Ranking Features for Landslide Prediction

Alessio Poloni [1]

Abstract

Landslide prediction can be carried out thanks to knowledge in the field of machine learning. In this study, a Random Forest Classifier algorithm is applied to a dataset containing eight landslide prediction variables. The aim is to determine which of these variables are the most important for the prediction. Two different methods, Mean Decrease in Impurity and Feature Permutation, are applied and compared to assess the importance of the variables. The results are in line with the literature and show that three variables are significantly more important. Moreover, a third feature importance method intrinsic to another algorithm confirmed the robustness of the model.

1. Introduction

In the field of environmental science and physical geography, landslides are a very important and widely studied phenomenon. In Switzerland, this natural hazard is particularly widespread in mountainous regions. Some methods commonly used to study landslides include LiDAR monitoring, geophysical surveys and numerical modelling, in order to prevent damages at infrastructures, high costs and fatalities. Moreover, it is also possible to use artificial intelligence, and more specifically Machine Learning, to address landslides studies (Riese, 2021).

Random Forest is a supervised Machine Learning algorithm based on an ensemble of decision trees (Tonini et al., 2020), developed by Breiman (2001). It can be used either as a classification method or for solving regression problems (Beucler, 2022). There are several reasons for choosing to apply Random Forest in this case. From the theory, it's usually more accurate in terms of errors and more robust, it has more representation power than individual models and it can quantify uncertainties (Beucler, 2022). Moreover, Trigila et al. (2013) reported that this algorithm is often used in landslide prediction problems. Riese (2021) reported from the literature, that results in these kinds of studies were generally similar or better with Random Forest compared to other algorithms.

In this study, a classification conducted with Random Forest is applied to a dataset containing eight landslides predictor variables (independent variables) and the landslide occurrence or absence (dependent variable). The aim is to rank the features importance for landslide prediction. In other words, is to classify the eight environmental factors based on their relative relevance in landslide occurrence. This is done with two different methods, Mean Decrease in Impurity and Feature Permutation, in order to compare them. Moreover, an evaluation of feature importance for another algorithm (Gradient Boosting Classifier) is done to assure that results with the two previous computations are reliable. Generally, understanding which variables are the most important can help determine which models are most robust. In the case of Random Forest, that is a black box model, techniques like feature importance allow to understand a little more about the algorithm. In the next section, the methodology is described, also with details about the hyperparameters tuning. Then, bar plots will help to visualize the results.

2. Data and Methodology

2.1. Data

The study area is all the Canton of Vaud (in western Switzerland). The input rasters and landslides dataset were provided by Dr. Marj Tonini (IDyST) and were processed and elab-

[1] Institute of Earth Surface Dynamics (IDyST), University of Lausanne, Lausanne, Switzerland. Correspondence to: Alessio Poloni <alessio.poloni@unil.ch>.

orated in the context of the Master Thesis of Julien Riese (Riese, 2021). The input dataset contains 5188 samples (lines), each of them having coordinates (x and y). Half of the dataset is composed by landslides observations, whereas the other half is composed by landslides absence cases. This can overcome a potential problem of overestimation that happens if the lower class contains more observations than the other one. The dataset is composed from eight independent variables (columns), that are landslides predictor variables: Digital Elevation Model, Slope, Plan Curvature, Profile Curvature, Distance from Roads, Land Cover, Topographic Wetness Index, Lithology classes (more details are in the code). Most are numerical variables, whereas Land Cover and Lithology classes are categorical. The choice of these variables was made by Riese (2021) after literature researches. Finally, the dependent (target) variable is the landslide occurrence (1) or absence (0) for each of the samples (also determined by Riese, 2021).

2.2. Methodology

The algorithm used is Random Forest Classifier, as mentioned above. The methodology workflow consists of some main steps, all developed in Python environment (with scikit-learn library) using Google Colaboratory interface (link to the code at the end of the paper).

(1) First of all, the dataset is split into train (67%) and test (33%) sets. (2) Then, an hyperparameters research is made to improve the algorithm performance. The most important for Random Forest are the number of decision trees (n-estimators) and the number of randomly sampled variables as candidates for each split (max-features) (Tonini et al., 2020).

(3) The features importance is evaluated for Random Forest. For this, two different methods are used: Mean Decrease in Impurity (intrinsic to Random Forest) and Feature Permutation (more generalist method). MDI evaluates which variable best divides the dataset into the two values that the dependent variable takes. The more this split take on a value of 50%, the more important the variable is. FP degrade one variable at a time and sees how much worse the prediction is. The largest variable is the one with the greatest worsening. A comparison of the two results is done to see eventual differences.

