# Supporting Material

- **Section 2.2.1. The Brief Principle of the Machine Learning Algorithm Applied**

**(1) Logistic Regression Model (LR)**

LR is applied to portray the relationship between the probability of landslide occurrence (0 for no occurrence and 1 for occurrence) and multiple affecting factors (Equation (1)). The parameters in the model are estimated by the maximum likelihood method, and the independent variables can be continuous or discrete:

$$P = 1/[1 + e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i)}] \quad (1)$$

where $P$ is the probability of hazard occurrence, $P \in [0,1]$, $\alpha$ is the constant term, $\beta_i$ is the partial regression coefficient reflecting the influence of affecting factor $x_i$ on $P$, and $i$ is the number of affecting factors. $P/(1-P)$ is the probability ratio of occurrence and non-occurrence of a hazard, and taking the natural logarithm of it can yield Equation (2):

$$\ln[P/(1-P)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \quad (2)$$

**(2) Support Vector Machine (SVM)**

As a binary classification model, the main idea of a SVM is to find a maximum-margin hyperplane which can make the two classes of data points as far from it as possible. Given the training data set $D = \{(x_i, y_i)\}, y_i \in \{-1,1\}$, maximizing the distance to the hyperplane $2/||w||$ is equivalent to minimizing $||w||^2/2$ and then obtaining the basic SVM model (Equation (3)):

$$\min (||w||^2/2), s.t. y_i(w^T x_i + b) \geq 1 \ (i = 1,2, \dots, m) \quad (3)$$

Equation (3) is a convex quadratic regularization problem with linear constraints. The optimal classification hyperplane (Equation (4)) can be found by solving its dual problem derived through Lagrange multipliers:

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b \quad (4)$$

where $\alpha_i$ is the Lagrange multipliers, $x_i$ is the feature vector composed of each affecting factor, $K(x_i, x)$ is the kernel function, and $b$ is the deviation coefficient. The output probability value has no fixed range, and thus in order to facilitate the comparison of different models, it needs to be converted into $[0,1]$ according to Equation (5) [1]:

$$p(x) = 1/(1 + e^{Af(x)+B}) \quad (5)$$

where $p(x)$ is the landslide susceptibility value, $p(x) \in [0,1]$, and $A, B$ are coefficients to be determined based on the maximum likelihood method.

**(3) Gradient-Boosting Decision Tree (GBDT)**

The basic idea of the GBDT is to build a strong classifier with multiple weak classifiers. Through the iterative process of fitting the residuals of the previous round of base learners according to the negative gradient direction of the loss function and building a new model to fit, the residuals of each round are gradually reduced, and thereby the outputs of the base learners approximate the true values step by step [2] (Equation (6)):

$$f(x) = f_T(x) = f_0(x) + \sum_{t=1}^{T} \sum_{j=1}^{J} c_{tj} I \ (x \in R_{tj}) \quad (6)$$

where $f(x)$ is the final strong classifier, $T$ is the maximum number of iterations, $c_{tj}$ is the best fit value of the negative gradient direction of the loss function at leaf node $j$, $R_{tj} (j = 1,2, \dots, J)$ is the leaf node region corresponding to the regression tree $t$, and $J$ is

the number of leaf nodes of regression tree $t$. The final output is correlated with the log odds and therefore is mapped to [0,1] by Equation (7) to keep the consistency of comparison:

$$p(x) = 1/(1 + e^{Af(x)}) \tag{7}$$

**(4) Random Forest (RF)**

RF is essentially an integrated algorithm consisting of substantial decision trees, and its brief process is as follows. Draw $k$ samples with $m$ features from the training set $D$ based on the bootstrap method and choose $n$ $(n \leq m)$ features randomly from each sample to form the feature space. Then, establish $k$ decision tree models for $k$ samples. Growing with the optimal feature as much as possible for the decision trees, the $k$ classification results can be obtained as $\{h_1(X), h_2(X), \ldots, h_k(X)\}$. The final classification result is determined by winning a majority vote or taking the mean value of $k$ results.

