Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 10, October 2021: 2124-2135

A Comparative Analysis of the Machine Learning Techniques for Finding the Relationship Between Soil and Its Appropriate Sub Parameters in Landslide Susceptibility Mapping

Sonam Lhamu Bhutia[1], Samarjeet Borah[2], Indira Swaminathan[3], Amisha Singh[4]
[1,3,4]Department of Information Technology / [2]Department of Computer Applications
Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, India
sonam.bh@smit.smu.edu.in, samarjeet.b@smit.smu.edu.in, indira_201800271@smit.smu.edu.in, amisha_201800281@smit.smu.edu.in

ABSTRACT

Landslides are known to be one of the most common natural disaster in hilly areas. These landslides have caused a huge loss to both property and life [2]. Hence, there arises a need for adoption of proper measures for minimizing the risk of landslides by zoning the areas as per its vulnerability to landslides.

Zonation of landslides can be done using a landslide susceptibility Map[LSM]. Thus, there arises the need for an efficient model for the development of an LSM. Landslides causal factors are very imperative in determining the accuracy of an LSM model as these factors will be considered as input parameters for the model. Some parameters may have its own sub parameters, which may be taken into consideration for the development of a model. Hence, finding the relationship between parameter and its sub parameters may be helpful in the development of an efficient and effective model for an LSM. The role of various machine learning techniques has become very vital in the area of geotechnical applications. The main contribution of the paper is to make a comparative analysis of various machine learning techniques to find the relevance between the main parameter and its sub parameter. Here, soil has been considered as the main parameter. The techniques used are Word2Vec, Sequence Matcher and Term Frequency-Inverse Document Frequency (TF-IDF). The correlation between the parameter and its sub parameter is based on the ratio obtained between the sub parameters and input parameter. The study showed that SequenceMatcher showed better results as compared to Word2Vec and TF-IDF.

Keywords – *Landslide, Word2Vec, SequenceMatcher, Term Frequency-Inverse Document Frequency (TF-IDF), Parameters, Sub Parameters*

1. INTRODUCTION

Every year, during monsoon season, most of the places having hilly and mountainous regions as their geographical structure are at a high risk of landslide. These landslides may pose a threat to both life and property.

Landslide is thus a natural phenomenon, hence methods for accurate detection of the landslide has always caught the attention researchers. However, proper been Thus, development of suitable model for the an accurate LSM has always been a significant area of research. LSM has always been vital for effective land use planning and risk assessment [12].

An extensive research in the quench of development of an accurate model for LSM has been done using qualitative and quantitative methods. However, quantitative methods are most preferred method as these analysis and assessments in this methods are based on mathematical models [7]. However, the qualitative models are subjective and are mostly based on the knowledge of the researcher [8]. In the recent years, machine learning (ML)and deep learning techniques have played a very significant role in the development of significant and robust models for landslide susceptibility mapping [9]. These techniques are not only limited to the development of model for LSM but it is also very vital for other geotechnical applications [9,10].

With the application of ML and deep leaning tools and techniques, appropriate selection of parameters for the model and its efficient assessment has been very effective and accurate [10,11]. Though finding the appropriate and relevant parameters have always attracted the researchers, however, consideration of appropriate sub parameters may yield better results. In this study, the main parameter that has been considered for study is soil. This paper aims to make a comparative analysis of the the ML techniques for finding the relationship between soil and its appropriate sub parameters. The selection of the sub parameters has been done on the basis of the calculated scores achieved by the application of each of the ML techniques. The techniques used are Word2Vec, SequenceMatcher and Term Frequency-Inverse Document Frequency (TF-IDF).

This manuscript is structured as follows – section 2 discusses some related works from literature, section 3 highlights methods and equipment, section 4 contains the results and the work is concluded in section 5.

2. RELATED WORK

Today, machine learning tools and techniques are in high demand for analysis of the data and development of accurate models [13]. Decision making using machine learning tools and techniques has brought about significant changes in the field of research [14]. For the development of an efficient model for LSM, the input parameters play a very essential role to an extent that it can bring about significant changes in the output with a small change in the input parameter [15]. Thus, it is imperative to understand the parameters and its influences on the model. Also, it is equally essential to understand the attributes of these parameters, this will help us understand the influence of sub parameters on its parameters [15,16] and the extent of collinearity between a parameter and its sub parameter [16]. Also, it is imperative to understand that the parameters that are redundant should be ruled out [18].

