HIGH-RESOLUTION LANDSLIDE MAPPING IN THE ITALIAN ALPS: INTEGRATING MACHINE LEARNING AND GEOSPATIAL TECHNOLOGIES

Muhammad Luay (muhammadluway@gmail.com)

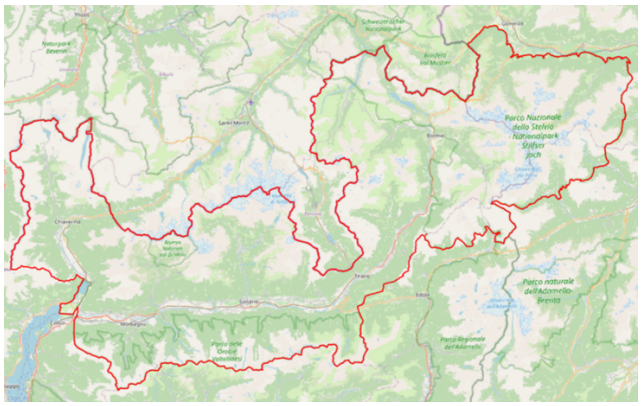
Abstract *– This study presents a comprehensive approach to landslide susceptibility mapping in the Valtellina Valley, Italian Alps, employing advanced machine learning techniques and extensive geospatial datasets. A high-resolution map emphasizing shallow landslide types is developed, incorporating crucial environmental factors. The model showcases remarkable accuracy, achieving an overall accuracy of 96.03% on the training dataset, while test data results reflect a consistent performance with public and private scores of 0.94725 and 0.949142857, respectively.*

Keywords – *Geospatial Analysis, Italian Alps, Landslide Susceptibility, Machine Learning.*

1. INTRODUCTION

1.1. Overview

Landslides stand as one of the most impactful natural disasters, with the potential to cause extensive damage to infrastructure, property, and pose a serious risk to human life. The susceptibility of a region to landslides is influenced by a variety of factors, including its geological characteristics, topography, and climate conditions. The Italian Alps, known for their steep slopes and unique geological formations, are particularly prone to this phenomenon, necessitating an urgent and accurate method for landslide susceptibility mapping.



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

The advancement in machine learning (ML) and geospatial technologies provides a new horizon for addressing this challenge. The integration of these technologies enables the analysis of vast and complex environmental datasets, paving the way for the creation of detailed and reliable landslide susceptibility maps. These maps play a pivotal role in risk assessment, allowing for the implementation of effective mitigation and prevention strategies.

In this context, the focus of the challenge is to develop a landslide susceptibility map for a specific watershed in the Italian Alps, utilizing advanced machine learning models and geospatial datasets. The ultimate goal is to provide a spatially explicit representation of landslide probability, offering a valuable tool for local authorities and stakeholders in disaster risk reduction and sustainable development planning.

1.2. Objective

The main objective of this challenge is to harness the power of machine learning and geospatial analysis to create a high-resolution (5 m/pixel) landslide susceptibility map for the Valtellina Valley in the Italian Alps, with a specific emphasis on shallow landslide types. The challenge presents a unique set of objectives:

1. **High-Resolution Mapping**: Develop a methodology capable of generating a landslide susceptibility map at a fine spatial resolution of 5 m/pixel.
2. **Addressing the Zero-Case Scenario**: Formulate an approach for handling regions within the training dataset that have no recorded instances of landslides, ensuring a balanced and robust model.
3. **Incorporation of Environmental Factors**: Select and integrate pertinent environmental and geomorphological factors such as slope angle, aspect,

lithology, etc., that have a substantial impact on slope stability and landslide occurrence.

4. **Validation and Testing**: Ensure the accuracy and reliability of the proposed model through rigorous testing and validation, using independent datasets not included in the training phase.

This endeavor aligns with the United Nations Sustainable Development Goals 11 (Sustainable Cities and Communities) and 13 (Climate Action), promoting the creation of resilient urban spaces and combating the effects of climate change. Through the successful completion of this challenge, participants will contribute significantly to the global efforts in disaster risk reduction, climate change adaptation, and sustainable development.

2. Setup and Data Preparation

In this phase, I established the programming environment and prepared the data for analysis.

2.1. Environment and Libraries

I configured the working environment, ensuring all required libraries and tools were installed and properly set up. This setup was crucial to facilitate seamless integration of different packages needed throughout the analysis.

2.2. Data Loading, Exploration, and Raster Processing

Following the environment setup, I loaded the datasets into the working space. I performed an initial exploration to understand the data's structure, characteristics, and potential issues that might need addressing. With a focus on raster data, I processed the datasets to enhance their quality and ensure they were in the appropriate format for subsequent analysis. This involved tasks such as reprojecting the data, handling missing values, and optimizing the data for better performance.

