218. https://doi.org/10.3390/geotechnics3040066

[6] Trinidad, M., Momayez, M. (2025). Machine learning in slope stability: A review with implications for landslide hazard assessment. GeoHazards, 6(4): 67. https://doi.org/10.3390/geohazards6040067

[7] Kassa, S.M., Wubineh, B.Z., Geremew, A.M., Kumar, N.D., Kacprzak, G. (2025). Domain-driven teacher–student machine learning framework for predicting slope stability under dry conditions. Applied Sciences, 15(19): 10613. https://doi.org/10.3390/app151910613

[8] Khajehzadeh, M., Keawsawasvong, S. (2023). Predicting slope safety using an optimized machine learning model. Heliyon, 9(12): e23012. https://doi.org/10.1016/j.heliyon.2023.e23012

[9] Leoni, L., BahooToroody, A., Abaei, M.M., Cantini, A., BahooToroody, F., De Carlo, F. (2024). Machine learning and deep learning for safety applications: Investigating the intellectual structure and the temporal evolution. Safety Science, 170: 106363. https://doi.org/10.1016/j.ssci.2023.106363

[10] López Vega, J.W., Torres Lázaro, J.C., Herrera Quispe, J. (2021). Modelo predictivo basado en machine learning para la estimación de vulnerabilidades de riesgo de inundación y deslizamiento. Caso de estudio: Instituciones educativas del Perú. In Actas del Congreso Internacional de Ingeniería de Sistemas, pp. 206-207. https://doi.org/10.26439/ciis2021.5637

[11] Speiser, J.L., Miller, M.E., Tooze, J., Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications, 134: 93-101. https://doi.org/10.1016/j.eswa.2019.05.028

[12] Zhang, Z., Liu, K., Su, Q., et al. (2026). A novel physics–data hybrid approach for slope stability assessment considering future rainfall patterns. Computers and Geotechnics, 192: 107853. https://doi.org/10.1016/j.compgeo.2025.107853

[13] Gong, W., Zekkos, D., Clark, M. (2026). A physics-based machine learning-informed model for predicting regional earthquake-induced landslides. Engineering Geology, 361: 108488. https://doi.org/10.1016/j.enggeo.2025.108488

[14] Dam, N.D., Manh, N.D., Prakash, I., Dung, V.T., Binh, P.T. (2026). Factor of safety analysis for deeply excavated slopes in road construction using machine learning models. In International Conference on Sustainability in Civil Engineering, pp. 175-183. https://doi.org/10.1007/978-981-95-1072-6_17

[15] Rajan, K.C., Aryal, M., Sharma, K., Bhandary, N.P., Pokhrel, R., Acharya, I.P. (2025). Development of a framework for the prediction of slope stability using machine learning paradigms. Natural Hazards, 121(1): 83-107. https://doi.org/10.1007/s11069-024-06819-3

[16] Salunke, R., Khan, S. (2025). Evaluating machine learning models for identifying at-risk highway slope assets. In Geotechnical Frontiers 2025, pp. 83-94. https://doi.org/10.1061/9780784485989.010

[17] Li, T.F., Xiao, X.P., Yan, R.H., et al. (2024). The spatial-temporal probability assessment for slope instability based on uncertainty of machine learning-based prediction. Results in Engineering, 24: 103545. https://doi.org/10.1016/j.rineng.2024.103545

[18] Bansal, V., Sarkar, R. (2024). Comparative analysis of slope stability for Kalimpong region under dynamic loading using limit equilibrium method and machine benchmark learning classifiers. Iranian Journal of Science and Technology, Transactions of Civil Engineering, 48(4): 2785-2807. https://doi.org/10.1007/s40996-024-01344-6

[19] Sharma, N., Saharia, M., Ramana, G.V. (2024). High resolution landslide susceptibility mapping using ensemble machine learning and geospatial big data. Catena, 235: 107653. https://doi.org/10.1016/j.catena.2023.107653

[20] Zheng, B., Wang, J., Feng, T., et al. (2023). Improving the landslide susceptibility prediction accuracy by using genetic algorithm optimized machine learning approach. International Journal of Intelligent Systems, 2023(1): 5525793. https://doi.org/10.1155/2023/5525793

[21] Vizcaíno Zúñiga, P.I., Cedeño Cedeño, R.J., Maldonado Palacios, I.A. (2023). Metodología de la investigación científica: Guía práctica. Ciencia Latina Revista Científica Multidisciplinar, 7(4): 9723-9762. https://doi.org/10.37811/cl_rcm.v7i4.7658

[22] Manterola, C., Quiroz, G., Salazar, P., García, N. (2019). Metodología de los tipos y diseños de estudio más frecuentemente utilizados en investigación clínica. Revista Médica Clínica Las Condes, 30(1): 36-49. https://doi.org/10.1016/j.rmclc.2018.11.005

[23] Oliveira, M.F.D., Ortiz, B.V., Morata, G.T., Jiménez, A.F., Rolim, G.D.S., Silva, R.P.D. (2022). Training machine learning algorithms using remote sensing and topographic indices for corn yield prediction. Remote Sensing, 14(23): 6171. https://doi.org/10.3390/rs14236171

[24] Intriago, M.G., Cevallos, R.A. (2025). Impacto de la normalización de datos en la precisión de modelos de aprendizaje supervisado. Revista de Investigaciones en Energía, Medio Ambiente y Tecnología, 10(2): 59-79. https://doi.org/10.33936/riemat.v10i2.7853

[25] Martínez Pérez, J.A., Pérez Martín, P.S. (2024). Regresión logística. Medicina de Familia. Semergen, 50(1): 102086. https://doi.org/10.1016/j.semerg.2023.102086

[26] Imani, M., Beikmohammadi, A., Arabnia, H.R. (2025). Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels. Technologies, 13(3): 88. https://doi.org/10.3390/technologies13030088

[27] Harle, S.M., Wankhade, R.L. (2025). Machine learning techniques for predictive modelling in geotechnical engineering: A succinct review. Discover Civil Engineering, 2: 86. https://doi.org/10.1007/s44290-025-00224-w