NEPAL OPEN UNIVERSITY
नेपाल खुला विश्वविद्यालय

Faculty of Science Health & Technology Masters in Geoinformatics

Landslide Susceptibility Analysis Using Machine Learning Models: A Case Study Of Baglung District, Nepal

Project Members

Ashish Chalise (77242007)

Bipin Sinjali (77242050)

Smaran Dahal (77242011)

November 2021

Table of Contents

List of figures

List of maps

List of tables

1. Abstract:

The classification method is used to identify the sensitivity of the landslides in Baglung district. Landslide inventory is prepared from landslide data collected from google earth. Along with, the SRTM (30 m) data is used to map the slope, elevation and curvature. Other landslide factors are collected from data available from ICIMOD portal. In the approach, the landslide and non landslide points are sampled from the normalized Frequency Ratio(FR) classified raster file that present safe and hazard i.e; 0, and 1 respectively. Moving forward, the result are prepared with two different decision tree. A landslide location map with in the ROI has studied from the Classification And Regression Tree (CART) decision tree model and Random Forest model. The model has help to train, test, and accuracy check of datasets. Moving forward the accuracy of the result prepared from both model are compared with each other.
Keywors: Landslide Susceptibility, CART, FR, decision tree, sampling, hazard.

2. Introduction:

The unstable slope land of the earth surface that make movement of rocks, soils, earth, or debris of the sloped area can be say landslide. Landslides can cause, or occur due to various factors i.e; earthquakes, rainfall, soil type, climate change, geological, hydrological, geomorphological conditions, and other (Geographic). Also, it can be say landslide is movement of earth materials from slow to rapid downslope that triggered by a wide variety of natural process (Health, 11 May 2020). Beside that landslide occur due to human activities. Massive human migration, urban development, industries and many more activities are doing, which has led to landslide. Landslides are more responsible for much more economic damage and loss of life each year, and millions of people have to reallocated, that has led towards poverty, low income source, mental health, and forceful crime (Health, 11 May 2020).

Thousand and thousand square kilometres area are destroyed by the landslide in the world. Every country in the world are facing landslide as a major natural disaster. The top five countries with the highest risk of landslides are Italy, Austria, China, The Philippines, and Ethopia with more than 7,500, 6,000, 5,600, 4,800, 4,800 square miles respectively (Watch, 2021).

Likewise, dozens of natural hazards and human induced disasters has exposed in the Nepal. Every year thousands of people have lost their lives and millions of properties have been damaged due to landslides occurring around the Nepal. Thus, major incidents for death are flood, landslide, thunderbolt, fire, cold wave, high altitude, and heavy rainfall (Affairs, 2019). Nepal Disaster Risk Reduction Portal data shows, in a decade period of time 2,386 times landslides incident has occurred in Nepal and has led to third highest natural disaster incidents.

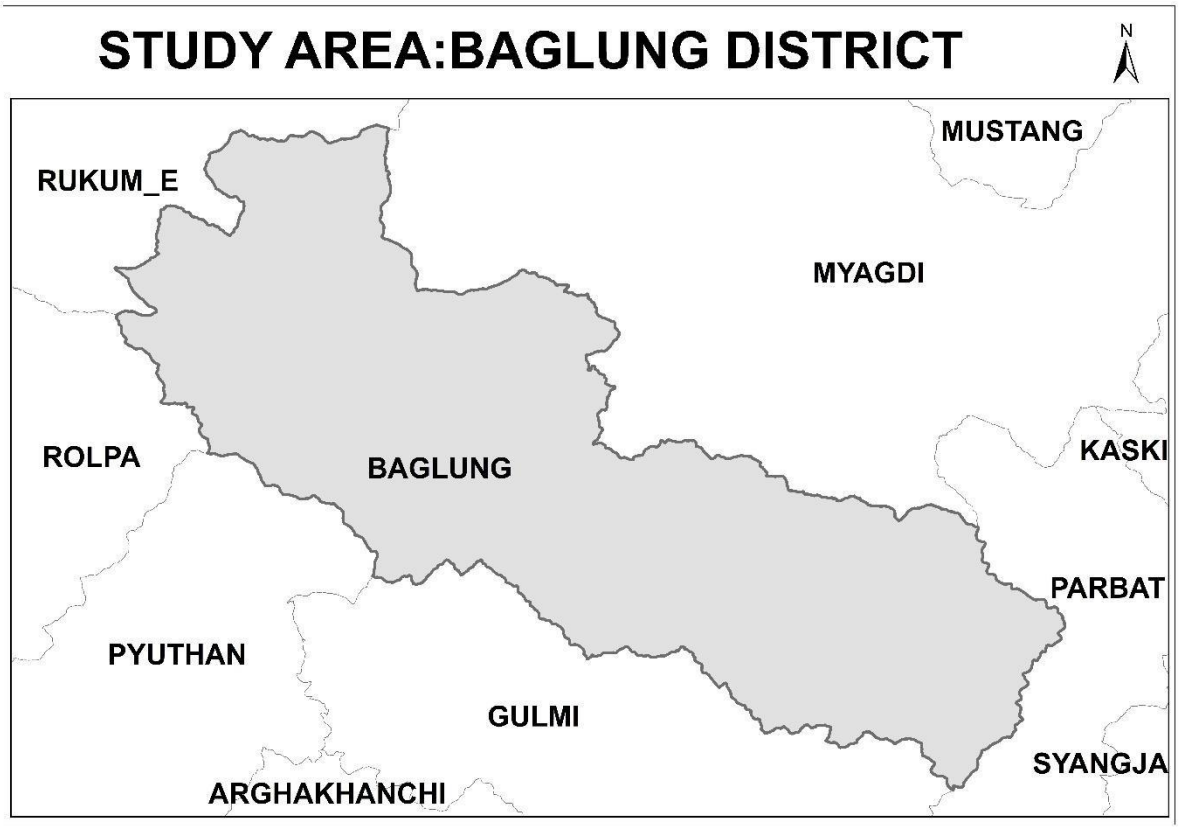
Landslide is one of the very common natural hazards in the hilly region of Nepal. In Nepal, where two third of the total area falls in hilly and mountainous region, landslides represent a major constraint on development, causing high levels of economic loss and substantial numbers of fatalities each year. Rugged and step topography, unstable geological structures, soft and fragile rocks, along with concentrated and prolonged heavy rainfalls during monsoon periods collectively cause severe land sliding problems and related phenomena in the mountainous part of Nepal (Acharya, 2018).

Analysing the Nepal Disaster Risk Reduction portal dataset, during the last two years 8 types of major disaster were recorded in Nepal with a total number of 6,541 disaster incidents. Of the total incidents, the total number of landslide occurred was recorded 758 and has recorded as second highest natural disaster incident in Nepal. Landslides has taken 383 lives, and has damaged estimated loss of rupees 85.51 million (Portal, 2021).

Researcher are doing and applying various technic to minimize the landslide disaster, and reduce impact on human lives, properties. Despite doing that, there have rare standard method that can produce reliable prediction of landslides incident. Thus, it has become singular issue for the proper study. Therefore, we have taken as a mini project and it will play a role to find out the sensitivity areas and will predict with the help of two different type of machine learning algorithms namely: Classification And Regression Tree (CART) and Random Forest(RF) .

3. Study area:

The study area is Baglung district of Gandaki Province, Nepal. From the geological aspect, it lies on the Himalayan range which has multiple characteristics. The Nepal Disaster Risk Reduction Portal shows that, Baglung district has recorded as the highest number of landslides affected in past decade (Portal, 2021). In today's world new technology software and resources for the analysis of spatial and non-spatial data can be done in the sophisticated functionality for the integration. Therefore, with this mini project and use of model for the landslide prediction it will help to identify most haphazard landslide areas and could help to minimize the property loss, human lives, and management of the landslides. Likewise, this district depend upon the agriculture can be helpful for the management and planning of disaster and increase the livelihood of local people (Development, 2020). The study Area map is shown below:



4. Objective:

The overall goal of the mini project is to implement the CART and RF models to identify the highly possible landslides hazard areas in the district Baglung district of Nepal. Going on the specific objectives of the research are:

- To prepare the landslide inventory map.
- To develop and apply models for identification of highly sensitive area.
- To compare the result from two different model.

5. Statement of problems:

Every project have some limitation, it is almost impossible to explore the full project without any issues or limitation. Since, our project is about landslides susceptibility and depend upon natural phenomena. The following are the statement of problems for the mini project:
First statement: The primary data collection would have been better for landslide inventory, but physically it is less chances to visit the Baglung district for primary data collection due to time limitation.
Second statement: Multiple factors plays a role during the modelling of landslide susceptibility. To gain the different factors we have to extract and use the data from various source and may not available latest data in the data source.

