GEO-AI Challenge for Landslide Susceptibility Competition

1st Place

Technical Report

By

Jacob Ojumu

(Zindi Username : Enigmatic)

(B. Sc Chemical Engineering)

Introduction

Landslides represent a substantial risk to infrastructure, property, and human lives on a global scale. The Italian Alps, characterized by their rugged terrain and geological attributes are especially susceptible to these dangers. Consequently, the objective of this undertaking is to generate a map illustrating the likelihood of landslides within a specific watershed. This was accomplished by harnessing geospatial environmental datasets and employing advanced machine learning models.

The resulting outcome offers a comprehensive visualization of the probability of landslide occurrences in the area. This valuable information can greatly aid local authorities in the implementation of effective strategies for preventing and mitigating landslide-related damages. Importantly, this product aligns with the United Nations Sustainable Development Goals 11 and 13, which center on the establishment of sustainable, resilient urban areas and the mitigation of climate change impacts.

Geospatial Data Sources for Competition

The competition provided impressive array of geospatial data sources on the zindi platform. These sources provided the foundational information that allowed dissection and understanding of the complex dynamics of the Italian Alps, a region that stands as a microcosm of geospatial intricacies. Here, the sources that fueled the data-driven exploration and landslide prediction efforts are discussed:

1. Digital Terrain Model (DTM):

The Digital Terrain Model was a fundamental component of our dataset arsenal. At a high resolution of 5 meters per pixel, this source allowed us to capture

the minute details of the topography in the Italian Alps. With a wealth of elevation information, insights were gained into the terrain's physical characteristics, including slope angles and elevation variations. Understanding the terrain was crucial, as it plays a pivotal role in influencing landslide susceptibility.

2. Training Dataset:

The training dataset (Train.gpkg) contained the IDs of these points and the Target. The Target field assigned a value of 1 to these landslide-prone areas, also the No Landslide Zone which indicated regions with a low probability of shallow landslide occurrence were represented as 0. This showed that it was a binary classification problem which could be solved with machine learning.

3. Road Network

The road network at a 1:10,000 scale, helped to understand the broader landscape. Roads can act as triggers for landslides, as excavation and construction can destabilize the terrain.This dataset enriched the analysis by highlighting factors contributing to landslide susceptibility.

4. River Network

The river network at a 1:10,000 scale also helped to understand landscape of the problem. The river network is often associated with erosion, a precursor to landslides, this was also very important in the modelling.

5. Geological Fault Zones Map

Geological fault zones are critical to understanding the geological context of the region. At a 1:10,000 scale, this dataset was a crucial addition as geological characteristics significantly impact landslide occurrences.

Land Use/Land Cover Map: The land use/land cover map provided insights into the human influence on the landscape. Changes in land use and land cover can affect the stability of the terrain. This dataset, too, was at a 1:10,000 scale, helping to consider local-level variations in land use patterns.

6. Meteorological Data

Meterological data comprised of interpolated yearly averaged hour precipitation and the 90th percentile of the hour precipitation for the year 2020, this source was pivotal in understanding the climatic variables influencing landslide occurrences. Precipitation is a well-documented trigger for landslides, and these datasets provided a detailed picture of the meteorological conditions.