Given the fact that the forest is constituted of multiple differentiated decision trees, the performance of the whole model improves as the classification accuracy of each tree increases. Here, the importance of each affecting factor is measured as the percentage of the average reduction in the Gini coefficient of the factor in the sum of that of all factors (Equation (8)):

$$I_i = \frac{\sum_{h=1}^{t}\sum_{j=1}^{N} D_{Gihj}}{\sum_{i=1}^{m}\sum_{h}^{t}\sum_{j=1}^{N} D_{Gihj}} \tag{8}$$

where $m$、$t$、$N$ are the total number of affecting factors, decision trees, and nodes in a single tree, respectively, $D_{Gihj}$ is the reduction in the Gini coefficient of affecting factor $i$ at the node $j$ of tree $h$, and $I_i$ is the importance of affecting factor $i$.

- **Section 2.3.1.**

✧ **The Introduction of the OPTICS Algorithm**

Spatial clustering analysis aims to divide the spatial data set into several different classes of clusters and maximize the similarity between the same classes and the difference between various classes. A density-based algorithm [3], like a spatial clustering algorithm, has the advantage of avoiding the interference of noise and discovering arbitrarily shaped clusters, which is more applicable to the non-uniformly distributed data such as landslides. The core idea is to calculate the density according to the minimum number of other points (namely $MinPts$) in the neighborhood (with $\varepsilon$ as the range) where the target point is located. Points that are not identified as part of any cluster will be labeled as noise.

The Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm is a typical density-based clustering algorithm. The principle of it can be summarized as follows. Start from a selected core point and continuously expand to the density-reachable region so as to obtain a maximum region containing the core and border point, where any two points in the region are density-connected. However, it is sensitive to the input parameters of $\varepsilon$ and hence only appropriate in the case of having a very clear search distance. That aside, it cannot cluster data sets well with large differences in densities.

The Ordering Point to Identify the Cluster Structure (OPTICS) algorithm, as an improved version of the DBSCAN algorithm, is similar in its idea but differs in that it prioritizes the search for high density by selecting a finite number of neighborhood parameters $\varepsilon_i (0 \leq \varepsilon_i \leq \varepsilon)$ and uses an ordered list of objects for clustering. In short, it

uses the distance between neighboring points to create an accessibility plot and then separates the clusters of different densities from the noise based on the plot, ultimately achieving class clustering. This method is better adapted to the spatial heterogeneity feature of data distribution and improves on the shortcomings of the DBSCAN algorithm. See the literature for the specific algorithm principle [4].

✧ **The Procedure for Identifying the Clustering Attribute Factor**

**Step 1:** Input the search distance and the minimum number of points and obtain the clustering result of the hazard points, depending on the OPTICS algorithm (Figure 4a);

**Step 2:** Construct Thiessen polygons according to each point and assign the clustering attribute of each point to the corresponding Thiessen polygon (Figure 4b);

**Step 3:** Transfer the vector layer of the Thiessen polygon to the raster as the clustering attribute factor and input it into the machine learning models with the other 14 affecting factors in Table 1 for training;

**Step 4:** Collate the training accuracy of each model under different combinations of parameters and confirm the optimal scheme (Figure 3).

- **Section 3.1. The Chosen Basis of the Affecting Factors and the Corresponding Influence Rationale**

The topographic factors, as the free face condition-controlling slopes, determine the development and distribution of slopes to a large extent. Elevation is one of the important topographic breeding factors. Landslides mostly occur in the middle of mountains and rarely at the top due to the long-term weathering effect that has hardened the rocks on top of mountains [5]. This is corroborated by the fact that fewer hazards are at higher elevations in the western part of Bijie (Figure 6a). The slope directly determines the stress distribution of slopes. Generally, the steeper the slope, the more unfavorable it is to the stability of the slope, and the greater the possibility of a landslide. Various aspects have different solar radiation intensities, resulting in diverse water evaporation, vegetation cover, and slope erosion conditions. Therefore, shady slopes are more prone to hazards than sunny slopes. The profile and plan curvature are critical parameters for characterizing the complexity of the terrain. For the profile curvature, a negative value indicates the surface is upwardly convex at that cell, while a positive value is upwardly concave[1]. A value of zero indicates the surface is flat. The concavity and convexity of the plan curvature is expressed in the opposite manner.

The rocks along the fault zone are more fragmented and severely weathered, contributing to hazard forming. Research has shown that landslides usually occur near the vicinity of dense fracture structures, and more earthquake landslides tend to happen at the place closer to the fault [6]. As an important internal factor affecting landslide development, the lithology of the strata is closely related to the destabilization mode of the slope, and the lithology of Bijie is mainly sedimentary rocks. The soil type effectively reflects the background characteristics of the geological conditions, and there are significant differences in hazard susceptibility and disturbance resistance between soil types.

