Analysis of Landslide dataset using various Machine Learning Algorithms

Aliyah Kabeer
PES University
Bangalore, India
aliyahk888@gmail.com

Serena A Gomez
PES University
Bangalore, India
serenagomez1304@gmail.com

Prachi Sengar
PES University
Bangalore, India
prachisengar.7@gmail.com

Paul John
PES University
Bangalore, India
pauljohn2802@gmail.com

Abstract*— Landslides are one of the most dangerous hazards in the world, causing a large number of fatalities every year. This paper discusses the analysis of landslides that have been triggered due to intense and prolonged rainfall in North America and South America. The data set used in this study is provided by the Global Landslide Catalog as uploaded by NASA which records landslides that have occurred between the years 2007 and 2016 [1]. In this paper we aim at identifying and predicting some of the factors involving landslides. One of the problems we have considered is to classify the fatality of a landslide given its geolocation. Rainfall data has been extracted from the respective geolocation in order to better analyze the data. We have also tried to classify the size of the landslide based on its geolocation. The various Machine Learning algorithms used in this study are K-Nearest Neighbors and Random Forests for classifying the size of the landslide and, Multi-Layer Perceptron classifier and Naive Bayes algorithm for classifying the fatality of the landslide. The predictions generated provide meaningful insights into the adverse effects of various geological factors and occurrences, as well as the number of casualties due to landslides.

Keywords—Landslides, K-Nearest Neighbors, Random Forests, Multi-Layer Perceptron, Naïve Bayes, Landslides.

I. Introduction

Landslides are one of the most pervasive hazards in the world, causing more than 11,500 fatalities in 70 countries since 2007. Saturating the soil on vulnerable slopes, intense and prolonged rainfall have been found to be some of the most frequent triggers of landslides. In general, landslides can be classified by type of material (rock, debris, or earth) and type of movement (slides, spreads, flows, etc.) Occasionally, because of complex temporal and spatial relationships of more than one type of movement within a single landslide, their intricate study and analysis often requires detailed interpretation of both land forms and geological sections, or cores.

Landslides have proven to be an extremely difficult event to analyse and work on as there are many factors that could trigger these events. In our paper we have dived into a more detailed and intricate method by using external APIs to gain more information on the weather conditions and other possible factors that could influence landslide activity. We have extracted additional rainfall and weather data from Visual Crossing, which provides weather data for areas all over the world. The several problems we have chosen to solve include, the prediction of the fatality of a given landslide affected area depending on its geolocation, which helps us to identify the severity of the landslide and also classifying the size of a landslide given the geolocation of the area, which helps us analyse which areas are more prone to large landslides.

The study area of this paper includes areas in the North American and South American continent. In the current study, several advanced machine learning techniques including K-Nearest Neighbors, Random Forests, Multi-Layer Perceptron classification and Naive Bayes, were adapted to perform both classifications and prediction problems.

II. Previous Work

The landslide dataset is a popular dataset that has attracted a lot of interest from data scientists all over the world. We have referred to various papers that have been published and focused on making ours unique and tried to improve on it as much as possible. Most papers involved describing their algorithm without much practical testing and analysis.

Some papers like 'Rainfall-Induced Landslide Prediction Using Machine Learning Models: The Case of Ngaruroro District, Rwanda' by Martin Kuradusenge, Santhi Kumaran, and Macro Zennaro use the Random Forest algorithm to predict those landslides that are caused due to heavy rainfall. This paper however does not take into account the soil moisture content. Additionally, the presence of a large number of trees in the Random Forest makes the model extremely slow. In our paper, Random Forests have been implemented on attributes that are fixed such as precipitation cover and wind direction, hence increasing the efficiency of the algorithm.

In the paper 'Mapping landslide susceptibility and types using Random Forest' by Khaled Taalab, Tao Cheng and Yang Zhang, the Random Forest algorithm has again been used. In this particular study, this algorithm is applied as a data mining approach to produce landslide susceptibility maps and classification maps. However, there is no way of knowing what these Random Forest variables represent. This particular model is overfitted to the specific scenario and hence it cannot be used in neighboring regions. This model did not consider rainfall as a trigger for landslides. The accuracy of this model is 88%. While in our study we have used the Random Forest algorithm to predict the size of the landslide. This prediction is done based on multiple attributes from data based on rainfall in that particular region. Hence this provides a more accurate outlook on the factors that trigger a landslide.