6. Literature Review:

Nepal covers almost one third of total Himalayan mountain ranges with 83% low to high mountainous area. Most of the population of Nepal lives in the Midlands, and for this reason, this zone is intensively cultivated [1]. Every year, impact of natural hazard results in huge loss of economy, environment and also human lives get affected. Floods and landslides are among the most destructive natural hazard [2].
Landslide susceptibility mapping can be used for efficient planning and management of natural resources and can be a useful tool for a region's sustainable development. Landslide susceptibility mapping has been implemented in recent decades as the subject of research around the world. The factors contributing to the occurrence of landslides include lithology, climatic, morphometric, and human factors, whereby road construction, settlements and landuse changes are the main anthropogenic factors for landslide. In the last three decades, machine learning ensemble approaches have been applied in different fields due to their superior performance capabilities and the ability to deal with complex and varying data. The geographic information system (GIS), combined with R programming, is a strong and effective tool for hazard mapping that has provided appropriate and meaningful results in landslide susceptibility mapping The benefits and advantages of machine learning ensemble models are not only that these techniques provide transparent computations but they also lead to more accurate models. Thus, the discovery of new landslide modeling technologies, processes, and models are essential [3].
Classification and Regression Trees (CART) have been used for predictive modelling in Machine Learning. ART algorithm provides a foundation for important algorithms like bagged decision trees, random forest and boosted decision trees.

7. Methodology:

7.1. Data:

The landslides occurs due to the function of direct and indirect natural and human factors. The SRTM DEM file with 30 m resolution file is used from the Landsat 8 – EarthExplore, and from other source like Google Earth is used to preparation landslide inventory data. The data of study area were gathered and updated with the available information from digitization. Also, Nepal Disaster Risk Reduction Portal (DRR)1 data were used to make cross validation of landslide inventory (Portal, 2021).

7.2. Software:

Since, this project falls under the Geospatial Programming course. Almost all task are done within the R programming and its extension tools. However, ArcGIS is also used for prepare data distance to river, distance to road and distance to fault. Likewise, Microsoft Excel and Word were used for data management and report preparation.

7.3. Methods:

The first task is to review literatures regarding the landslide susceptibility. Afterwards different thematic layers, and data pre-processing are prepared from the parameters selected for the study. These different causative factors are analysed. Likewise a landslide inventory of the study area is prepared. After that the Landslide Susceptibility Index (LSI) are prepared using Random Forest and classification And Regression Tree (CART) classification decision tree model with the training and test datasets i.e; 70% and 30 % respectively. Afterwards, the result are reclassified to a category map from which the landslide susceptibility zonation map were derived. The map contains four classes of relative landslide hazards: low, medium, high, and very high risk. At last to check the accuracy of the model two different approaches namely: the confusion matrix data frame, and Receiver Operating Characteristic Curve (ROC) on the test data dataset are done.

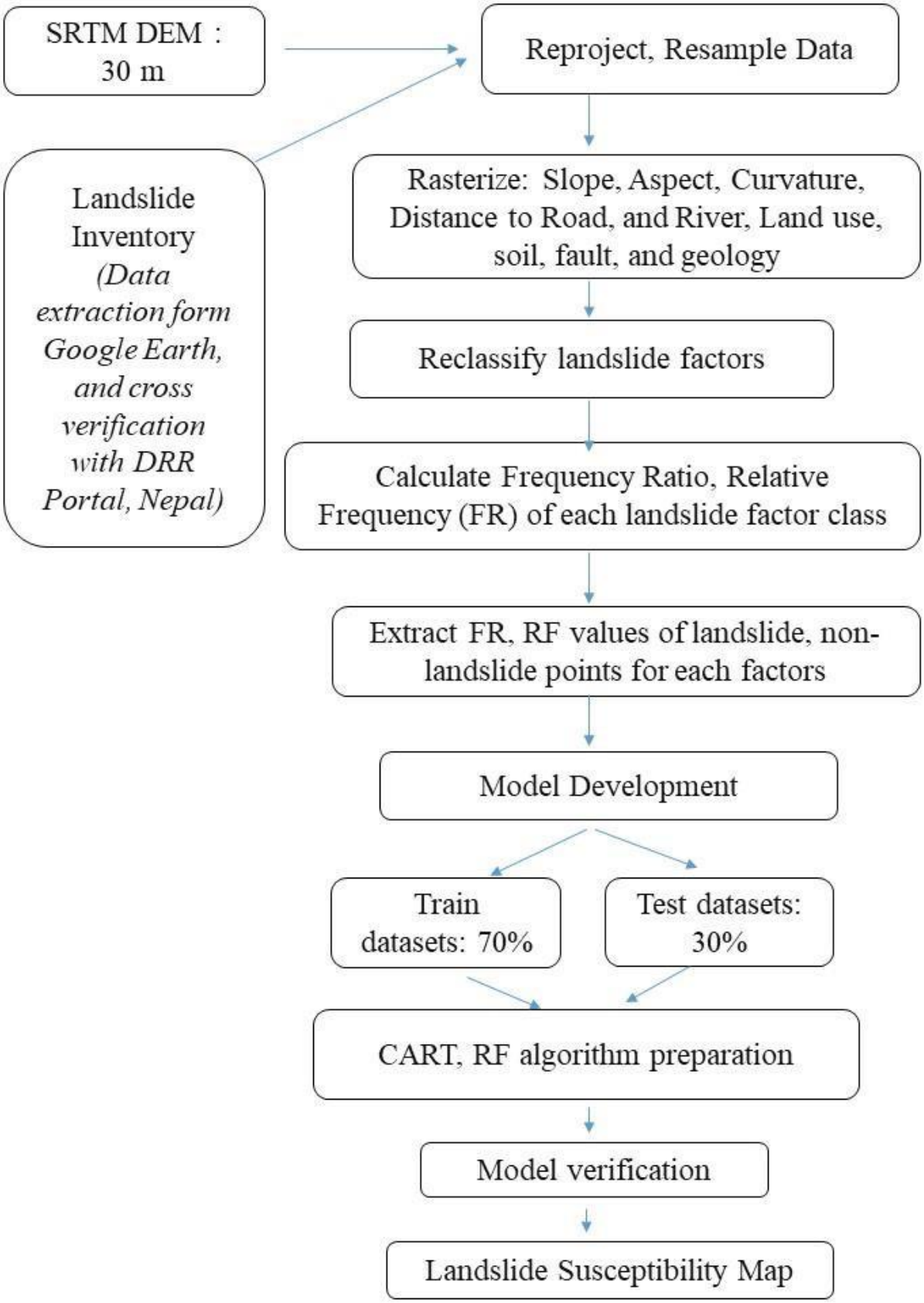
7.3.i. Landslide inventory map:

The landslide inventory mapping is a systematic mapping of existing landslides in a Region of Interest (ROI) using different technique such as field survey, satellite images, air phots, and historical landslide records (literature). Listing out the landslide inventory, it given a spatial distribution of landslides. With respect to different technique we have use online Google Earth to extract the spatial distribution of landslide. Along with that, for the verification of the inventory, the literature record managed by the Nepal Disaster Risk Reduction Portal (DRR) has used. In this mini project, 89 spatial points of landslide is prepared form the online for the study of landslide susceptibility map.

Multiple map are pre-processed from the DEM file of the Baglung district and has reclassified to each susceptibility factor. The following maps shows the reclassified of the individual susceptibility factors.

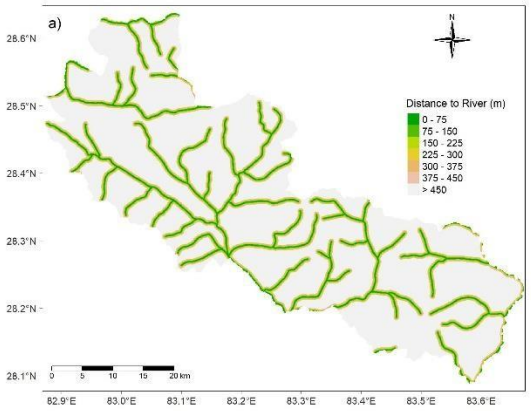
1 The DRR portal data are as off October 6, 2021.