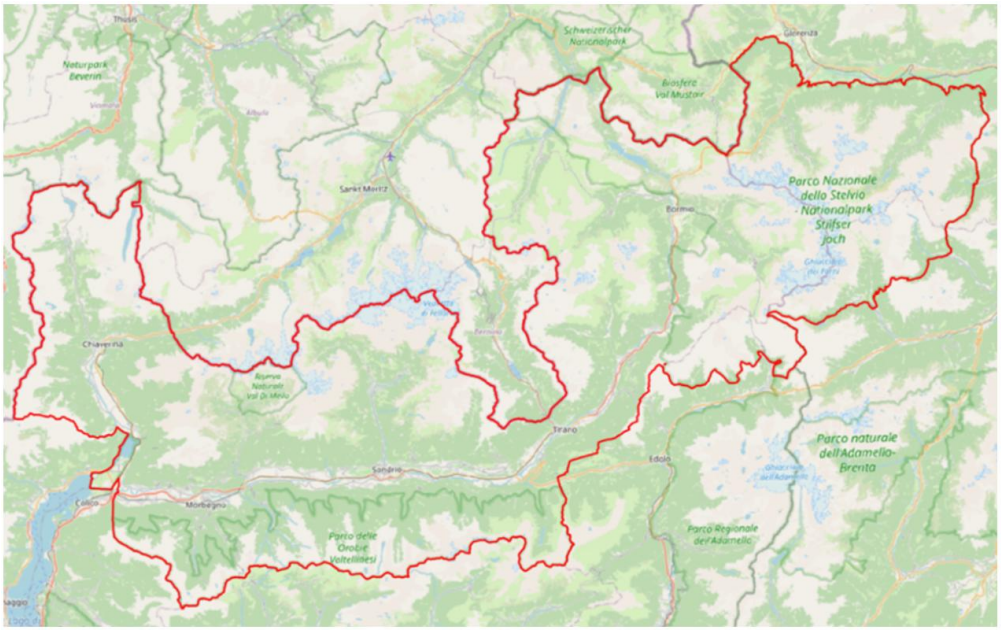

Figure 1 : Plot of the Region of Interest for the Competition

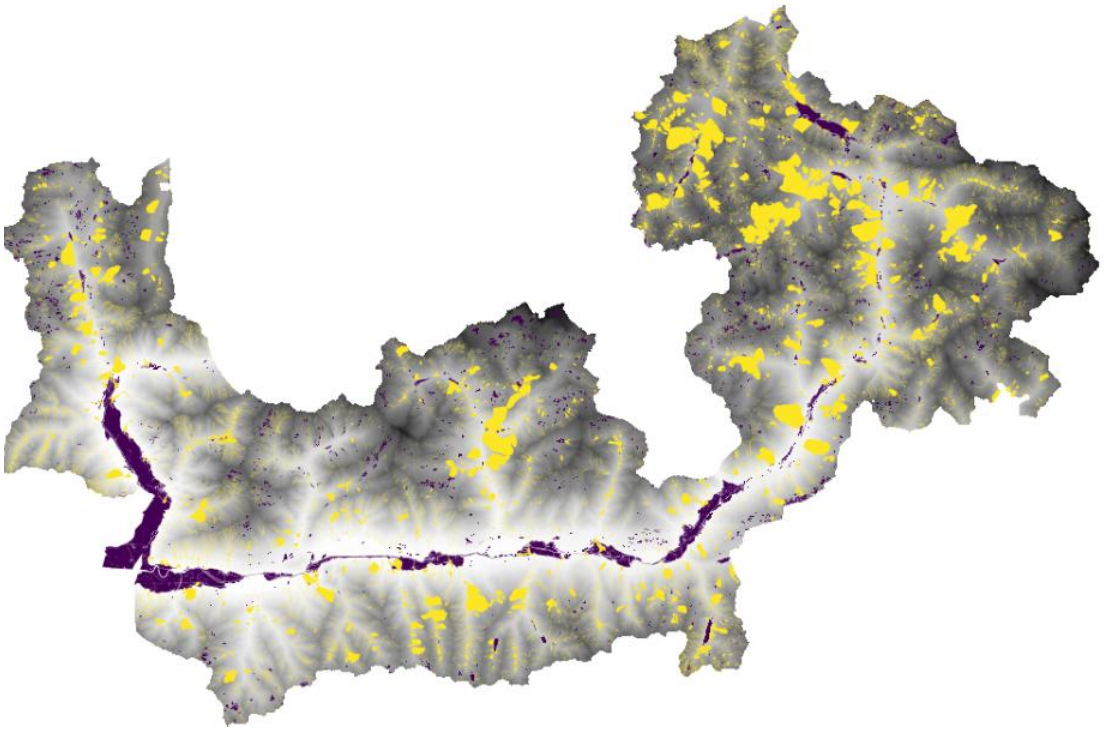

Figure 2 : Plot of the Digital Terrain Model(yellow : 1, purple : 0)

Feature Engineering

In this competition and any machine learning modelling problem, it is always important to create new features which would help the machine learning model. I employed the use of the gdal library to extract out some features from the Digital Terrain Model (DTM). They are listed and explained further below. After getting the features, the tiff files were converted to the geopackage format(gpkg) so that it can be easier to use. Also, a spatial join was done after getting the features with the provided datasets e.g river network, road network etc. The spatial join took a lot of time to run on a 8gb ram laptop.

1. Aspect Degrees:

Aspect degrees serve as a compass for understanding the orientation of slopes and terrain faces in a landscape. Expressed in degrees, they convey critical

information about whether a slope faces north, south, east, west, or falls within an intermediate direction. This data is essential for characterizing terrain features and is particularly pertinent in regions like the Italian Alps, known for their complex topography. Aspect degrees played a pivotal role in landslide prediction, as slope orientation can landslide susceptibility. Beyond this, aspect degrees find utility in land use planning, guiding decisions in agriculture, urban development, and conservation. The figure 3 below shows the aspect degrees from the digital terrain model.

