SLOPE STABILITY THROUGH PREDICTABILITY USING MACHINE LEARNING ON SYNTHETIC DATA

Data de aceite: 03/07/2023

Tallys Celso Mineiro

Carlos Rodrigues Pontes

ABSTRACT: The study of slope stability benefits the society in several aspects, whether in the context of road infrastructure, urban slopes or even artificial slopes. Several methodologies can be used to analyze the stability of a slope, usually finite element computational tools are used for such analysis from simple to complex study cases. The predictability of slope stability through a database using Machine Learning methodology is an alternative of initial analysis that promotes an early understanding of the needs in synthetic pre-defined study cases in order to obtain predictable conditions for the stability of a slope, defining its constituent soil and slope geometry. The results are effective since the goodness of fit is 0.95 for Linear Regression Model and 0.94 for Random Forest Regression Model. Thus, it can be verified that the use of Machine Learning in this type of study helps in the decision-making process, as well as being effective in pre-dimensioning for future projects.

KEYWORDS: Slope stability, Machine Learning, predictability, programming.

1 I INTRODUCTION

The process of analyzing and predicting slope stability, according to Lin (2018), is extremely important for geotechnical engineering. This is due to the fact that to reduce and prevent accidents caused by slope ruptures, whether artificial or not, an exquisite study of slope stability analysis and structural stabilization processes is required. Certainly, the structural complexity of slopes makes the study of predictability in this field of study challenging. Therefore, computer simulations are very categorical tools for thriving in analysis processes such as this one.

In these last decades, the application of data mining and computational predictability tools has grown voraciously due to the adaptability capacity of establishing non-linear relationships between input and output

data of statistical methods (Bui, 2019). Although Lin (2018) points out that, depending on the regression or classification method used, these may not be enough to solve the problem.

Many regression methods implemented in programming have been developed, among these two that are interesting for the study of slope stability are Linear Regression, due to the similarity with the model proposed by Mohr-Coulomb, and Random Forest Regression, which has the ability to unify statistical regressivity and decisive classification to propose a selectivity of correlative data.

Thus, the present study aims to apply the regression models implemented in Machine Learning to analyze the data, provided by them, to the parameters present in a synthetic database reproduced and generated by the GeoStudio SLOPE/W software in order to compare and validate if the models adopted will be able to adapt to the real database in order to enable better conditions for dimensioning slope projects and facilitate decision-making inherent to questions on the subject.

2 | METHODOLOGY

2.1 Slope geometry and database

The case studies analyzed here range from simple to complex slopes. In these, several variables can interfere with regard to the stability of the structure, ranging from the slope of the slope, its height and its constituent materials. Figure 1 exemplifies an outline of the slope geometry that can be different on the hight.

The synthetic database is composed of 240 simulated cases with slopes varying from 3 to 6 meters in height (H) with slope conditions (f) varying between 45°, 60° and 75°. The slopes are composed of only one layer of soil to simplify the analysis and modeling process. According to the modeling methodology of Silva et al (2013), the cohesions (C) vary between 0, 2, 4, 6 and 8 KPa and friction angles (Φ) vary among 25°, 30°, 35° and 40°. Regarding the specific weight (Y) of the constituent soils, they vary from 13 kN/m³, to 20 kN/m³.

Therefore, using combinations of these parameters, 40 synthetic soil types can be simulated and with the 6 types of geometric combinations, a result of 240 data from simulated case studies is obtained as mentioned.