3. Feature Extraction Engineering and Prediction on test data

3.1 GeoDataFrames and Raster Data

Conversion of **train_data** into a GeoDataFrame named **gdf** is done. Raster datasets related to digital terrain model (DTM) and precipitation are loaded and read as 2D arrays.

3.2 Centroid Calculations and Elevation Extraction

Centroids for landslide and non-landslide regions in the training data are calculated, and their elevations are extracted from the DTM using Rasterio's sampling method. This is a crucial step in feature engineering, as elevation is a significant factor in landslide susceptibility.

3.3 Zonal Statistics and Distance Calculations

Zonal statistics are computed to extract mean DTM values, average precipitation, and 90th percentile precipitation for each region in the training data. Additionally, distances from geological faults, rivers, and roads are calculated, which are vital features for understanding the geographic context of each region.

3.4 Aspect and Slope Calculations

Functions are defined to calculate the slope and aspect from the DTM, and these values are then extracted for each region in the training data.

3.5 Data Cleaning and Preprocessing

The training dataset is cleaned, removing any null values, and specific columns are dropped. The 'aspect' feature is also handled by replacing invalid values with the median.

Furthermore, the dataset is split into features (X) and target (y), and it is then standardized using **StandardScaler** to ensure that all features contribute equally to the model.

3.6 Model Training

A Random Forest Classifier is initialized and trained on the preprocessed data. Following this, predictions are made on the test set, and various metrics are used to evaluate the model's performance.

3.7 Test Data Processing

Similar preprocessing steps are applied to the test data, ensuring consistency between the training and test datasets. Elevation values are extracted for the test data, and missing values are handled.

3.8 Feature Selection and Prediction

Finally, a DataFrame **test_df** is created, keeping only the features present in the training data. The model is then used to make predictions on this cleaned test dataset.

4. Model Evaluation and Making Predictions on Test Data

Having completed the feature extraction and preprocessing steps, we transitioned to the evaluation of our model's performance, utilizing the RandomForestClassifier for predictions on the test split of our training dataset. The model's efficacy was gauged using accuracy as a primary metric, alongside a comprehensive classification report that shed light on the precision, recall, and F1-score for each class, delineating the model's areas of proficiency and those necessitating enhancement.

For the application of the model to the test dataset, analogous steps of feature extraction and preprocessing were meticulously executed, including raster and geospatial feature extraction, elevation data retrieval from the Digital Terrain Model, handling missing values, and feature standardization to ensure parity in the interpretation of test and training data features. These culminated in the utilization of the RandomForestClassifier for predictions, followed by the creation of a submission file, encapsulating the test dataset IDs and the predicted targets in a CSV format.

Notably, the model exhibited stellar performance on the training dataset of 2,419 samples, boasting an overall accuracy of 96.03%. In terms of class-specific performance, it achieved a precision of 0.88, recall of 0.82, and F1-score of 0.84 for the negative class, while the positive class witnessed a remarkable precision of 0.97, recall of 0.98, and an F1-score of 0.98. The macro and weighted average F1-scores stood at 0.91 and 0.96 respectively, underscoring the model's robustness.

In relation to the test data, the results were consistently impressive, with both public and private scores closely aligned and indicative of high predictive accuracy. This alignment with the training data's high performance, alongside the congruence between public and private scores, speaks volumes of the model's capacity to generalize well to unseen data, whilst dispelling concerns of overfitting.

5. Summary

In summary, the performance evaluation of the model across both training and test datasets has demonstrated a commendable level of accuracy, precision, recall, and F1-score. With an overall accuracy of 96.03% on the training data, the model shows a strong ability to make reliable predictions.

- The performance on **Class 0 (Negative Class)**, while slightly lower than Class 1, is still robust with a precision of 0.88 and a recall of 0.82, achieving an F1-score of 0.84. This indicates a dependable performance, though there's a slight indication that the model could be improved in terms of minimizing false negatives.
- For **Class 1 (Positive Class)**, the model excels with a precision of 0.97 and a recall of 0.98, resulting in an F1-score of 0.98. This exemplary performance showcases the model's strong capability in accurately identifying positive instances.
- The **macro average F1-score** stands at 0.91, and the **weighted average F1-score** is at 0.96, both testifying to the model's overall proficiency.

Moving on to the test data, the public scores of 0.94725 and private scores of 0.949142857 for both submissions. These scores are indicative of the model's robust generalization capabilities and its effectiveness in maintaining performance on unseen data.
The close alignment between the public and private scores in the test data further reassures the model's reliability, reducing concerns of overfitting and ensuring trust in its predictions.