National University of Computer and Emerging Sciences, Lahore

Landslide Insight: Exploring Data Patterns and Predictive Modeling

Sophia Razzaq	21L-5607
Idrees Arshad	21L-5632
Burhan Anwar	21L-5676
Aroob Khalid	20L-1313

Instructor: Miss Esha Tur Razia Babar

FAST School of Computing

National University of Computer and Emerging Sciences

Lahore, Pakistan

May 2024

Contents

Introduction

In this project, the objective was to develop a landslide prediction system using geographical coordinates (longitude and latitude) and the corresponding landslide category. The project began by utilizing the NASA Global Landslide Catalog dataset, which provided a comprehensive collection of landslide occurrences worldwide. Through initial data visualization and analysis, the focus was narrowed down to the region of Pakistan, where the dataset was further explored. Several machine learning algorithms were employed, including DBSCAN, SVM, Bagging, Gradient Boosting Classifier, AdaBoost Classifier, FP Growth, Logistic Regression, Linear Regression, Naive Bayes, Decision Tree, and Random Forest. Each algorithm was evaluated using various metrics to determine its effectiveness in predicting landslide types. The results obtained from these evaluations showcased promising performance, with several algorithms achieving high accuracy percentages. This report presents a detailed account of the methodology, findings, and insights gained from this landslide prediction project, highlighting the potential for utilizing machine learning techniques in improving landslide risk assessment and management.

Challenges

One of the primary challenges was the quality and completeness of the dataset, requiring thorough cleaning and preprocessing. Dealing with missing values, inconsistencies, and outliers was crucial for accurate analysis and modeling. Additionally, imbalanced data distribution posed a challenge, requiring exploration of techniques like oversampling and undersampling for balanced predictions. Selecting relevant features and engineering new ones, considering factors like topography, climate, vegetation, and soil properties, required domain knowledge and careful validation.

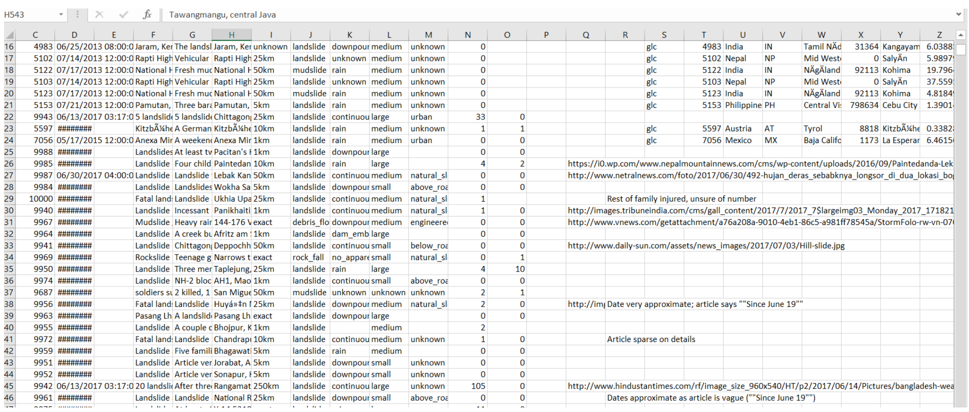


Figure 1: poor quality of dataset .

Getting to Know the Dataset

To begin the analysis, we first explored the dataset to gain a better understanding of its structure and contents. This section presents the key findings related to the shape of the data and the column names.

Shape of the Data

```
# display shape of the data

df.shape
```

```
(11033, 31)
```

Figure 2: initial shape of dataset .

Column Names and Data Types

```
# display features data types

df.dtypes
```

```
source_name                  object
source_link                  object
event_id                      int64
event_date                   object
event_time                  float64
event_title                  object
event_description            object
location_description         object
location_accuracy            object
landslide_category           object
landslide_trigger            object
landslide_size               object
landslide_setting            object
fatality_count              float64
injury_count                float64
storm_name                   object
photo_link                   object
notes                        object
event_import_source          object
event_import_id             float64
country_name                 object
country_code                 object
admin_division_name          object
admin_division_population   float64
gazeteer_closest_point       object
gazeteer_distance           float64
submitted_date               object
created_date                 object
last_edited_date             object
longitude                   float64
latitude                    float64
dtype: object
```

Figure 3: initial columns of dataset .