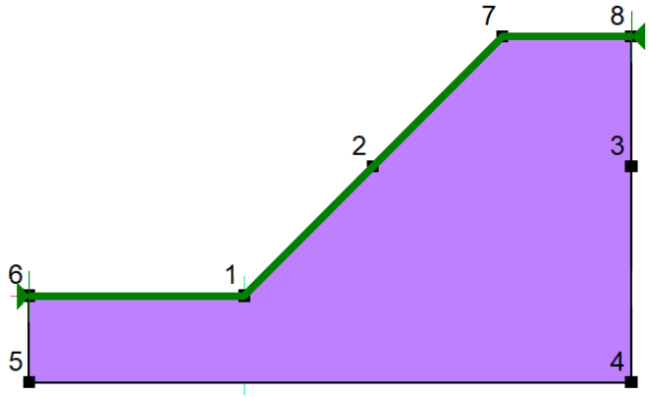

Figure 1. Geometric conformation of slopes.

Concerning the database, it is formed by the variables described above, these data being implemented in the GeoStudio software and obtaining their respective safety factors (SF) according to the analysis model proposed by Fellenius. Table 1 shows the database with the case studies, in which the index shown corresponds to the height of the slope and its slope with the horizontal, respectively. Therefore, SF3 A45, indicates the factor of safety simulated in a slope of 3 meters of height with inclination of 45°.

When it comes to safety coefficients to evaluate the stability conditions or susceptibility to slope failure, according to the balance of acting forces and resistive forces, the SF index can be classified according to Carvalho (1991) as unstable (SF<1), condition stability limit associated with imminent failure (SF=1) and stable (SF>1).

SOIL	COHESION	FRICTION ANGLE	SPECIFIC WEIGHT	SF3 A45	SF6 A45	SF3 A60	SF6 A60	SF3 A75	SF6 A75
Soil 1	0	25	13	0.66	0.494	0.382	0.301	0.268	0.201
Soil 2	0	30	13	0.818	0.612	0.474	0.372	0.331	0.249
Soil 3	0	35	13	0.992	0.742	0.574	0.451	0.402	0.302
Soil 4	0	40	13	1.188	0.89	0.688	0.541	0.481	0.362
Soil 5	2	25	13	1.029	0.81	0.818	0.603	0.64	0.469
Soil 6	2	30	13	1.187	0.944	0.91	0.687	0.708	0.536
Soil 7	2	35	13	1.361	1.092	1.01	0.779	0.784	0.61
Soil 8	2	40	13	1.557	1.26	1.124	0.884	0.868	0.694
Soil 9	3	25	15	1.14	0.884	0.949	0.667	0.747	0.525
Soil 10	3	30	15	1.297	1.018	1.04	0.763	0.815	0.592
Soil 11	3	35	15	1.471	1.166	1.141	0.855	0.89	0.666
Soil 12	3	40	15	1.668	1.334	1.255	0.96	0.975	0.75

Soil 13	4	25	15	1.3	0.991	1.085	0.749	0.873	0.606
Soil 14	4	30	15	1.457	1.125	1.218	0.85	0.968	0.673
Soil 15	4	35	15	1.631	1.273	1.33	0.96	1.044	0.747
Soil 16	4	40	15	1.828	1.441	1.444	1.069	1.129	0.831
Soil 17	4	25	17	1.225	0.941	1.023	0.711	0.817	0.568
Soil 18	4	30	17	1.382	1.075	1.14	0.811	0.896	0.635
Soil 19	4	35	17	1.556	1.223	1.241	0.913	0.971	0.709
Soil 20	4	40	17	1.753	1.39	1.355	1.017	1.056	0.793
Soil 21	6	25	19	1.418	1.059	1.182	0.809	0.99	0.657
Soil 22	6	30	19	1.575	1.204	1.315	0.91	1.086	0.733
Soil 23	6	35	19	1.749	1.352	1.462	1.021	1.192	0.807
Soil 24	6	40	19	1.946	1.52	1.583	1.146	1.282	0.891
Soil 25	6	25	18	1.46	1.08	1.216	0.831	0.959	0.674
Soil 26	6	30	18	1.617	1.232	1.349	0.932	1.055	0.754
Soil 27	6	35	18	1.791	1.38	1.497	1.043	1.157	0.828
Soil 28	6	40	18	1.988	1.548	1.633	1.168	1.242	0.912
Soil 29	2	25	19	0.913	0.732	0.681	0.524	0.528	0.409
Soil 30	2	30	19	1.07	0.866	0.772	0.607	0.595	0.477
Soil 31	2	35	19	1.244	1.014	0.873	0.7	0.668	0.548
Soil 32	2	40	19	1.441	1.182	0.987	0.804	0.747	0.608
Soil 33	8	25	17	1.789	1.245	1.487	0.988	1.227	0.804
Soil 34	8	30	17	1.947	1.407	1.62	1.099	1.328	0.889
Soil 35	8	35	17	2.121	1.586	1.767	1.211	1.434	0.983
Soil 36	8	40	17	2.317	1.769	1.934	1.337	1.554	1.079
Soil 37	8	25	20	1.62	1.16	1.348	0.912	1.107	0.737
Soil 38	8	30	20	1.777	1.322	1.481	1.013	1.203	0.822
Soil 39	8	35	20	1.951	1.488	1.628	1.125	1.309	0.909
Soil 40	8	40	20	2.148	1.655	1.794	1.25	1.429	0.993