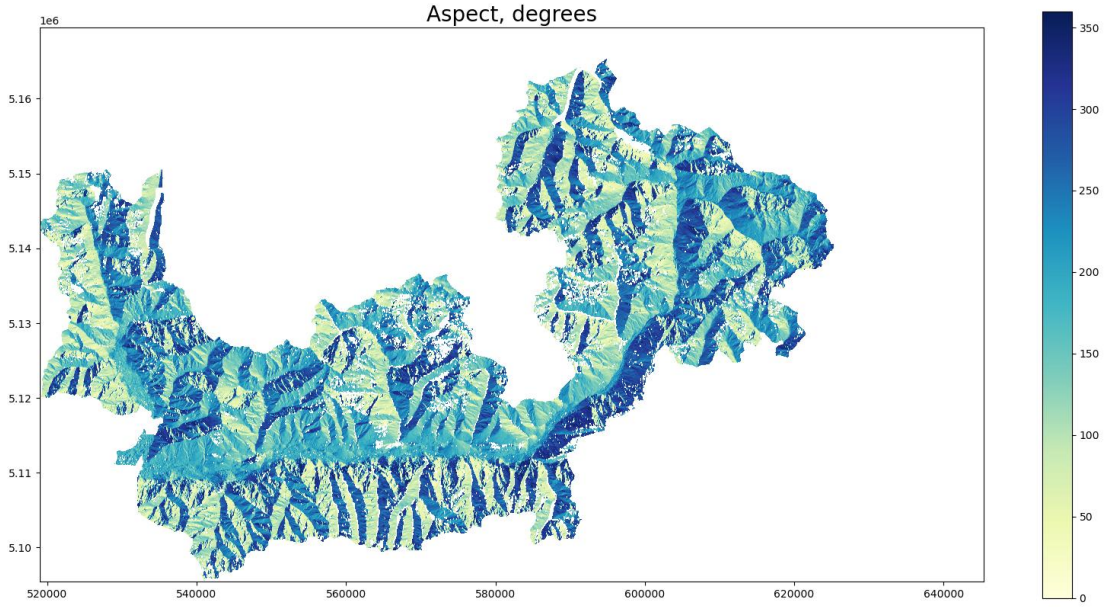

Figure 3 : Plot of the Aspect Degrees map

2. Northerness

Northerness refers to the characteristics and influences associated with the northern aspect of a location. It plays a pivotal role in climate, as northern regions typically experience colder temperatures due to the angle of sunlight. This features was used in the modelling for the landslide susceptibility. It is between -1 to 1. The figure 4 below shows the northerness.

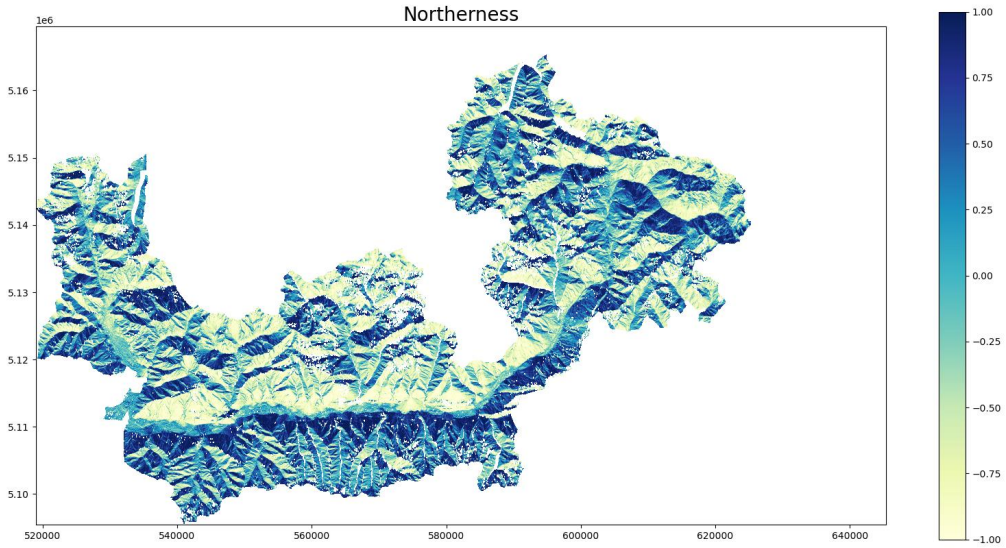

Figure 4 : Plot of the Northerness map

3. Roughness

Surface roughness is a critical geospatial parameter used to characterize the topography and terrain of landscapes. It quantifies the irregularity or variations in elevation across a given area. In geospatial analysis, understanding surface roughness is essential for a range of applications, including geological studies and land use planning The GDAL(Geospatial Data Abstraction Library) was also used to extract this feature out. The figure 5 below shows the Roughness from the digital terrain model.

