RandomForest for Landslide Susceptibility Mapping

ITU GeoAI Challenge

Isaac Oluwafemi Ogunniyi

Abstract

This work addresses the challenge of landslide susceptibility within North Italy, a region which is prone to landslides because of its steep slopes, heavy rainfall, and geological instability. The methodology used is a machine learning approach, involving the training of a RandomForest model on geological faultline and land-use-land-cover data to classify the susceptibility or otherwise of specific areas within the region. This approach yielded an impressive result of 77.66% accuracy on a held-out set of test data.

1. Introduction

A landslide is a mass movement of material, such as rock, earth or debris, down a slope occurring suddenly or more slowly over a long period of time. They occur when the force of gravity acting on a slope exceeds the forces resisting the movement of the slope [1].

Landslides pose a significant threat to infrastructure, property, and human life on a global scale. Thus the aim of this work is to create a landslide susceptibility map that will assist local authorities in implementing effective mitigation measures such as monitoring, early warning, evacuation, stabilization, and restoration to prevent and minimize damages caused by landslides.

The task of mapping landslide susceptibility is approached as a machine learning classification problem where the features of the geological faults and the nature of land cover and use of a specific geographical point and its immediate surrounding is used to classify that geographical point as susceptible to landslides or not.

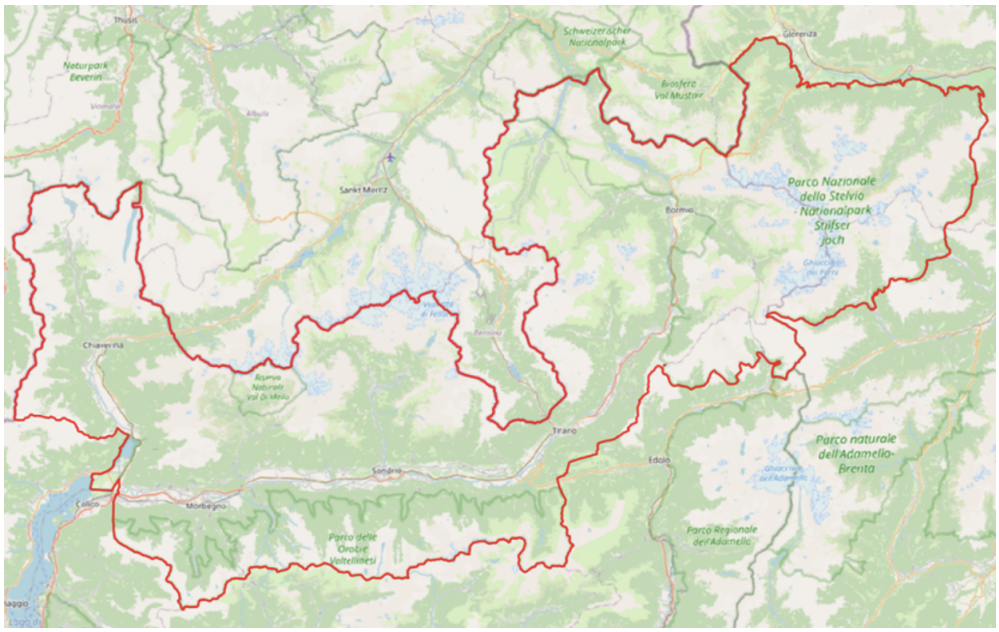

Figure 1. Location of the case area of Valtellina Valley, Northern Italy

1. Dataset

The data used in this work are the sole property of Zindi and ITU [2]. It is made up of:

a. geological fault zones map at a 1:10,000 scale in vector format (source: Lombardy Region)