Table 1. Database.

2.2 Machine Learning process

Initially, the analysis algorithm reads a base file, through the Pandas library (pd), with the case study database, that is, the table with the variables inherent to the slope conditions and their respective minimum safety factor. Therefore, through the use of the Seaborn (sns)

and Matplotlib.pyplot (plt) libraries, heat and correlation graphs can be plotted between all the variables in the database, in order to get an idea of which variables have the highest rate of interference in the safety factor.

The data to be analyzed by the Machine Learning model must be randomly subdivided into training and test data, since, for the most part, artificial intelligences need to be trained or go through a simulation state to effectively be tested and verified with the other data. Therefore, in the vast majority of cases, the Sklearn library is used, as it is one of the most important Artificial Intelligence libraries in Python. The data splitting process is done with the aid of *sklearn.model_model* and *its train_test_split* function, which splits the database into input data to simulate predictability (usually called x) and result data that is expected to be achieved (usually called y), in this case study being the slope safety factor.

Resolutely, it is preferred that the training data be between 20 and 35% of the general data, both for x and for y. This functionality is determined by *train_test_split* through *test_size* and additionally randomly selected through *random_state*. Thus, in this work, the value of 35% was adopted for such division.

The training must be simulated using regressions, which in this study will use the linear regression imported into the code by the LinearRegression resource from the *sklearn.linear_model* library and the decision tree method imported into the code by the RandomForestRegressor resource from the Sklearn.ensemble library. The f*it* function fits the data for training in the Machine Learn methodology in this set of libraries, while the *predict* function performs the predictions with the test data of the set x.

Finally, to evaluate the functionality of the predictability, the *metrics.r2_score* function of the Sklearn library will be used to calculate the R^2 between the test set y and the regressions used separately. Figure 2 shows a schematic flowchart of the algorithm calculation process.