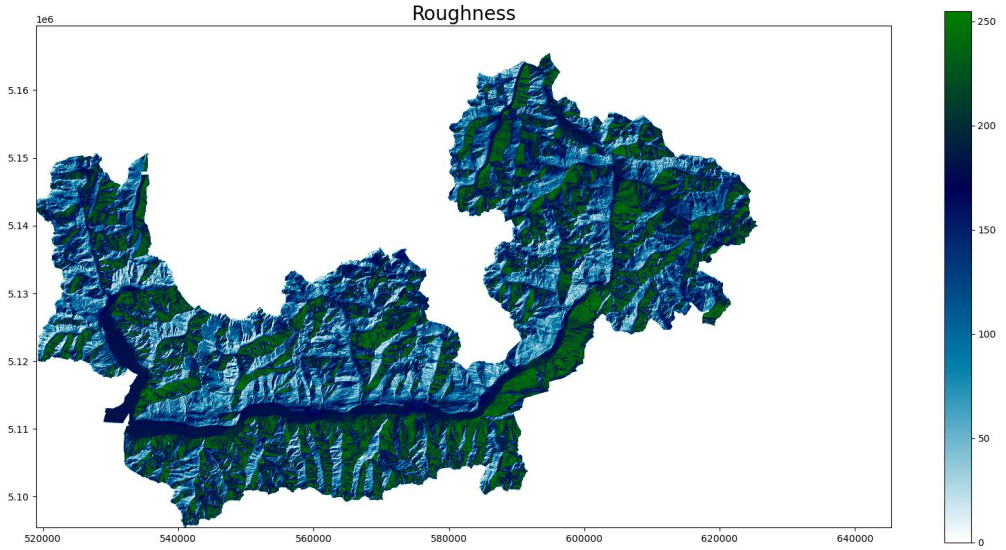

Figure 5 : Plot of the Roughness

4. Topographic Position Index:

The Topographic Position Index (TPI) serves as a key geospatial tool, offering insights into the local topographic position of a point within a landscape. It aids in terrain classification, environmental assessments, geological studies, land use planning, and conservation efforts, providing valuable information about relative elevations and spatial context. The gdal package was also used to extract it from the digital terain model. The figure 6 below shows the topographic position index from the digital terrain model.

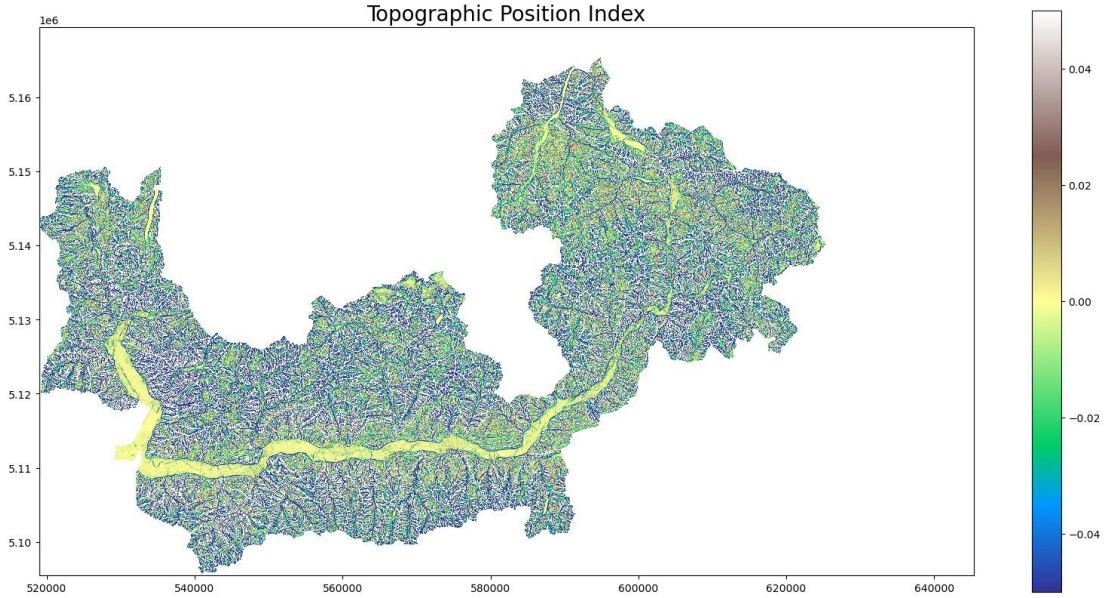

Figure 6 : Plot of the Topographic Position Index(TPI)