In terms of hydrology, heavy rainfall is the main causal factor triggering landslides.

[1] https://desktop.arcgis.com/zh-cn/arcmap/10.6/tools/3d-analyst-toolbox/curvature.htm

Different rainfall patterns and mounts have distinct induction mechanisms with the disaster and tend to cause different degrees of damage [7]. During the evolution of gullies, the downward erosion of flowing water produced by river scouring, erosion, and hollowing on the bank can soften the geotechnical body and significantly reduce its strength. Thus, we selected the annual average rainfall, flow accumulation, and distance to the river as affecting factors.

Land use reflects the exploitation state of land resources, and an unreasonable land use layout indirectly contributes to landslides. Vegetation roots can remarkably improve soil resistance to scouring and erosion, thereby reducing the hazard risk, and the Normalized Difference Vegetation Index (NDVI) is a vegetation index commonly used to quantify the extent of surface vegetation cover. As a road destroys the original geological environment of a slope through excavation, the place closer to a road is more likely to be hit by a landslide.

- **Section 3.1. Spearman's Correlation and Distance Correlation Analysis of Variables**

Spearman's rank correlation coefficient is appropriate for both continuous and discrete ordinal variables [8]. Meanwhile, it is insensitive to outliers and may be effective in some cases when the Pearson correlation coefficient is inapplicable. Therefore, we further detected linear relationships by calculating the Spearman's correlation coefficient matrix of 15 affecting factors and the output variable (Figure S1). The output variable represents whether the landslide occurred or not at the sample point, with 1 labeled as occurrence and 0 as no occurrence. The coefficient values between all variables were less than 0.7, indicating that there was no significant linear relationship.

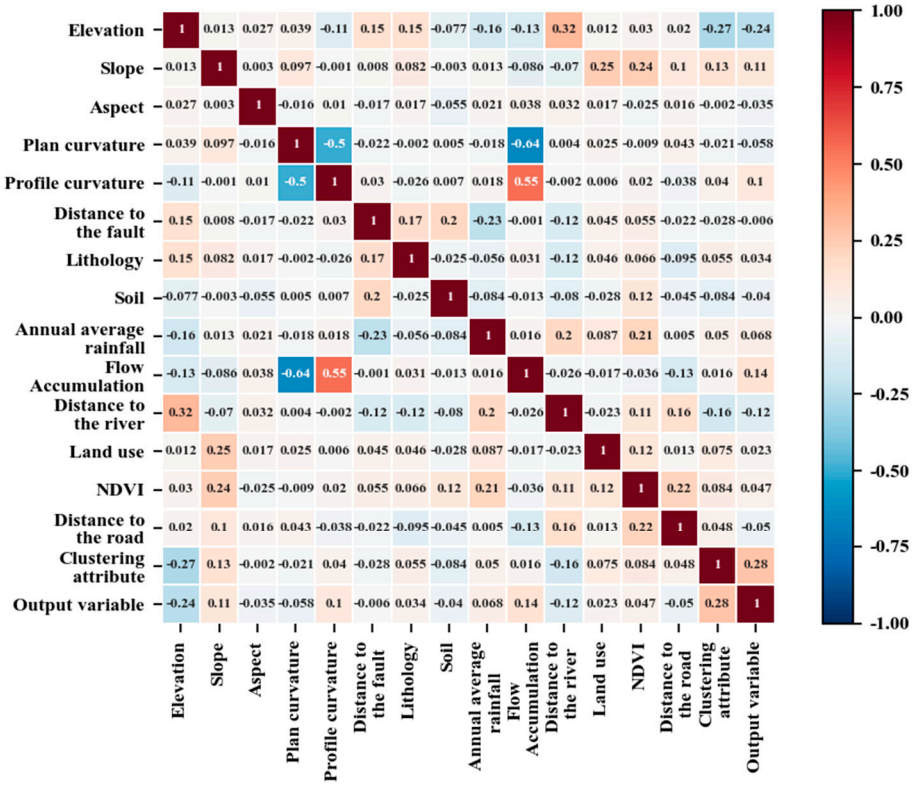


**Figure S1.** Spearman's correlation coefficient matrix of affecting factors and output variable.

Distance correlation can reflect both linear and nonlinear associations, which overcomes the disadvantage of the Pearson correlation [9]. Similarly, all variables are used to calculate the distance correlation to further detect other possible types of relationships or correlations (Figure S2). The coefficient values lay between 0 and 1, with higher values indicating stronger correlations. The results in the figure also show no significant linear or nonlinear association between the input and output variables.

