Landslide trigger classification using Support Vector Machine and Random forest.

Coding language: R

Machine learning models: SVM and Random Forest

Visualization: Tableau

Varun Sundaram

20th Feb 2019

Email @ varunsundaram@live.in

Landslide trigger classification using Support Vector Machine and Random forest.

Varun Sundaram

Abstract— **Landslide susceptibility mapping and modelling that deal directly with the forecast of landslides have been researched in the past. However, this study presents landslide trigger classification as a novel approach in determining areas that are potentially prone to landslides. Random forest (RF) is used on the multi-class category on the landslide dataset. Average performance was observed and when the SMOTE sampling was done to combat class imbalance before implementing RF, the performance was observed to be better than without. Kernelised SVM with four different kernels were tried out with 25 combinations of hyper-parameters. The kernels help define the hyperplane that is used for the classification problem. The hyper-parameters that brought out the optimal performance is with gamma value of 0.1 and cost function value of 100. The radial kernel outperformed the other kernels with a better recorded accuracy. Landslide trigger classification are used in drawing up warning zones prior to the occurrence of landslides by tracking the landslide triggers.**

Keywords: Landslide, Trigger classification, SMOTE, Support Vector Machine, Random Forest, Hyper parameter tuning.

I. INTRODUCTION

Landslide is a term used to describe mass displacement or movement of soil, rock and other debris, along a slope due to gravitational forces. There are many reasons for this to occur, changes in climate, human activities or geological triggers. Landslides are classified based on these triggers that cause them and the extent of damage caused. Studies have shown that landslides cause by different triggers tend to have different patterns. These triggers can be dynamic in nature, making it very difficult to analyze. Due to this reason, often they are not considered during the study of susceptibility of an area for landslides.

Analyzing the historical records of landslide could be key to understanding predicting landslides in the future. Therefore, study of landslide classification can be used in studies to understand and predict the occurrence of landslides and their impact on susceptible areas. Landslides are extremely hazardous disasters as they cause loss of lives, property and have major impact on economy of the area affected. Landslides often times than not have a multitude of causes for its occurrence. These causes may include geological, morphological, physical, and human [1] but just a single trigger [2] The trigger can be defined as an external stimulus that produces a very rapid response by generating a landslide. This landslide is caused as a result of increasing the stresses or by the minimizing of the strength of slope materials. The external stimuli can be grouped in the form of several categories such as earthquake, rainfall, volcanoes, rapid stream or storms. Some cases of landslide activities may not occur with an obvious attributable trigger. A group of combined causes, like chemical or physical weathering of materials result in the failure of the slope eventually leading to landslides [3].

Landslides are at times caused by human intervention, as in the cases of construction of dams, highways, mining, quarrying etc. These factors result in the alteration of slope grade, and thereby the morphology of the landscape. Taking a case study from the Eastern Himalayas tells us that landslides were caused as a result of recent settlements in the region accounting for causing damage to the livelihood of the local residents [4]. Some of the most common landslide triggers are rainfall and earthquakes those of which do not directly have any influence or interference from humans. Out of all the trigger reasons, most of the landslides occurring all over the world are driven by these two factors. Since the amount of loss including human lives, personal belongings and damage of infrastructure are involved in landslides were very high, the prediction of landslides plays an important role in saving people lives.

This is very crucial in areas where the landslide susceptibility is very high. The idea of this research is to classify the landslide triggers and identify the possibility of landslide occurrences when a particular trigger event is happening around a landslide susceptible area. This can be achieved by understanding the past landslide occurrences and the possible triggers that caused the landslide.

The data collected for this research involves the location and time parameters which is very vital in understanding the location aspect and trigger reason around the landslide area. Though, the main triggers were because of natural calamities, there are few man made mistakes and improper planning of infrastructure which has made the land surfaces to be more susceptible to landslides. Reasons like constant deforestation, mining activities and soil quarrying has made the land weaker and lose its grip to heavy rain and earthquakes and became more prone to erosion and landslides.