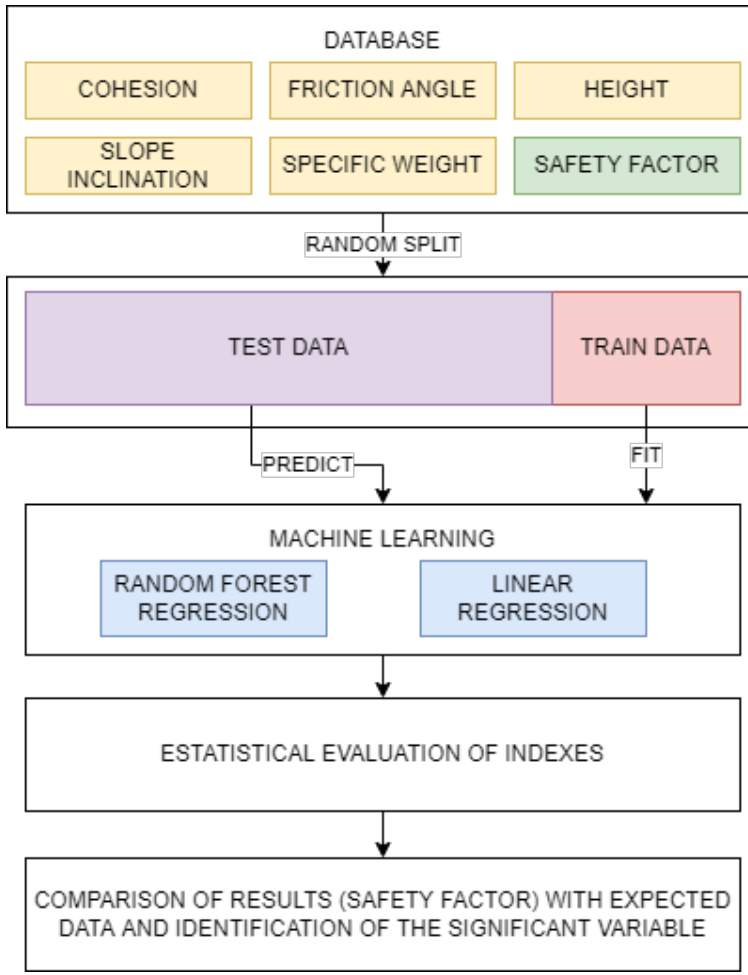

Figure 2. Statistical code analysis procedure.

2.2.1 Linear Regression

The Linear Regression model, according to Bui (2019), tries to correlate a linearized dependence tendency between a predicted variable and a specific parameter among the information in the database, thus being normally presented by Equation 1:

$$y = \alpha + \beta x + k \quad (1)$$

Wherein y is an independent variable and x is a dependent variable, α and β being structural parameters of linearization. The parameter k is relatively described as being a source of random error; however, in order to improve predictability, the analyzes are always associated with a normal distribution of error (Seber, 2012).

2.2.2 *Random Forest Regression*

The Random Forest Regression model proposed by Breiman (2001), as it is a supervised learning algorithm, has both decision-making features and regressive statistical methods. Suggestively by its name, the method executes by elaborating a combination tree of other decision trees mostly trained by the bagging method.

Additionally, the proposed method is effective in terms of defining the importance of the database parameter for the predictability of the data under study. Therefore, this is possible due to the fact that the classifier is composed of generated trees and the new dataset is discriminated by the classifier. So, the classification result is dependent on the number of votes per classification tree.

3 | RESULTS

After all the data processing in the database, a predictability analysis graph can be plotted between the synthetic data, that is, the test data and the related data already predicted. This graph can be seen in Figure 3.

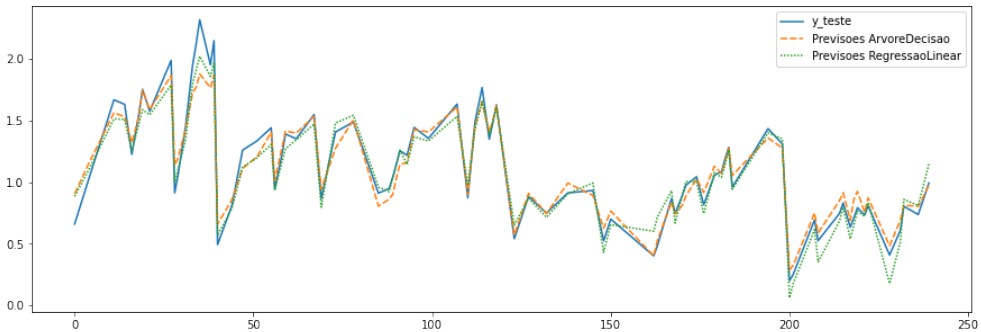

Figure 3. Comparison between test data and predictions by the models.