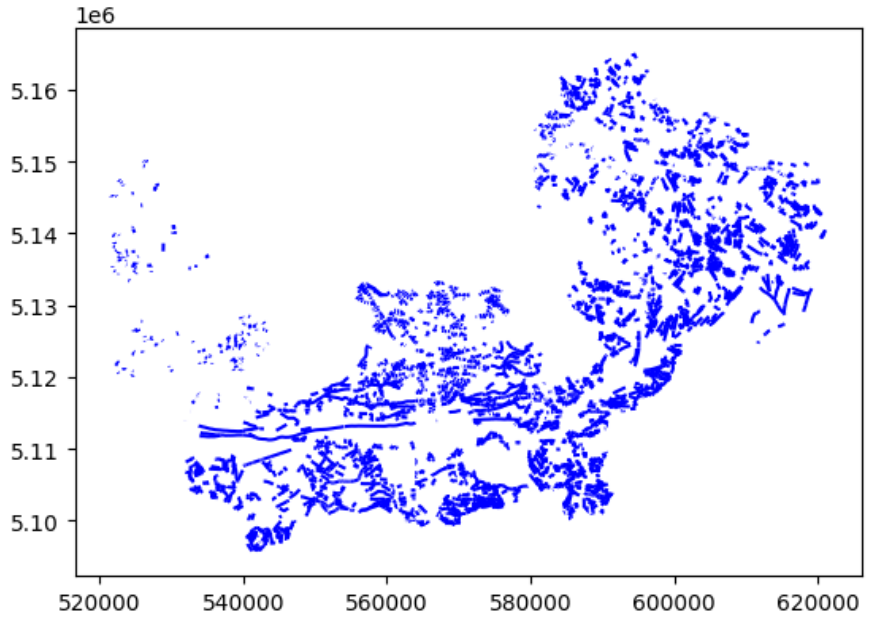

Figure 2. Map of geological Fault lines within the region

b. a land use/land cover map at a 1:10,000 scale in vector format (source: Lombardy Region)