In our paper, we have focused on using the right algorithms to ensure that our models provide maximum efficiency and high accuracy. The implementation of using

the Visual Crossing API helped us get more data for each landslide, hence giving us more attributes to work with.

III. Proposed Solution

A. Pre processing

We obtained our data from the landslide dataset provided by GLC of NASA and the rainfall/weather data obtained from the weather API [2].

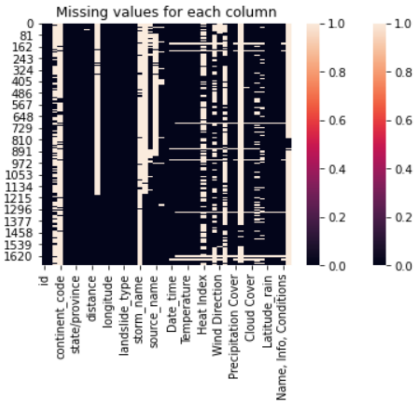

Fig. 1. Heatmap to detect the null values in the dataset.

For data cleaning we first constructed a heatmap to detect the presence of missing data in our dataset. These missing values are handled by filling the values present in the trigger, location description, time, landslide type, landslide size, source name and source type columns with string values (Example – 'Unknown', 'Other', etc.). The values present in these categorical attributes are then uniformly cased. Irrelevant attributes such as Continent code and Country code are dropped. Null entries present in the latitude and longitude columns are also dropped.
Computation of missing data in the numerical attributes such as distance and population are filled by the respective column means. While the numerical values present in the fatalities column is grouped based on landslide size and their respective means are computed and filled in place of the null entries. Following which these values are also binned and labelled as 'Few to None', 'Few', 'Moderate', 'Moderate to Many', 'Many', 'Too Many' depending on the number of fatalities. Dates present in the date column are parsed and added to a new column date_parsed. Categorical columns like landslide size and landslide type are encoded using map function.
Finally, missing values in the weather data are computed mainly using the means of the respective attributes and null entries in the temperature attribute are filled by grouping the data based on country name and then calculating their respective means.

B. Building a model

1) Landslide size prediction

a) K-Nearest Neighbours

K Nearest Neighbor is a popularly used Supervised Learning algorithm that is used for classification and regression. This algorithm follows instance based and lazy learning methods. We are using the KNN algorithm in our study to predict the size of the landslide.

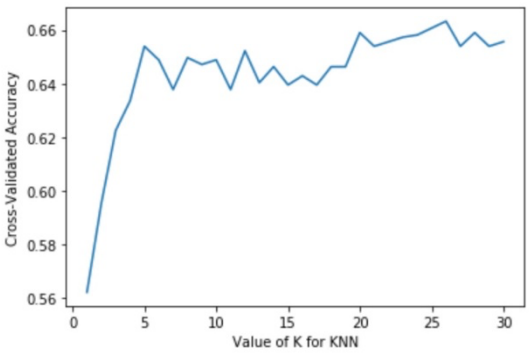

Fig. 2. Graph showing the selection of K value for the K-Nearest Neighbors Algorithm.

During the construction of our model, we used the 70-30 rule for train-test data split. The parameters considered during this construction are K neighbors, distance measure, p value and weights. The value for K neighbors is computed by using a range of hyperparameter values from 1-32 to fit into sklearn's GridSearchCV model in order to refine our parameter K to the most optimum one. Additionally, we set the cross-validation batch sizes cv equal to 10 and set accuracy as the preferred scoring metric. It returns the best value of K as 26, but to avoid taking an even number, we set K to 27 to fit our model as it proved to give a nearly equal accuracy as seen in the graph above (Fig. 2)For distance/proximity measure we are using the Manhattan distance measure because it was found to be the most optimum measure after performing trial and error. Weights considered for this algorithm depend on the distance measure.

The features that we considered to classify a new entry are Latitude, Longitude, Precipitation cover and Dew point where Precipitation cover and dew point are obtained from the weather data. The model classifies the landslide size of the new entry as either 'Small', 'Medium', 'Large', 'Very Large' or 'Other'.

b) Random Forest

Random Forest is also a Supervised Learning algorithm that is used for classification and regression. This model follows ensemble learning method which combines multiple classifiers to improve the performance of the overall model. We are using the Random Forest algorithm in our study to predict the size of the landslide.

