

*Project Report On*

# **Predicting Rainfall-Induced Landslides and Gender-Based Twitter Analysis of Joshimath Crisis**

*Submitted in requirement for the course*

**B.Tech. Project (CSN-400B)**

*of Bachelor of Technology in Computer Science and Engineering*

**Submitted By**

**Hardik Thami**

19114035

hardik\_t@cs.iitr.ac.in

**Devanshu Chaudhari**

19114027

devanshu\_c@cs.iitr.ac.in

**Purushottam**

19114065

purushottam@cs.iitr.ac.in

Under the supervision of

**Prof. Sudip Roy**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**INDIAN INSTITUTE OF TECHNOLOGY, ROORKEE**

**ROORKEE- 247667 (INDIA)**

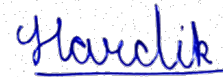
**May 12, 2023**

# Candidate's Declaration

We hereby declare that the work carried out in this dissertation entitled “**Predicting Rainfall-Induced Landslides and Gender-Based Twitter Analysis of Joshimath Crisis**” is presented on behalf of partial fulfilment of the requirements for the award of degree of **Bachelor of Technology in Computer Science and Engineering** submitted to the **Department of Computer Science and Engineering, Indian Institute of Technology (IIT) Roorkee** under the supervision of **Prof. Sudip Roy**, Assistant Professor, Dept. Of CSE, IIT Roorkee. The work presented in the report is authentic to the best of our knowledge and has been done from **July 2022 to May 2023**.

DATE: 18/05/2023

SIGNED:

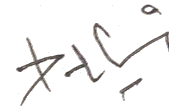


PLACE: IIT Roorkee

HARDIK THAMI  
(19114035)

DATE: 18/05/2023

SIGNED:

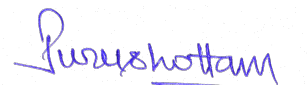


PLACE: IIT Roorkee

DEVANSHU CHAUDHARI  
(19114027)

DATE: 18/05/2023

SIGNED:



PLACE: IIT Roorkee

PURUSHOTTAM  
(19114065)

# Certificate

This is to certify that the statement made by the candidate is correct to the best of my knowledge and belief.

DATE: 18/05/2023

SIGNED: .....

PROF. SUDIP ROY  
(ASSISTANT PROFESSOR)  
DEPT. OF CSE, IIT ROORKEE

# Acknowledgement

First and foremost, we would like to express our sincere gratitude towards our guide **Prof. Sudip Roy**, Assistant Professor, Department of Computer Science and Engineering, IIT Roorkee for his ideal guidance throughout the entire period. We want to thank him for the insightful discussions and constructive criticisms which certainly enhanced our knowledge as well as improved our skills. His constant encouragement, support and motivation were key to overcome all the difficult and struggling phases.

We would also like to thank **Department of Computer Science and Engineering, IIT Roorkee** for providing resources for the project work.

We would also like to thank **Prof. Roopam Shukla**, Assistant Professor, Centre of Excellence in Disaster Mitigation and Management, IIT Roorkee for her guidance throughout the entire period.

We also extend our gratitude to **Tanu Gupta** and **Tamal Mandal**, for peer-reviewing our work and providing us with valuable suggestions/feedback.

We humbly extend our sincere thanks to all concerned persons who co-operated with us in this regard.

# Abstract

India is one of the countries in the world with the highest landslide hazard. The management of landslide and its mitigation is made even more difficult by high population density, huge area, and rainfall. Hence, there is a need for understanding of landslide occurrences and the factors which can be beneficial for predicting it. Here, multiple factors that are the cause of rainfall-induced landslides have been used to predict such landslides using machine learning techniques. The results prove that with better datasets, these systems can find great use in early warning systems for forecasting landslides caused due to rainfall.

The advent of social media in recent times has taken people to express their views and emotions on the social media platforms. At the time of natural disasters, social media provides large amount of information on relief efforts and the emotions of affected people. We have studied people's response on Twitter to the recent Joshimath crisis. The way people responded on social media is likely to be influenced by many factors including gender. This study classifies Twitter users based on their gender and analyzes their response to the Joshimath Crisis.

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Rainfall-Induced Landslide Prediction . . . . .	2
1.2	Gender-Based Twitter Analysis of Joshimath Crisis . . . . .	3
<b>2</b>	<b>Motivation, Objectives and Problem Statements</b>	<b>4</b>
2.1	Motivation and Objectives . . . . .	4
2.2	Problem Statements . . . . .	5
2.2.1	Rainfall-Induced Landslide Prediction . . . . .	5
2.2.2	Gender-Based Twitter Analysis of Joshimath Crisis . . . . .	5
<b>3</b>	<b>Rainfall-Induced Landslide Prediction</b>	<b>6</b>
3.1	Literature Survey . . . . .	6
3.2	Methodology . . . . .	7
3.2.1	Data Collection . . . . .	7
3.2.1.1	Global Landslide Inventory (GLC) . . . . .	8
3.2.1.2	Rainfall Data . . . . .	9
3.2.1.3	Digital Elevation Model (DEM) . . . . .	9
3.2.2	Data Preprocessing . . . . .	9
3.2.2.1	Data Cleaning . . . . .	10
3.2.2.2	Data Reduction . . . . .	11
3.2.2.3	Data Integration . . . . .	11
3.2.3	Machine Learning Techniques for Landslide Prediction . . . . .	14
3.2.3.1	Creating Training and Test Data . . . . .	14
3.2.3.2	Logistic Regression . . . . .	15
3.2.3.3	Decision Tree . . . . .	16
3.2.3.4	Support Vector Machine . . . . .	17
3.2.3.5	Gaussian Naive Bayes . . . . .	18

<b>4</b>	<b>Data Analytics for Landslide Prediction</b>	<b>20</b>
4.1	Results . . . . .	20
4.1.1	ROC Curve . . . . .	21
4.1.2	Logistic Regression . . . . .	22
4.1.3	Decision Tree . . . . .	23
4.1.4	Support Vector Machine . . . . .	25
4.1.5	Gaussian Naive Bayes . . . . .	26
4.1.6	Landslide Visualization . . . . .	28
4.1.7	Source Code Link . . . . .	29
<b>5</b>	<b>Gender-Based Twitter Analysis of Joshimath Crisis</b>	<b>30</b>
5.1	Methodology . . . . .	30
5.1.1	Tweet Extraction and Translation . . . . .	31
5.1.2	Gender Estimation of Twitter User . . . . .	31
5.1.3	Unigrams and Bigrams extraction from the tweets text . . . . .	32
5.1.4	Sentiment Analysis of Tweets . . . . .	32
<b>6</b>	<b>Data Analytics for Gender-Based Twitter Analysis of Joshimath Crisis</b>	<b>34</b>
6.1	Results . . . . .	34
6.1.1	Tweet Extraction and Translation . . . . .	34
6.1.2	Gender Estimation of Twitter User . . . . .	34
6.1.3	Unigrams and Bigrams from tweets text . . . . .	36
6.1.4	Sentiment Analysis of Tweets . . . . .	38
6.1.5	Source Code Link . . . . .	39
<b>7</b>	<b>Conclusions and Future Work</b>	<b>40</b>
7.1	Landslides Prediction . . . . .	40
7.2	Gender-Based Twitter Analysis of Joshimath Crisis . . . . .	40
	<b>Bibliography</b>	<b>42</b>

# List of Figures

3.1	Flowchart for Rainfall-Induced Landslide Prediction . . . . .	7
3.2	Landslide Features Present in NASA GLC Dataset [1]. . . . .	8
3.3	Python Snippet for Data Cleaning. . . . .	10
3.4	Short-term rainfall (mm) for landslides. . . . .	12
3.5	Long-term rainfall (mm) for landslides. . . . .	13
3.6	Elevation Relief (m) of area affected by landslides. . . . .	13
3.7	Python Snippet for balancing dataset using SMOTE. . . . .	15
4.1	An example ROC curves with a scale of AUC values . . . . .	21
4.2	Logistic Regression ROC Curve. . . . .	22
4.3	Logistic Regression Confusion Matrix. . . . .	23
4.4	Decision Tree ROC Curve. . . . .	24
4.5	Decision Tree Confusion Matrix. . . . .	24
4.6	Support Vector Machine ROC Curve. . . . .	25
4.7	Support Vector Machine Confusion Matrix. . . . .	26
4.8	Gaussian Naive Bayes ROC Curve. . . . .	27
4.9	Gaussian Naive Bayes Confusion Matrix. . . . .	27
4.10	Shaded Regions Affected By Landslides. . . . .	29
5.1	Flowchart of Gender-Based Twitter Analysis of Joshimath Crisis . . . . .	31
6.1	Gender Estimation Probability Bar Chart. . . . .	35
6.2	Top 100 most common words. . . . .	36
6.3	Percentage sentiments distribution across all entity types. . . . .	38



# List of Tables

4.1	Comparison of performance metrics of trained models. . . . .	28
6.1	Gender Estimation Probability Distribution. . . . .	35
6.2	Top-20 Unigrams and Top-10 Bigrams from all entity types. . . . .	37

# List of Abbreviations

DMS	Disaster Management Support
GLC	Global Landslide Inventory
DEM	Digital Elevation Model
SMOTE	Synthetic Minority Over-sampling Technique
ROC	Receiver Operating Characteristic Curve
CNN	Convolutional Neural Network
FR	Frequency Ratio
SVM	Support Vector Machine

# Chapter 1

## Introduction

### 1.1 Rainfall-Induced Landslide Prediction

Landslides are a widespread problem across the globe whose occurrences and effects are increasing sharply due to unforeseeable heavy rains, earthquakes, or volcanic eruptions.

A study by Froude and Petley (2018) on deadly global landslides shows that worldwide spatial distribution of landslides is heterogeneous among which, Asia (mainly south and east) represents the most affected geographical region by landslide disaster.

India ranks first among nations in terms of deadly landslides [2]. Hence, it is becoming increasingly imperative to have a better understanding of landslide occurrences and the factors which can be beneficial for predicting it.

When it comes to factors causing landslides, rainfall is one of the most common factors, among various other factors such as elevation, vegetation, soil and bedrock etc. Hence, predicting rainfall-induced landslides is not only extremely beneficial to reduce number of deaths due to landslides, but is also useful when planning disaster rescue operations.

Generally, forecasting landslides caused due to rainfall is done using a metric called rainfall threshold, which is essentially the approximate minimum amount of rainfall that needs to occur in a region for it to cause a landslide in that region. This requires the use of rain gauge equipment to measure the amount of rainfall over a pre-defined area, which can be very costly and even then the rainfall measured by the rain gauge maybe erroneous.

In this study, we present a possible solution to the above problem by using various machine learning techniques that can be used to predict landslides caused due to rainfall.

## **1.2 Gender-Based Twitter Analysis of Joshimath Crisis**

The recent natural disaster in Joshimath, Uttarakhand has attracted widespread attention and concern from people around the world. As news of the crisis spread, people took to social media to express their reactions, share updates, and offer support. However, as with any social issue, the way people respond to the Joshimath crisis on social media is likely to be influenced by a range of factors, including their gender.

The intersection of gender and crisis communication highlights the significance of analyzing gender-specific responses to natural disasters. It is expected that gender influences both the content and tone of social media discussions during such crises. Different topics may have different importance to males and females.

This report aims to provide a gender based analysis of tweets related to the Joshimath crisis. By examining tweets, we seek to identify patterns and differences in the ways that different genders discuss the crisis on social media. Through this analysis, we hope to gain a deeper understanding of how gender shapes public discourse around natural disasters and crisis response on social media.

## **Chapter 2**

# **Motivation, Objectives and Problem Statements**

### **2.1 Motivation and Objectives**

India is prone to many natural disasters like floods, landslides, cyclones, forest fires, earthquakes, drought, etc. ISRO Disaster Management Support (DMS) Programme, provides near real time information support of such disasters and for landslides it offers it's zonation and inventory [3].

However, considerable time taken by both pre-processing and post-processing of satellite imagery causes delay in rescue efforts by disaster management teams. Therefore, an alternative method is required and hence we propose machine learning based techniques to predict a landslide caused by rainfall over a long/short period of time.

Machine learning algorithms have been increasingly used for landslide prediction due to their ability to extract patterns and relationships from large datasets. Classification algorithms in machine learning can be effective here, due to their ability to automatically categorize or classify data based on patterns and characteristics.

The following were the objectives for this study:

- Build a rainfall-based landslide forecasting model using machine learning algorithms.
- Analyse Twitter data for Joshimath land subsidence incident to understand the commonality and differences in views and sentiments of a large audience.