Most of the models developed for LSM are data dependent models [17], and these data dependent models have adopted Logistic Regression models to find the collinearity between its causative factors by selecting only those parameters with strong coherence. The technique has made a significant contribution in studying the collinearity between the input parameters. The scope of research in the area of LSM demand the selection of optimal and most relevant parameters [18]. Contribution towards the selection of appropriate parameters and for finding the correlation between these parameters using ML techniques like Artificial Neural Network(ANN), Random Forest(RF) [18]. Incorporation of Analytical Hierarchy Process (AHP) along with Geographical Information

System(GIS) have shown significant results in the selection of the appropriate parameters for LSM [19]. Application of hybrid techniques which included the application of Convolution Neural Network with other techniques like Support Vector Machine(SVM) and RF also yielded a much significant result in the selection of appropriate conditioning factors of landslide as input parameters. Application of various ML techniques have been very vital for effective management of landslides, it has been very significant in making proper selection of landslide conditioning factors, where redundant conditioning factors and collinearity between the conditioning factors have been taken into consideration [20]. Application of ML ensemble modelling have also better model performance with better efficacy [21] in the management of landslides.

ML techniques have been very vital for efficient decision making for selection of most appropriate selection of the parameters. However, there is no set guidelines for efficient and effective selection of landslide conditioning factors for a model for LSM [21]. In this study, an attempt using theoretical approach has been made to find the relevance of parameters with its sub parameters by calculating the ratio between the sub parameters and the main parameter. For this experiment, soil has been considered as the main parameter for the study. The concept of text mining has been applied for the theoretical analysis for the study. Text mining techniques used for the study are Word2Vec, SequenceMatcher and TF-IDF.

3. MATERIALS AND METHOD

The experiments for the study has been conducted in an Intel i5 machine with 16 GB of RAM and 4 GB graphics card. Python 3.6 has been used with IDLE as an integrated development environment (IDE). A set of files containing 52 pdf files have been manually created as dataset. The files in the data consists of relevant research papers in the area of landslide susceptibility mapping from various high repute conferences and journals. Only those papers where the LSM models portrayed an accuracy of 70% and more has been taken into consideration [23].

For this experiment, soil has been considered as the main parameter, hence the experiment is conducted using the three techniques to find the most relevant sub parameters of soil. The relevance between soil and its sub parameters has been implemented using Word2Vec, SequenceMatcher TF-IDF. In Word2Vec both Cosine Similarity and Euclidean distance has been applied for calculation of relevance between soil and its sub parameters. Figure.1 shows the flow diagram as per the experiment conducted. Table 1 indicates various python libraries during the conduction of the experiment.

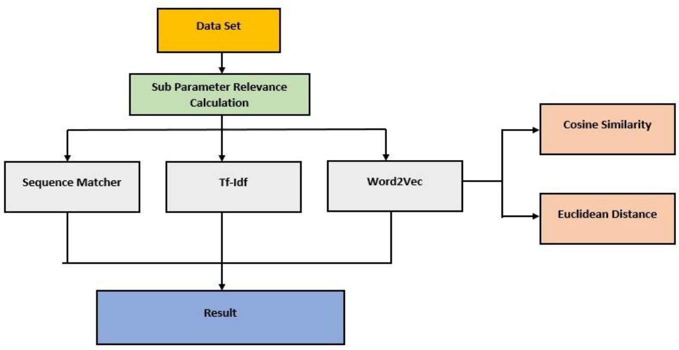

Fig.1 *Flow Diagram for the Proposed Methodology*

Table 1. *Python Libraries Applicable to the experiment*

Library	**Purpose**
NumPy	To support large and multi-dimensional arrays and matrices mathematical functions.
Pandas	To read data from CSV
Gensim	For topic modelling, document indexing and similarity retrieval .
CSV	Save data as CSV
Difflib	For comparison of sets of data
Scikit-Learn	For supporting machine learning algorithms
Matplotlib	For plotting the required plots
PYPDF	To read the Pdf file

3.1 SequenceMatcher

The concept of SequenceMatcher has been applied in the experiment for theoretical analysis for calculations of sub parameters of soil. This technique used to compare pairs of input sequences [25]. This method does not yield minimal edit sequence but is useful to use when we need to find similarity between two words on a character level [26]. Application of the technique to calculate the ratio for finding the sub parameters of the soil has been indicated.