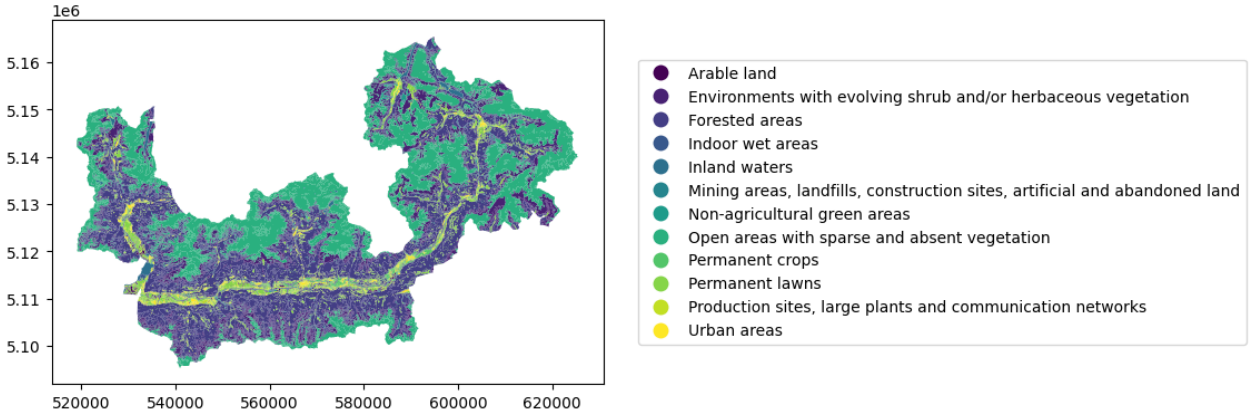

Figure 3. Map of the land use/ land cover categories that make up the region

c. Train dataset: 12,140 observations with 3 columns including the target

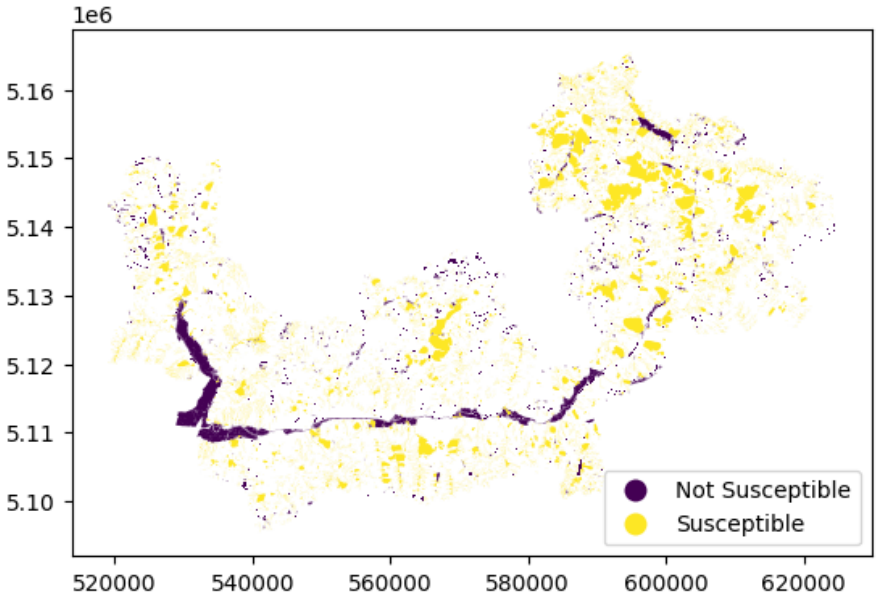

Figure 4. Map showing the train data geographies and their susceptibility to landslides

d. Test dataset: 40,000 observations with 2 columns

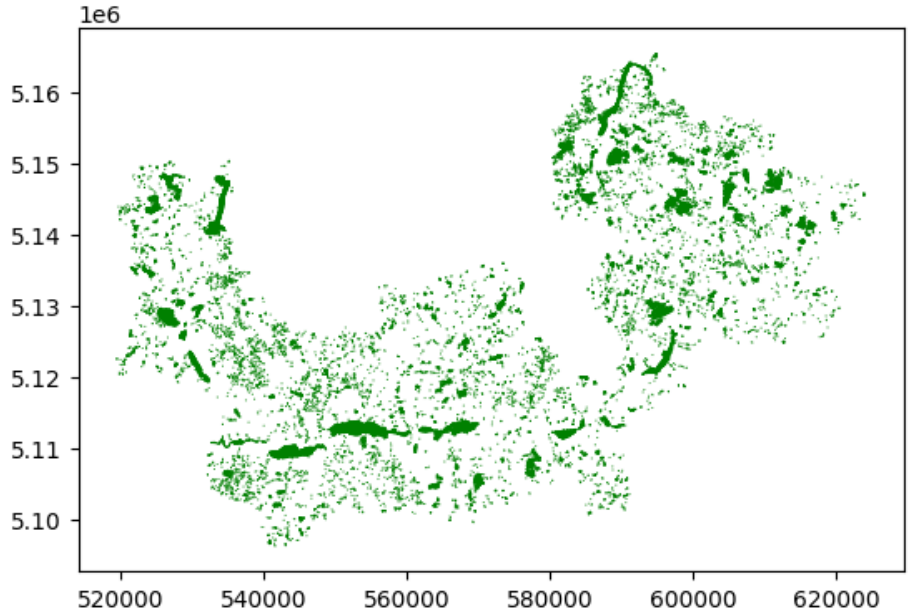

Figure 5. A map of the test geographical point with a 116 units buffer

3. Data Preprocessing

The train dataset was transformed through a series of preprocessing steps together with the geographical fault lines data and land use/ land cover data into a 12,140 by 39 dataframe. The test data also went through a similar transformation.

3.1 Data Transformation

The preprocessing steps are as follows:

1. There was a disparity between the geometry types of the train and test datasets. To address this, we converted the test points into polygons by adding a buffer of 116 units.

2. We matched the train and test datasets with the fault lines that either partially or fully lie within them. Given that multiple fault lines can fall within a single geometry, we grouped the matched data. This grouping was based on unique IDs from the original train and test data.

3. For each group, we aggregated its multiple observations and represented the total as their sum.

4. Similarly, we matched the train and test datasets with land use/land cover details. These details pertained to portions of land within the datasets. As with the geological fault lines, we aggregated the multiple land use/land cover observations for each instance.

5. After matching, we combined the data from geological fault lines with the data from land use/land cover via a column-wise concatenation.