## **2.2 Problem Statements**

In this study, we have focused on two independent problem statements that are as follows:

### **2.2.1 Rainfall-Induced Landslide Prediction**

This study aims to build a rainfall-induced landslide prediction model using various machine learning classification algorithms such as, Logistic Regression, Decision Tree, Support Vector Machine and Gaussian Naive Bayes.

The problem statement can be further divided into many sub problems such as:

- Data Collection
- Data Preprocessing
- Creating Training and Test Data
- Machine Learning Techniques for Landslide Prediction

### **2.2.2 Gender-Based Twitter Analysis of Joshimath Crisis**

The sentiments on Joshimath land subsidence are identified using a twitter data set consisting of tweets from various Twitter account. The problem statement can be further divided into sub problems such as:

- Tweet Extraction and Translation
- Gender Estimation of Twitter User
- Sentiment Analysis of Tweets

## Chapter 3

# Rainfall-Induced Landslide Prediction

### 3.1 Literature Survey

Machine learning algorithms have been increasingly used for landslide prediction due to their ability to extract patterns and relationships from large datasets. This literature review aims to provide an overview of the recent advances in landslide prediction using machine learning/deep learning techniques.

Faraz et al. (2019) used a Logistic Regression algorithm to predict rainfall-induced landslides. The model was trained on NASA GLC Dataset containing landslides till 2017, and achieved a AUC of 0.93 [1].

Deepak et al. (2017) used a Support Vector Machine (SVM) algorithm to predict landslides in the Mandakini Valley of Uttarakhand, India. The SVM model was trained on elevation, slope, aspect, drainages, geology/lithology, buffer of thrusts/faults, buffer of streams and soil along with the past landslide data and achieved a AUC of 0.829 [4].

Yuvaraj et al. (2021) used a Frequency Ratio (FR) and Binary Logistic Regression algorithm to predict landslides in the Nilgiris District of Tamil Nadu, India. Several factors such as Lithology, NDVI, Rainfall, Lineament density, Lineament buffer, Slope, Soil, Depth of the soil, Aspect and Land use/Land cover were used to train the models, and achieved a AUC of 0.863 and 0.866 for respective machine learning algorithms [5].

Sansar et al. (2021) used a Convolutional Neural Network (CNN) to predict landslides in the Western Ghats of India. The mean accuracies of correctly classified landslide values was found to increase from 65.5% to 78% by using slope data [6].

## 3.2 Methodology

Fig. 3.1 shows the steps taken in the landslide prediction study. Since we need to predict landslides caused by rainfall, we need to prepare a dataset consisting of both rainfall-induced landslides and landslides caused due to other factors. Our work can be further dividing into following steps:

- Data Collection
- Data Preprocessing
- Machine Learning Techniques for Landslide Prediction

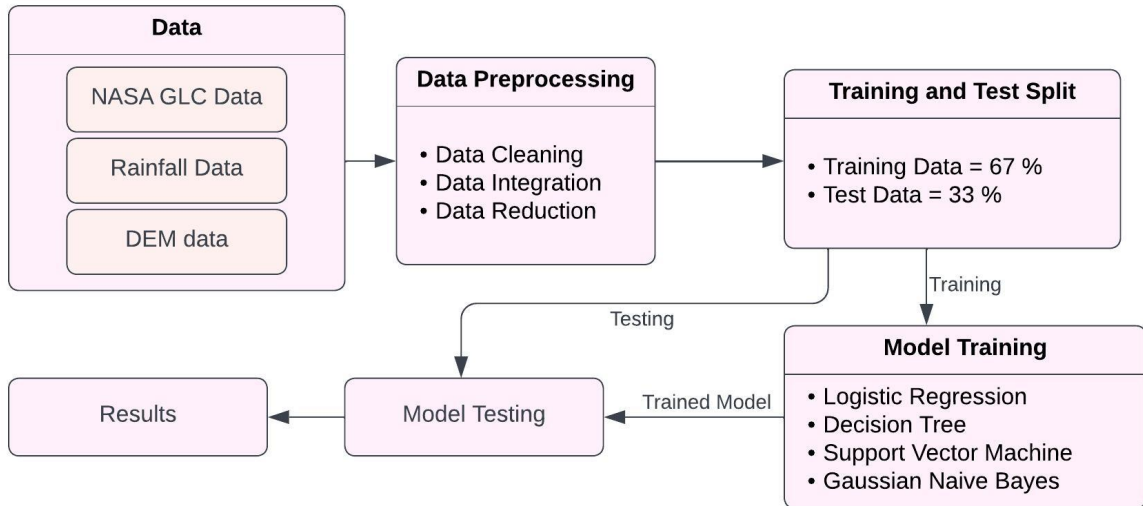


Figure 3.1: Flowchart for Rainfall-Induced Landslide Prediction

### 3.2.1 Data Collection

When making a landslide prediction model based on machine learning / deep learning, the lack of sufficient landslide data presents a major challenge for the successful implementation of machine learning / deep learning techniques.

To overcome this limitation, we integrated the most recent Global Landslide Catalogue (GLC) of NASA (consists of landslides upto to the year 2021) with global rainfall datasets and publicly available datasets of landslide controlling factors [7].



Furthermore, studies show that rainfall, temperature, elevation, soil moisture, slope, aspect, and land cover have been identified as important predictors of landslide occurrence. However, for the scope of the study, we have only considered rainfall and elevation for landslide prediction.

### 3.2.1.1 Global Landslide Inventory (GLC)

The Global Landslide Catalogue (GLC) is a database of landslide events that have occurred worldwide from the year 2007 to 2021. The database is maintained by NASA's Earth Science Division and it is a publicly available dataset [7].

The Fig. 3.2 shows the features present in this dataset. It contains information on the location, date, type, and size of landslide events, as well as the trigger mechanisms and impacts of each event.

It is based on a variety of sources, including media reports, government reports, and scientific publications. The database is continuously updated as new landslide events are reported or discovered.

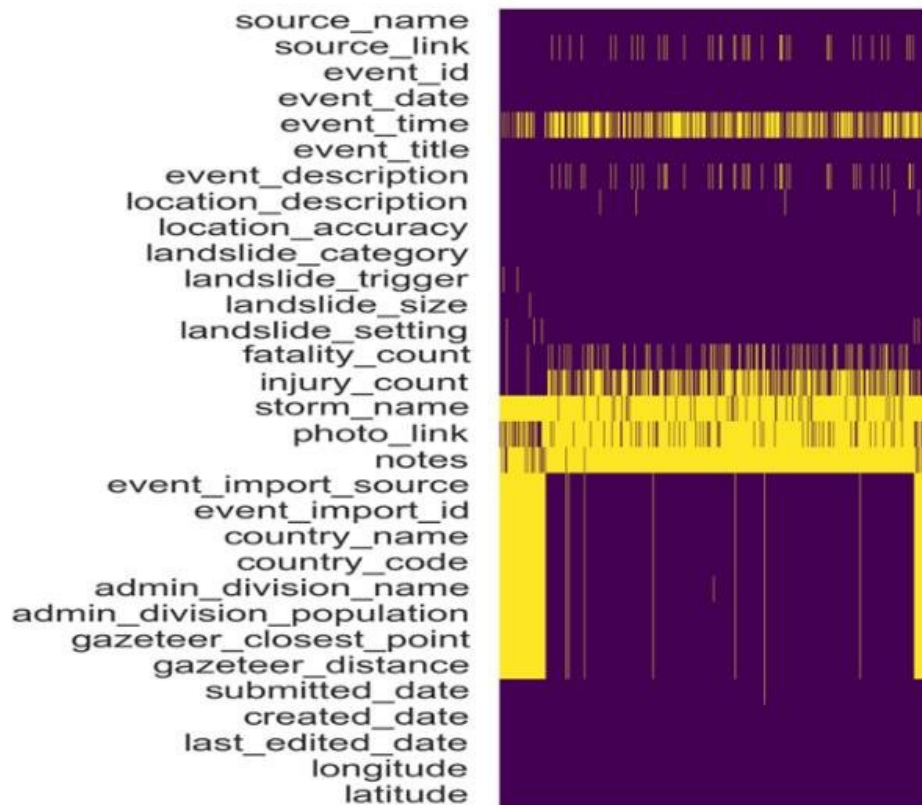


Figure 3.2: Landslide Features Present in NASA GLC Dataset [1].

### **3.2.1.2 Rainfall Data**

Since we are predicting rainfall-induced landslides, historical rainfall data for training model is of paramount importance. For this, we have used daily rainfall data from Open-Meteo Historical Weather API [8] to estimate accumulated rainfall from day of the landslide to eleven days before the landslide occurred to account for landslides caused due to short-term/long-term rainfall.

By integrating data from weather stations, aircraft, buoys, radar, and satellites, the Historical Weather API uses reanalysis datasets to provide a comprehensive historical record of weather conditions. These datasets also employ mathematical models to estimate missing data, enabling the API to offer detailed weather information for locations that may not have had weather stations nearby, such as rural areas or the open ocean.

The Open-Meteo Historical Weather API uses the ERA5 reanalysis model developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) [9]. ERA5 stands for "Fifth generation of ECMWF atmospheric reanalyses of the global climate," and it provides hourly estimates of a wide range of weather variables, including temperature, precipitation, wind speed, and atmospheric pressure, with a spatial resolution of 31km.

### **3.2.1.3 Digital Elevation Model (DEM)**

Digital Elevation Models (DEM) play a critical role in landslide prediction, as they provide information on the topography and surface characteristics of the terrain that can influence the occurrence and behavior of landslides. DEMs are digital representations of the Earth's surface, typically created using remote sensing techniques.

This study uses NASA SRTM1 [10], which is a digital elevation model of the Earth's surface created using radar data collected by the Space Shuttle Endeavour in 2000. It provides global coverage at a spatial resolution of approximately 30 meters and is widely used for topographic mapping and environmental modeling. The dataset is freely available to the public.

## **3.2.2 Data Preprocessing**

This study aims to predict landslides caused due to rainfall in India only. We performed data preprocessing to remove landslides not caused in India, remove unnecessary features from dataset and integrate rainfall and elevation data from various APIs as discussed in detail below:

### 3.2.2.1 Data Cleaning

The GLC dataset consists of information on the location, date, type, and size of landslide events, as well as the trigger mechanisms and impacts of each event.

Although the GLC is a useful source of information on rainfall-induced landslides, it has some limitations:

- Includes a small number of landslides that are triggered by factors other than rainfall, such as human actions and earthquakes.
- Each landslide is associated with a confidence radius (which is an estimate of the potential area over which the landslide event could have occurred)

As of April 2023, the GLC had data on just 1742 landslides that occurred in India. Landslides not occurring in India have been ignored for the purposes of this study.

In order of our model to be reasonably accurate, we have considered only those landslides with radius of confidence less than or equal to 5km. We have also removed landslides whose triggering factor is unknown.

Finally, we are left with 833 landslide events, out of which 52 are non rainfall-induced landslides. The python snippet that summarizes the data cleaning process is shown in Fig. 3.3

```
1 """
2 Remove landslides on the basis of landslide_trigger (triggering
   factor), location_accuracy, country_code
3 """
4 df_cleaned = df[
5     (
6         df['country_code'] == 'IN'
7     ) &
8     (
9         df['landslide_trigger'] != 'unknown'
10    ) &
11    (
12        (df['location_accuracy'] == 'exact') |
13        (df['location_accuracy'] == '1km') |
14        (df['location_accuracy'] == '5km')
15    )
16 ]
```

Figure 3.3: Python Snippet for Data Cleaning.

### 3.2.2.2 Data Reduction

Since the GLC is based on a variety of sources, including media reports, government reports, and scientific publications, it consists of such information about the landslide event that is not needed for the purposes for this study.

Following are the features/information that we have kept in the reduced dataset for each landslide (and discarded the remaining features):

- **landslide\_trigger**
- **location\_accuracy**
- **event\_date**
- **latitude**
- **longitude**

### 3.2.2.3 Data Integration

Data integration in data mining is the process of combining data from multiple sources into a single, unified dataset. In this study, rainfall data and elevation data was integrated into the cleaned and reduced dataset.