Algorithm

Input: Set of 52 Pdf file manually created for the study
Output: A set of sub-parameters
Step-1: Pre-processing of Data.
Step-2: Calculate the similarity ratio by passing input parameter and one sub parameter at a time into predefined ratio function of class SequenceMatcher
Step-3: Store the output in the csv file

3.2 Word2Vec

Word2Vec is a one of the natural language processing techniques in which algorithm uses simple neural network model which predict the nearby words for each and every word in a sentence [27]. It creates vectors that are distributed numerical representations of words. In this experiment, to calculate the relevance of sub parameters to it parameter, this technique has been applied using the following techniques:

- Euclidean Distance, which represents the shortest distance between two points or vectors and most of the machine learning algorithms including K-Means uses this method to find or measure the similarity between two vectors or words and Cosine Similarity.
- Cosine Similarity, it is a measure which calculates the cosine of the angle between two vectors. This method take comparison between documents on a normalized space as it does not take

magnitude into consideration. It generates the output on basis of how related two vectors are by just looking at the angles between them instead of their magnitudes.

Algorithm

Input: Set of 52 Pdf file manually created for the study
Output: A set of sub- parameters
Step-1: Pre-processing of Data. **Step-2**: Convert all the words into 2D vectors
Step-3: Compute Euclidean Distance. D = $\sqrt{(x_1 - x_2)^2 + (y_2 - y_1)^2}$ **Step-4**: Calculate Cosine Similarity Similarity (A, B) = $\frac{A \,.\, B}{\|A\| \, \|B\|}$
Step-5: Store the output in respective csv file.

3.3 Term Frequency-Inverse Document Frequency(TF-IDF)

This technique is used to quantify a word in a document. In this method usually computation of a weight to each word is done which signifies the importance of the word in a document or corpus [28]. This method is very useful for information Retrieval and text Mining [29]. Application of the technique has been done in the experiment to calculate the ratio between soil and its sub parameters.

Algorithm

Input: Set of 52 Pdf file manually created for the study **Output**: A set of sub- parameters
Step-1: Pre-processing of Data. **Step-2:** Compute TF-IDF vectors using TF-IDF algorithm **Step-3:** Calculate Cosine Similarity Similarity (A, B) = $\frac{A \,.\, B}{\|A\| \, \|B\|}$ **Step-4:** Select the most relevant sub parameters

4. RESULTS & DISCUSSION

The role of parameter selection has always been vital for the development of an efficient model for LSM. The role of sub parameters to parameters are also very significant for the development of the model. In this experiment, the relevant sub parameters of soil have been calculated using SequenceMatcher, Word2Vec, TF-IDF.

4.1 Sub-Parameter Selection using SequenceMatcher

Sub-parameters relevant to soil were calculated using SequenceMatcher, Table 2 shows the calculated sub parameters.

***Table 2**. Calculated sub-parameters using SequenceMatcher*

Sub Parameters	Ratio SM
slope	0.4
soil texture	0.5
soil drainage	0.470588235
slope	0.444444444
soil texture	0.470588235
soil material	0.470588235
soil type	0.615384615
soil tropography	0.4
soil thickness	0.444444444
soil	1
spi	0.571428571
sti	0.571428571
soil,lithology	0.444444444
soil depth	0.571428571
(spi)	0.4
stoniness	0.461538462

4.2 Sub-parameter selection using Word2Vec

Here, results obtained for calculation of the relevant sub parameters of soil using Word2Vec. Here, the calculation has been using both Cosine Similarity and Euclidean distance. Table 3 indicated the sub parameters obtained using Euclidian Distance where the parameters having lesser Euclidian distance are considered to be a better sub parameter and Table 4 indicated the calculation of sub parameters using Cosine Similarity where the sub parameters having higher cosine similarity are considered to be more relevant to its corresponding parameter.