Preprocessing

Removing unwanted Columns

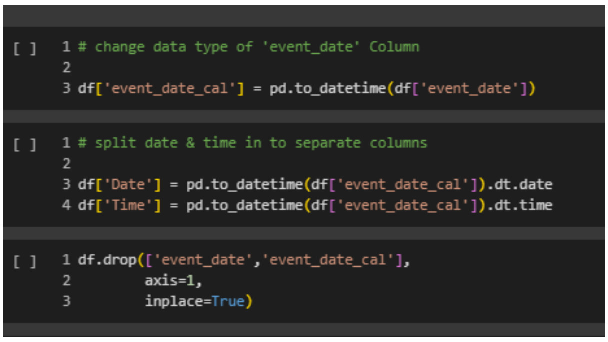
Following columns served no meaning to our prediction and analysis.

```
# drop unwanted columns

df.drop(['event_id','event_time','location_description','event_title','event_description','photo_link', 'notes',
         'event_import_source'  ,'event_import_id','country_code','submitted_date', 'created_date', 'last_edited_date'],
        axis=1,
        inplace=True)
```

Figure 4: Dropped the unwanted columns

Dealing with date column



```
# change data type of 'event_date' Column

df['event_date_cal'] = pd.to_datetime(df['event_date'])
```

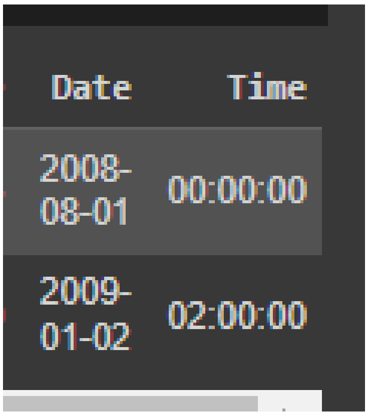
```
# split date & time in to separate columns

df['Date'] = pd.to_datetime(df['event_date_cal']).dt.date
df['Time'] = pd.to_datetime(df['event_date_cal']).dt.time
```

```
df.drop(['event_date','event_date_cal'],
        axis=1,
        inplace=True)
```

Figure 5: Date Column.

Extracted Date and time from the event column.



Date	Time
2008-08-01	00:00:00
2009-01-02	02:00:00

Figure 6: Updation.

Further advance prepossessing has been done along side when it where it was necessary. At this stage we have completed a generalized prepossessing.

Exploratory Data Analysis and Visualization

GLOBAL DATASET

At first we will see trends around the globe and make interpretations accordingly.

Event Reported Source

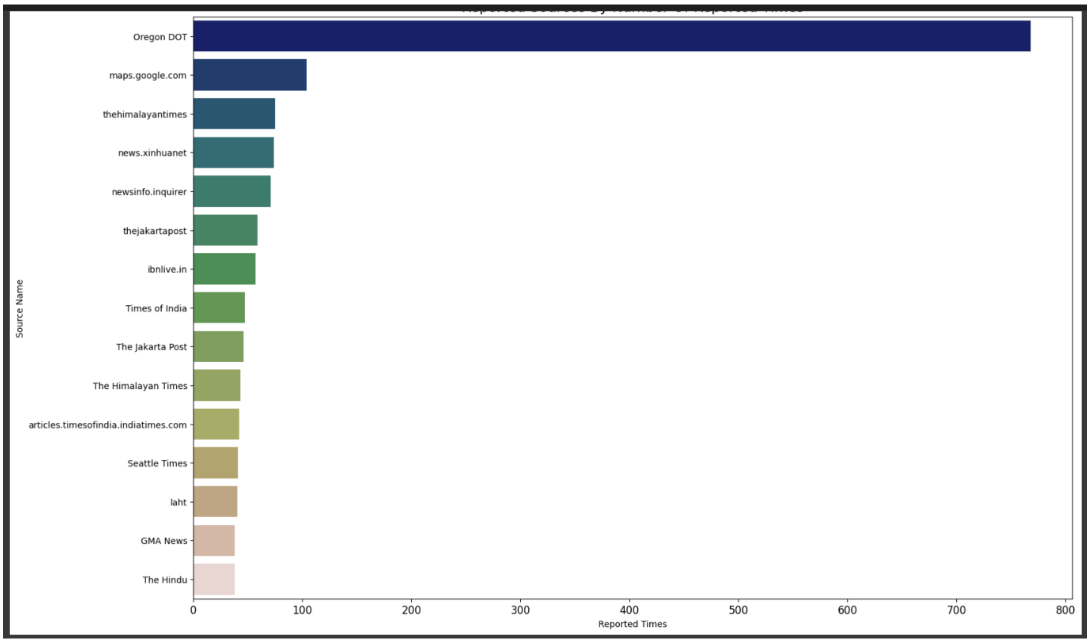


Figure 7: Number of reports made by each source

Origon DOT Have Reported Huge Number Of Events During This Time Period (1988 - 2017)