Following are the features/information that were added in the cleaned and reduced dataset for each landslide:

- **short\_term\_rainfall:** Accumulated rainfall on the day of landslide and a day before landslide occurred [11]. Fig. 3.4 shows the bar chart between frequency and short\_term\_rainfall where short\_term\_rainfall is shown in ranges such 0-50, 50-100 so on upto 200+. Here frequency is the number of landslides that have short\_term\_rainfall in a given range.
- **long\_term\_rainfall :** Accumulated rainfall from the day of landslide occurrence to 10 days prior to landslide occurrence [11]. Fig. 3.5 shows the bar chart between frequency and long\_term\_rainfall where long\_term\_rainfall is shown in ranges such 0-50, 50-100 so on upto 750+. Here frequency is the number of landslides that have long\_term\_rainfall in a given range.

- **elevation\_relief** : Difference between highest elevation and lowest elevation in an area. Fig. 3.6 shows the bar chart between frequency and elevation\_relief where elevation\_relief is shown in ranges such 0-500, 500-1000 so on upto 3000-3500. Here frequency is the number of landslides that have elevation\_relief in a given range.
- **isRainfallInducedLandslide** : Binary Value depicting whether it is rainfall-induced landslide (1) or not (0)
- **RainfallInducedLandslideProbability** : Number representing the probability of it being rainfall-induced landslide. If probability  $\geq 0.5$ , **isRainfallInducedLandslide** = 1 else **isRainfallInducedLandslide** = 0

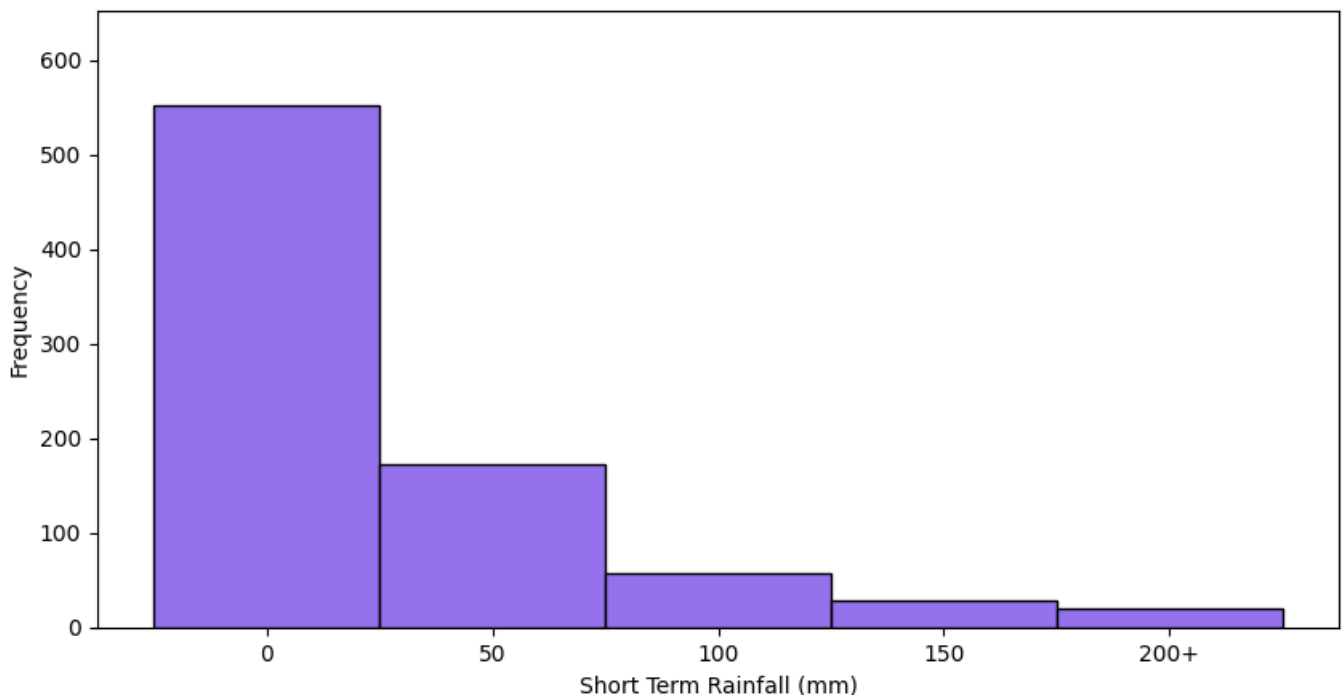


Figure 3.4: Short-term rainfall (mm) for landslides.

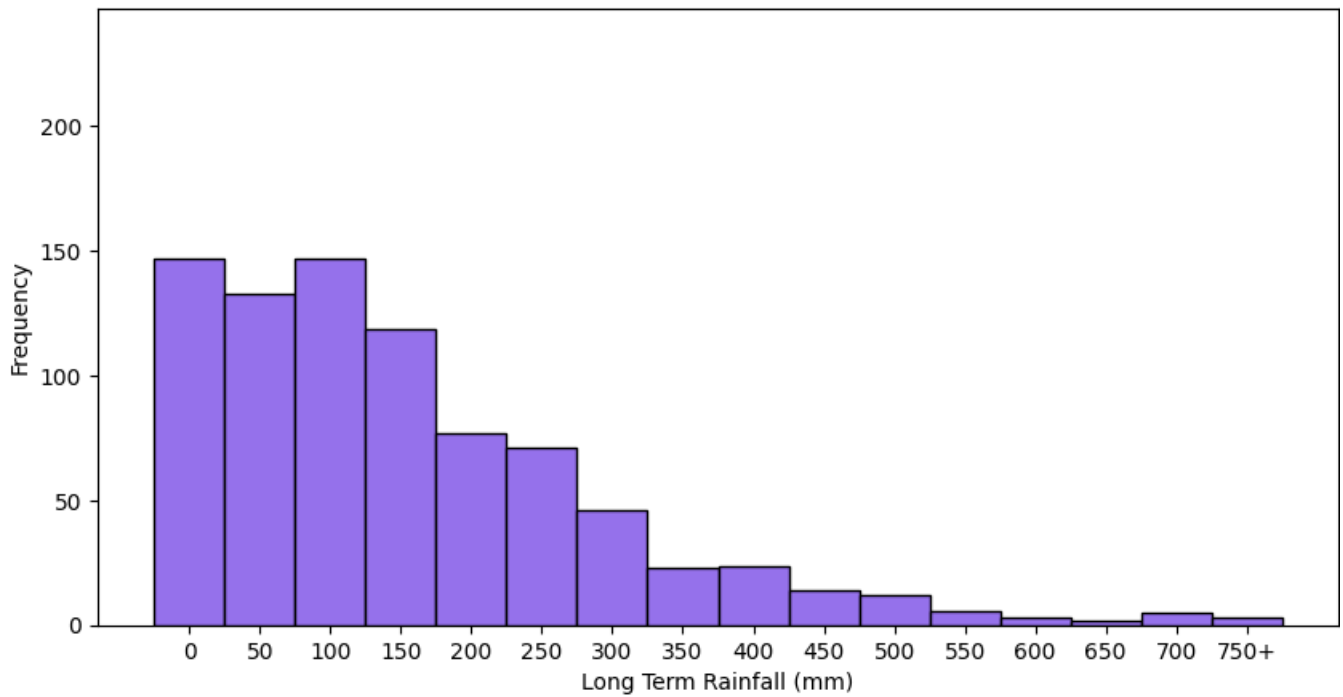


Figure 3.5: Long-term rainfall (mm) for landslides.

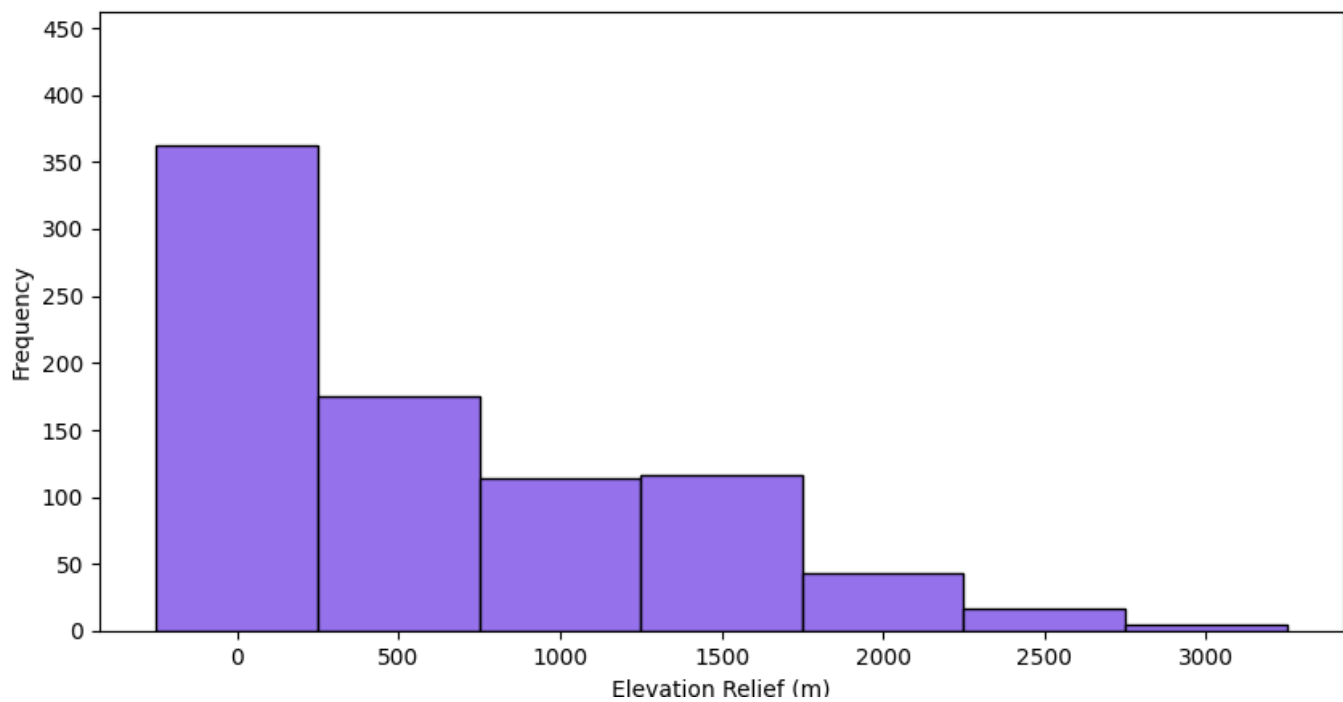


Figure 3.6: Elevation Relief (m) of area affected by landslides.

### 3.2.3 Machine Learning Techniques for Landslide Prediction

Machine learning is a type of artificial intelligence that enables computers to learn and make predictions or decisions without being explicitly programmed. Essentially, machine learning involves training a computer program on a large amount of data so that it can recognize patterns and make predictions based on new data.

In this study, we used machine learning techniques such as Logistic Regression, Decision Tree, Support Vector Machine, and Gaussian Naive Bayes to differentiate between landslide/non-landslide instances. But before doing so, we have to create separate training and testing data which then can be used for training and assessing our models.

#### 3.2.3.1 Creating Training and Test Data

In order to train our model, landslide cases that are not triggered by rainfall must be added to our training and test sets. As of now, our dataset includes about 52 such cases and 781 rainfall-induced landslides.

In this state, our dataset is imbalanced, in other words, the number of observations in one class is much larger or much smaller than the number of observations in another class.

To overcome this, we have applied Synthetic Minority Over-sampling Technique (SMOTE) to increase the instances of minority class in the dataset. The python snippet for balancing our dataset is shown in Fig. 3.7. The SMOTE algorithm works as follows:

- For each minority instance  $x_i$  in a dataset  $D$ , we find its  $k$  nearest neighbors.
- For each minority instance  $x_i$ , we generate synthetic instances by randomly selecting one of its  $k$  nearest neighbors  $x_j$ , and creating a new instance  $x_k$  as shown in Eqn. 3.1:

$$x_k = x_i + (x_j - x_i) * r \quad (3.1)$$

where  $r$  is a random number between 0 and 1.

- We repeat the above steps for each minority instance in the dataset  $D$ , and combine the original and synthetic instances to create a new balanced dataset  $D'$ .

The SMOTE algorithm has several parameters that can be adjusted to control its behavior. The most important parameters are the number of synthetic samples to be generated and the number of nearest neighbors ( $k\_neighbors$ ) to consider. In this study,

the number of synthetic samples to be generated were **729** (781 - 52) and we have kept *k\_neighbors* as **5**.

```
1 from imblearn.over_sampling import SMOTE
2
3 X = df[['short_term_rainfall', 'long_term_rainfall', '
         elevation_relief']]
4 Y = df['isRainfallInducedLandslide']
5
6 over_sampler = SMOTE(sampling_strategy='minority', random_state=42)
7 X_NEW, Y_NEW = over_sampler.fit_resample(X, Y)
```

Figure 3.7: Python Snippet for balancing dataset using SMOTE.