5. Topographic Ruggedness Index (TRI)

The Topographic Ruggedness Index is a geospatial metric used to quantify the roughness and ruggedness of terrain. TRI provides a numerical representation of the variation in elevation across a given area. TRI is particularly valuable in geospatial analysis, aiding in the characterization of terrain features, identification of landscape dynamics and suitability assessments for various applications. It is calculated by examining the differences in elevation between neighboring data points within a specified window, thus allowing for the assessment of terrain ruggedness on

a local scale. The figure 7 below shows the topographic ruggedness index from the digital terrain model.

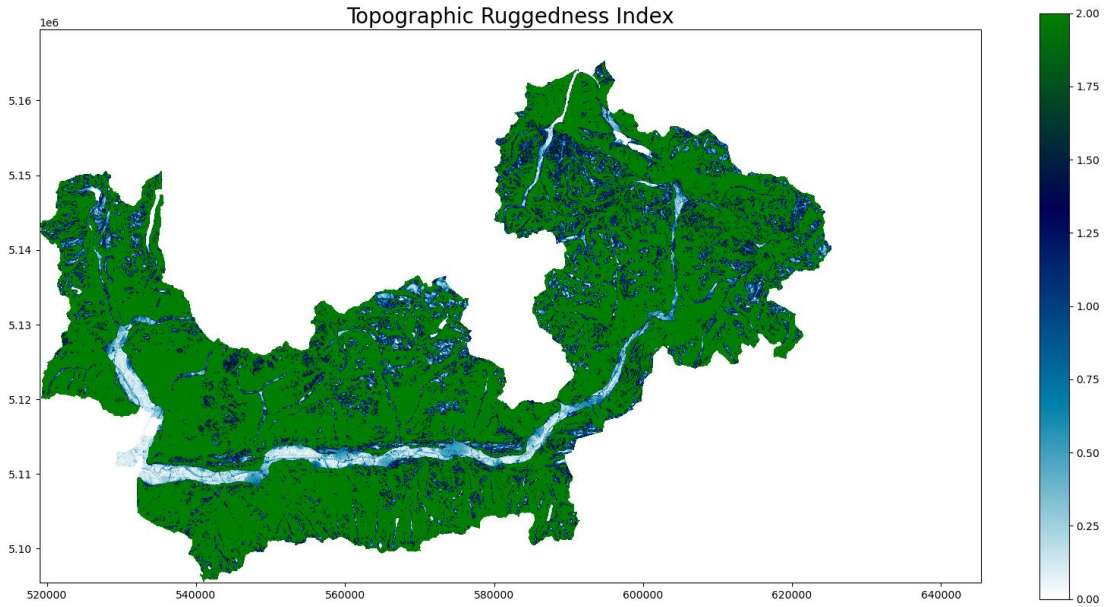

Figure 7 : Plot of the Topographic Ruggedness Index=

6. Distance to Fault zones

The distance to fault zones was also calculated, its important to find out out close the faultszones are to the locations. As seen in the figure 8 below, most of the points were close to the fault zones, the histogram is right skewed.

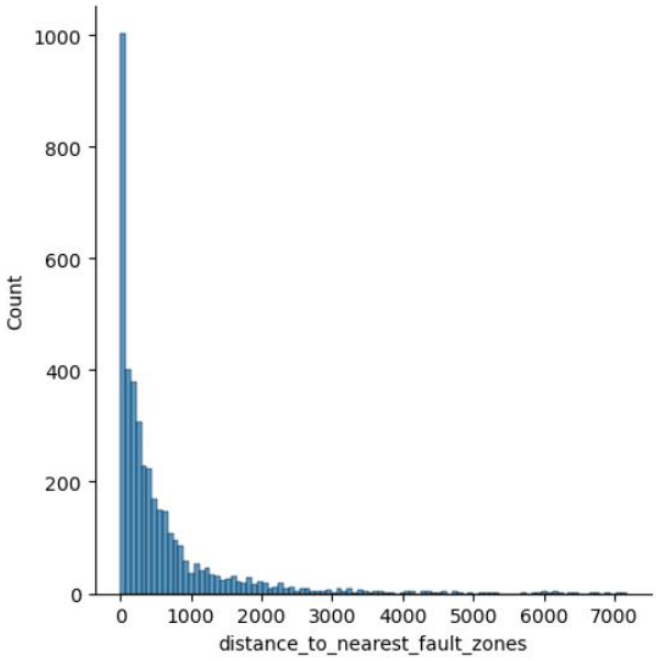

Figure 8 : Pllot of the distance to the nearest fault zones

7. Distance to nearest river:

I also calculated how close the geometries are to the nearest rivers given to us in the river network data. This is a good feature because points close to a river have a high chance of having a landslide. The figure 9 below also shows the distannce to the rivers and it was also right skewed with a splike around zeor.