***Table 3**: Calculation of sub-parameters using Euclidean Distance*

Sub Parameters	Ratio Euclidean	Sub Parameters	Ratio Euclidean
slope	0.10863869	subwatershed basin	0.12892127
soil texture	0.09498497	spi	0.13496713
soil drainage	0.09823995	tws	0.07279116
soil effective thickness	0.13560557	planar curvature	0.12742288
soil material	0.10522808	sly view factors	0.11504687
forest map	0.1342415	twi	0.13198629
altitude	0.088549025	catchment area	0.09772854
rock type	0.1340773	slope length (ls)	0.07759537
trophographical	0.056647435	tri	0.12270633

elevation			
water conditions	0.10933321	(twi)	0.12220944
soil tropography	0.12830718	geology	0.07753857
wood density	0.1410393	distance to river and drainage	0.08846421
forest type	0.12684208	distance to lineaments	0.14672063
lithplogy	0.06168561	ndvi,land use	0.14728284
fault buffer	0.08617982	precipitation	0.1157411
valley buffer	0.03537658	(ndvi)	0.1427703
soil	0.002184791	inner texture	0.048746128
profile curvature	0.12841083	stream sediment transport index (sti)	0.14003477
slope gradient	0.044172134	faults and folds	0.10313589
distance to drainage	0.05290663	density of geological boundary	0.0942927
surface area ratio	0.03879488		

Table 4*: Calculation of Sub Parameters using Cosine Similarity*

Sub Parameters	**Ratio Cosine**	**Sub Parameters**	**Ratio Cosine**
soil texture	0.545	distance from lineament	0.928
Slope	0.603	distance from roads	0.521
Aspect	0.59	spi	0.899
Drainage	0.917	tws	0.693
Altitude	0.936	sly view factors	0.586
slope angle	0.85	tpi	0.98
mean water shed	0.995	catchment area	0.995
tree density	0.45	convergence index	0.686
rock type	0.996	slope length (ls)	0.548
ground water	0.439	and normalized difference vegetation index (ndvi)	0.968
trophographical elevation	0.88	tri	0.966
water conditions	0.937	normalized difference vegetation index (ndvi) values	0.988
wood density	0.911	peak ground acceleration (pga)	0.441
soil topographic type	0.663	the slope angle	0.949
forest type	0.979	sediment transport index (sti)	0.809
forest diameter	0.655	topographic wetness index(twi)	0.423

Lithplogy	0.795	(twi)	0.792
fault buffer	0.702	slope aspect	0.423
valley buffer	0.931	Drainage	0.97
Soil	1.004	geology	0.951
profile curvature	0.947	distance to river and drainage	0.565
slope gradient	0.883	overburden depth	0.959
Ndvi	0.644	distance to lineaments	0.961
distance to drainage	0.986	soil,lithology	0.47
road density	0.484	ndvi,land use	1
surface area ratio	0.941	distance tofaults	0.553
subwatershed basin	0.801	Tropography	0.723

4.3 Sub-parameter selection using TF-IDF

Table 5 shows the selection of sub parameter of soil using TF-IDF. One of the main advantage of using this technique is that it helps rule out stop words from the documents, thus making the calculations more efficient.

***Table 5**: Final Calculation of Sub Parameters for TF-IDF*

Sub parameters	**Ratio**
soil texture	0.62039295
soil drainage	0.649350101
soil effective thickness	0.467607999
soil texture	0.62039295
soil material	0.550241648
soil type	0.635866499
soil topography	0.579751013
soil thickness	0.62039295
soil topographic type	0.495595581
Soil	1
soil,lithology	0.602151998
soil depth	0.635866499

4.4 Analysis

In this experiment, calculation of the ratios to find the relevance of soil to its sub parameters was done using Tf-Idf, SequenceMatcher and Word2Vec. Word2Vec was applied using Euclidean distance and Cosine Similarity. As per, the results obtained, SequenceMatcher technique as considered to be efficient among the other techiques used in the experiment. Table 6 gives a summary of the techniques used along with the inferences of each of technique. Figure 2 indicates a comparative analysis of the results obtained using various ML techniques mentioned in the study.