(4) Finally, it's insightful to evaluate the features importance for another algorithm, to interpret correctly the results and avoid biased interpretations. For this, Gradient Boosting Classifier has been chosen, because it can perform well for landslides prediction problems, and feature importance is evaluated with Permutation Importance method (intrinsic to Gradient Boosting), that is the most appropriate method (A. Trucchia, personal communication, 23.5.22).

3. Results

3.1. Random Forest performance and hyperparameters research

For the initial run of the Random Forest Classifier, the accuracy score is 83.2%, and the mean squared error is 0.168. Then, the hyperparameters research is done to improve the model performance. In Table 1 are presented the best parameters found with GridSearchCV method. Generally, in classification problems, max-features is set around the square root of the number of predictors (Tonini et al., 2020).

	n-estimators	max-features
Range of search	50, 100, 150, 200, 250, 300, 350, 400, 450, 500	2, 3, 4
Hyperparam chosen	300	2

Table 1. Range of hyperparameters research and hyperparameters chosen.

The run of Random Forest Classifier with these improved parameters generates an accuracy score of 83.8%, and a mean squared error of 0.162. Looking at Table 2, it is possible to say that there has been a little improvement: the accuracy and true positives/true negatives increased, whereas the error and false positives/false negatives decreased.

	Before tuning	After tuning
Accuracy score	83.2%	83.8%
Mean squared error	0.168	0.162
Confusion matrix	[680 164] [123 746]	[684 160] [118 751]

Table 2. Performance metrics evaluated and confusion matrix.

3.2. Feature importance for Random Forest

The feature importance computed for Random Forest with Mean Decrease in Impurity (intrinsic method) is represented in Figure 1. Black lines represent the error bars.

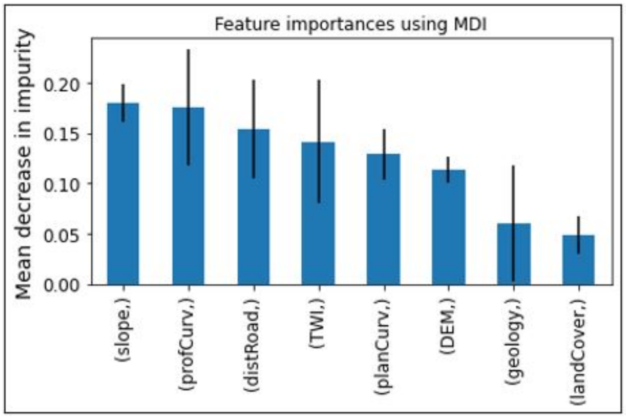

Figure 1. Feature importance with Mean Decrease in Impurity approach for RandomForestClassifier.

On the other hand, the feature importance computed for Random Forest with Feature Permutation (generalist method) is represented in Figure 2.

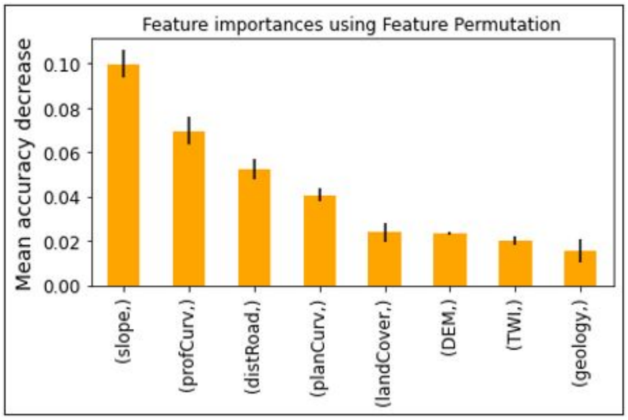

Figure 2. Feature importance with Feature Permutation approach for RandomForestClassifier.

3.3. Feature importance for Gradient Boosting

The feature importance computed for Gradient Boosting with Permutation Importance (intrinsic method) is represented in Figure 3. This is useful to check results obtained above.

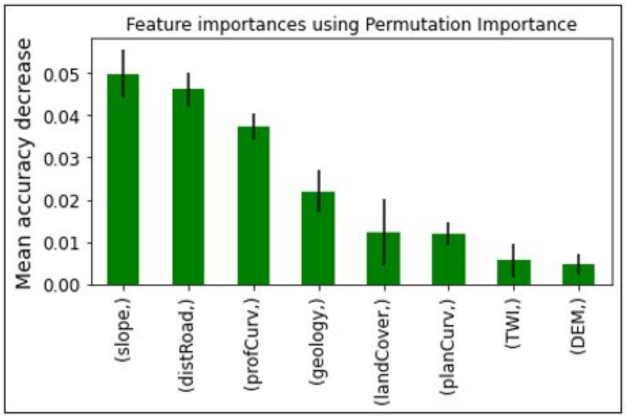

Figure 3. Feature importance with Permutation Importance approach for GradientBoostingClassifier.