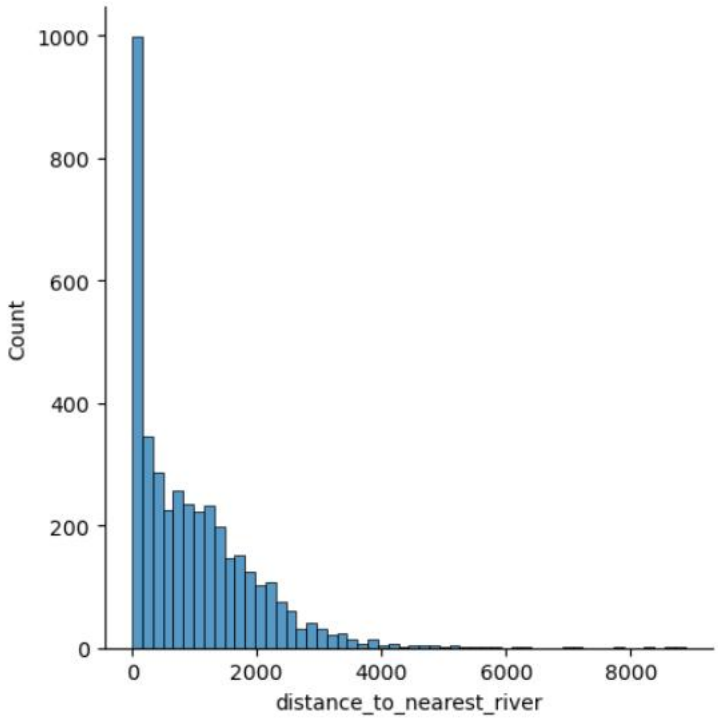

Figure 9 : Pllot of the distance to the nearest riiver

8. Distance to nearest road:

It is also necessary to find distance to nearest road. Human activities via roads could also be a factor in the landslides. The figure 10 below for the distance to neareast road is also right skewed with a spike around zero also.

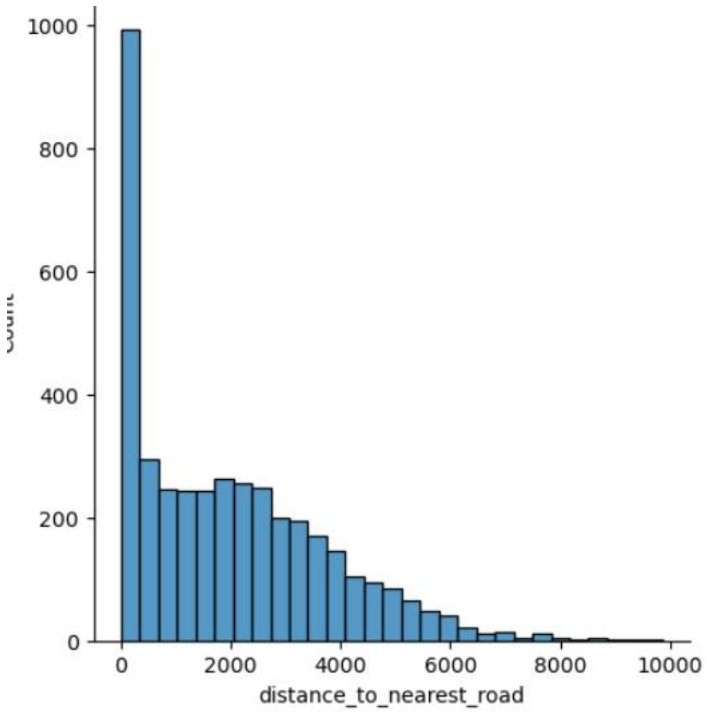

Figure 10 : Plot of the distance to the nearest road

Other features include; dtm, average_precipation, perc_precipitation, easterness, distance to nearest road, distance to nearast river, distance to nearest fault zone, hillshade, centroid, latitude, longitude,area, perimeter, bounding box, aspect ratio and convex hull.

Modelling

The processed train data was used for the modelling which contained all the features necessary for modelling. The targets were initially plotted based on the latitude and longitude to see the spread of the susceptible points to landslide. As seen in the figure 11 below, the target ‘1’ which is represented with orange and signifies the points susceptible to landslide dominated the plot. It appears that only a few points were not susceptible to landslide(blue color). To buttress and validate this finding, the value counts was checked for

the target column and it is shown below. It shows there was a class imbalance and the method used for the modelling was to use the imblearn package to resample the data.