As a result, the dataset consists of 781 rainfall-induced landslides and 781 landslides not induced by rainfall. This data was split into training (67%) and test (33%) sets which then were used for training and assessing our models.

### 3.2.3.2 Logistic Regression

In machine learning, logistic regression is a supervised learning algorithm used for binary classification problems. The goal of logistic regression is to learn a function that maps input variables to a binary output variable. The output variable represents the probability of the input belonging to one of the two possible classes.

The mathematical formula for logistic regression is shown in Eqn. 3.2 and 3.3:

$$p = 1/(1 + e^z) \quad (3.2)$$

$$z = -(w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n) \quad (3.3)$$

where:

$p$  is the predicted probability of the binary output variable (i.e., the probability of the input belonging to the positive class)

$e$  is the mathematical constant

$w_0$  is the intercept or bias term

$w_1, w_2, \dots, w_n$  are the weights or coefficients associated with the input variables

$x_1, x_2, \dots, x_n$

$x_1, x_2, \dots, x_n$  are the input variables



The equation represents a logistic function that maps the input variables to a probability value between 0 and 1. The weights or coefficients are learned during the training phase using an optimization algorithm that minimizes the difference between the predicted probabilities and the actual binary output values in the training data.

To make a binary classification prediction using the logistic regression model, a threshold value is chosen, typically 0.5, above which the input is classified as the positive class, and below which it is classified as the negative class.

### 3.2.3.3 Decision Tree

The decision tree algorithm is a supervised machine learning technique that uses a tree-like model to make decisions based on multiple input variables. The basic idea behind the algorithm is to create a model that splits the data into subsets, and then recursively splits each subset again until a stopping criterion is met. Each split is based on a specific input variable, and the goal is to find the split that maximizes the separation of the data into classes. Below is the high-level overview of how the decision tree algorithm works.

- **Choosing a Split variable:** The algorithm starts by selecting an input variable that is most useful in dividing the data into different classes. There are several methods for selecting the best variable, including the **Gini Index**, **Entropy**, and **Information Gain**. The selected variable is used to split the data into two or more subsets.
- **Data Splitting:** Based on the selected variable, the data is split. For example, if the variable is age, the data might be split into two subsets: one for people under 30 and one for people over 30. Each subset contains only the data that meets the criteria for that subset.
- **Building Tree:** The process of splitting the data is repeated recursively for each subset until a stopping criterion is met. The stopping criterion might be a maximum depth of the tree or a minimum number of data points in each leaf node. As the algorithm proceeds, a tree-like model is built, with nodes representing the input variables and edges representing the possible outcomes of those variables.
- **Pruning Tree:** Once the tree is built, it may be too complex and overfit to the training data. To avoid this, the algorithm can prune the tree by removing branches that do not improve the accuracy of the model. This is done using a validation set or cross-validation.

- **Prediction:** Once the tree is built and pruned, it can be used to make predictions on new data. To do this, the input variables for the new data are evaluated at each node of the tree, following the edges that correspond to the input values, until a leaf node is reached. The class label associated with that leaf node is then used as the predicted class for the new data point.

In this study, we have used **DecisionTreeClassifier** from the **scikit-learn** library for classifying rainfall-induced and non rainfall-induced landslides. In decision trees, the cutoff point can be defined as a condition or stopping criterion that determines when to stop splitting the tree and create a final set of leaf nodes.

The most common criterion for the cutoff point is based on the number of samples or instances at each node in the tree. Specifically, the cutoff point can be defined as a function of the number of samples or instances  $n$  at a given node.

The following are some of the ways a cutoff criteria can be defined mathematically:

- *max\_depth*: If the depth of the tree reaches a certain threshold called *max\_depth*, the tree stops growing and creates a final set of leaf nodes. This criterion ensures that the tree does not become too deep and avoids overfitting. It's value was kept as 3.
- $n < min\_samples\_split$ : If the number of samples at a node is less than a certain threshold called *min\_samples\_split*, the node is not split further, and a leaf node is created. This criterion ensures that the tree does not create nodes that are too specific to the training data and avoids overfitting. It's value was kept as 2.
- $n < min\_samples\_leaf$ : If the number of samples at a leaf node is less than a certain threshold called *min\_samples\_leaf*, the node is considered as overfitting and not used. This criterion ensures that the tree does not create too many small leaf nodes and ensures that each leaf node has enough samples to generalize well. It's value was kept as 1.

### 3.2.3.4 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm that can be used for both classification and regression tasks. In this algorithm, a hyperplane is constructed to separate the input data into different classes, and the goal is to maximize the margin between the classes.

Let's consider a simple binary classification problem, where we have two classes labeled as  $+1$  and  $-1$ . We have a set of training data points, represented as a set of feature vectors  $x_i$  and their corresponding labels  $y_i$ . SVM seeks to find a hyperplane in the feature space that can best separate the two classes. The Eqn. 3.4 show the definition of hyperplane:

$$w^T x + b = 0 \quad (3.4)$$

where  $w$  is a weight vector that determines the orientation of the hyperplane,  $b$  is a scalar that determines its position, and  $x$  is a feature vector. The goal is to find the optimal values of  $w$  and  $b$  such that the hyperplane can separate the two classes with maximum margin.

To find the optimal hyperplane, SVM tries to solve the optimization problem depicted by Eqn. 3.5 and 3.6:

$$\text{minimize } (1/2) \|w\|^2 \quad (3.5)$$

$$\text{subject to } y_i(w^T x_i + b) \geq 1, \text{ for all } i = 1, 2, \dots, n \quad (3.6)$$

where  $\|w\|$  is the Euclidean norm of the weight vector, and  $n$  is the number of training data points. The constraint ensures that all data points are correctly classified and lie on the correct side of the hyperplane.

### 3.2.3.5 Gaussian Naive Bayes

Gaussian Naive Bayes is a probabilistic algorithm used for classification tasks. It is based on Bayes' theorem, which is a fundamental concept in probability theory that states the probability of an event occurring given some prior knowledge about the event.

In the context of classification, the goal is to determine the probability of an input belonging to a particular class.

The Gaussian Naive Bayes algorithm can be mathematically formulated as follows:

Let  $X = (x_1, x_2, \dots, x_n)$  be a vector of  $n$  input features, and let  $C = c_1, c_2, \dots, c_k$  be a set of  $k$  possible class labels.

#### Training

- Estimate the prior probability of each class  $P(c_i)$  by calculating the proportion of training instances that belong to each class.

- Estimate the mean and variance of each input feature for each class using the training instances for that class.

### Prediction

Given a new instance  $x$ , calculate the posterior probability  $P(c_i|x)$  for each class  $c_i$  using Bayes' theorem shown in Eqn. 3.7

$$P(c_i|x) = P(c_i) * P(x|c_i) / P(x) \quad (3.7)$$

where  $P(x|c_i)$  is the likelihood of the instance  $x$  given class  $c_i$ , which can be calculated using the Gaussian probability density function shown in Eqn. 3.8:

$$P(x|c_i) = (1/(sqrt(2\pi) * \sigma_{c_i})) * exp(-(x - \mu_{c_i})^2 / (2 * \sigma_{c_i}^2)) \quad (3.8)$$

where  $\mu_{c_i}$  and  $\sigma_{c_i}$  are the estimated mean and standard deviation of the input feature distribution for class  $c_i$ .

$P(x)$  is a normalizing constant that ensures that the posterior probabilities sum to 1 and it is calculated as shown in Eqn. 3.9:

$$P(x) = \sum_i P(c_i) * P(x|c_i) \quad (3.9)$$

The predicted class for the instance  $x$  is the class with the highest posterior probability and it can be deduced using Eqn. 3.10:

$$\text{argmax } P(c_i|x) \text{ for } c_i \text{ in } C \quad (3.10)$$

Note that the assumption of independence between the input features is taken into account by assuming that the joint probability distribution of the input features can be decomposed into the product of the individual probabilities of each feature, given the class label. This assumption of independence can be shown mathematically as Eqn. 3.11.

$$P(x|c_i) = \prod_j P(x_j|c_i) \quad (3.11)$$

where  $x_j$  is the value of the  $j$ -th input feature of the instance  $x$ .

## Chapter 4

# Data Analytics for Landslide Prediction

### 4.1 Results

At this stage, we have used logistic regression, decision tree, support vector machine and gaussian naive bayes algorithms to decide between landslide and non-landslide cases. For this, the data was split into training (67%) and test (33%) sets which then were used for training and assessing our models.

Several metrics can be used to evaluate the accuracy of machine learning models. Here, the metrics used are Accuracy Score, and Receiver Operating Characteristic (ROC) Curve.

The mathematical formula for Accuracy metric is shown in the Eqn. 4.1:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (4.1)$$

where:

True Positives (TP) = Number of cases where the model predicted positive and the actual value was positive

False Positives (FP) = Number of cases where the model predicted positive and the actual value was negative

True Negatives (TN) = Number of cases where the model predicted negative and the actual value was negative

False Negatives (FN) = Number of cases where the model predicted negative and the actual value was positive

### 4.1.1 ROC Curve

The Receiver Operating Characteristic (ROC) Curve is a graphical representation of the performance of a model. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) for different threshold values.

The true positive rate is the proportion of actual positive examples that are correctly classified as positive by the model, while the false positive rate is the proportion of negative examples that are incorrectly classified as positive.

To calculate the TPR and FPR for a given threshold, first we need to count the number of true positives, false positives, true negatives, and false negatives.

The mathematical formulation for true positive rate and false positive rate is shown in Eqn. 4.2 and 4.3:

$$TPR = TP / (TP + FN) \quad (4.2)$$

$$FPR = FP / (FP + TN) \quad (4.3)$$

We can repeat this calculation for different classification thresholds to get a set of TPR and FPR values. These values can be plotted on a graph with TPR on the y-axis and FPR on the x-axis to create the ROC curve.

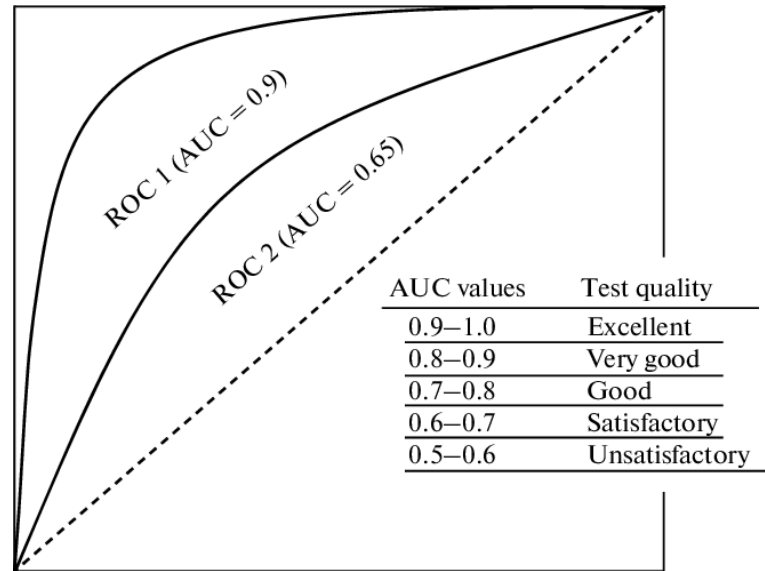


Figure 4.1: An example ROC curves with a scale of AUC values

AUC (Area Under the ROC Curve) is a commonly used metric to evaluate the performance of a model, with an AUC of 1 indicating a perfect classifier and an AUC of 0.5 indicating a random classifier as shown in Fig. 4.1.

### 4.1.2 Logistic Regression

Fig. 4.2 and Fig. 4.3 show the ROC Curve and the Confusion Matrix respectively for Logistic Regression Model. From the confusion matrix, we can see that we obtain an accuracy of 0.79462. The area under the ROC Curve is observed to be 0.91, which signifies the model has great discriminatory ability.

A macro-average will compute the metric independently for each class and then take the average hence treating all classes equally, whereas a micro-average will aggregate the contributions of all classes to compute the average metric. A macro-average and micro-average are generally used in multi-class or imbalanced datasets. In our case, macro-average roc and micro average roc will come to be similar to the roc curve of individual classes as our dataset was not imbalanced.