6. Finally, to help with the performance of the model, balance was introduced into the distribution of observations in the train dataset which are susceptible to landslide versus the observations which are not susceptible. This was accomplished with the help of the Synthetic Minority Over-sampling Technique (SMOTE).
   The train dataset was further divided into train and validation dataset. The oversampling operation affected only the train portion of this data division.

3.2 Feature Engineering

In the course of data preprocessing, a number of new features were engineered as follows:

- An ‘area’ column representing the area of the geometry whose susceptibility is being predicted. For the test data, this is a fixed value of 42205.396489 squared units

- ‘lulc_area’ representing the sum of the land use/ land cover areas which lie partially or fully within the geometry of interest.

- A series of columns with names such as ‘cod_11’ and ‘cod_12’ which represent the number of land use/ land cover of type 11 and 12 respectively that fall within the geometry of interest.
  Below is a description of the land use/ land cover category codes:

Table 1. Description for the different land use/ land cover type codes

Land use/ land cover code	Description
11	Urban areas
23	Permanent lawns
31	Forested areas
13	Mining areas, landfills, construction sites, artificial and abandoned land
32	Environments with evolving shrub and/or herbaceous vegetation
22	Permanent crops
33	Open areas with sparse and absent vegetation
51	Inland waters
21	Arable land
14	Non-agricultural green areas
12	Production sites, large plants and communication networks
41	Indoor wet areas

- Another series of columns with names such as ‘area_31’ and ‘area_23’ which represent the total area of a particular land use/ land cover category found within a geometry of interest. The same number code definition for land use/ land cover categories in the table above applies.

- The following set of features are one-hot encoded versions of the geographical fault lines features: 'tipoel_01', 'tipoel_02', 'tipoel_03', 'dtipoel_faglia', 'dtipoel_frattura principale', 'dtipoel_sovrascorrimento', 'tipofag_451', 'tipofag_452', 'dtipofag_presunto', 'dtipofag_sicuro'

4. Proposed Model

The structure of the proposed model is a scikit-learn pipeline made up of 3 steps:

1. A StandardScaler object for scaling the data and making it more suitable for machine learning algorithms.
2. A SelectKBest object to select the most important features for predicting landslide susceptibility
3. A RandomForestClassifier to fit and learn the complex relationships between the processed features and the target.

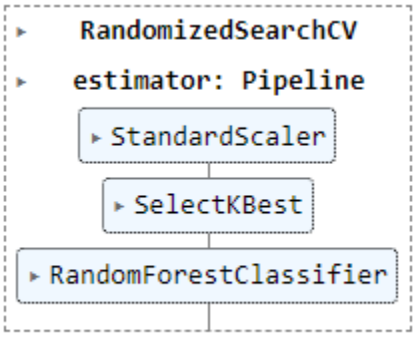

Figure 6. Structure of the Proposed Model

4.1 Training the Model

The data available for training was randomly split (80-20) into a train set of 16,772 observations (post oversampling) and a validation set of 2,428 observations.

The training process involved fitting the model to the data and tuning the hyperparameters by use of RandomizedSearchCV.

The table below shows the hyperparameters and the range of values over which they were tuned:

Table 2. Hyperparameter search space for RandomizedSearchCV

Pipeline Step	Hyperparameter	Minimum value	Maximum value
RandomForestClassifier	Number of estimators	30	500
RandomForestClassifier	Maximum tree depth	1	50 or None
RandomForestClassifier	Minimum samples split	2	30
RandomForestClassifier	Minimum samples leaf	1	10
RandomForestClassifier	Maximum features	10%	90%
RandomForestClassifier	Bootstrap	False	True
SelectKBest	Number of features	5	37

4.2 Evaluating the Model

The evaluation of the model's performance yielded an accuracy score of 93.46% on the train set and 88.63% on the validation data set.
The tuned model was then used to predict the susceptibility of a held-out test data and its performance was an accuracy of 77.66%.