Table 1 : Table showing unique values in the train data

Target	Number of Occurences	Pecentage
0	1657	14%
1	10483	86%

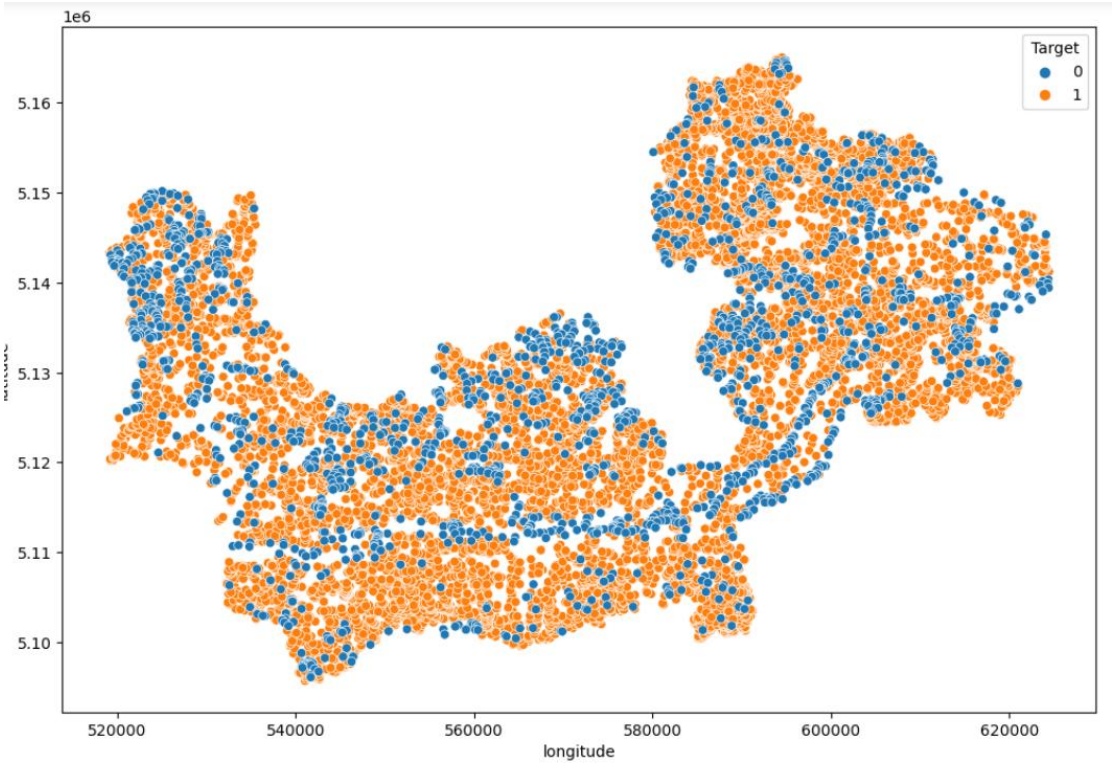

Figure 11 : Map showing the Targets in the Train data

Resampling:

Several resampling techniques for the undersampling and oversampling methods were investigated but the final method settled for was using the random over sampler method which is a type of over sampling. The purpose was to ensure that the model wasn't predicting the dominant class more than the non-dominant class.

Data Normalization :

For the data normalization, I decided to use the standard scaler which was better than the minmax scaler in this case. It is sometimes a practice to use standard scaler when there are negative values and non-negative values present in a distribution while a min max scaler when dealing with totally positive numbers, so I decided to use this knowledge for the project.

Validation Mechanism :

I used the Stratified K-fold method to perform a 3 fold training, validation and prediction. After training on a fold, the validation metrics were obtained and the model was also applied to the test data to obtain the test predictions for the final submission

Ensembling/Weighting:

After getting the results for each fold for each model type, the results were calculated and the median was calculated. Also after getting the model results for each model, the results were also gathered and the median was calculated which was then submitted for the challenge.

Models

The models used include catboost classifier, lightgbm classifier, xgboost classifier and histgradient boosting classifier. I did not do too much hyperparameter tuning for the models but I believe with more hyper-parameter tuning the results could have been better. In the long run, I ensured that they were giving good results relative to the Zindi Public leaderboard scores to ensure that there wasn't overfitting of any sort. I also compared the results of the models.