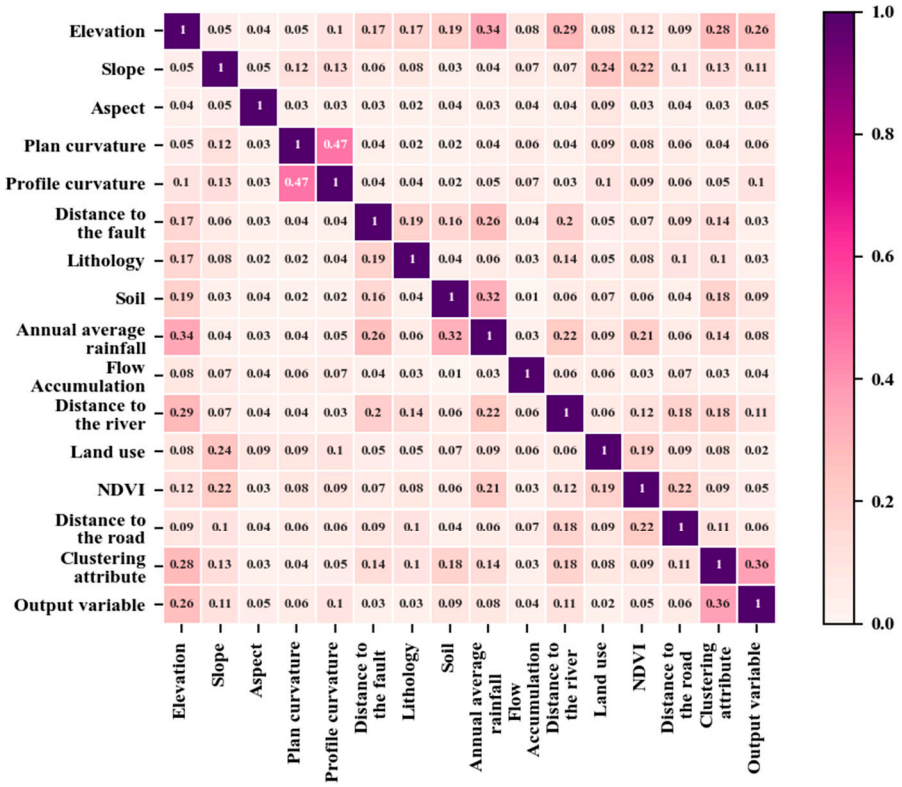


**Figure S2.** Distance correlation matrix of affecting factors and output variable.

## References:


1. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **1999**, *10*, 61-74.
2. Shi, J.; Zhang, J.; Shen, C. Construct and evaluate the classification models of six types of geological hazards in Bijie city, Guizhou province, China. *Natural Hazards and Earth System Sciences Discussions* **2020**, 1-28.
3. Ester, M.; Kriegel, H.; Sander, J.; Xu, X., **1996**; p 226-231.
4. Ankerst, M.; Breunig, M. M.; Kriegel, H.; Sander, J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record* **1999**, *28*, 49-60.
5. Liu, C.; Li, W.; Wu, H.; Lu, P.; Sang, K.; Sun, W.; Chen, W.; Hong, Y.; Li, R. Susceptibility evaluation and mapping of China’s landslides based on multi-source data. *Nat. Hazards* **2013**, *69*, 1477-1495.
6. Fan, X.; Scaringi, G.; Korup, O.; West, A. J.; Westen, C. J.; Tanyas, H.; Hovius, N.; Hales, T. C.; Jibson, R. W.; Allstadt, K. E.et al. Earthquake‐Induced Chains of Geologic Hazards: Patterns, Mechanisms, and Impacts. *Rev. Geophys.* **2019**, *57*, 421-503.
7. Bai, S.; Wang, J.; Thiebes, B.; Cheng, C.; Yang, Y. Analysis of the relationship of landslide occurrence with rainfall: a case study of Wudu County, China. *Arab. J. Geosci.* **2014**, *7*, 1277-1285.
8. Lehman, A. *JMP for basic univariate and multivariate statistics: a step-by-step guide*; SAS Institute, 2005.
9. Székely, G. J.; Rizzo, M. L.; Bakirov, N. K. Measuring and testing dependence by correlation of distances. *The annals of statistics* **2007**, *35*, 2769-2794.