Figure 1 : Method for the preparation of landslides susceptibility

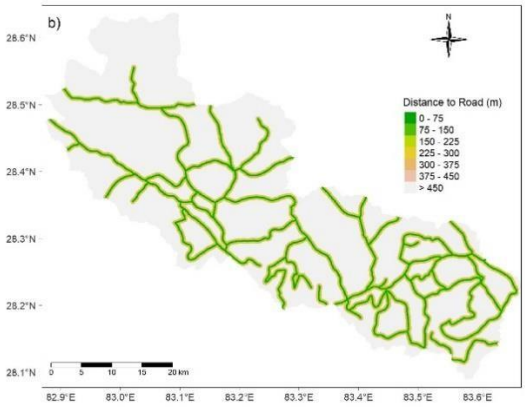


Landslide Susceptibility Factor Maps

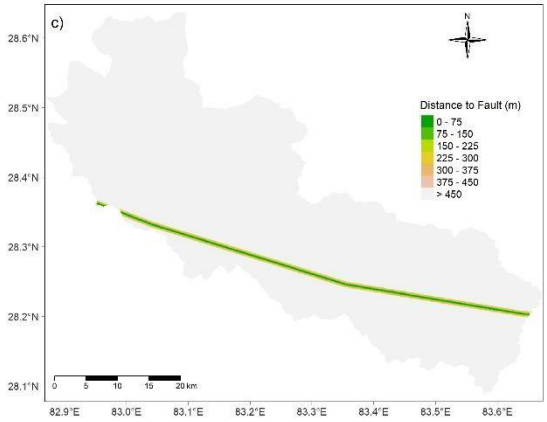
Map 1: River reclassified



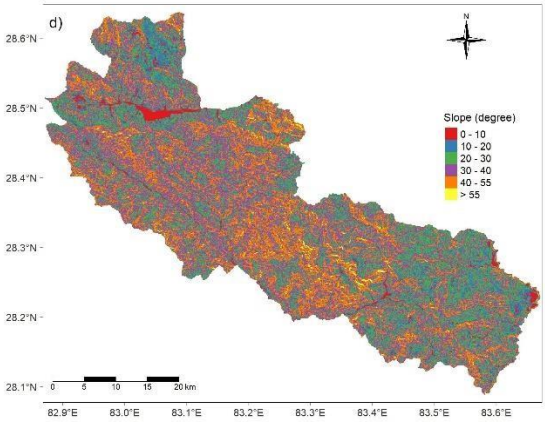
Map 2: Road reclassified



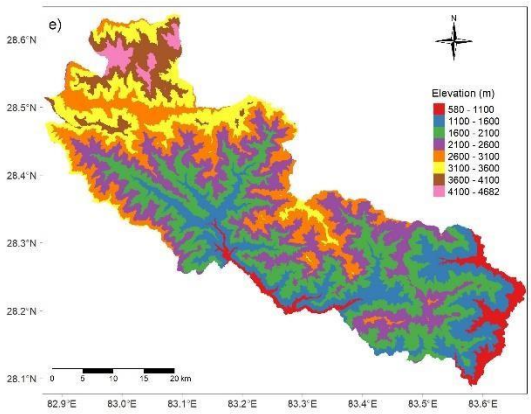
Map 3: Fault reclassified



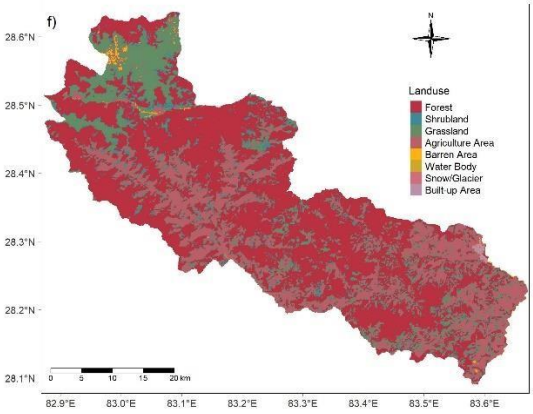
Map 4: Slope reclassified



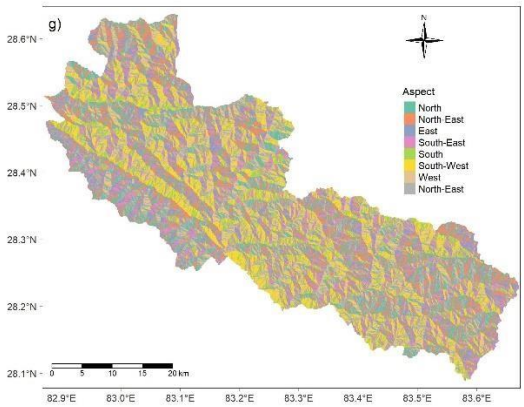
Map 5 : Elevation reclassified



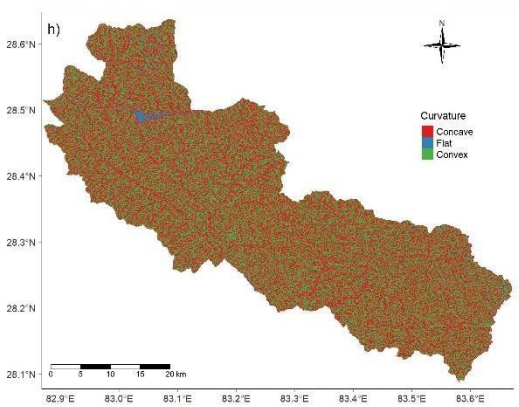
Map 6 : Landuse reclassified



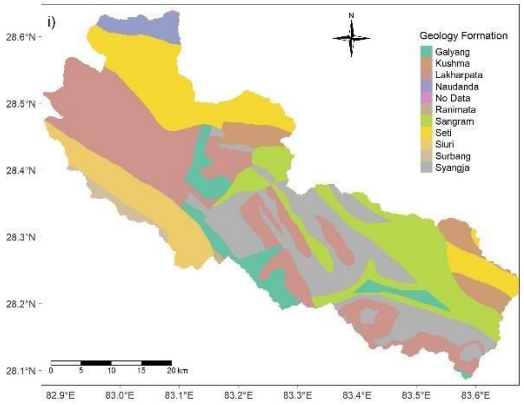
Map 7 : Aspect reclassified



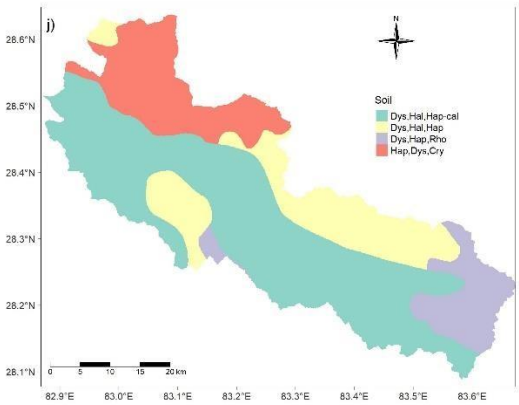
Map 8 : Curvature reclassified



Map 9 : Geology reclassified



Map 10 : Soil reclassified



7.3.ii. Slope map:

The slope map does plays crucial role to develop landslide susceptibility because it directly related to slope angle, and he major parameters of slope stability analysis if the slope angle (Lee and Min, 2001). In the slope map, six slope categories are classified and the pixel values in the each classes are determined. The below table shows the highest landslide occurrence is in between 40 to 55 degree with the number of 43.

Table 1 : Frequency ratio of slope to landslide occurrences

Data layers	Class	Pixel	Number of landslide pixel	FR	RF
Slope (Degree)	0-10	71389	1	0.3196	0.0468
	10-20	233987	2	0.1950	0.0286
	20-30	539474	8	0.3383	0.0496
	30-40	755211	33	0.9970	0.1460
	40-55	413001	43	2.3755	0.3480
	>55	17545	2	2.6008	0.3810
Sum				6.8262	

7.3.iii. Aspect map:

It is believed, aspect is an important factor in preparation of landslide susceptibility map. Also, it is connected with the various factor such as exposure to sunlight, drying winds, rainfall, and discontinuities may affect the occurrence of landslides (Carrara A and Cardinali M, 1991). The relation between the aspect and landslide is shown in the below table (Table 2). The aspect has classified into eight classes. South part has the hight number of landslide with 30 and followed by 19 in South-east part of Baglung district.

Table 2 : Frequency ratio of aspect to landslide occurrences

Data layers	Class	Pixel	Number of landslide pixel	FR	RF
Aspect	North (0-22.5 and 337.5-360)	247024	4	0.3695	0.0593
	North-east (22.5-67.5)	290227	4	0.3145	0.0505
	East (67.5-112.5)	273535	10	0.8341	0.1338
	South-east (112.5-157.5)	268795	19	0.0000	0.0000
	South (157.5-202.5)	254324	30	2.6913	0.4319
	South-west (202.5-247.5)	260876	17	1.4868	0.2386
	West (247.5-292.5)	210055	4	0.4345	0.0697
	North-west (292.5-337.5)	225771	1	0.1011	0.0162
Sum				6.2317	

7.3.iv. Curvature map:

The commonly used parameter in landslide hazard analysis curvature. Curvature of the hillside in a horizontal plane or the curvature of the contours on a topographic map. Hillsides can be subdivided into regions of concave outward plan curvature called hollows, convex outward plan curvature called noses, and straight contours called planar regions. Also, hollows have a slightly higher probability for landslides than noses (Gregory C, 2006). As mention, the below table proved that concave have the highest number of landslides with 52 (Table 3).

Table 3 : Frequency ratio of curvature to landslide occurrences

Data layers	Class	Pixel	Number of landslide pixel	FR	RF
Plan Curvature	Concave (<-0.05)	960281	52	1.2432	0.4920
	Flat (-0.05-0.05)	98185	2	0.4676	0.1851
	Convex (<0.05)	984750	35	0.8160	0.3229
Sum				2.5268	

7.3.v. Elevation map:

Elevation map will help to determine the minimum and maximum heights of landslide occurrence within the ROI. To identify, at what elevation have the highest number of landslide, the below table 4 shows the frequency of landslide occurrence. We can see elevation between 1600 – 2100 m has the highest frequency with 35 and followed by the elevation class of 580 – 1100 m. On the lowest frequency, the class of 4100 m and above does not have any landslide in the Baglung district.