During the construction of this model, we are using the 70-30 rule for the train-test data split. The parameters considered for this algorithm are n estimators, minimum samples split and random state. The value of n estimators which is the number of trees in the forest, is limited to 50. The value for minimum samples split is the number of splits and is limited to 10. The value for random is set to 3.

The features that we considered to classify a new entry are Latitude, Longitude, Precipitation cover and Wind direction where precipitation cover and wind direction are obtained from the weather data. This model classifies the size of the landslide as either 'Small', 'Medium', 'Large', 'Very Large' or 'Other'.

Evaluation - These models provided an accuracy of approximately 70% for predicting the size of a landslide.

2) Fatality predictions

a) Naïve Bayes

The Naïve Bayes algorithm is a form of Supervised learning algorithm and is based on the Bayes' Theorem. It is a classification method.

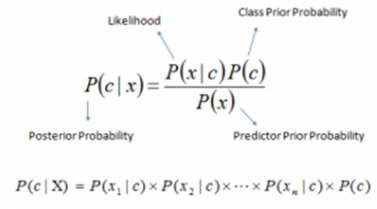

Fig. 3. The Bayes Theorem Formula

During the construction of this model, we are using the 70-30 rule for the train-test data split. We have used the Gaussian Naïve Bayes algorithm to classify a new entry.

The features that we have considered for this model are Latitude, Longitude, Precipitation cover, Wind direction and size_ordinal which is the size of the landslide. This model classifies the number of fatalities as either 'Few to None', 'Few', 'Moderate', 'Moderate to Many', 'Many' or 'Too Many'.

b) Multi-layer Perceptron

Multi-Layer Perceptron is a class of feed-forward artificial neural network that comes under Supervised Learning algorithms that learn a function by training on a dataset. This model is used for either classification or regression. There can be multiple linear layers between the input and output layers.

During the construction of this model, we are using the 70-30 rule for the train-test data split. The parameters considered for this algorithm are hidden layer size, activation, solver and max_iter. The size of hidden layer is limited to (16,16). The activation function that we are using is the identity function. The solver parameter is for weight optimization and it is set to 'adam'. The value of max_iter is number of iterations and it is set to 500.

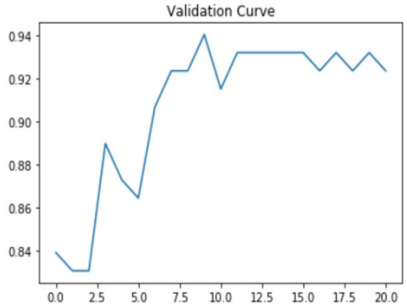

Fig. 4. Validation curve for the MLP algorithm

The features that we have considered to classify a new entry are Latitude, Longitude, Precipitation cover, Wind direction, Precipitation, Wind speed and size_ordinal which is the size of the landslide. This model classifies the number of fatalities as either 'Few to None', 'Few', 'Moderate', 'Moderate to Many', 'Many' or 'Too Many'.

Evaluation - These models provided an accuracy of approximately 95% for predicting the fatalities for a landslide.

IV. Experimental Results

Through series of testing and analysing the data we were able to create models that were able to classify the fatalities of each landslide event into different groups as well as predict the landslide size according to the geolocation. We were able to successfully implement four models on the dataset used and gained a lot of insight as to how these said factors were able to influence and relate to our problems.

Our first problem involved the classification of the size of each landslide into various groups. The size of the landslide was a factor that we considered to be very important as it could in turn help in predicting and classifying other various factors regarding landslides. The after-effects of a landslide can be long-lasting and extreme, including loss of life and natural resources, mass destruction, and extensive damage to land, to name a few. Deep landslides, that are triggered as a result of natural calamities like volcanic activity or major earthquakes can kill thousands of people and destroy thousands of square kilometres of land. The material from a landslide can also increase the chance of floods occurring in an area. Landslides also have devastating effects on the livelihood of farmers due to their massive, sometimes irreparable, environmental impacts. [4] This makes the size of landslides an important factor as the larger the landslide the more adverse the effects could be. We used the classic K-Nearest Neighbours algorithm and the Random Forests algorithm to develop two models that helped us achieve the same. The models worked relatively well, and we were successfully able to classify majority of the landslides. While predicting the size of the landslide, we noticed that our models were susceptible to incorrectly predicting the landslides that were generally bigger than the rest. This was due to lack of data and information on very large landslides in the dataset we used. The models did however work very well for small and medium-sized landslides as the data that was available turned out to be sufficient for the same. Our models proved to be efficient and provided an impressive accuracy of close to 70%, considering the amount of data that was available for us to make these predictions.