To give an example of what the landslide trigger classification can do, a city X is considered. According to the past records, when construction of a heavy infrastructure project took place in the months of September, October, and November in city X, landslide activities have occurred. This can be used as an indicator in the landslide occurrence.
This research aims to build a classification model with multiple algorithms which would understand and observe a pattern from the past landslide data and mark the exact trigger that has caused the landslide to happen. This study is divided into 2 main sections, which involves application of proven machine learning techniques in classification problems to precisely classify the landslide triggers. Support vector machines are one of the fundamental machine learning techniques available for classification problems. It provides for regression and classification solutions that are non-linear in nature. The input variables provided for the SVM model occupies a large dimension space with its inner product provided by kernel functions. Optimization in the form of hyper-parameters is involved for the best performance of the model [5].

II. RELATED WORK

Random forest (RF) works by making use of a group of decision trees [6] The selection of training samples is performed randomly with the help of bootstrap aggregation for the final prediction. Majority voting is used for deciding the output. The idea behind RF lies in the strength of the decision trees which is maintained by minimizing the correlation present among the trees in the forest.
The implementation of this algorithm needs the setting up of parameters (number of variables, number of trees). The significance of spatial prediction of landslides to occur in the future from the past and present landslide activities is discussed with the help of advanced techniques involving landslide susceptibility mapping. Training samples are first selected randomly for each individual tree by using bootstrap sampling. Next step lies in the selection of the best split of data which is randomly sampled into a subset of attributes by the tree inducer [7] This paper puts forth an application of machine learning techniques such as bagging and support vector machines along with random forest with their optimal parameters for the best sought out performance [8].

Support vector machine is used [9] in the study for the classification of land susceptibility by making use of linear separable case analysis with low-dimensional input space. The classification is performed over the landslide pixels and no-landslide pixels. The aim of the SVM model lies in finding the optimal separating hyper plane that can distinguish between the two classes, landslides, and no-landslides [10]. There have been papers [11] that have used confusion matrices and receiver operating characteristic (ROC) curve to compare performances among different functions and models. In this paper, McNemars test statistic was also used to explore the statistical significance among the differences in the different models that are selected. Multivariate (i.e., SVR, LR, and DT) approaches are found to outperform the bivariate methods (i.e., FR, SI and WOE) by an overall of 13%. Among the multivariate approaches, SVR technique performed with the highest accuracy, with FR outperforming the other techniques when it came to bivariate methods.

Some research studies [12] adopt clustering algorithms in classifying landslide susceptible areas from those that are not. [12] uses image classification along with clustering techniques. The most commonly used k-means clustering algorithm may account for failures in arranging the data to the target groups accordingly. This possibility of failure is influenced by the ability of the cluster centres in guessing different results. This

study uses the bacterial foraging algorithm (BFA) in resolving the image data for clustering problems. The second part of the clustering involves the constrained clustering, which is nothing but the technique that is used in times when there are only few label data available. The constrained BFA classification result was found to do better than the k-means and regular BFA classification.

Landslide conditioning factors such as altitude, slope, aspect, distance to road, distance to stream, distance to fault, rainfall, plan and profile curvature, and seismicity were used for the construction of hybrid models in prediction landslide susceptibility. Statistical index (WI) is used in creating the model [13] Bivariate statistics is chosen as the category under WI to compare each factor with the inventory map [14] What this means, is that all of the classes in conditioning factors are categorized in reference to their landslide density. Statistical correlation, map crossing and landslide inventory map and attribute factors are used for the WI method on different parameters [15].

The author performs landslide risk assessment using 12 landslide conditioning factors built from various data sources [16] Son La hydropower basin was used for this research and the predictive capability were assessed using Information Gain ratio using 10-foldcross validation technique and those factors with no predictive ability were removed. Five different models were then built (Support Vector Machines, Multi- layer perceptron Neural nets, Logistic model trees, Kernel Logistic regression, and Kernel basis function neural networks) followed by assessing their performance using metrics such as kappa index, ROC and other measures.

It was observed that the MLP neural nets performed with the highest prediction ability of 90.2%, followed by SVM and Kernel Logistic Regression model. Factors such as VIF, tolerances, and Pearsons correlation were quantified to determine that there was no multicollinearity among the 12 factors. As a next step, the highest influencing factors of these 12 were identified and 4 factors that did not give any value to the predictive ability were removed. The author suggests that LMT model with a positive predictive value of 80% can be a potential technique was landslide susceptibility.