__Table 6__: Summary of Findings and Analysis of Techniques

<table>
<tr><th>Sub-Parameter</th><th colspan="4">Results</th><th colspan="4">Inference</th></tr>
<tr><th></th><th>Euclidean Distance</th><th>Cosine Ratio</th><th>Sequence Matcher</th><th>TF-IDF</th><th>Euclidean Distance</th><th>Cosine Ratio</th><th>Sequence Matcher</th><th>TF-IDF</th></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Slope</td><td>0.121</td><td>0.756</td><td>0.4</td><td>0</td><td rowspan="10">Sub parameters having ratio below 0.15 are considered to be more relevant to soil. However, it was observed that some of the ratios obtained were irrelevant.</td><td rowspan="10">Sub Parameters having ratio above0.4 are considered to be more relevant sub parameters. It was also observed that some of the some sub parameters have shown negative values, which can thus be inferred that these sub parameters may be discarded.</td><td rowspan="10">It was observed that the most relevant sub parameters obtained ratio more than 0.4. The results obtained showed that the ratio of all the sub parameters had value more than 0.4.</td><td rowspan="10">The ratio of the sub parameters more than 0.4 was considered as relevant ones. However, some of the sub parameters had values more than 0.4, which indicated irrelevant results.</td></tr>
<tr><td>Soil Texture</td><td>0.279</td><td>-0.057</td><td>0.5</td><td>0.620</td></tr>
<tr><td>Soil Drainage</td><td>0.097</td><td>0.946</td><td>0.470</td><td>0.649</td></tr>
<tr><td>Soil Material</td><td>0.132</td><td>-0.0463</td><td>0.47</td><td>0.550</td></tr>
<tr><td>Soil Type</td><td>0.263</td><td>-0.646</td><td>0.62</td><td>0.636</td></tr>
<tr><td>Soil Topography</td><td>0.114</td><td>-0.94</td><td>0.4</td><td>0.579</td></tr>
<tr><td>Soil Thickness</td><td>0.421</td><td>-0.56</td><td>0.44</td><td>0.620</td></tr>
<tr><td>Soil</td><td>0.001</td><td>0.994</td><td>1</td><td>1</td></tr>
<tr><td>Soil Depth</td><td>0.113</td><td>0.715</td><td>0.57</td><td>0.636</td></tr>
<tr><td>Stoniness</td><td>0.285</td><td>-0.999</td><td>0.46</td><td>0</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