Table 4 : Frequency ratio of elevation to landslide occurrence

Data layers	Class	Pixel	Number of landslide pixel	FR	RF
Elevation (m)	580 - 1100	94694	16	3.8790	0.4612
	1100 - 1600	317971	12	0.8664	0.1030
	1600 - 2100	479825	35	1.6746	0.1991
	2100 - 2600	448409	12	0.6144	0.0730
	2600 - 3100	333729	8	0.5503	0.0654
	3100 - 3600	223530	4	0.4108	0.0488
	3600 - 4100	110658	2	0.4149	0.0493
	4100 - 4681.106	34400	0	0.0000	0.0000
Sum				8.4105	

7.3.vi. Landuse/landcover map:

The increase in soil strength due to root reinforcement has great potential to reduce the frequency of landslide occurrence. Also, the vegetation cover introduces some mechanical changes through soil reinforcement and slope loading (Sivakami C, 2014). In this study, an image has eight classes of the landslide map. The agriculture area have the highest number of landslide frequency with 44 followed by the forest area with 22 landslides.

Table 5 : Frequency ratio of landuse to landslide occurrence

Data layers	Class	Pixel	Number of landslide pixel	FR	RF
Land Use	Forest	1093710	22	0.4629	0.0967
	Shrubland	152931	6	0.9028	0.1886
	Grassland	274459	17	1.4253	0.2977
	Agriculture area	507153	44	1.9963	0.4170
	Barren area	12506	0	0.0000	0.0000
	Water Body	3278	0	0.0000	0.0000
	Snow/glacier	396	0	0.0000	0.0000
	Built-up area	3485	0	0.0000	0.0000
Sum				4.7872	

7.3.vii. Geology map:

The Lakharpata Formation of geology has the highest number of landslides in the study area, and followed by the Syangja Formation of geology. The following table shows in more detail.

Table 6 : Frequency ratio of geology to landslide occurrence

Data layers	Class	Pixel	Number of landslide pixel	FR	RF

	Galyang Formation	135590	8	1.3968	0.1672
	Kushma Formation	102478	8	1.8482	0.2212
	Lakharpata Formation	579831	22	0.8983	0.1075
	Naudanda Formation	48656	0	0.0000	0.0000
Geology	No Data	721	0	0.0000	0.0000
	Ranimata Formation	2697	0	0.0000	0.0000
	Sangram Formation	338525	12	0.8392	0.1004
	Seti Formation	311068	8	0.6089	0.0729
	Siuri Formation	166479	11	1.5643	0.1872
	Surbang Formation	26640	0	0.0000	0.0000
	Syangja Formation	394349	20	1.2007	0.1437
Sum				8.3563	

7.3.viii. Soil map:

Landcover with different soil characteristics has diverse effects in the occurrence of landslides. It does not only affects the development degree of landslides in the areas, but also determines the type and scale of landslide (XIanyu Yu, 2021). The soil map has four categories and the highest number frequency of landslide is in soil type of Dystrochrepts, Halpumbrepts, Haplustalfs-calcarious Materials with 46. The least number of landslide is in the soil type of Haplumbrepts, Dystrochrepts, Cryumbrepts.

Table 7 : Frequency ratio of soil to landslide occurrence

Data layers	**Class**	**Pixel**	**Number of landslide pixel**	**FR**	**RF**
Soil	Dystrochrepts, Halpumbrepts, Haplustalfscalcarious Materials	1059531	46	0.9957	0.2401
	Dystrochrepts, Haplumbrepts, Haplustalfs	407886	23	1.2932	0.3118
	Dystrochrepts, Haplustalfs, Rhodustalfs	218159	14	1.4718	0.3548
	Haplumbrepts, Dystrochrepts, Cryumbrepts	355591	6	0.3870	0.0933
Sum				4.1477	

7.3.ix. Distance to rivers map:

The distance to river map shows the buffer zones with seven different classes. It does not effect in occurrence of landslide directly. Despite that, the proximity of the slope to the drainage structures is important factor in terms of stability because, it may affect stability of slopes or by saturating the lower part of material until the water level increase (Sivakami C, 2014). Several buffer zones the table 7 shows greater than 450 m has the highest number of landslide our study area, and the buffer zone between 150 to 225 has the lowest landslide occurrence.

Table 8 : Frequency ratio of distance to river to landslide occurrence

Data layers	**Class**	**Pixel**	**Number of landslide pixel**	**FR**	**RF**

	0-75	99306	9	2.1218	0.2494
	75-150	95611	4	0.9794	0.1151
	150-225	77017	2	0.6080	0.0715
Distance to Rivers	225-300	89641	7	1.8282	0.2149
	300-375	75279	4	1.2440	0.1462
	375-450	85158	3	0.8248	0.0969
	>450	1585022	60	0.9010	0.1059
Sum				8.5071	

7.3.x. Distance to roads map:

This is one off the factor amongst the multiple factors for landslide occurrence. In the study, the area are divided into seven categories. The number of landslide were determined with the buffer zone form the road to landslide inventory. The buffer zone greater than 450 m has the highest number of landslide with 58, and can found one landslide occurrence in the buffer zone between 300 to 375 m.

Table 9 : Frequency ratio of distance to roads to landslide

Data layers	**Class**	**Pixel**	**Number of landslide pixel**	**FR**	**RF**
	0-75	96387	9	2.1415	0.2350
	75-150	92265	7	1.7400	0.1910
	150-225	74656	7	2.1504	0.2360
Distance to Roads	225-300	87254	3	0.7885	0.0865
	300-375	72868	1	0.3147	0.0345
	375-450	82670	4	1.1097	0.1218
	>450	1535067	58	0.8665	0.0951
Sum				9.1114	

7.3.xi. Distance to fault:

In our study area we have only one fault line. Despite having one fault line we have buffered the zone into seven classes. The total number of 86 landslides are the recorded in distance of greater than 450 m from fault to landslide inventory.

Table 10 : Frequency ratio of distance to fault to landslide

Data layers	**Class**	**Pixel**	**Number of landslide pixel**	**FR**	**RF**
	0-75	12707	0	0.0000	0.0000
	75-150	12150	0	0.0000	0.0000
	150-225	10930	1	2.1004	0.2965
Distance to Fault	225-300	11657	1	1.9694	0.2780
	300-375	11404	1	2.0131	0.2842

	375-450	11380	0	0.0000	0.0000
	>450	1972988	86	1.0007	0.1413
Sum				7.0836	

8. Model preparation, result, and validation

8.1. Frequency Ratio

Frequency ratio is a quantitative technique for landslide susceptibility assessment using spatial data. It is frequently and effectively used for landslide susceptibility mapping. As it is quantitative method so it quantified between the landslide inventory, and causative factors. (Hawas Khana, 2019). The frequency ratio are get for each classes of causative factors type or range were calculated from their relationship with landslide occurrence. Likewise, the ratio was calculated for sub criteria of parameter, and then the frequency ratios were summed to calculate the Landslide Susceptibility Index (LSI). The Frequency Ratio of each classes were calculated with the following formula.

Frequency Ratio calculation: (Mi/M)/(Ni/N),
Where,
Mi = The number of pixels with landslides for each subclass conditioning factor,
M = The total number of landslides in the study area, Ni =
The number of pixels in the subclass area of each factor, N
= The number of total pixels in the study area.

Relative frequency: FR of class / sum total of all FR value in that factor

The relative frequency is calculated to normalize value within 0 to 1

From the calculated FR values of each class of each reclassified raster their respective relative frequency is also calculated with the help of following formula:
Relative frequency : FR of class / sum total of all FR value of that factor
Thus the FR values of each factor is normalized from 0 to 1. Finally the calculated relative frequency of FR values for each landslide and non-landslide points are extracted to prepare landslide susceptibility model.

8.2. Random Forest

The Frequency Ratio values are fitted to run the Random Forest. The importance values plays the primary role to create a confusion matrix of the datasets. Also we can say slope and aspect are main important factors for Landslide susceptibility in the RF. With the confusion matrix, it will help to know how many has agreed on the same categories. Figure 2 has given, slope and aspect are the importance factor for the occurrence of landside because the importance factor are 14.281, and 14.043 respectively. Looking only importance factor won't support for the landslide occurrence, the confusion matrix that has predicted 24 times a safe zone (value 0), and on the other side again 24 times it has predicted a landslide. Along with that, the confusion has given the accuracy level of landslides in our study area is 0.88 i.e, 88 percentage. Furthermore, we will use the ROC for the accuracy measurement under the Frequency ratio.