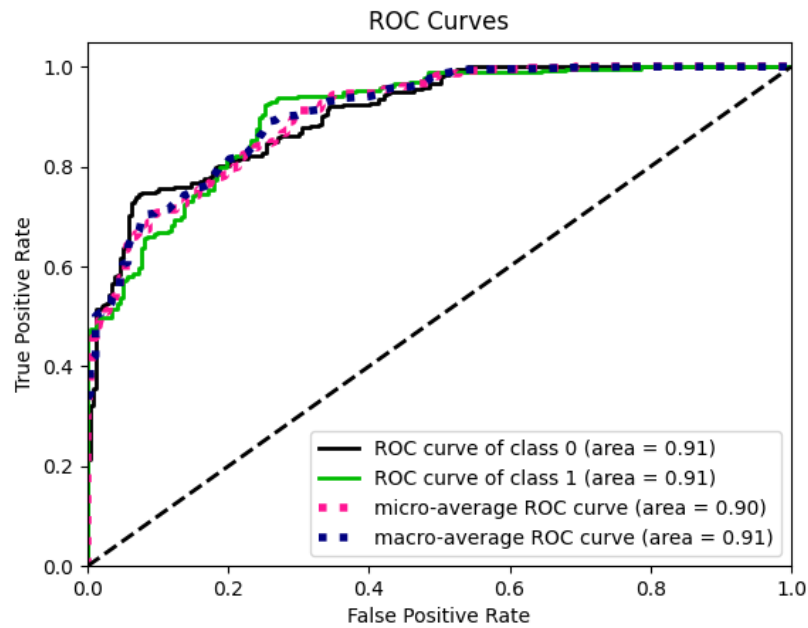


Figure 4.2: Logistic Regression ROC Curve.

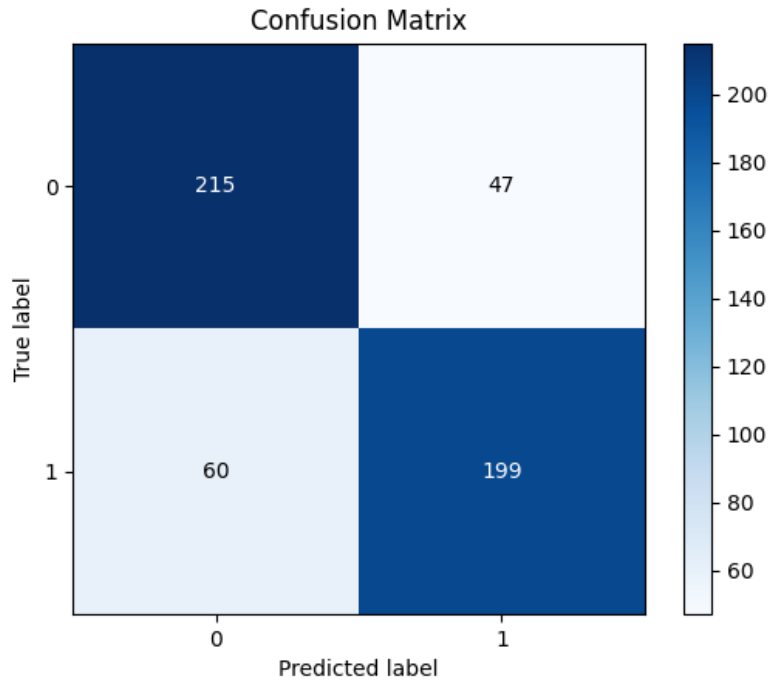


Figure 4.3: Logistic Regression Confusion Matrix.

### 4.1.3 Decision Tree

Fig. 4.4 and Fig. 4.5 show the ROC Curve and the Confusion Matrix respectively for Decision Tree Classifier. From the confusion matrix, we can see that we obtain an accuracy of 0.85604. The area under the ROC Curve is observed to be 0.91, which signifies the model has great discriminatory ability.

A macro-average will compute the metric independently for each class and then take the average hence treating all classes equally, whereas a micro-average will aggregate the contributions of all classes to compute the average metric. A macro-average and micro-average are generally used in multi-class or imbalanced datasets. In our case, macro-average roc and micro average roc will come to be similar to the roc curve of individual classes as our dataset was not imbalanced.



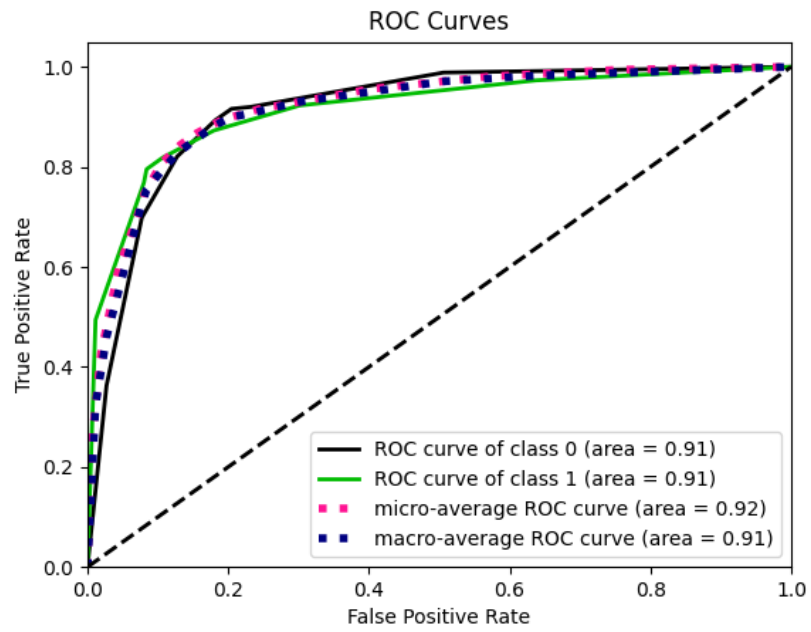


Figure 4.4: Decision Tree ROC Curve.

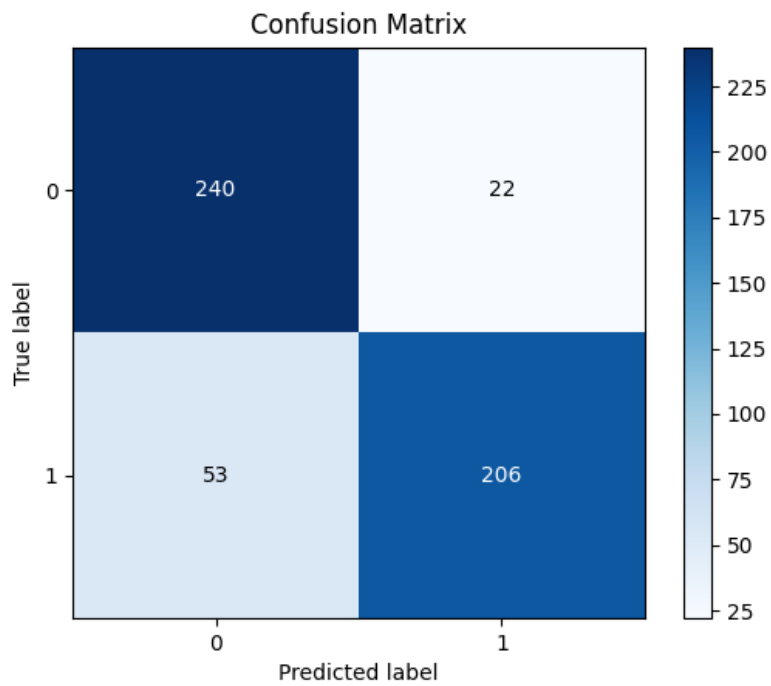


Figure 4.5: Decision Tree Confusion Matrix.

#### 4.1.4 Support Vector Machine

Fig. 4.6 and Fig. 4.7 show the ROC Curve and the Confusion Matrix respectively for Support Vector Machine Model. From the confusion matrix, we can see that we obtain an accuracy of 0.82725. The area under the ROC Curve is observed to be 0.91, which signifies the model has great discriminatory ability.

A macro-average will compute the metric independently for each class and then take the average hence treating all classes equally, whereas a micro-average will aggregate the contributions of all classes to compute the average metric. A macro-average and micro-average are generally used in multi-class or imbalanced datasets. In our case, macro-average roc and micro average roc will come to be similar to the roc curve of individual classes as our dataset was not imbalanced.

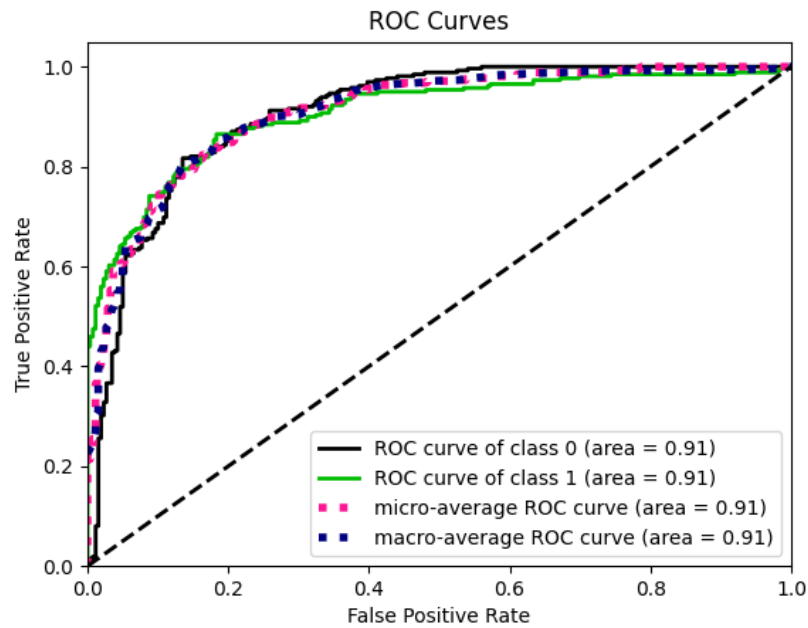


Figure 4.6: Support Vector Machine ROC Curve.

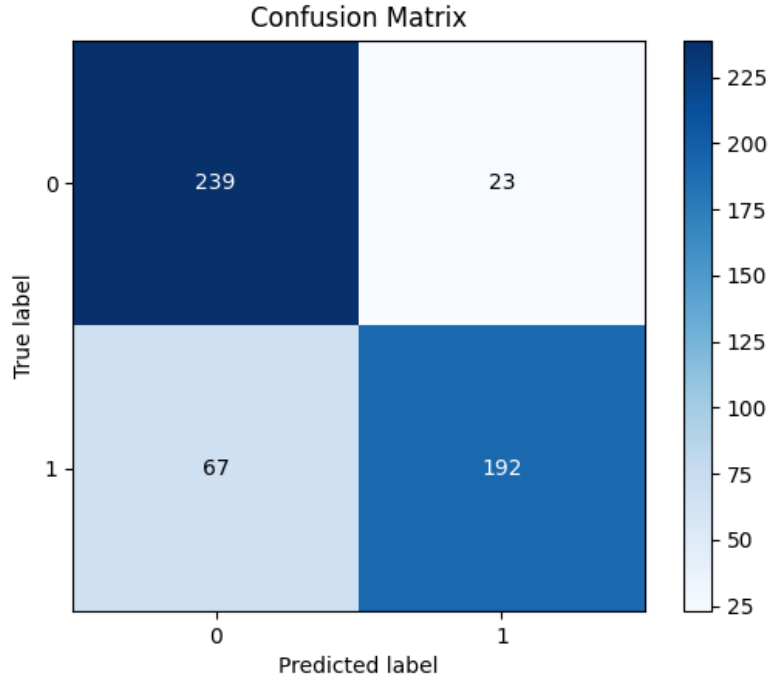


Figure 4.7: Support Vector Machine Confusion Matrix.

### 4.1.5 Gaussian Naive Bayes

Fig. 4.8 and Fig. 4.9 show the ROC Curve and the Confusion Matrix respectively for Gaussian Naive Bayes Model. From the confusion matrix, we can see that we obtain an accuracy of 0.77351. The area under the ROC Curve is observed to be 0.89, which signifies the model has great discriminatory ability.

A macro-average will compute the metric independently for each class and then take the average hence treating all classes equally, whereas a micro-average will aggregate the contributions of all classes to compute the average metric. A macro-average and micro-average are generally used in multi-class or imbalanced datasets. In our case, macro-average roc and micro average roc will come to be similar to the roc curve of individual classes as our dataset was not imbalanced.