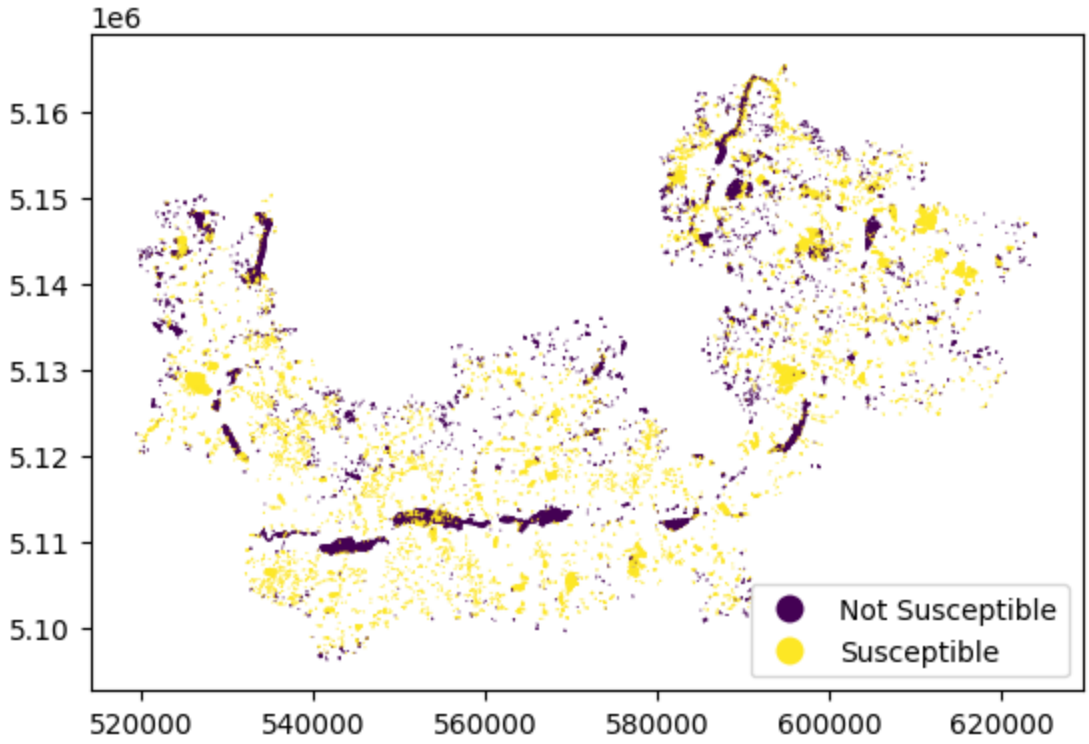

Figure 7. Landslide susceptibility mapping of the trained Model on the final test data.

5. Conclusion and Future Work

This solution demonstrates an effective approach to mapping landslide susceptibility using machine learning. By utilizing geological fault lines data, land use/ land cover data, feature engineering, and a RandomForestClassifier model, accurate predictions can be made for the susceptibility of a specific geography.

The code of this project runs in a Python environment with the main packages being:

- pandas (1.5.3) [3]
- numpy (1.23.5) [4]
- scikit-learn (1.2.2) [5]
- scipy (1.11.3) [6]
- geopandas (0.13.2) [7]
- imbalanced-learn (0.10.1) [8]

. The notebook can be executed on Google Colab with a CPU runtime, and the entire process takes less than 40 minutes. No paid subscriptions or additional resources are required.

The nature of the test data in this work differed from the train data in that train data contained multipolygon geometries whereas the test data contained point geometries. In future work, instead of approaching the problem by considering a fixed area surrounding the test points, the problem can be tackled by sampling and using points from the area of the train geometries so the train and test data will both be point geometries. This approach holds the promise of increasing the data available for training significantly which could potentially enhance the model's performance.

References

1. Understanding landslides - British Geological Survey.
2. GEO-AI Challenge for Landslide Susceptibility Mapping - Zindi
3. What's new in 1.5.3 (January 18, 2023) — pandas 2.1.1 documentation
4. NumPy 1.23.5 Release Notes
5. Version 1.2.2 — scikit-learn 1.3.1 documentation
6. Release Notes — SciPy v1.11.3 Manual
7. User guide — GeoPandas 0.13.2+37.g8279cc3.dirty documentation
8. Imbalanced-learn