Figure 2 : Confusion matric and accuracy of FR

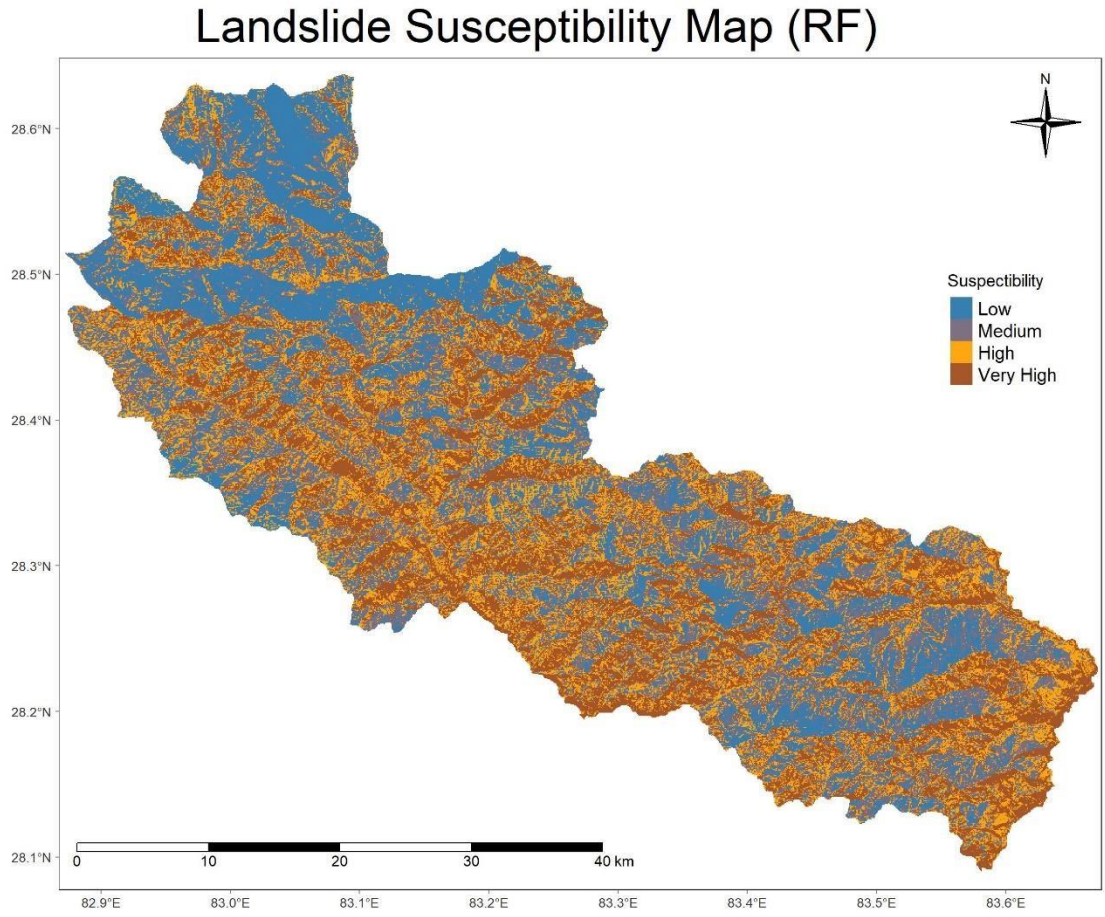
```
> rf$importance
             MeanDecreaseGini
rf_river             2.2364691
rf_road              4.1075731
rf_fault             0.8207446
rf_geology           3.8538238
rf_soil              4.1437370
rf_elevation         5.5454643
rf_landuse           8.4776509
rf_slope            14.2814362
rf_aspect           14.0435311
rf_curvature         3.0321441
> pred1=predict(rf,test,type="class")
> (confusion = table(test$hazard, pred1))
   pred1
     0  1
  0 24  3
  1  3 24
> sum(diag(confusion))/sum(confusion)
[1] 0.8888889
```

The final outcome of the Landslide Susceptibility by using the Random Forest is generated by the importance value of each causative factors.

Calculation:

Model = river*0.22364691+road*0.41075731+fault*0.08207446+geology*0.38538238+soil*0.41437370+elevation*0.55454643+landuse*0.84776509+slope*1.42814362+aspect*1.40435311+curvature*0.30321441

Map 11 : Landslide Susceptibility Map using Random Forest (RF) model



The Random Forest has calculated of each classed of causative factors, and then the Landslide Susceptibility Map has been created from the importance factor. Doing so has given us the brief idea of landslide hazard and safe area of the study area. The map is classified into four groups (i.e. Low, Medium, High, Very High) to see the susceptibility level from the FR method. But, checking an accuracy or making data validation gives more correct result. Alternatively, it gives support to our model. Thus, the Receiver Operator Characteristics (ROC) curve has used as accuracy assessment to check the performance of model.

In simply, the Area Under Curve AUC calculate the AUC for multiple logistic regression models because it allows us to see which model is best at making predictions. The interpretation of the ROC curves move to the top left corner of the plot, thus in this categories it does better accuracy or it does better classification of the data. Likewise, the AUC is calculated to quantify and tells us how much of the plot is located under the curve. Thus, we can say, closer of AUC to 1, the better the model. Moving toward the graph representation, the ROC curve places the True Positive Rate (Sensitivity) in the Y-axis, and on the X-axis, it will be the False Positive Rate (1- specificity).

The prediction and success ratio curve are developed from the test and train datasets respectively. The graphs are under;

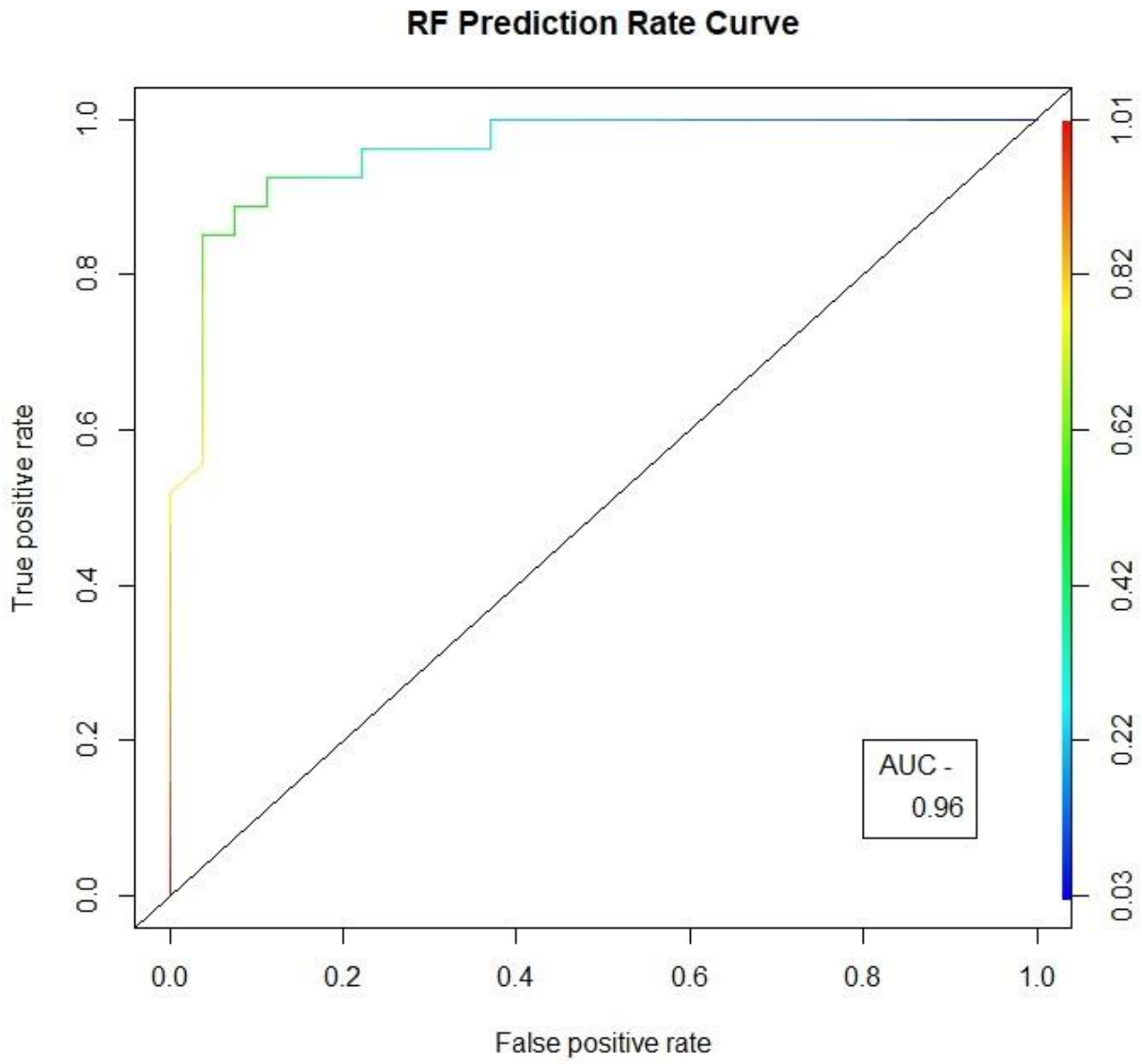
8.2.i. Prediction Ratio Curve under Random Forest

The datasets are module and the prediction ratio curve is calculated with the test datasets. The accuracy level from the prediction ratio curve under RF model and the AUC is 0.96, which shows the area under the curve is 96% i.e, accuracy. The figures 3 shows the calculation of the curve and figure 4 shows the graph of prediction ratio;

Figure 3 : Method for prediction ratio curve calculation under RF

```
> #rf auc
> pred1=predict(rf,test,type="prob")[,2]
> pred2<-prediction(pred1,test$hazard)
> roc=performance(pred2,"tpr","fpr")
> plot(roc)
> abline(a=0,b=1)
> #auc
> auc=performance(pred2,"auc")
> auc=unlist(slot(auc,"y.values"))
> auc
[1] 0.9595336
```

Figure 4 : Graph of Prediction Ratio Curve under RF



8.2.i. Success Ratio Curve under Random Forest

The train datasets were taken for the development of Success Ratio Curve in the Random Forest model. The curve has given the accuracy of 1 which mean 100 percentage the model has success

on the classification of datasets. The figure 5 and 6 shows the calculation method, and graph of the success ratio curve respectively.