Geospatial Visualization Of Globally Events

Figure 8: Open Streat Map Style

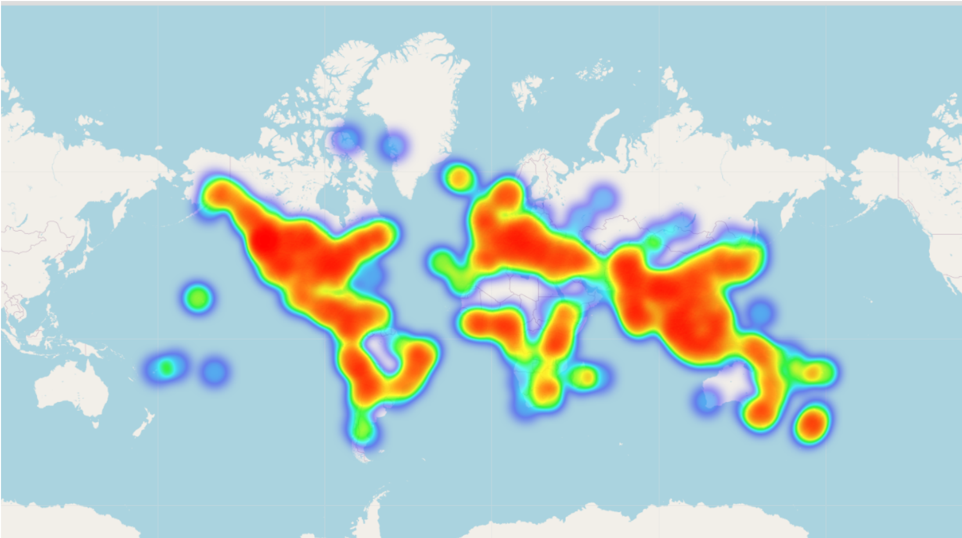


Figure 9: Heatmap.

Accoring to the above maps we can determine that lot of land slide events happend in Indial Ocean North America South America

Events by year and month

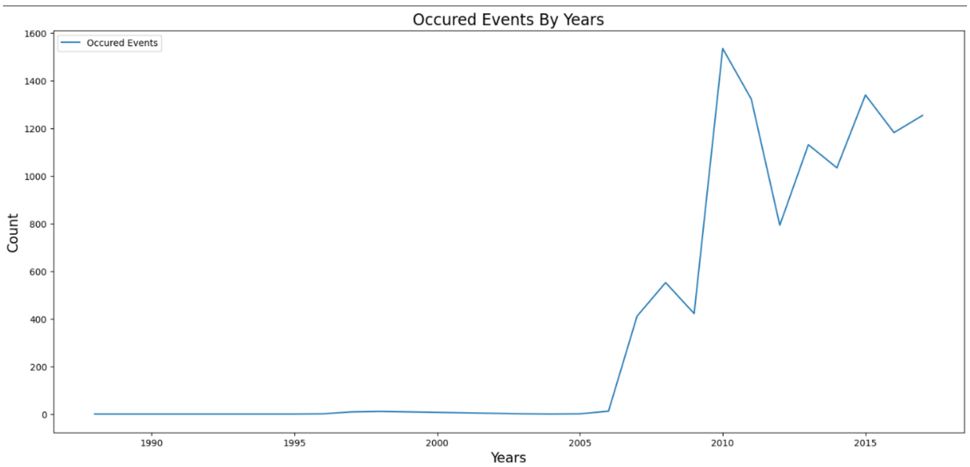


Figure 10: Events by year.

We Can See There Are Most Event Occured During 2010.