Our second problem involved classifying the fatalities of each landslide into different groups. Landslides are an extremely dangerous event that has taken the lives of a large number of people all over the world. Between 1998-2017, landslides have caused more than 18,000 deaths worldwide and have affected an estimated 4.8 million people. In mountainous areas with snow and ice especially, rising temperatures and climate change are expected to trigger more and more landslides. As permafrost melts, rocky slopes can become more unstable resulting in a landslide [3], thus making landslides even more dangerous and difficult to predict. The number of fatalities could also be on the rise unless necessary

precautions and safety measures are taken. We used a multi-layer perceptron with an artificial neural network and also the naïve Bayes algorithm to proceed with the classification. We were able to notice that both the models worked particularly similar for all cases and had no specific issues other than the fact being that the dataset took into account only the number of fatalities and not the number of people affected by the landslide. The number of fatalities being binned into different groups also helped improve the accuracy of both the models as we noticed that the number of fatalities were similar in many cases. Predicting the exact number of fatalities proved to be a near impossible task and hence grouping them up improved the accuracy of our model. Our models worked very well, and we were successfully able to classify the landslides into various categories of fatality with an accuracy of around 94% for both the algorithms.

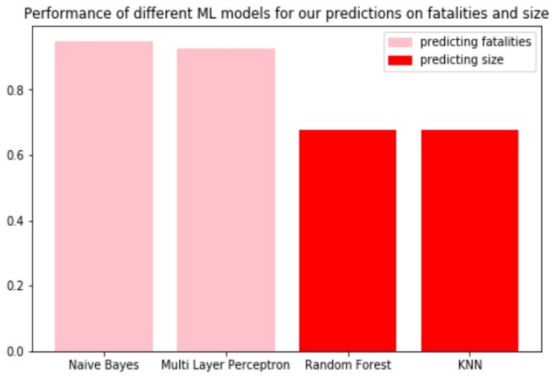

Fig. 5. Performance comparison of the various ML algorithms used.

V. Conclusion

The analysis of the landslide dataset has proven to be an interesting challenge and we gained plenty of insight into what landslides are and how various factors are involved in these events. Our models provided an accuracy of approximately 70% for predicting the landslide size and around 94% for predicting the fatalities. We also found it to be insightful that landslides cause around 25 to 50 deaths a year and cause upto 3.5 billion USD in damages, in the American continents alone. Being such a disastrous calamity, we felt that it was only right, being a part of the next generation, to help provide as much information as possible so that the number of fatalities and cost can be minimized. Aliyah Kabeer and Serena Gomez worked on the algorithms and perfecting the model. Prachi Sengar and Paul John worked on the exploratory data analysis and developing the paper.

References

[1] Kaggle. NASA “Landslides after Rainfall, 2007-2016”, available at https://www.kaggle.com/nasa/landslide-events

[2] Free Weather API Visual Crossing, https://www.visualcrossing.com/

[3] Landslides – World Health Organization, https://www.who.int/health-topics/landslides#tab=tab_1

[4] Landslides: FAO in Emergencies – Food and Agriculture, https://www.fao.org/emergencies/emergency-types/landslides/en/

[5] Martin Kuradusenge, Santhi Kumaran, and Macro Zennaro, “Rainfall-Induced Landslide Prediction Using Machine Learning Models: The Case of Ngororero District, Rwanda”, year 2020, https://www.mdpi.com/1660-4601/17/11/4147

[6] Khaled Taalab, Tao Cheng, and Yang Zhang, “Mapping landslide susceptibility and types using Random Forest”, year 2018, https://www.tandfonline.com/doi/full/10.1080/20964471.2018.1472392