The results indicated that multivariate approaches (i.e., SVR, LR and DT) outperformed the bivariate methods (i.e., FR, SI and WOE) by about 13%. Within the multivariate approaches, SVR method performed the best with the highest accuracy, while FR method was the most effective and accurate bivariate method. Interpretation of AUC values and the McNemars statistical test results revealed that the SVR method was superior in modeling landslide susceptibility compared with the other multivariate and bivariate methods. Multivariate (i.e., SVR, LR, and DT) approaches are found to outperform the bivariate methods (i.e., FR, SI and WOE) by an overall of 13%. Among the multivariate approaches, SVR technique performed with the highest accuracy, with FR outperforming the other techniques when it came to bivariate methods.

III. METHODOLOGY

The research problem is addressed by using Knowledge Discovery and Data mining technique (KDD). Based on KDD approach, the following steps were followed in sequence to explore the datasets in order to observe a pattern to find the suitable algorithm to categorize the landslide trigger factors from the obtained dataset.

1) Data collection Extraction
2) Exploratory Data Analysis (EDA).
3) Linearity and outlier test.

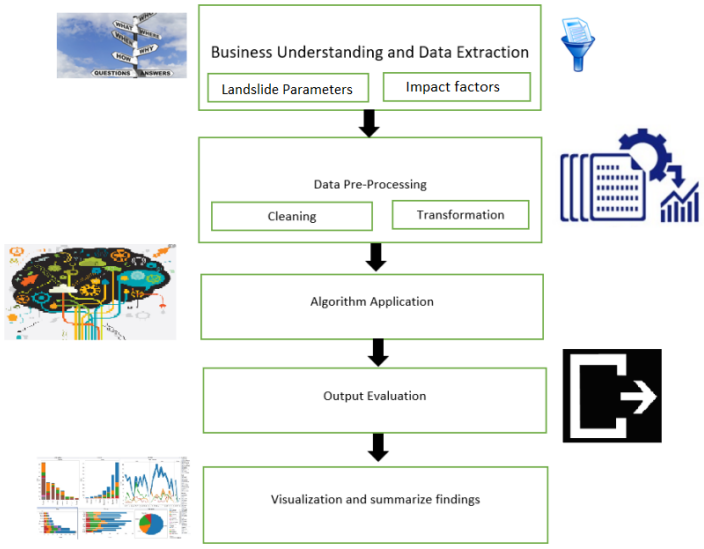

Fig. 1. KDD Methodology followed

A. Data collection & Extraction

Landslide occurrence dataset was gathered from https://data.nasa.gov/. This dataset approximately contains about 10000 rows which has labelled landslide

trigger classification data. The obtained data set has features like landslide occurrence time, month, type of landslide, impact of the landslide, country, location in the form of latitude and longitude, fatality count, injury count, population, etc. The redundant columns were chosen and removed from the original dataset to avoid the problem of multicollinearity.

B. Exploratory Data Analysis

In the very first step, the data was pre-processed and transformed to the needs of the algorithm. The objective of this step was to remove the missing values, special or unwanted characters present in the data. All the operation was performed using R.

- The missing values were imputed by using rfimpute function under Random Forest package.
- Few rows were removed from the dataset as it had NAs in almost all the columns.
- Some categories of the dependent variable were grouped under same class. For example, large rain, heavy rain, mild rain was grouped as one category named Rain.
- The country column was transformed into continent as there was more than 100 categories, whereas Random Forest can handle up to 53 categories only.
- Variables such as continent, event month, event time, landslide category and size were coded as factor variable.

The dataset gathered has a major class imbalance problem, where one of the landslide trigger categories called Rain was dominant. This problem was managed up to some extent by over sampling the data using SMOTE, which uses the techniques of k- nearest neighbours and bootstrapping to synthetically generate the observations to normalise the class imbalance problem in the data. The usage of SMOTE will be further explained in implementation section.

C. Lineartity and Outlier test.

Since the dependent variable in this research is of categorical, Chi-Square test has been conducted with other predictor categorical variables to understand their co-relation. As we know that, the null hypothesis for a Chi-Square test is that the variables under test are independent to each other and the alternative hypothesis proposes that the variables are co-related.