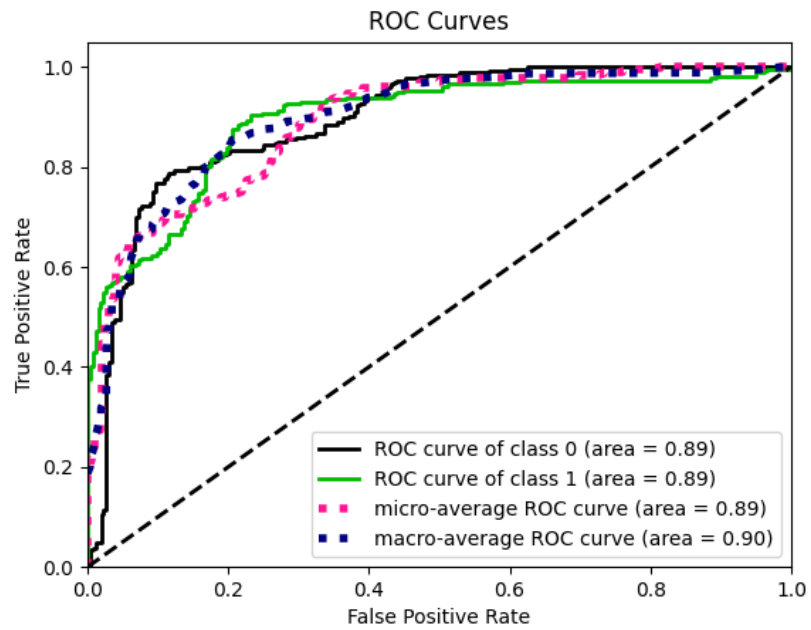


Figure 4.8: Gaussian Naive Bayes ROC Curve.

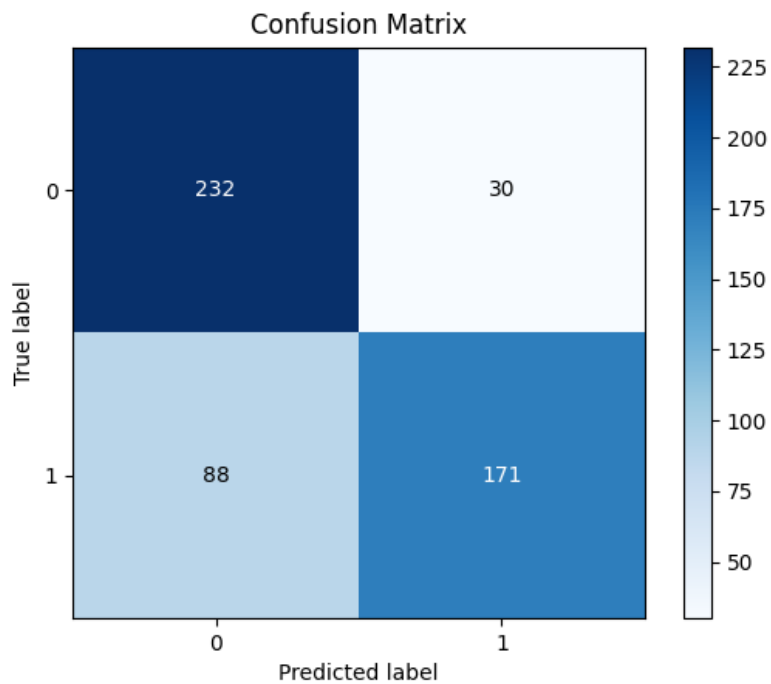


Figure 4.9: Gaussian Naive Bayes Confusion Matrix.

S.No	Algorithm	Accuracy	Area Under ROC Curve (AUC)
1	Logistic Regression	0.79462	0.91
2	Decision Tree	0.85604	0.91
3	Support Vector Machine	0.82725	0.91
4	Gaussian Naive Bayes	0.77351	0.89

Table 4.1: Comparison of performance metrics of trained models.

Table 4.1 shows the performance of all the models taken into consideration for this study. It can be seen that from all 4 models, **Decision Tree** Classification Model shows the **best results**.

#### 4.1.6 Landslide Visualization

To visualise the data we have obtained, we use OpenStreetMap (OSM) which is a free and open-source mapping platform that allows users to access and contribute to a wide range of geographic data.

To create interactive maps and add features like hovering dialogue, shapefile shading, we have used the Folium library in python. Folium is a Python library that allows users to create customized and interactive maps using data from various sources, including GeoJSON files and Pandas data frames.

The map in Fig. 4.10 displays the predicted probability of landslides in various districts of India based on a machine learning model.

The severity of the landslide probability in each district is indicated by the darkness of the color, with darker shades representing higher probabilities. The map was created using the Python libraries Folium and GeoJSON, which allowed for the overlaying of the landslide probability data onto a geographic map of India.

On hovering over the highlighted areas we can see the name of the particular district with the probability of landslide occurring in it. The probability we obtained comes from the decision tree classification model mentioned earlier.

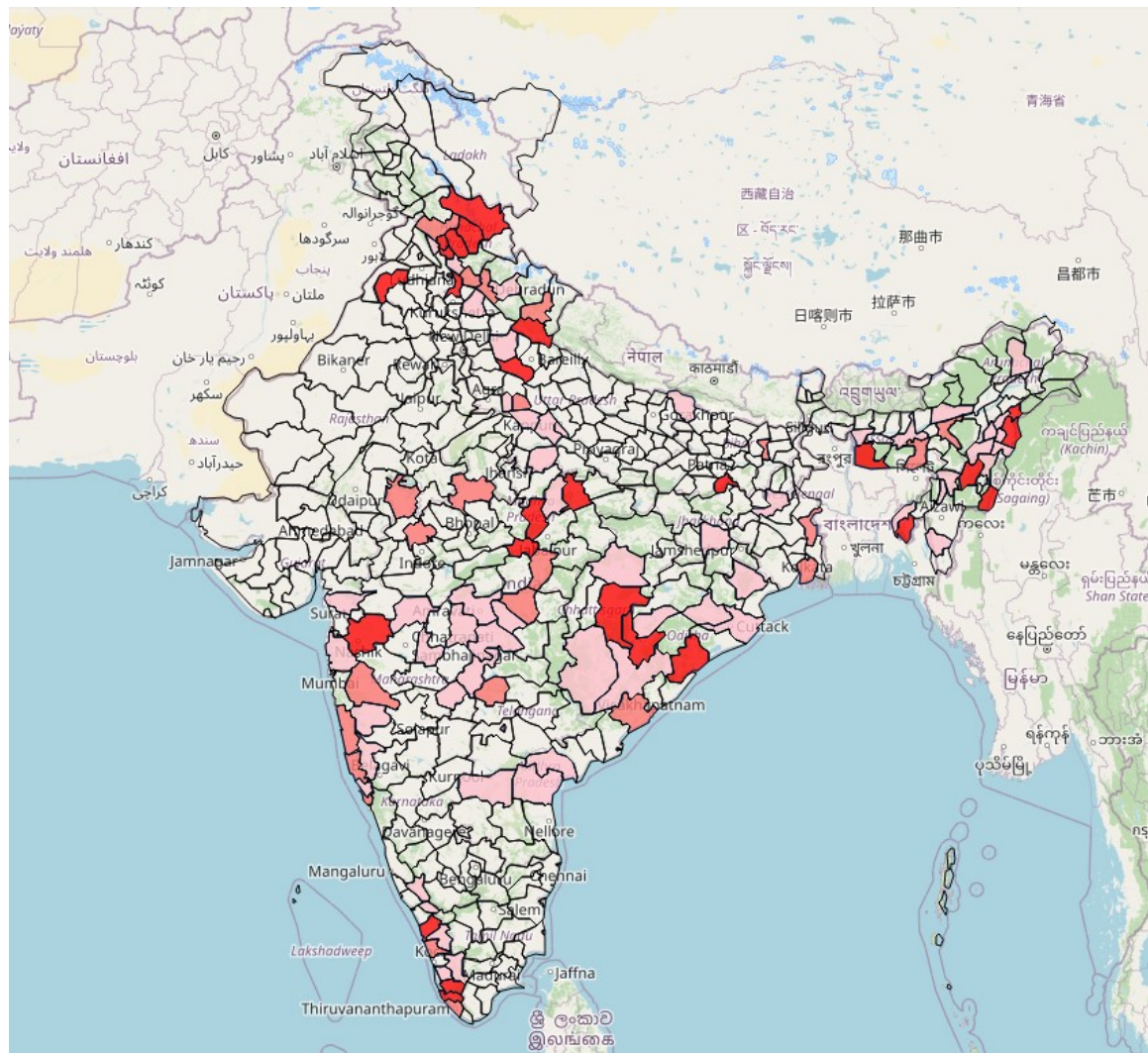


Figure 4.10: Shaded Regions Affected By Landslides.

### 4.1.7 Source Code Link

GitHub Repository Link for Code and Plots: [Link](#)

## **Chapter 5**

# **Gender-Based Twitter Analysis of Joshimath Crisis**

Joshimath, a town situated in Chamoli district of the state of Uttarakhand, is a renowned hiking and pilgrimage destination. The town is located on a fragile mountain slope in a region that is prone to landslides. Several reports came afore in which houses and roads started developing cracks as the land beneath the town continued to sink gradually. The land subsidence intensified starting from the month of December 2022, that got the whole country concerned over the fate of the town.

According to images released by the National Remote Sensing Centre of the Indian Space Research Organisation, Uttarakhand's Joshimath has witnessed a rapid subsidence of nearly 5.4 cm from Dec 27, 2022 to Jan 8, 2023 and the report stated that a subsidence of nearly 9 cm was recorded between April-November 2022 [12].

In this study, we have focused on gender analysis and sentiment analysis of tweets related to Joshimath Disaster. This has helped us in determining which sets of words are most commonly used by males, females, and organizations which may reflect the unique perspectives and priorities of each group, and may have implications for the types of actions and solutions that each group advocates for in response to the disaster.

### **5.1 Methodology**

In the study of twitter data on Joshimath crisis, the steps mentioned in the flow chart shown in the Fig. 5.1 are followed. The steps are elaborated in detail in the following parts.

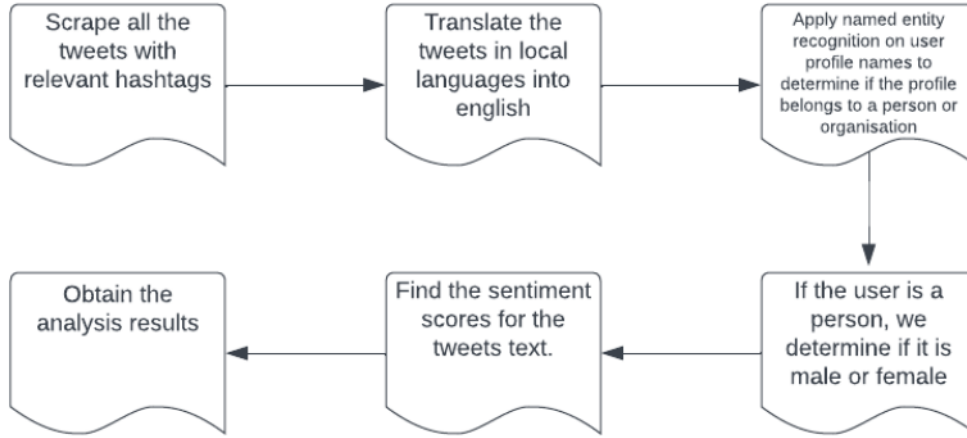


Figure 5.1: Flowchart of Gender-Based Twitter Analysis of Joshimath Crisis

### 5.1.1 Tweet Extraction and Translation

Twitter changes its API specifications frequently, so many open source libraries are outdated as a result. After carefully testing several libraries and noticing their issues, we decided to use the Scweet [13] library for tweet extraction.

- Information such as Timestamp, Text, Embedded\_text, Emojis, Comments, Likes, Retweets, Image link, URL for tweets, UserScreenName, UserName, Profile Description, Join Date for a specific user can be extracted reliably.
- The following hashtags have been considered for extracting tweets from 1st December, 2022 to 8th February, 2023: #Joshimath, #SaveJoshimath, #JoshimathIsSinking, #Joshimathcrisis, #joshimathsinking.
- Some of the extracted data such as tweets' text and name of the users are in local languages. These data needed to be translated into English to be useful for our study. We use Google Translate API for this purpose.