Figure 5 : Method for Success Ratio Curve calculation under RF

```
> #rf auc_train
> pred1=predict(rf,train,type="prob")[,2]
> pred2<-prediction(pred1,train$hazard)
> roc=performance(pred2,"tpr","fpr")
> plot(roc)
> abline(a=0,b=1)
> #auc
> auc=performance(pred2,"auc")
> auc=unlist(slot(auc,"y.values"))
> auc
[1] 1
```

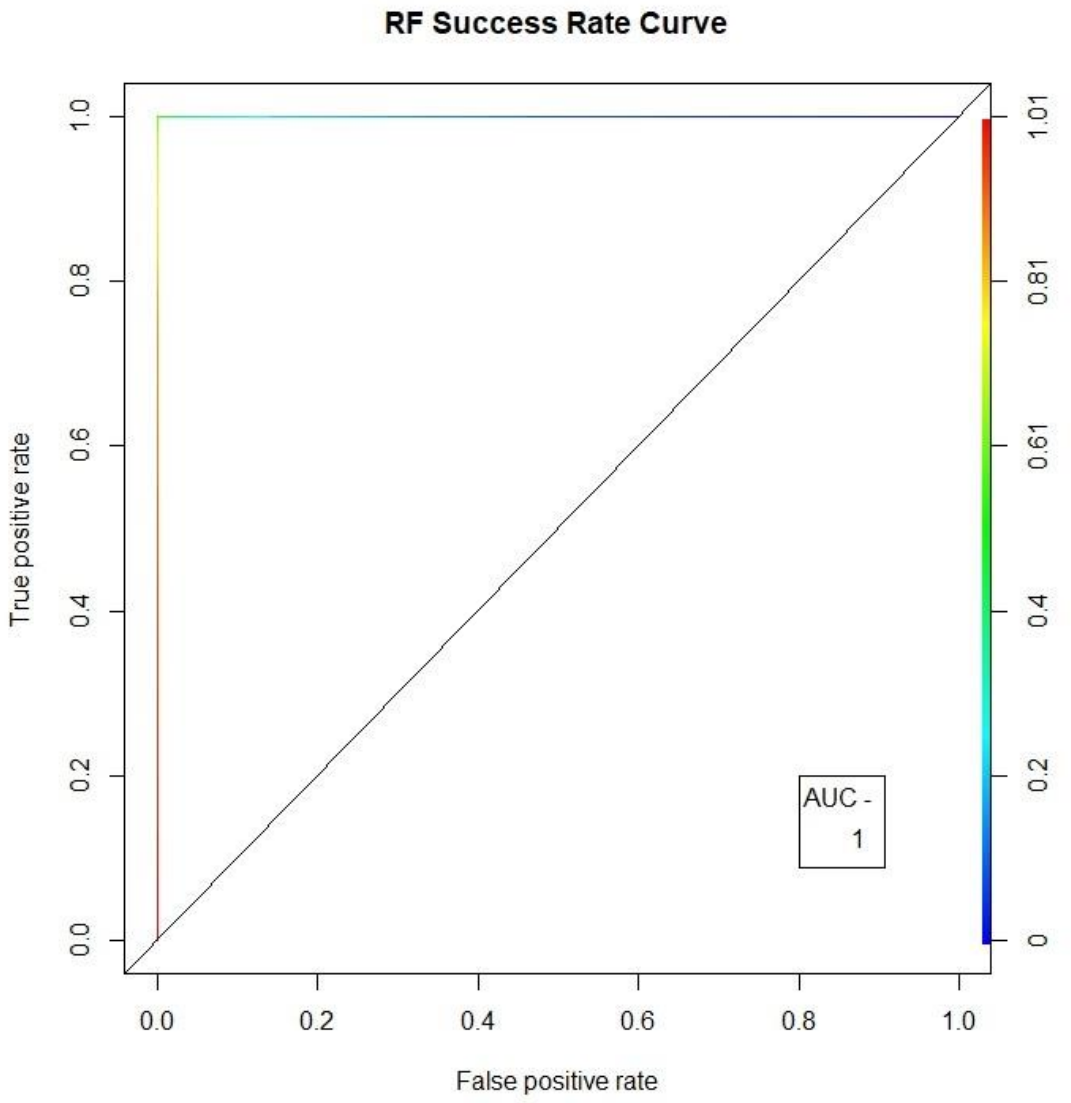


Figure 6 : Graph of Success Ratio Curve under RF

8.3. Classification And Regression Tree (CART):

Since, CART is a machine learning with present of classification and prediction tree model. The model would be appropriate to use for decision tree making with the classification and predict model. Further explanation, the CART term is used to describe decision tree algorithms that are

used for classification and regression learning tasks (Ninja, n.d.). Thus, to explain the CART we have to understand the classification, and regression decision tree individually; (i) Classification tree: Basically classification decision tree is used to classify the datasets into multiple groups. Alternatively, the process of splitting the datasets into classes according to its response variable (homogeneity). i.e; training and test dataset. (ii) Regression tree: It process of predicting the problems with response to the continuous variable. Its main task is to split the datasets for each independent variable by fitting the target variable by using the independent variables.

The pre-processed and modelled data has carried out for the final stage of the landslide susceptibility. The principle of the CART decision tree is to classify and run the regression among the data. Thus, in the classification of the data we will split or assign the data into training and test datasets with respective to 70% and 30%. The assigned percentage of the datasets will be used to fit in model. Now, making a decision tree using R where, hazard has used a dependent variable in the training datasets, and other variable as independent variables. Then, class method is used to classify the datasets. Not only that we predict our outcome on the test dataset, and has predicted those classes into either 0 or 1.

The major task is to identify the safe and hazard (landslide) cases from our fit model, and it is calculated from the importance value of the each landslide occurrence factors. The confusion matrix has given us that 24 times a safe zone is predicted (value 0). And 21 times predicted a landslide. The confusion matrix does to provide future insight into where our prediction model will success and where it will fail. The accuracy fact is 0.83 which mean it says the variables are related to each other. Since the value are closer to 1. The regression value present between -1 to 1 (Ng & Soo). The confusion matrix and result value:

Figure 7 : Confusion matrix and accuracy under CART

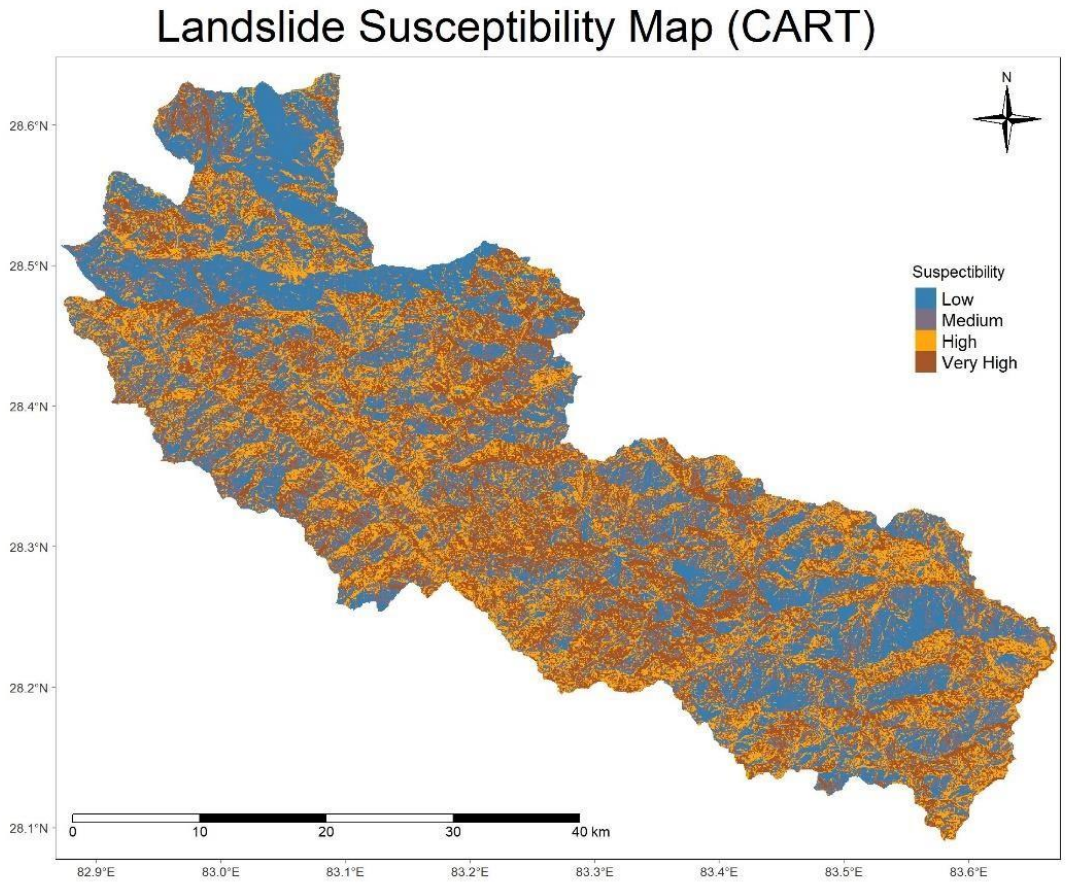
```
> cart_train$variable.importance
   rf_slope    rf_aspect     rf_road     rf_soil  rf_landuse    rf_river rf_curvature    rf_fault rf_elevation  rf_geology
  16.985024    12.517241    7.556933    4.016218    3.923891    3.157572    1.908085    1.294887    1.241790    0.381617
> pred = predict(cart_train, test, type = "class")
> (confusion = table(test$hazard, pred))
   pred
     0  1
  0 24  3
  1  6 21
> sum(diag(confusion))/sum(confusion)
[1] 0.8333333
```