For the alternative hypothesis to be true, the p-value must be significant i.e. 0.5. From the test results it can be seen that the p-value for each test is very lesser

```
        Pearson's Chi-squared test

data:  Trigger and Landslide_Size
X-squared = 1841.2, df = 21, p-value < 2.2e-16


        Pearson's Chi-squared test

data:  Trigger and Landslide_Category
X-squared = 411.06, df = 56, p-value < 2.2e-16


        Pearson's Chi-squared test

data:  Trigger and Continent
X-squared = 663.58, df = 28, p-value < 2.2e-16
```

Fig. 2. Pearson Chi-Square test

than 0.5. This implies that all the predictor categorical variables in the dataset has very high co-relation with the dependent variable i.e., Landslide Trigger and can contribute significantly in the models outcome.

```
          Rainbow test

data:  datanew$population ~ datanew$fatality_count
Rain = 0.57421, df1 = 4565, df2 = 4562, p-value = 1


        Rainbow test

data:  datanew$population ~ datanew$injury_count
Rain = 0.57345, df1 = 4565, df2 = 4562, p-value = 1


        Harvey-Collier test

data:  datanew$population ~ datanew$fatality_count
HC = 3.0512, df = 9126, p-value = 0.002285
```

Fig. 3. Rainbow and Harvey-Collier test

The data is then checked for linearity. Initially, the required variables are fed into the model and the residuals were plotted against the fitted values. Since the interpretation of this output was hard, two popular methods were chosen to check the linearity of the data. A. Harvey-Collier Test B. Rainbow Test The data is considered to be linear if the p-value is significant i.e., lesser than 0.5. From the picture, it is observed that the p-value obtained from the rainbow test is equal to 1, which means the data is not acceptable. But, the results of Harver-Collier tests were completely opposite

in which the p-value is significant (0.002285).

From the results obtained by using the above methods, it can be seen that usage of either Neural network or Kernelised SVM models, would be an appropriate choice of algorithm for this particular data. Another important test that was required to be done is detection of outliers in the dataset.

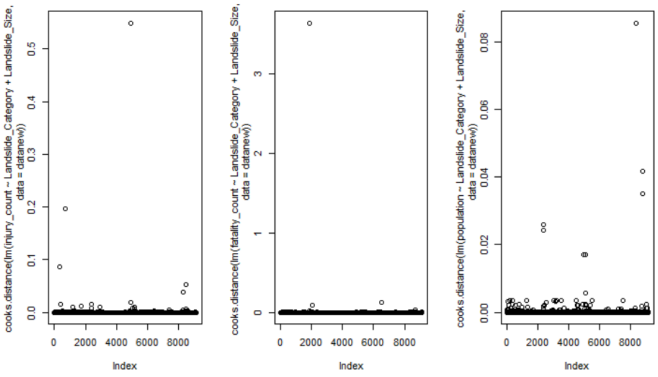

Fig. 4. Outlier Test

Outlier test was performed for the numerical variables such as fatality count, injury count and population. A separate function was written to identify the outliers from the data. This function would consider all the data above 75th quartile and below 25th quartile as outliers. The function is then applied to all numeric columns and the outliers were detected. Since the dataset is already synthesised, instead of removing the outliers it has been adjusted to fit into the 1st and 3rd quartiles.

IV. IMPLEMENTATION

This section explains about the steps involved in building the model from the starting stages which can be covered under below subsections.

- Feature Engineering.
- Support Vector Machine.
- Hyper parameter tuning.
- Random Forest.

A. Feature Engineering

After understanding the data, it is now important to consider the factors like linearity, outliers and multicollinearity to meticulously choose the features which are essential for the application of algorithm. In order to choose the appealing variables to the model, VarImpPlot was plotted to understand the importance of each variable to the model.