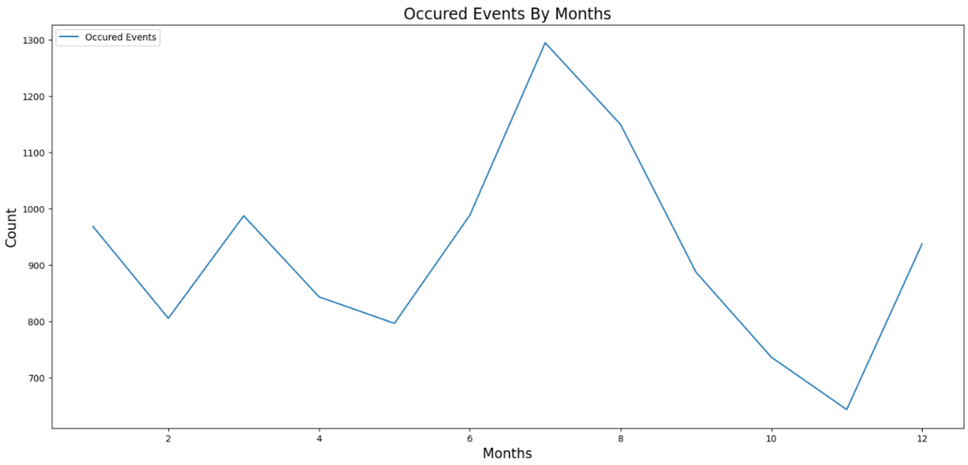


Figure 11: Events by month.

We Can See There Are Most Event Occured During 3 Quater Of The year.

Events in 2010

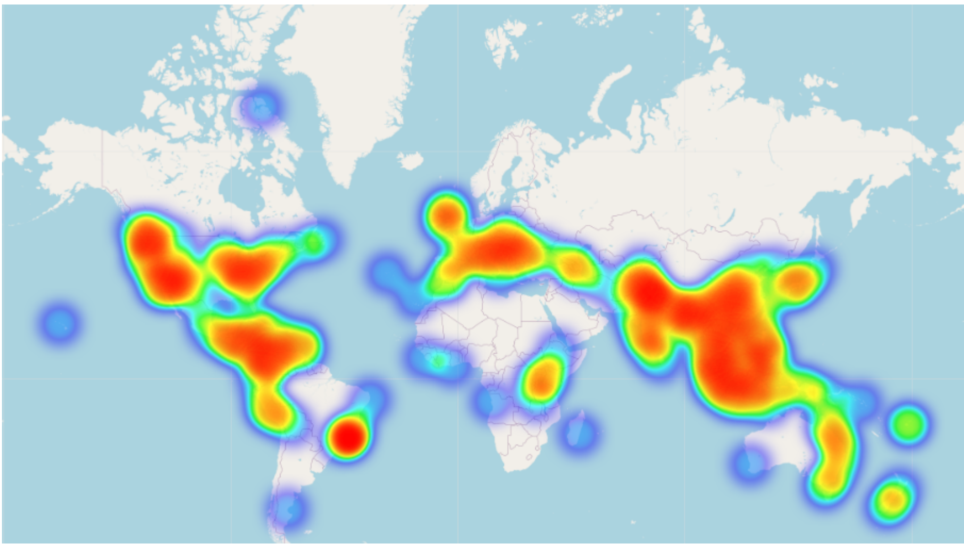


Figure 12: Heatmap of events occured in 2010

In 2010 More Events Occured In Indian Ocean

Landslide categories

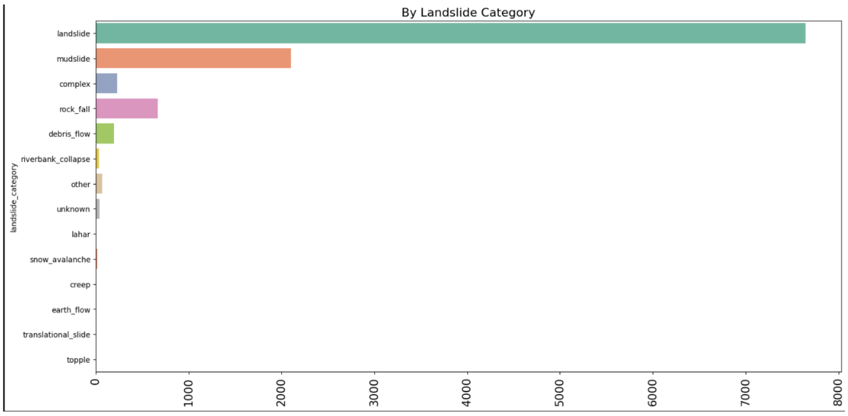


Figure 13: Landslide categories

We can see mostly the news reports were of Landslide.