The Landslide Susceptibility by using the CART model is generated by the importance value of each causative factors. Calculation of the factors;

Calculation:

Model = river*0.3157572+road*0.7556933+fault*0.1294887+geology*0.0381617+ soil*0.4016218+elevation*0.1241790+landuse*0.3923891+slope*1.6985024+aspect* 1.2517241+curvature*0.1908085

Map 12 : Landslide Susceptibility Map using CART



The map 12 has four categories of the susceptibility which are Low, Medium, High, and Very High. The map has created from the importance factor under the CART model. However, making a classified map does not validate the model. Thus, the Receiver Operator Characteristics (ROC) curve, and Area under Curve (AUC) has used to find out how perfectly the data are classified under the model.

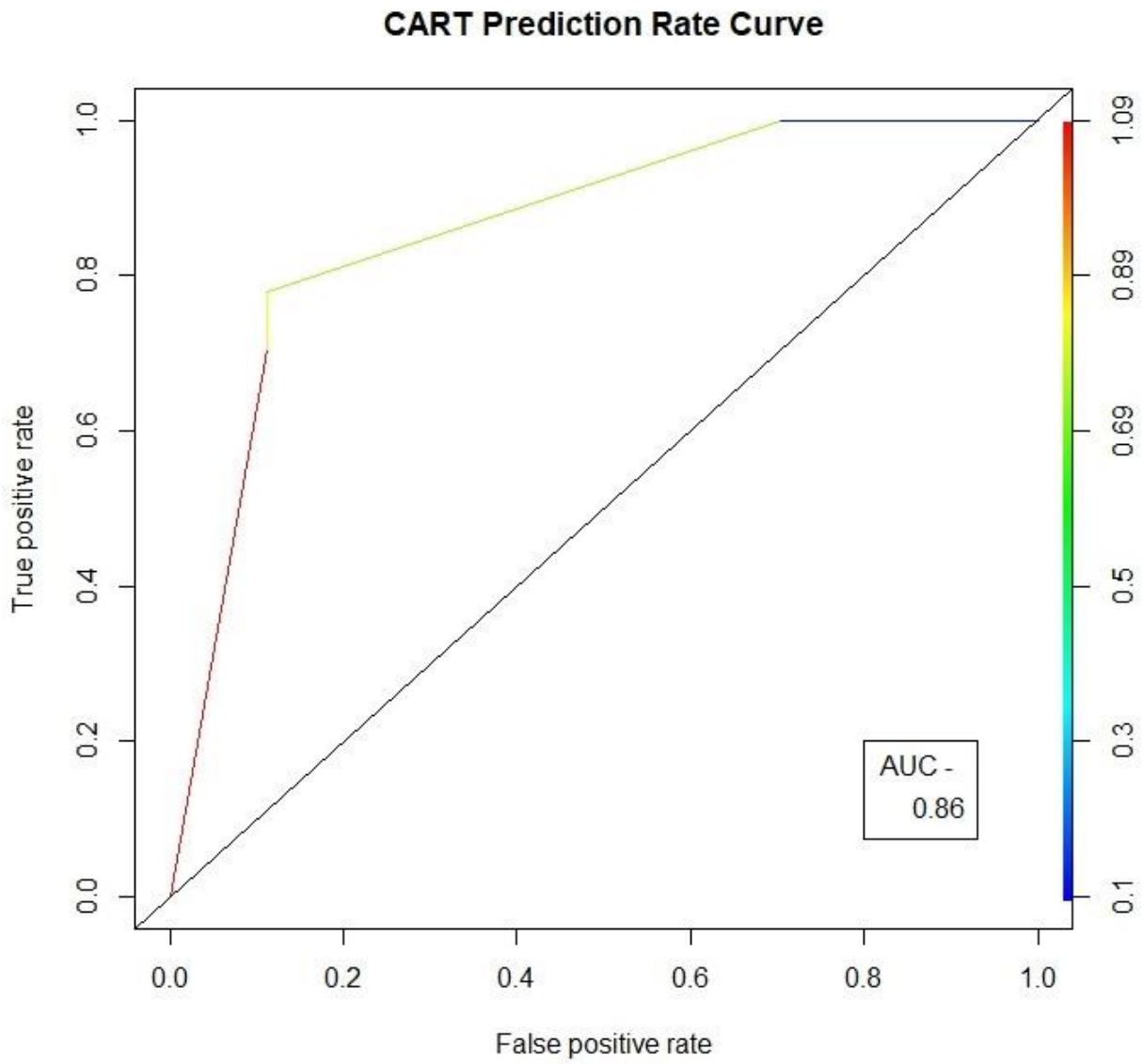
8.2.i. Prediction Ratio Curve under CART

The prediction ratio curve is used to check the performance of model by using a test datasets. The accuracy level from the prediction ratio curve under CART model and the AUC is 0.86 which means 86 percentage are more accurate. The figures 8 shows the calculation of the curve and figure 9 shows the graph of prediction ratio;

Figure 8 : Method for prediction ratio curve calculation under CART

```
> #cart auc
> pred1=predict(cart_train,test,type="prob")[,2]
> pred2<-prediction(pred1,test$hazard)
> roc=performance(pred2,"tpr","fpr")
> plot(roc,colorize=T)
> plot(roc)
> abline(a=0,b=1)
> #auc
> auc=performance(pred2,"auc")
> auc=unlist(slot(auc,"y.values"))
> auc
[1] 0.8621399
```

Figure 9 : Graph of Prediction Ratio Curve under CART



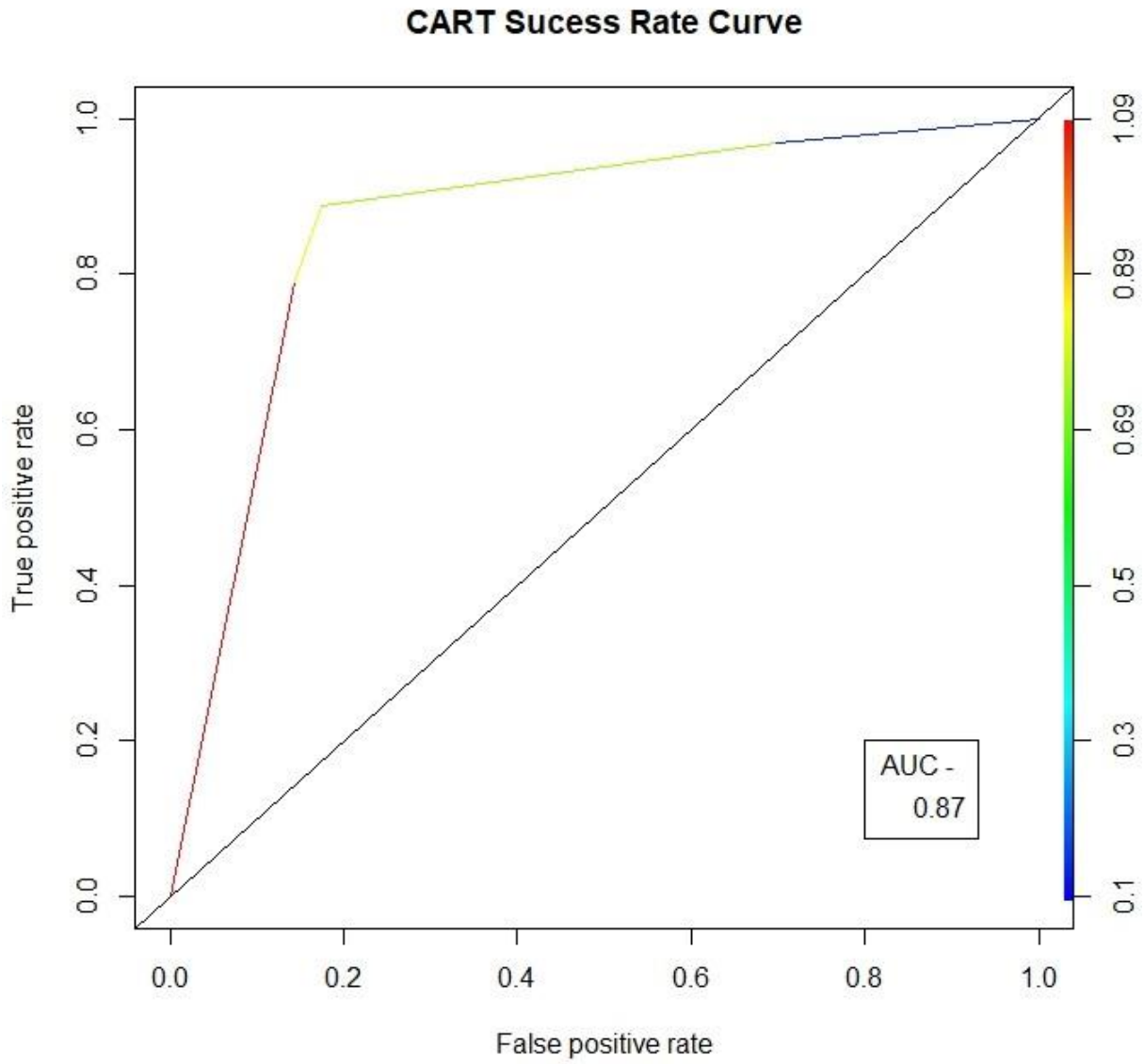
8.2.i. Success Ratio Curve under CART

The Success Ratio Curve in the Random Forest model has taken the training datasets during the model. The curve has given the accuracy of 0.8655914 which mean 87 percentage the model has success on the classification of datasets. The figure 10 and 11 shows the calculation method, and graph of the success ratio curve respectively.