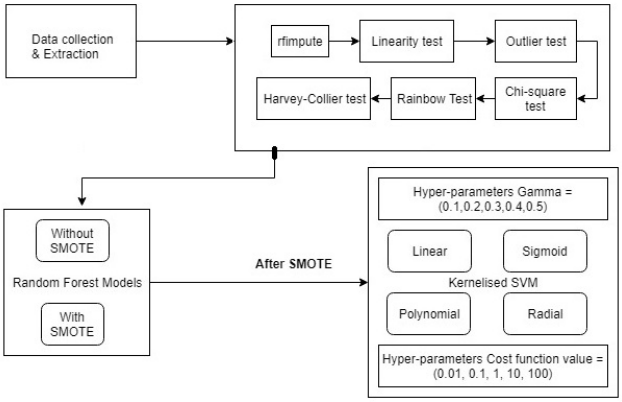

Fig. 5. Model Flow

This model helps to identify the important variables to build the model. It is evident from the graph that time and location are the most important variables in classifying the trigger reason for landslide occurrence. Also, all variables are plotted in order, as per their importance to the model which explains the change in performance of the model when some variable is removed or included. The Gini graph explains how pure each leaf nodes would be depending on the inclusion or exclusion of independent variables in the algorithm.

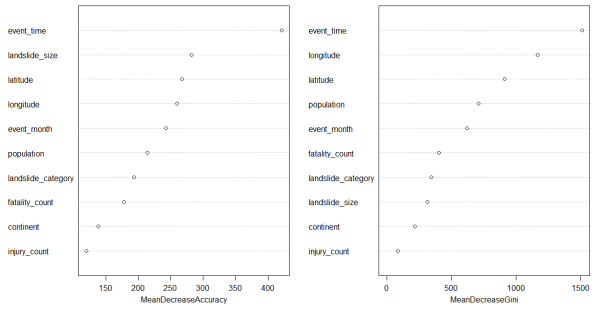

Fig. 6. Variable importance analysis

B. Support Vector Machine

Support vector machine (SVM) is a discriminative classifier that separates classes through linear separation. SVM can be used in classification and regression problem as well, but it is often used in classifying problems. It works by construction of hyperplanes in multidimensional space which splits the classes from one another. The main concept is that choosing the right hyperplane which has the maximum distance between the nearest data point between the classes.

After choosing the required features by using

varImpPlot, the predictor variables are transformed into a matrix. The data was split into 75% and 25% for training and testing respectively. Before feeding the training data into the SVM model, the data is synthesised using SMOTE to handle the class imbalance problem. Since the data collected has dominance of classes like Rain and Flood, the data is oversampled by creating more minority classes and reducing majority classes to address the class imbalance problem. The SMOTE function uses techniques of kNN and bootstrapping to create the new observations.
The perc.upper parameter was set to 400 and perc.under was set to 200 and the k value is set to 6 which takes the value for number of ks in kNN. This overall setting has created an additional observation of 2376 which makes the total number of observations in the dataset as 11505 with classes being normalised.

C. Hyper parameter Tuning

The SVM model performs differently when ran with different kernel, gamma and cost setting. It is important to find the appropriate parameters. In order to identify the right model for SVM, four different kernels were tested with numerous values of gamma and cost function. Four for loops were constructed for each of the kernels to test the SVM model for different values of gamma and cost functions. The kernels chosen in for models are Radial, Sigmoid, Linear and Polynomial. The mathematical expression for each of the kernels can be seen below.

$$\text{Radial basis function} : K\ (x_i, y_i) = \left(-\gamma ||X_i - X_j||\right),\ \gamma > 0,$$
$$\text{Polynomial} : \mathrm{K}\ (x_i, y_i) = \left(\gamma X_i^T X_j + r\right)^d,\ \gamma > 0,$$
$$\text{Sigmoid} : K\ (x_i, y_i) = \tanh\left(\gamma X_i^T X_j + r\right),$$
$$\text{Linear} : K\ (x_i, y_i) = X_i^T\ X_j,$$

Fig. 7. SVM Kernel equations

where γ is the gamma term for all kernels except linear in the kernel functions.
d is the polynomial degree term for Polynomial kernel in the kernel function.
r is the bias term for Polynomial and Sigmoid kernels in the kernel functions.

γ, d, and r are user-controlled parameters, which can be used to increase the accuracy of the SVM models[].

The performance of each kernels was analysed for different gamma terms and cost functions by using a for loop in R program. The results were tabulated to compare the performance of one kernel over the other with varying gamma and cost function values.