Table 2 : Table showing the metrics obtained from each model

Model	Mean Accuracy	Mean Precision	Mean Recall	Mean f1-score	Mean roc-auc score
Catboost Classifier	0.9773	0.9941	0.9602	0.9769	0.9772
LGBMClassifier	0.9863	0.9986	0.9739	0.9861	0.9863
XGBoostClassifier	0.9847	0.9983	0.9710	0.9845	0.9845
HistGradientBoostingClassifier	0.9812	0.9977	0.9945	0.9808	0.9812

I noticed that the classes predicted for the competition test data seemed to be balanced. So I also check for the amount of 1 and 0s predicted and submitted by the models before and after finding the median. The result is shown in table 3 below.

Table 3 : Table showing the ration of targets in the submission file for competion.

Model	Target 0	Target 1	% Target 0	% Target 1
Catboost Classifier	20945	19055	52.4	47.6
LGBMClassifier	20132	19868	50.3	49.7
XGBoostClassifier	20273	19727	50.7	49.3
HistGradient BoostingClassifier	20663	19337	51.7	48.3
Final Submission	20661	19339	51.7	48.3

From the table above, it appears that in the test data to be predicted, the data points or targets were not imbalanced. Probably a 50-50 split but the model was still able to precisely give good predictions with a few false positives and false negatives. More results could be seen with a classification report and confusion matrix.

The test data targets were also plotted with the seaborn package in python with the x as the longitude and the y as the latitude. From the figure 12 below, it is easier to navigate which parts are more susceptible to landslide.

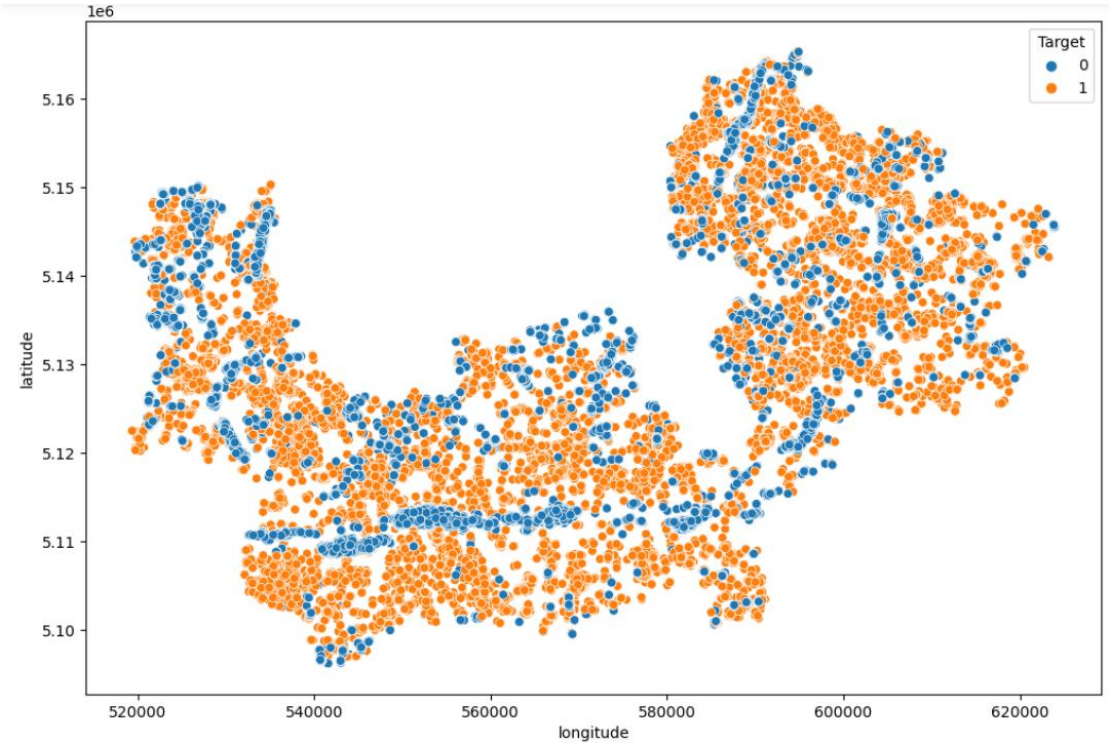

Figure 12 : Map showing the Targets in the Submission file for competition

Conclusion:

This competition leveraged machine learning models and geo-location datasets to identify locations susceptible to landslide. From the analysis, it appears that a lot of regions in this region of interes seem to be susceptible to landslide. My model is able to help the authorities in this region of interest to make data driven informed solutions. More model optimization and feature engineering can be done to achieve a more robust solution. I believe that this will be a stepping stone to greater things that can be done for the people in the region of interest.