### 5.1.2 Gender Estimation of Twitter User

The profile name of a twitter account gives an idea of the gender of the user. But the twitter profile making the tweet might not belong to a person. In our study of gender-wise analysis of tweets, there needs to be a third pool of tweets apart from male and female users. This third pool of users are the twitter profiles that belong to news channels, government and non-government organisations etc. So, the process of estimating a user's gender can be divided into two steps:



- **Person Identification:** Identify whether a tweet is made a person or an entity like news channels, government bodies etc. To accomplish this, we perform Named Entity Recognition on user's name on twitter. For performing Named Entity Recognition, we use spacy [14] library. The result of Named Entity Recognition is classification of users into PERSON, ORG (organisations) etc. The Named Entity Recognition is carried out by spacy model called `en_core_web_trf` with accuracy of 89.8% [15].
- **Gender Estimation:** If the result of Named Entity Recognition on the user's name is PERSON, we use the name-based gender classification API provided by `genderize.io` [16] to determine their gender.

### 5.1.3 Unigrams and Bigrams extraction from the tweets text

Analysis of the words that are used by the users in their tweets can also tell us about the topics that they want to comment upon. For this, we extract unigrams and bigrams from all the tweets text. N-grams are n continuous sequence of words or tokens in a document. One word is called unigram and two words sequence is called bigram.

The text content of the tweets need to be pre-processed before we proceed with extraction of n-grams. This includes converting to lowercase, removing punctuations and unwanted characters, tokenizing and lemmatizing. We have extracted unigrams and bigrams along with the frequency of their occurrence for all the tweets and also male, female and non-persons separately.

### 5.1.4 Sentiment Analysis of Tweets

Sentiment analysis is a useful tool in disaster management and crisis response. By analyzing social media content, such as Twitter, sentiment analysis can provide insights into the public's reactions and emotions during a disaster. For example, sentiment analysis can identify areas where people are expressing fear, anger, or frustration and help prioritize the allocation of resources and assistance. This information can help emergency responders, relief organizations, and government agencies make better-informed decisions and respond more effectively to the needs of affected communities.

For sentiment analysis, we use VADER [17] library. It classifies tweets into positive, negative and neutral based upon the sentiment scores of the tweets' text. The sentiment score ranges from -1 (highly negative) to 1 (highly positive). For example, a tweet is : *"Although this was posted just 10 days ago on Dec 27, it feels like a long time.*

*It seems unbelievable that the situation has deteriorated so rapidly in #Joshimath in #Uttarakhand. In our own ways, let us express solidarity & hope for the best for the people of the battered town.”*

The sentiment score of the above text is 0.967, which shows that the sentiment of the text is highly positive.

## Chapter 6

# Data Analytics for Gender-Based Twitter Analysis of Joshimath Crisis

### 6.1 Results

#### 6.1.1 Tweet Extraction and Translation

From 1st December, 2022 to 8th February, 2023, a total of **5856** Tweets have been extracted using hashtags mentioned above. The complete dataset in CSV format is available here [18].

#### 6.1.2 Gender Estimation of Twitter User

Table 6.1 and Fig. 6.1 show the gender estimation probability distribution and its bar chart respectively. Apart of that, following are some statistics based on the results of gender estimation:

- Total Number of Tweets Extracted: 5856
- Total Number of Male Tweets: 1965
- Total Number of Female Tweets: 464
- Total Number of Male Users: 1154
- Total Number of Female Users: 259
- Total Number of Persons: 1732 (1154 + 259 + people whose gender is unidentified)

- Among male and female users, their gender has been determined with:
  - minimum probability of 50%
  - maximum probability of 100%
  - average probability of 98% (weighted average)

<b>Probability Range (%)</b>	50-59	60-69	70-79	80-89	90-99	100
<b>No. of users</b>	9	11	16	52	260	1065

Table 6.1: Gender Estimation Probability Distribution.

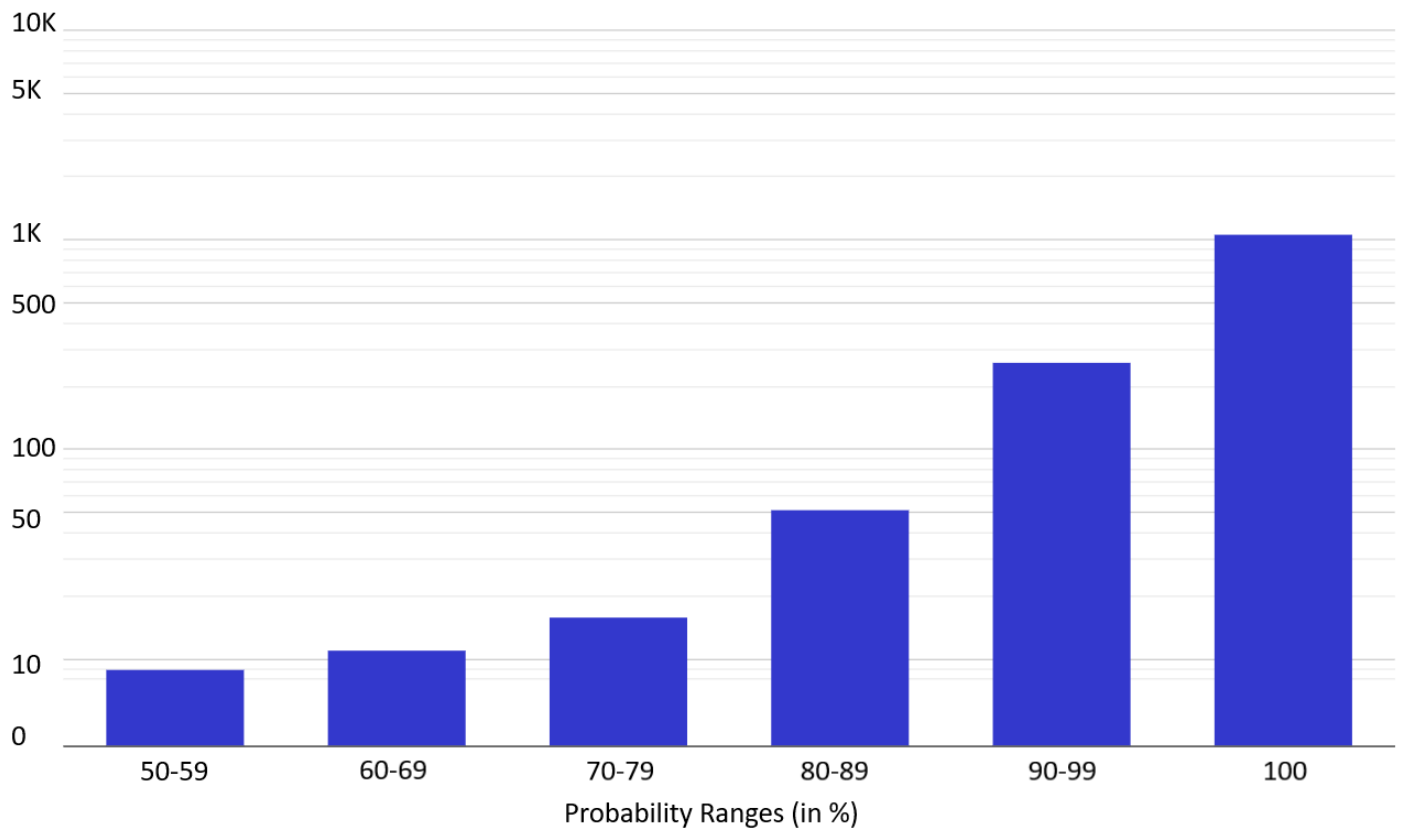


Figure 6.1: Gender Estimation Probability Bar Chart.



unique perspectives and priorities of each group, and may have implications for the types of actions and solutions that each group advocates for in response to the disaster.

Unigrams	Bigrams
<b>All Tweets</b>	
joshimath, sinking, uttarakhand, people, crack, house, crisis, government, landslide, land, disaster, save, subsidence, news, family, cm, also, affected, dhami, town	joshimath sinking, joshimath crisis, save joshimath, land subsidence, joshimath uttarakhand, sinking joshimath, chief minister, supreme court, pushkar singh, crack house
<b>Males</b>	
joshimath, sinking, people, uttarakhand, crisis, house, government, crack, save, landslide, disaster, also, land, development, family, city, due, town, replying, report	joshimath sinking, joshimath crisis, save joshimath, joshimath uttarakhand, sinking joshimath, land subsidence, landslide joshimath, people joshimath, joshimath landside, supreme court
<b>Females</b>	
joshimath, sinking, uttarakhand, house, people, save, crisis, crack, land, government, landslide, development, disaster, town, city, subsidence, report, himalaya, nature, also	joshimath sinking, joshimath crisis, save joshimath, sinking joshimath, land subsidence, joshimath uttarakhand, people joshimath, joshimath landslide, crack house, uttarakhand joshimath
<b>Organisations</b>	
joshimath, sinking, uttarakhand, crack, crisis, house, people, land, landslide, government, disaster, dhami, cm, subsidence, affected, news, family, minister, due, also	joshimath sinking, joshimath crisis, land subsidence, joshimath uttarakhand, sinking joshimath, chief minister, pushkar singh, cm dhami, supreme court, singh dhami

Table 6.2: Top-20 Unigrams and Top-10 Bigrams from all entity types.

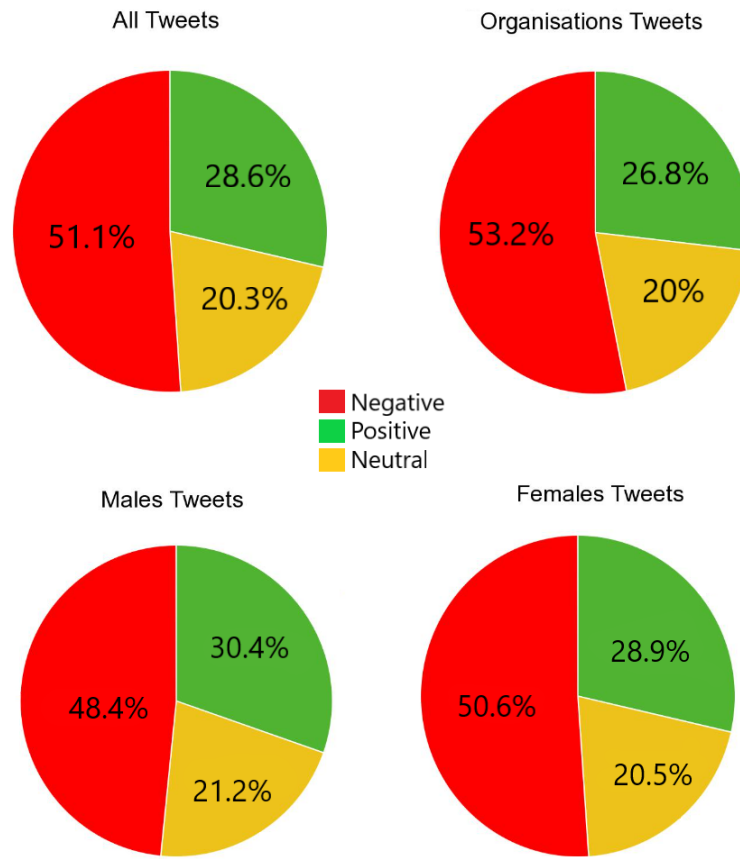


Figure 6.3: Percentage sentiments distribution across all entity types.

#### 6.1.4 Sentiment Analysis of Tweets

On combining the gender estimation results with the sentiment analysis results, we can deduce gender-wise sentiments of the people and also see how organisations are perceiving this incident. This can help us get a better understanding of the commonalities and differences in views and sentiments of different stakeholders.

Pie chart representations of the percentage of tweets made from different types of twitter profiles are shown in Fig. 6.3. Pie charts for sentiment analysis for gender-wise tweets are shown along with those of tweets from organisations. On analysing the charts, the following statements can be made:

- We can say that all groups have a predominantly negative sentiment in response to the disaster. The fact that more than half of the tweets are negative suggests that the event had a significant impact on the collective emotional state of the individuals and organizations.
- The fact that organizations, which are typically perceived as more stoic and less

emotional than individuals, have a similar sentiment distribution to individuals suggests that the crisis elicited a strong emotional response from all stakeholders.

### **6.1.5 Source Code Link**

**GitHub Repository Link for Code and Plots:** [Link](#)



# Chapter 7

## Conclusions and Future Work

### 7.1 Landslides Prediction

Landslides being a widespread issue across the globe, it is becoming increasingly necessary to have a better understanding of landslide occurrences and the factors which can be beneficial for predicting it.

While our results provide valuable insights into the relationship between landslides vs rainfall/elevation relief, there are several areas that could be explored in future research. One important avenue for future research would be to examine the impact of other factors such as Vegetation, soil and bedrock on occurrence of landslides. Additionally, further research is needed to enhance the datasets like accommodating more landslides cases from different sources in current dataset. Finally, it would be valuable to investigate the performance/accuracy of other machine learning and deep learning algorithms.