D. Random Forest

A random forest is another dual-purpose algorithm which can be used for regression as well as for classification problems. It is simply a collection of decision tree which is bootstrapped and bagged to produce an ensembled decision tree output. This improves the accuracy and reduces the error as multiple samples were taken from the original data and different features were selected each during each sampling because of bootstrapping approach.

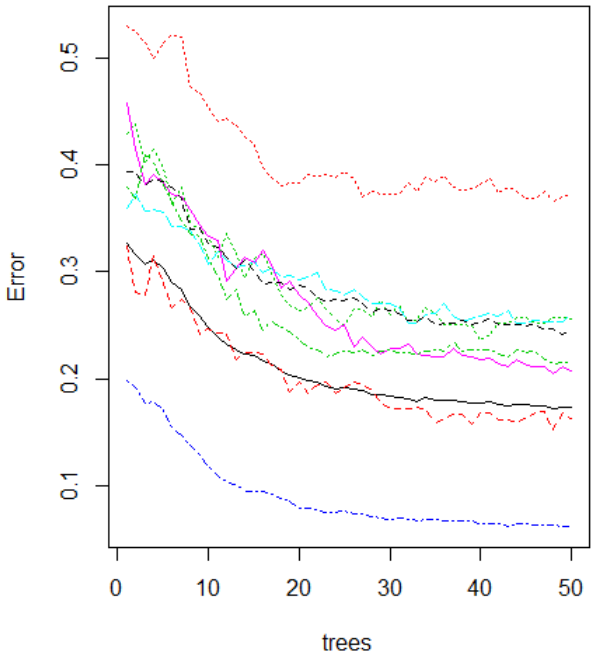

Fig. 8. Random Forest model output

The performance of the random forest can be highly influenced by the setting of mtry which decides the number of variables to be selected at each split of the tree and ntree which decides the number of trees which will be constructed for the model.
In the random forest model built for this research, 2000 trees have been built. This number is selected after series of trial and error to find the best fitting model of random forest algorithm. Again, SMOTE technique was applied to adjust the data for class imbalance problem.

The dataset was split into 75:25 for training and testing purposes. The train data was fed into the model to classify the trigger factor which caused landslides. The model is then evaluated against the test data to analyse the performances.

V. EVALUATION

Overall, forty-one (103) models have been developed in this research with Random Forest (3) and Support Vector Machine (100) with different hyperparameter settings. The results are captured and tabulated.

NUMBER OF TREES	SMOTE	ACCURACY (in %)
1000	No	65.97
1000	Yes	70.88
1500	No	62.13
1500	Yes	71.21
2000	No	70.79
2000	Yes	75.06

Fig. 9. Random Forest performance with & without SMOTE

The random forest model ran with ntree value as 1000 has achieved accuracy of about 58% in classifying the landslide triggers at 95% confidence interval. The kappa value for this model 0.50 which is not a reasonable score, this means that most of the classification has occurred by chance itself. Another random forest model with ntree set as 1500 was run, this model has obtained accuracy of 71.21% and kappa value of 0.51. The last random forest model was run with ntree set as 2000 which has produced a much better results than other two RF models with 75.06% accuracy and kappa value of 0.57.
Even though the last RF model has performed better than the other two models, the overall performance of the model however, was not satisfactory in comparison.

This has motivated to carry out the research using Support Vector Machine, which is expected to classify the triggers better than random forest models.

Hyper Parameters	Kernel	Accuracy in %
Gamma = 0.1	Linear	76.76
	Polynomial	73.71
Cost = 100	Sigmoid	79.46
	Radial	82.16

Fig. 10. SVM performance with different Kernel settings

The SVM model was initially run without synthesising the data. The model generated with this data has performed very poorly with almost all the classes were grouped under the majority classes. Also, the majority classes were overfitted because of lack of other classes in the training data. Hence, the data was synthesised using SMOTE technique to create additional observations for minority classes.

VI. CONCLUSION FUTURE WORK

In the present study, the performance of state of the art machine learning model is tweaked with the optimal hyper-parameters, with the appropriate kernel function in defining the hyperplane for classification. Landslide trigger classification is put forward as a sustainable workaround in drawing up warning zones for landslides that may occur in the future. The tracking of landslide triggers will conversely allow the tracking of landslide events that are to occur, based on the records from past data on such events. Support vector machine algorithm with Radial kernel, a gamma value of 0.1 and cost function value of 100 performs the best among the 106 models that were run for this study.