In observance, the models fit the problem very well. Additionally, the linear regression model tends to have more conservative values compared to the test data, while the predictions referring to the random forest regression model tend to have a slightly higher standard deviation. In statistical terms, the R^2 factor calculated for the models are 0.95 and 0.94 for linear regression and random forest regression, respectively, confirming what was observed in the graph.

As far as predictability is concerned, the random tree regression model manages to describe the importance for such a process, as it is a decisive classification model as commented. However, it is significant to remark that the degree of parameter importance does not correspond to the theoretical calculation model to determine the respective safety indexes, but to the forecasting process. Thus, Figure 4 shows the percentage of the degree of impact on predictability.

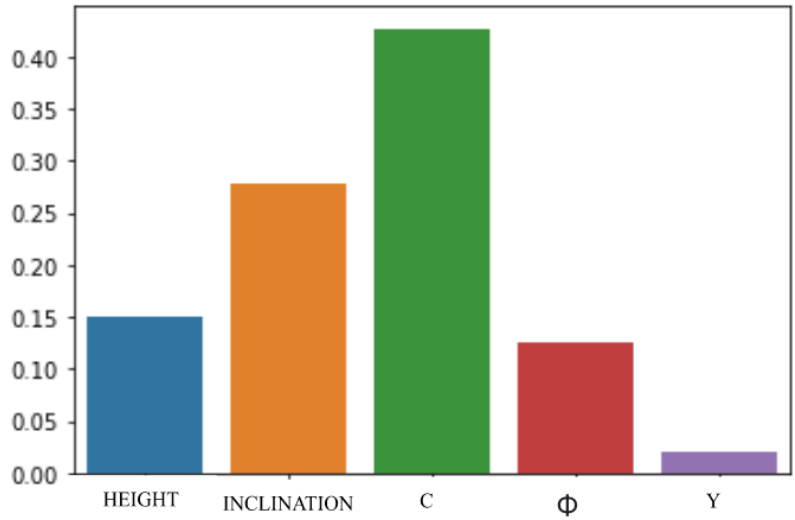

Figure 4. Importance of each parameter in the random forest model.

In a way, analyzing the parameters of cohesion and friction angle, intrinsic to the soil, for the predicted data, it can be inferred that the prediction models provide what is elucidated in the Mohr-Coulomb analysis theory. This fact can be seen in Figure 5, while the values increase the corresponding shear index, indirectly represented by the safety factor, representatively also increases. In this way, the reliability of being able to make pre-dimensions from predicted data is relevant for an initial decision making. Figure 6, in turn, shows that the slope height parameters and its respective slope, belonging to the slope geometry and conformation, are inversely proportional to its stability condition.

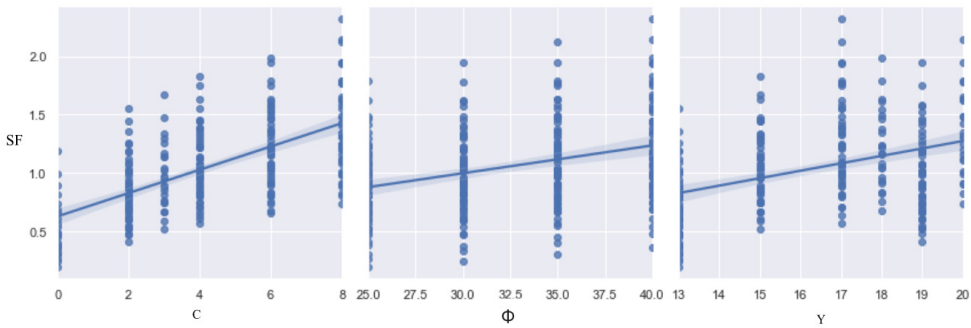

Figure 5. Relationship between stability and soil parameters.