Figure 10 : Method for Success Ratio Curve calculation under CART

```
#cart auc_train
pred1=predict(cart_train,train,type="prob")[,2]
pred2<-prediction(pred1,train$hazard)
roc=performance(pred2,"tpr","fpr")
plot(roc)
abline(a=0,b=1)
#auc
auc=performance(pred2,"auc")
auc=unlist(slot(auc,"y.values"))
auc
] 0.8655914
```

Figure 11 : Graph of Success Ratio Curve under CART



9. Conclusion and Recommendation

The landslide susceptibility map (CART) shows that almost 24.44 % and 17.83% of the total baglung is susceptible to high and very high risk zones. Similarly, the landslide susceptibility map(RF) shows that almost 30.41% and 17.64% of the total area is susceptible to high and very high risk zones. From both the model it can be seen that more that 40% of the total area is susceptible to landslide risk.

The accuracy assessment shows Random Forest model in comparison to CART algorithm performs better as the accuracy assessment shows 88 % and 96% accuracy from CART and RF respectively. The main reason behind the more accuracy from random forest model can be that its randomized feature selection method. Unlike CART algorithm which depends specifly on a feature and than creates child trees , the RF algorithm randomly selects a feature which makes this method more accurate than the other. Thus it can be concluded that baglung district is one of the most risk prone area for landslide. Also machine learning models can be effective

methods for landslide susceptibility analysis with RF being more accurate than the CART. This study recommends that for a better landslide susceptibility results high accurate data is preferred. Similarly, the more landslide conditioning factor the more result is expected. The accuracy of project is highly dependent upon the landslide factors and models used to predict the landslide.

10. References

Dahal, Ranjan. (2012). Rainfall-induced Landslides in Nepal. International Journal of Erosion Control Engineering. 5. 1-8. 10.13101/ijece.5.1.

Pokhrel, D., Bhandari, B.S. and Viraraghavan, T. (2009), "Natural hazards and environmental implications in Nepal", Disaster Prevention and Management, Vol. 18 No. 5, pp. 478-489.

Arabameri, Alireza & Saha, Sunil & Roy, Jagabandhu & Chen, Wei & Blaschke, Thomas & Bui, Dieu. (2020). Landslide Susceptibility Evaluation and Management Using Different Machine Learning Methods in The Gallicash River Watershed, Iran. Remote Sensing. 12. 475. 10.3390/rs12030475

Acharya, T. D. (2018). Regional Scale Landslide Hazard Assessment using Machine Learning Methods in Nepal.

Affairs, M. o. (2019). *Nepal Diaster Report.* Kathmandu, Nepal: Ministry of Home Affairs.

Baral, M. (2009). *Water induced disasters, Flood Hazard Mapping and Koshi Flood Disaster of Nepal.*

Carrara A and Cardinali M, D. R. (1991). *GIS techniques and statistical models in evaluating landslide hazard.* Earth Surface Processes and Landforms.

Dangol, S., & Bormudoi, A. (2015). FLOOD HAZARD MAPPING AND VULNERABILITY ANALYSIS OF BISHNUMATI RIVER, NEPAL. *Nepalese Journal on Geoinformatics, Survey Department*.

Derdous, O. L. (2015). A GIS Based Approach for the Prediction of the Dam Break Flood Hazard - A Case Study of Zardezas Reservoir ‘Skikda, Algeria. *ournal of Water and Land Development J27(1): .*

Development, M. o. (2020). *Statistical Information on Nepalese Agriculture.* Kathmandu, Nepal: Ministry of Agriculture and Livestock Development.

EHA. (2012). *Avalanche and floods in Seti River.* Retrieved from http://www.searo.who.int/entity/emergencies/crises/2012.05.07.Nepal_SitRep-2_May12.pdf

Gee, D. M. (2010). USE OF BREACH PROCESS MODELS TO ESTIMATE HEC-RAS DAM BREACH. Las Vegas.

Geographic, N. (n.d.). Landslide.

Gregory C, O. (2006). *Plan curvature and landslide probability in regions dominated by earth flows and earth slides.* USA: Elsevier.

Hawas Khana, M. S. (2019). *Landslide susceptibility assessment using Frequency Ratio, a case study of northern Pakistan.* Pakistan: ScienceDirect.

Health, E. R. (11 May 2020). Landslide Susceptibility Mapping Using Machine. *Landslide Susceptibility Mapping Using Machine*, 2.

Kumar, S., Jaswal, A., Pandey, A., & Sharma, N. (2017). Literature Review of Dam Break Studies and Inundation Mapping Using Hydraulic Models and GIS. *International Research Journal of Engineering and Technology (IRJET)*.

Kumar, S., Jaswal, A., Pandey, A., & Sharma, N. (2017). Literature Review of Dam Break Studies and Inundation Mapping Using Hydraulic Models and GIS. *International Research Journal of Engineering and Technology*, 55-61.

Lee and Min, L. S. (2001). *Statistical Analysis of Landslide Susceptibility at Yongin.* Korea.

MGS Engineering Consultants, I. (2007). DAM BREAK INUNDATION ANALYSIS AND DOWNSTREAM HAZARD CLASSIFICATION.

Ng & Soo, A. N. (n.d.). *Numsense! Data Science for the Layman.*

Ninja, C. (n.d.). *Classification & Regression Trees*. Retrieved from Coading Ninja: https://www.codingninjas.com/blog/2020/12/02/what-is-classification-regressiontrees/

Portal, N. D. (2021). *Reported number of Incidents.* Kathmandu, Nepal: Nepal Diaster Riask Reduction Portal.

Sachin. (2014). *DAM BREAK ANALYSIS USING MIKE11 FOR LOWER NAGAVALI DAM AND RUKURA DAM* . Rourkela: Department Of Civil Engineering National Institute Of Technology.

Sharma, P., & Mujumdar., S. (2017). Dam Break Analysis Using HEC-RAS and HEC-GeoRAS – A Case Study of Ajwa Reservoir. 108–13.

Shrestha, B. B., & Nakagawa, H. (2016). *Hazard assessment of the formation and failure of the Sunkoshi landslide dam in Nepal.* Retrieved May 2018, from https://doi.org/10.1007/s11069-016-2283-3.

Sivakami C, S. A. (2014). *Landslide susceptibility zone using FR model.* IISTE.

USACE. (2014). *Using HEC-RAS for Dam Break Studies.* US Army Crops of Engineers, Hydrologic Analysis Centre.

Wahl, T. L. (2010). Dam Breach Modeling – an Overview of Analysis Methods. *Joint Federal Interagency Conference on Sedimentation and Hydrologic Modeling.*

Watch, R. (2021, October 6). *The List of Countries at High Riask of Landslides*. Retrieved from Blog Resource Watch: https://blog.resourcewatch.org/2018/08/27/italy-austria-andchina-top-the-list-of-countries-at-high-risk-of-landslides-right-now/

XIanyu Yu, K. Z. (2021). *Study on landslide susceptibility mapping bases on rock-soil characteristics factors.* Scientific reports.

Xiong, Y. (. (2011). A Dam Break Analysis Using HEC-RAS . *Water Resource and Protection*, 370-379.

1. Dahal, Ranjan. (2012). Rainfall-induced Landslides in Nepal. International Journal of Erosion Control Engineering. 5. 1-8. 10.13101/ijece.5.1.

2. Pokhrel, D., Bhandari, B.S. and Viraraghavan, T. (2009), "Natural hazards and environmental implications in Nepal", Disaster Prevention and Management, Vol. 18 No. 5, pp. 478-489.

3. Arabameri, Alireza & Saha, Sunil & Roy, Jagabandhu & Chen, Wei & Blaschke, Thomas & Bui, Dieu. (2020). Landslide Susceptibility Evaluation and Management Using Different Machine Learning Methods in The Gallicash River Watershed, Iran. Remote Sensing. 12. 475. 10.3390/rs12030475

0.09
5105