Landslide trigger classification can potentially bring a low cost prediction model that can perform as well as landslide susceptibility modelling and mapping. The trigger classification can be used in building up maps that can categorize areas into landslide prone areas by making use of landslide triggers and by tracking their occurrences. Superlearner packages [17] can be used to fit multiple machine learning models to build an ensemble model. The hybrid model could be built by choosing 1 or more different machine learning algorithms each for screening and prediction stages while using the superlearner packages.

REFERENCES

[1] D. Alexander, "On the causes of landslides: Human activities, perception, and natural processes," *Environmental Geology and Water Sciences*, vol. 20, no. 3, pp. 165–179, 1992.

[2] D. J. Varnes, "Slope movement types and processes," *Special report*, vol. 176, pp. 11–33, 1978.
[3] G. F. Wieczorek, "Landslides: investigation and mitigation. chapter 4-landslide triggering mechanisms," *Transportation Research Board Special Report*, no. 247, 1996.
[4] O. Kjekstad and L. Highland, "Economic and social impacts of landslides," in *Landslides–disaster risk reduction*, pp. 573–587, Springer, 2009.
[5] X. Yao and F. Dai, "Support vector machine modeling of landslide susceptibility using a gis: A case study," *IAEG2006*, vol. 793, 2006.
[6] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[7] L. Rokach, "Decision forest: Twenty years of research," *Information Fusion*, vol. 27, pp. 111–125, 2016.
[8] T. Kavzoglu, I. Colkesen, and E. K. Sahin, "Machine learning techniques in landslide susceptibility mapping: a survey and a case study," in *Landslides: Theory, Practice and Modelling*, pp. 283–301, Springer, 2019.
[9] H. Hong, B. Pradhan, D. T. Bui, C. Xu, A. M. Youssef, and W. Chen, "Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at suichuan area (china)," *Geomatics, Natural Hazards and Risk*, vol. 8, no. 2, pp. 544–569, 2017.
[10] N. Micheletti, L. Foresti, S. Robert, M. Leuenberger, A. Pedrazzini, M. Jaboyedoff, and M. Kanevski, "Machine learning feature selection methods for landslide susceptibility mapping," *Mathematical Geosciences*, vol. 46, no. 1, pp. 33–57, 2014.
[11] T. Kavzoglu, E. K. Sahin, and I. Colkesen, "An assessment of multivariate and bivariate approaches in landslide susceptibility mapping: a case study of duzkoy district," *Natural Hazards*, vol. 76, no. 1, pp. 471–496, 2015.
[12] S. Wan, S.-H. Chang, T.-Y. Chou, and C. M. Shien, "A study of landslide image classification through data clustering using bacterial foraging optimization," *Journal of Chinese Soil and Water Conservation*, vol. 49, no. 3, pp. 187–198, 2018.
[13] I. N. Aghdam, M. H. M. Varzandeh, and B. Pradhan, "Landslide susceptibility mapping using an ensemble statistical index (wi) and adaptive neuro-fuzzy inference system (anfis) model at alborz mountains (iran)," *Environmental Earth Sciences*, vol. 75, no. 7, p. 553, 2016.
[14] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, "Decision tree regression for soft classification of remote sensing data," *Remote Sensing of Environment*, vol. 97, no. 3, pp. 322–336, 2005.
[15] W. Chen, W. Li, H. Chai, E. Hou, X. Li, and X. Ding, "Gis-based landslide susceptibility mapping using analytical hierarchy process (ahp) and certainty factor (cf) models for the baozhong region of baoji city, china," *Environmental Earth Sciences*, vol. 75, no. 1, p. 63, 2016.
[16] D. T. Bui, T. A. Tuan, H. Klempe, B. Pradhan, and I. Revhaug, "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree," *Landslides*, vol. 13, no. 2, pp. 361–378, 2016.
[17] E. Polley, E. LeDell, C. Kennedy, S. Lendle, and M. van der Laan, "Package superlearner," 2018.