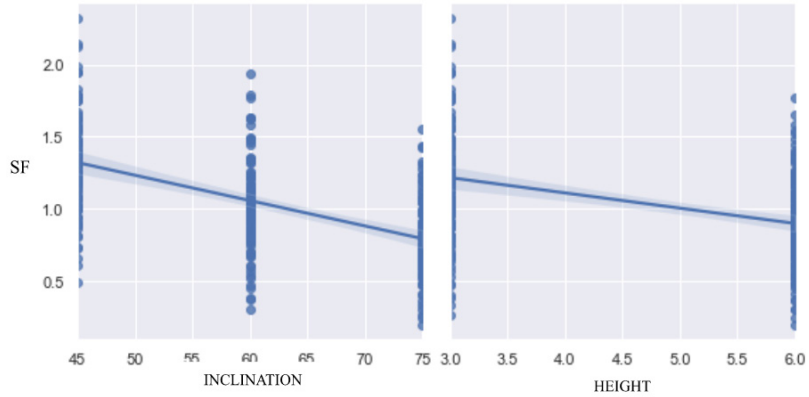

Figure 6. Relationship between stability and slope geometry.

Analyzing the database in terms of qualitative classification, in Figure 7 it can be seen that more data from unstable situations were simulated, this is due to the fact that the prediction was made for data with notable angulations and heights resulting in a reduction of the stability condition. As a result, the data training process prioritizes a reduction in most predictions. This can be noticed in some parts of the graph in Figure 3, to which most of the data are more conservative.

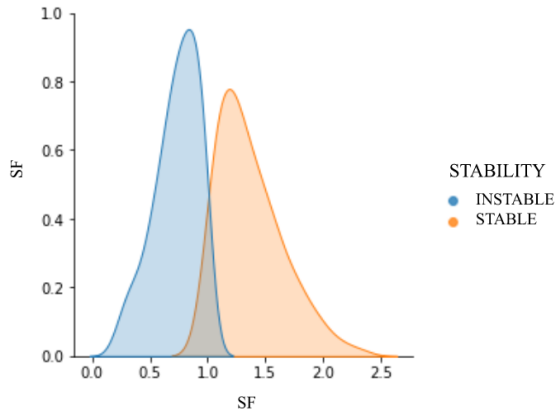

Figure 7. Data distribution diagram regarding classification.

4 | CONCLUSION

Studies such as this one are fundamental for the development of technologies for decision making or projects involving the scope of slope stability. As seen, the use of computational tools implemented with Artificial Intelligence are widely applicable and of considerable reliability; therefore, optimizing data processing and enabling greater interactivity between the real and predicted in order to solve problems like this.

Taking into account the statistical data, the models are effective and can be adapted to any database. Remembering that the choice of the model must be very well grounded in order to have a fit that best represents the real phenomena.

It is recommended for future work to analyze the predictability for other calculation models and observe the interactivity between them, apply the methodologies to a real data base or verify other slope conformation conditions.

REFERENCES

Bui, D. T.; Moayedi, H.; Gör, M.; Jaafari, A.; Foong, L. K. (2019) Predicting Slope Stability Failure through Machine Learning Paradigms. ***International Journal of Geo-Information***.

Breiman, L. (2001) Random forests. ***Machine Learning***, v. 45, n. 1, p. 5-32.

Carvalho, P. A. S. (1991) ***Manual de geotecnia: taludes de rodovia: orientação para diagnóstico e soluções de seus problemas***. São Paulo: IPT.

Lin, Y.; Zhou, K.; Li, J. (2018) Prediction of Slope Stability Using Four Supervised Learning Methods. ***IEEE Access***, v. 6, p. 31169-31179.

Seber, G. A.; Lee, A. J. (2012) **Linear Regression Analysis**. John Wiley & Sons, Hoboken, NJ, USA, v. 329.

Silva, E. L.; Gomes, R. A.; Guimarães, R. F.; Carvalho Júnior, O. A. (2013) Emprego de modelo de suscetibilidade a escorregamentos rasos para gestão de risco de desastres no município de Vitória-ES. ***Sociedade & Natureza***, v. 25, p. 119-132.