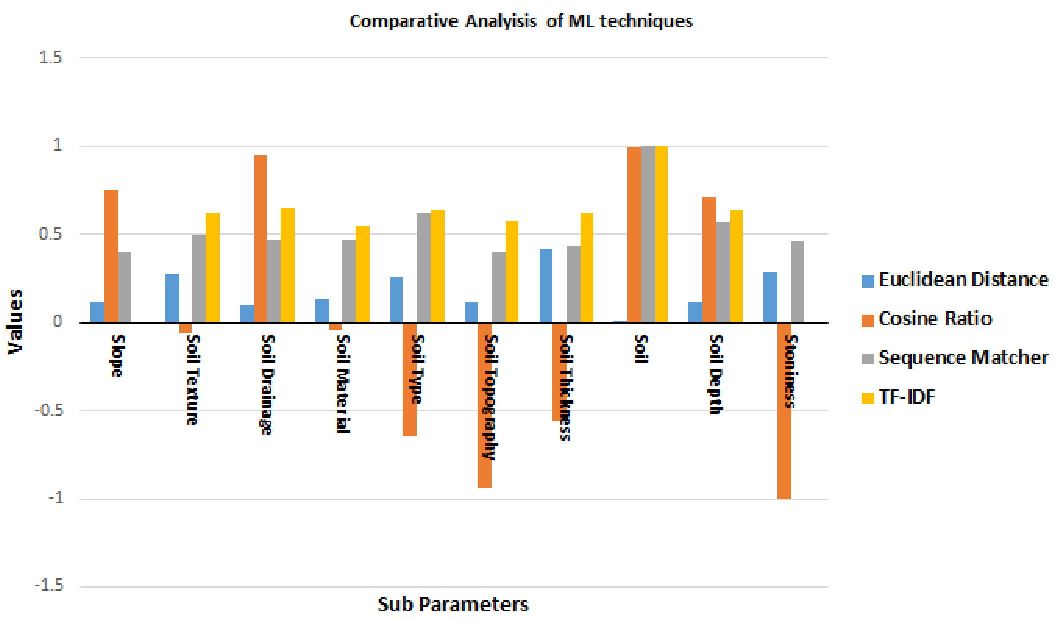

Fig.2. *Comparative Analysis of the ML Techniques used in the Experiment*

5. CONCLUSION

An attempt to find the relevant sub parameters of soil which may be considered for development of a suitable model fo**r** LSM. This LSM may help us in effective management of landslides and proper urban planning [24]. In the study, only the sub parameter relevant to soil has been taken into consideration for comparative analysis, the other sub parameters obtained has been discarded. The study indicated that the results obtained using SequenceMatcher was the most preferred no undesirable results were obtained using this technique and the results obtained were having the ratio more than 0.4. However, with Word2Vec and TF-IDF, along with the relevant sub parameters, some irrelevant sub parameters were also obtained. In Word2Vec, these sub parameters that were not relevant to the experiment were also observed to have the distance below 0.15 in Euclidean distance and above 0.4 in cosine similarity. In TF-IDF, it was observed that some irrelevant sub parameters were also obtained. Dataset for the study was manually prepared and its manual scraping was time consuming. As a large number of sub parameters were obtained, visualization of the output is challenging. The present experiment has been carried out only for one input parameter, soil.

Future scope may involve the calculation of relevance of sub parameters to its main parameter for other input parameters as well. This may help in better understanding of the dependency of the parameters and its relevance for the development of a suitable model for LSM. The future work may also include the automated scraping of the data which would be much efficient and effective. Other techniques may also be experimented which may help in obtaining better results

REFERENCES

1. Sikkim Physical Features, Geography of Sikkim, Physical Features of Sikkim. (n.d.). www.indovacations.net. Retrieved November 8, 2021, from https://www.indovacations.net/sikkim tourism/Sikkim_PhysicalFeatures.htm
2. Sharma, L. P., Patel, N., Ghose, M. K., & Debnath, P. (2014). Development and application of Shannon's entropy integrated information value model for landslide susceptibility assessment and zonation in Sikkim Himalayas in India. Natural Hazards, 75(2), 1555–1576. https://doi.org/10.1007/s11069-014-1378-y

3. Sonker, I., Tripathi, J. N., & Singh, A. K. (2021). Landslide susceptibility zonation using geospatial technique and analytical hierarchy process in Sikkim Himalaya. Quaternary Science Advances, 4, 100039. https://doi.org/10.1016/j.qsa.2021.100039.
4. Geography of Sikkim, Geographical Features of Sikkim, Sikkim Geography. (n.d.). Www.thisismyindia.com. Retrieved November 9, 2021, from https://www.thisismyindia.com/sikkim/sikkim-geography.html
5. Achour, Y., & Pourghasemi, H. R. (2020). How do machine learning techniques help in increasing accuracy of landslide susceptibility maps? Geoscience Frontiers, 11(3), 871-883. doi: 10.1016/j.gsf.2019.10.001
6. Wang, H., Zhang, L., Yin, K., Luo, H., & Li, J. (2021). Landslide identification using machine learning. Geoscience Frontiers, 12(1), 351-364. doi: 10.1016/j.gsf.2020.02.012
7. Panchal, S., & Shrivastava, A. Kr. (2021). Landslide hazard assessment using analytic hierarchy process (AHP): A case study of National Highway 5 in India. Ain Shams Engineering Journal. https://doi.org/10.1016/j.asej.2021.10.021
8. Shano, L., Raghuvanshi, T. K., & Meten, M. (2020). Landslide susceptibility evaluation and hazard zonation techniques – a review. Geoenvironmental Disasters, 7(1). https://doi.org/10.1186/s40677-020-00152-0
9. Wang, H., Zhang, L., Yin, K., Luo, H., & Li, J. (2021). Landslide identification using machine learning. *Geoscience Frontiers*, *12*(1), 351–364. https://doi.org/10.1016/j.gsf.2020.02.012
10. Youssef, A. M., & Pourghasemi, H. R. (2021). Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. *Geoscience Frontiers*, *12*(2), 639–655. https://doi.org/10.1016/j.gsf.2020.05.010
11. Sun, D., Wen, H., Wang, D., & Xu, J. (2020). A random forest model of LANDSLIDE Susceptibility mapping based on hyperparameter optimization using Bayes algorithm. Geomorphology, 362, 107201. https://doi.org/10.1016/j.geomorph.2020.107201
12. Kaur, H., Gupta, S., Parkash, S., Thapa, R., Gupta, A., & Khanal, G. C. (2019). Evaluation of landslide susceptibility in a hill city of Sikkim Himalaya with the perspective of hybrid modelling techniques. *Annals of GIS*, *25*(2), 113–132. https://doi.org/10.1080/19475683.2019.1575906
13. *Machine learning tools - javatpoint*. www.javatpoint.com. (n.d.). Retrieved November 15, 2021, from https://www.javatpoint.com/machine-learning-tools.
14. Jiang, H., Mao, H., Lu, H., Lin, P., Garry, W., Lu, H., Yang, G., Rainer, T. H., & Chen, X. (2021). Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. *International Journal of Medical Informatics*, *145*, 104326. https://doi.org/10.1016/j.ijmedinf.2020.104326
15. Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, *32*(2), 135–154. https://doi.org/10.1007/bf00547132
16. Ghanoum, T. (2021, July 4). *Why multicollinearity isn't an issue in machine learning*. Medium. Retrieved November 16, 2021, from https://towardsdatascience.com/why-multicollinearity-isnt-an-issue-in-machine-learning-5c9aa2f1a83a.
17. Süzen, M. L., & Kaya, B. Ş. (2012). Evaluation of environmental parameters in logistic regression models for landslide susceptibility mapping. *International Journal of Digital Earth*, *5*(4), 338–355. https://doi.org/10.1080/17538947.2011.586443
18. Alqadhi, S., Mallick, J., Talukdar, S., Bindajam, A. A., Van Hong, N., & Saha, T. K. (2021). Selecting optimal conditioning parameters for landslide susceptibility: An experimental research on Aqabat al-Sulbat, Saudi Arabia. *Environmental Science and Pollution Research*. https://doi.org/10.1007/s11356-021-15886-z
19. Senouci, R., Taibi, N.-E., Teodoro, A. C., Duarte, L., Mansour, H., & Yahia Meddah, R. (2021). GIS-based expert knowledge for landslide susceptibility mapping (LSM): Case of Mostaganem Coast District, west of Algeria. *Sustainability*, *13*(2), 630. https://doi.org/10.3390/su13020630
20. Pham, B. T., Pradhan, B., Tien Bui, D., Prakash, I., & Dholakia, M. B. (2016). A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of uttarakhand area (india). *Environmental Modelling & Software*, *84*, 240–250. https://doi.org/10.1016/j.envsoft.2016.07.005
21. Di Napoli, M., Carotenuto, F., Cevasco, A., Confuorto, P., Di Martire, D., Firpo, M., Pepe, G., Raso, E., & Calcaterra, D. (2020). Machine Learning Ensemble Modelling as a tool to improve landslide susceptibility mapping reliability. *Landslides*, *17*(8), 1897–1914. https://doi.org/10.1007/s10346-020-01392-9