Even then, our results prove that with better datasets, these systems can find great use in early warning systems for forecasting landslides caused due to rainfall.

### 7.2 Gender-Based Twitter Analysis of Joshimath Crisis

Twitter contains a large amount of raw data in the form of text, photos, videos and audios, that has been uploaded by users in the form of tweets. For the Joshimath land subsidence incident, a textual analysis on the tweets' texts highlights the commonalities and differences in views and sentiments of both genders as well as organizations. From the analysis, we find that although all the groups are discussing the same topics, there are some notable differences in their response.

- More concern on the economic and infrastructural impact of the disaster is shown by the male group.

- More concern to the environmental impact of the disaster is shown by the female group.
- The organisations show a greater concern to the political response to the disaster along with economic and environmental impact.

For the scope of further research, the gender estimation can be improved by graphical analysis of user's profile image. Also performing Social Network Analysis, a research method developed primarily in sociology and communication science, can provide a comprehensive understanding of the network structure, information diffusion, and gender-specific dynamics related to the Joshimath crisis on social media.

# Bibliography

- [1] F. Tehrani, G. Santinelli, and M. Herrera, “A framework for predicting rainfall-induced landslides using machine learning methods un cadre pour prédire les glissements de terrain induits par les précipitations à l’aide d’un apprentissage automatique,” 09 2019.
- [2] M. J. Froude and D. N. Petley, “Global fatal landslide occurrence from 2004 to 2016,” *Nat. Hazards Earth Syst. Sci.*, vol. 18, p. 2161–2181, 2018.
- [3] I. S. R. Institute, “Disaster management: National and international,” accessed: Nov 27, 2022. [Online]. Available: <https://www.isro.gov.in/DisaterManagementNationalInternational.html>
- [4] D. Kumar, M. Thakur, C. Dubey, and D. Shukla, “Landslide susceptibility mapping and prediction using support vector machine for mandakini river basin, garhwal himalaya, india,” *Geomorphology*, vol. 295, 06 2017.
- [5] Y. R M and B. Dolui, “Statistical and machine intelligence based model for landslide susceptibility mapping of nilgiri district in india,” *Environmental Challenges*, vol. 5, p. 100211, 07 2021.
- [6] S. Meena, O. Ghorbanzadeh, C. Westen, T. Gudiyangada, T. Blaschke, R. Singh, and R. Sarkar, “Rapid mapping of landslides in the western ghats (india) triggered by 2018 extreme monsoon rainfall using a deep learning approach,” *Landslides*, 01 2021.
- [7] D. Kirschbaum, R. Adler, Y. Hong, S. Hill, and A. Lerner-Lam, “A global landslide catalog for hazard applications: Method, results, and limitations,” *Natural Hazards*, vol. 52, pp. 561–575, 03 2009.
- [8] Open-Meteo, *Historical Weather API*, accessed: Apr 10, 2023. [Online]. Available: <https://open-meteo.com/en/docs/historical-weather-api>
- [9] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, and J.-N. Thépaut, “The era5 global reanalysis,” *Quarterly Journal of the Royal Meteorological Society*, 05 2020.
- [10] T. Farr, P. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, L. Roth, D. Seal, S. Shaffer, J. Shimada, J. Umland, M. Werner, M. Oskin, D. Burbank, and D. Alsdorf, “The shuttle radar topography mission,” *Rev. Geophys.*, vol. 45, 06 2007.
- [11] S. He, J. Wang, and S. Liu, “Rainfall event–duration thresholds for landslide occurrences in china,” *Water*, vol. 12, no. 2, 2020. [Online]. Available: <https://www.mdpi.com/2073-4441/12/2/494>
- [12] T. Hindu, “Joshimath sank by 5.4 cm in 12 days, says isro report,” accessed: Feb 25, 2023. [Online]. Available: <https://www.thehindu.com/sci-tech/energy-and-environment/isro-releases-satellite-images-showing-rise-in-joshimath-land-subsidence/article66373138.ece>
- [13] Y. A. Jeddi, *A simple and unlimited twitter scraper*, accessed: Feb 25, 2023. [Online]. Available: <https://github.com/Altimis/Scweet>
- [14] Explosion, *Industrial-strength Natural Language Processing (NLP) in Python*, accessed: Nov 27, 2022. [Online]. Available: <https://pypi.org/project/spacy/>

- [15] —, *Facts and Figures, The hard numbers for spaCy and how it compares to other tools*, accessed: May 12, 2023. [Online]. Available: <https://spacy.io/usage/facts-figures>
- [16] C. Strømgren, *genderize.io: A simple API to predict the gender of a person given their name*, accessed: Feb 25, 2023. [Online]. Available: <https://genderize.io>
- [17] C. Hutto, *VADER (Valence Aware Dictionary and sentiment Reasoner)*, accessed: Feb 25, 2023. [Online]. Available: <https://pypi.org/project/vaderSentiment/>
- [18] H. Thami, Purushottam, and D. Chaudhari, “Twitter dataset,” accessed: Feb 25, 2023. [Online]. Available: <https://docs.google.com/spreadsheets/d/1kg6jav48zBSfbXroRskB3rFiOWSvnG-TyQyipQj-0Q/edit?usp=sharing>

# BTP\_Hardik\_Final\_Report\_tin.pdf

## ORIGINALITY REPORT

17%

SIMILARITY INDEX

11%

INTERNET SOURCES

9%

PUBLICATIONS

10%

STUDENT PAPERS

## PRIMARY SOURCES

1	tinmarino.github.io Internet Source	2%
2	Submitted to Middle East College of Information Technology Student Paper	1%
3	Submitted to Napier University Student Paper	1%
4	ebin.pub Internet Source	1%
5	www.science.gov Internet Source	1%
6	Submitted to CSU, San Jose State University Student Paper	1%
7	www.thehindu.com Internet Source	1%
8	Submitted to University of Westminster Student Paper	1%
9	Submitted to Aston University Student Paper	<1%

10

Submitted to Liverpool John Moores  
University

Student Paper

<1 %

11

Submitted to University of Queensland

Student Paper

<1 %

12

[eprints.soton.ac.uk](https://eprints.soton.ac.uk)

Internet Source

<1 %

13

Submitted to Leeds Beckett University

Student Paper

<1 %

14

Priyom Roy, Tapas R. Martha, Kirti Khanna,  
Nirmala Jain, K. Vinod Kumar. "Time and path  
prediction of landslides using InSAR and flow  
model", Remote Sensing of Environment,  
2022

Publication

<1 %

15

Submitted to Birkbeck College

Student Paper

<1 %

16

R. Ramyea, S. Preethi, K. Keerthana, R.  
Keerthana, J. Kavivarman. "An Intellectual  
Supervised Machine Learning Algorithm for  
the Early Prediction of Hyperglycemia", 2021  
Innovations in Power and Advanced  
Computing Technologies (i-PACT), 2021

Publication

<1 %

17

Submitted to Colorado State University,  
Global Campus

Student Paper

<1 %

18	Pal, M.. "An assessment of the effectiveness of decision tree methods for land cover classification", Remote Sensing of Environment, 20030830 Publication	<1 %
19	Submitted to Indian Institute of Technology, Kanpur Student Paper	<1 %
20	Submitted to University of Surrey Student Paper	<1 %
21	<a href="http://www.researchgate.net">www.researchgate.net</a> Internet Source	<1 %
22	L.M. Mitnik, V.P. Kuleshov, M.K. Pichugin, M.L. Mitnik. "Sudden Stratospheric Warming in 2015–2016: Study with Satellite Passive Microwave Data and ERA5 Reanalysis", IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018 Publication	<1 %
23	<a href="http://doc.lagout.org">doc.lagout.org</a> Internet Source	<1 %
24	<a href="http://worldwidescience.org">worldwidescience.org</a> Internet Source	<1 %
25	Submitted to British University in Egypt Student Paper	<1 %

26	epdf.pub Internet Source	<1 %
27	Elena Rangelova, Mark Huiskes. "Chapter 3 Pattern Recognition for Multimedia Content Analysis", Springer Nature, 2007 Publication	<1 %
28	Submitted to Kwame Nkrumah University of Science and Technology Student Paper	<1 %
29	Submitted to Sheffield Hallam University Student Paper	<1 %
30	Sukarna Barua. "A Novel Synthetic Minority Oversampling Technique for Imbalanced Data Set Learning", Lecture Notes in Computer Science, 2011 Publication	<1 %
31	apprize.best Internet Source	<1 %
32	byjus.com Internet Source	<1 %
33	"Advancing Culture of Living with Landslides", Springer Science and Business Media LLC, 2017 Publication	<1 %
34	Mohammad Shahin, F. Frank Chen, Ali Hosseinzadeh, Neda Zand. "Using Machine	<1 %



Learning and Deep Learning Algorithms for  
Downtime Minimization in Manufacturing  
Systems: An Early Failure Detection  
Diagnostic Service", Research Square Platform  
LLC, 2023

Publication

35

Submitted to University College London

Student Paper

<1 %

36

Submitted to University of Abertay Dundee

Student Paper

<1 %

37

Submitted to University of Nottingham

Student Paper

<1 %

38

Yuvaraj R M, Bhagyasree Dolui. "Statistical  
and machine intelligence based model for  
landslide susceptibility mapping of Nilgiri  
district in India", Environmental Challenges,  
2021

Publication

<1 %

39

[www.sjsu.edu](http://www.sjsu.edu)

Internet Source

<1 %

40

"Proceedings of International Conference on  
Remote Sensing for Disaster Management",  
Springer Science and Business Media LLC,  
2019

Publication

<1 %

41

Ming Li, Hongyu Zhang, Rongxin Wu, Zhi-Hua  
Zhou. "Sample-based software defect

<1 %

# prediction with active and semi-supervised learning", Automated Software Engineering, 2011

Publication

42

[article.jccee.org](http://article.jccee.org)

Internet Source

<1 %

43

[bspace.buid.ac.ae](http://bspace.buid.ac.ae)

Internet Source

<1 %

44

[link.springer.com](http://link.springer.com)

Internet Source

<1 %

45

Shabtai, A.. "Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey",  
Information Security Technical Report, 200902

Publication

<1 %

46

[libraries.io](http://libraries.io)

Internet Source

<1 %

47

Submitted to University of Northumbria at  
Newcastle

Student Paper

<1 %

48

[iieta.org](http://iieta.org)

Internet Source

<1 %

49

[uclqkem.wixsite.com](http://uclqkem.wixsite.com)

Internet Source

<1 %

50

[www.scirp.org](http://www.scirp.org)

Internet Source

<1 %

51	<a href="http://www.mdpi.com">www.mdpi.com</a> Internet Source	<1 %
52	Kaleab Wondemu Nuri, Michee Sanza Kanda, Elikana Kulwa Justine, Amiya Ranjan Panda, Himanshu Sekhar Pradhan. "Predicting Contraceptive Usage for Married African Women Residing in Rural Areas: A Comparative Study of Deep Learning and Machine Learning Models with XAI Insights", Research Square Platform LLC, 2023 Publication	<1 %
53	<a href="http://kdd.cs.ksu.edu">kdd.cs.ksu.edu</a> Internet Source	<1 %
54	<a href="http://max-success.eu">max-success.eu</a> Internet Source	<1 %
55	<a href="http://mdpi-res.com">mdpi-res.com</a> Internet Source	<1 %
56	<a href="http://scholar.afit.edu">scholar.afit.edu</a> Internet Source	<1 %
57	<a href="http://webthesis.biblio.polito.it">webthesis.biblio.polito.it</a> Internet Source	<1 %
58	<a href="http://www.frontiersin.org">www.frontiersin.org</a> Internet Source	<1 %
59	Elias Garcia-Urquia. "Establishing rainfall frequency contour lines as thresholds for	<1 %

60

Hisham Al Majzoub, Islam Elgedawy, Öykü Akaydın, Mehtap Köse Ulukök. "HCAB-SMOTE: A Hybrid Clustered Affinitive Borderline SMOTE Approach for Imbalanced Data Binary Classification", Arabian Journal for Science and Engineering, 2020

Publication

<1 %

61

Rupesh Kumar Tipu, Suman, Vandna Batra. "Development of a hybrid stacked machine learning model for predicting compressive strength of high-performance concrete", Asian Journal of Civil Engineering, 2023

Publication

<1 %

Exclude quotes On

Exclude matches < 5 words

Exclude bibliography On