4. Discussion

For the hyperparameters research, it can be observed that there was an improvement, but it was somehow limited. The accuracy score is a metric used to see the model's accuracy in classification. Generally, an acceptable model should have a value greater than 80%. The model implemented in this study is therefore good. The mean squared error alone is not very significant, as it is not an absolute index of the reliability of the estimate made, but depends on the range of data variation. However, it is useful for comparing a model implemented with different sets of hyperparameters, as in this case. The confusion matrix is useful both on its own, to see how many true positives and true negatives have been identified, but also for comparison. These three metrics together showed that the model has improved, but not so much (accuracy score +0.6%) This could be caused by the already well-prepared dataset, specifically designed by Riese (2021) to be treated with Random Forest. In any case, the results confirm that the hyperparameters to be changed reported in the literature (Tonini et al., 2020) are sufficient to generate an improvement.

For the feature importance, in Figure 1 (the method intrinsic to Random Forest) the predictive variables with a $mean \geq 0.15$ are Slope, Profile Curvature and Distance from Roads. The importance of the variables decreases gradually and the relative importance of the first three variables compared to the other numerical ones is moderate.

This method seems to consider the two categorical variables as the least important. Even in Figure 2 (the more generalist method), the variables with the highest importance are the same three, in the same order. In this case, the importance of Slope is significantly higher than for the other variables. Moreover, the decrease in importance is more pronounced among the first four variables, while it stabilises in the last four.

The doubt is that the results obtained relate only to the specific Random Forest computation. This is why Figure 3 (the method intrinsic to Gradient Boosting) serves as a comparison as the algorithm is different. Again, the most important variables are the same three. Slope always remains the main factor. However, for Gradient Boosting, Distance from Roads is more important than Profile Curvature. Moreover, the difference in relative importance between the first three variables and the others is much more pronounced than in the two previous methods.

It's possible to observe that all methods underlined the same three variables as the most important factors for landslide occurrence prediction. The fact that Slope, Profile Curvature and Distance from Roads were identified as the most important factors by either a method intrinsic to Random Forest, a more generalist method also applied to Random Forest, and a method intrinsic to a different algorithm, shows how robust the results are. These results are also confirmed by findings reported in the literature (as synthetised for example by Riese, 2021). However, there are differences between the figures concerning the relative importance of these three variables compared to the others. Generally, the MDI method should be more appropriate for Random Forest in particular, the FP being to generalist. But MDI seems to be less likely than FP and PI to omit a feature, and the amplitude of error bars is also larger for MDI. On the other hand, the computation of FP is more costly because this method degrades one variable at a time and re-do the Random Forest, whereas MDI revise each split of Random Forest without re-doing it each time.

5. Conclusions

In this study, a ranking of feature importance was made for eight predictive variables for landslides occurrence. Two methods computed for Random Forest Classifier were used, and another method for a different algorithm served as a control. Slope, Profile Curvature and Distance from Roads were selected as the most relevant features for landslides prediction. The three methods computed gave the same results, but with different relative importance between variables. So, the model implemented with Random Forest was robust and the results were in line with the literature. This study focused on a certain number of variables, but in future studies it would be useful to try to include new environmental variables for the prediction of landslides.

6. Other resources

Link to the dataset used: click here

Complete Python code: click here

References

[1] Beucler, T. (2022). Machine learning for environmental science – Lecture 3 [PowerPoint presentation]. Found in the environment Moodle UNIL: https://moodle.unil.ch/course/view.php?id=21543

[2] Breiman, L. (2001). Random forests. Machine Learning, 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324

[3] Riese, J. (2021). *Landslide susceptibility mapping in the canton of Vaud using random forest - A focus on the marginal effect of different predictive variables.* (Master degree, Université de Lausanne).

[4] Tonini, M., et al. (2020). A machine learning-based approach for wildfire susceptibility mapping. The case study of the Liguria region in Italy. Geosciences 10.3: 105.

[5] Trigila, A., Frattini, P., Casagli, N., Catani, F., Crosta, G., Esposito, C., Iadanza, C., Lagomarsino, D., Mugnozza, G. S., Segoni, S., Spizzichino, D., Tofani, V., & Lari, S. (2013). Landslide susceptibility mapping at national scale: The italian case study. In C. Margottini, P. Canuti, & K. Sassa (Eds.), Landslide science and practice: Volume 1: Landslide inventory and susceptibility and hazard zoning (pp. 287–295). Springer. https://doi.org/10.1007/978-3-642-31325-738