22. Shahabi, H., & Hashim, M. (2015). Landslide susceptibility mapping using GIS-based statistical models and remote sensing data in tropical environment. *Scientific Reports*, *5*(1). https://doi.org/10.1038/srep09899
23. Bhutia, S. L., Borah, S., Pradhan, R., & Sharma, B. (2020, December). An Experiment on Parameter Selection for Landslide Susceptibility Mapping using TF-IDF. Journal of Physics: Conference Series, 012029. https://doi.org/10.1088/1742-6596/1712/1/012029
24. Gaidzik, K., & Ramírez-Herrera, M. T. (2021). The importance of input data on landslide susceptibility mapping. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-98830-y
25. Resna annual conference - 2020. Intuitive Sequence Matching Algorithm Applied To A Sip-And-Puff Control Interface For Robotic Assistive Devices. (n.d.). Retrieved November 20, 2021, from https://www.resna.org/sites/default/files/conference/2020/NewEmergingTechnology/70Schweitzer.html.
26. _classtype sequencematcher. difflib.SequenceMatcher. (n.d.). Retrieved November 20, 2021, from http://epydoc.sourceforge.net/stdlib/difflib.SequenceMatcher-class.html.
27. Johnson, D. (2021, October 29). *Word embedding tutorial: Word2vec model gensim example*. Guru99. Retrieved November 20, 2021, from https://www.guru99.com/word-embedding-word2vec.html#3.
28. *Introduction to term frequency - inverse document frequency (TF-IDF): Natural language processing*. INDIAai. (n.d.). Retrieved November 20, 2021, from https://indiaai.gov.in/article/introduction-to-term-frequency-inverse-document-frequency-tf-idf-natural-language-processing.
29. *Understanding TF-ID: A simple introduction*. MonkeyLearn Blog. (2019, May 10). Retrieved November 20, 2021, from https://monkeylearn.com/blog/what-is-tf-idf/.