Event Setting Types

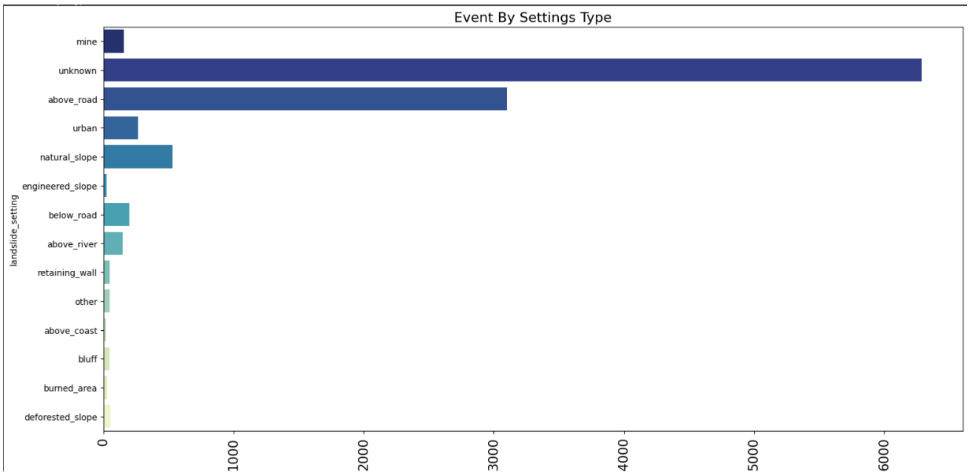


Figure 14: Event Setting Type.

Accoring to above map we can determine lots of events happend in above Roads setting. so roads contructions cause to these events most. if we can contruct roads with more safety and pre analysis we can reduce these events happening

Event Trigger Type

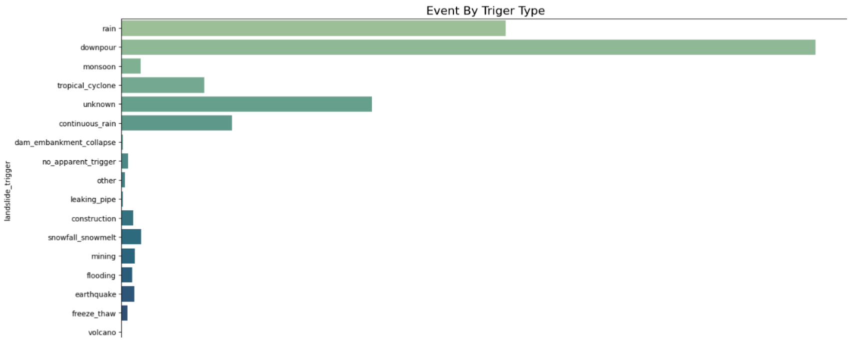


Figure 15: Event Trigger Type.

We can see downpour and raining has triggered landslides the most.

Top 10 countries with most events

Top 10 Countries By Events

```
[ ]  1 df['country_name'].value_counts().head(1

country_name
United States     2992
India             1265
Philippines        675
Nepal              481
China              426
Indonesia          355
United Kingdom     229
Brazil             214
Canada             174
Malaysia           171
Name: count, dtype: int64
```

Figure 16: Top 10.

Inisghts

The comprehensive analysis of the global dataset spanning from 1988 to 2017 sheds light on the patterns and trends of landslide events worldwide. Notably, the data highlights the significant reporting role of the origin DOT. Geospatial visualizations pinpoint hotspots primarily in the Indian Ocean, North America, and South America. The peak in landslide occurrences during 2010, particularly in the third quarter, underscores a crucial period for such events. Furthermore, the dominance of landslides, often triggered by downpour and rainfall, underscores the need for proactive measures, especially in road construction, to mitigate future incidents.While US and India remain the highest landslide reported countires by far.

PAKISTAN DATASET

Extracting Pakistan Data

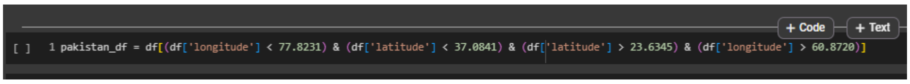


Figure 17: Pakistan coordinates.

After extracting Pakistan data we found out that there Are 596 Events Occurred In Pakistan During 1988 To 2017.

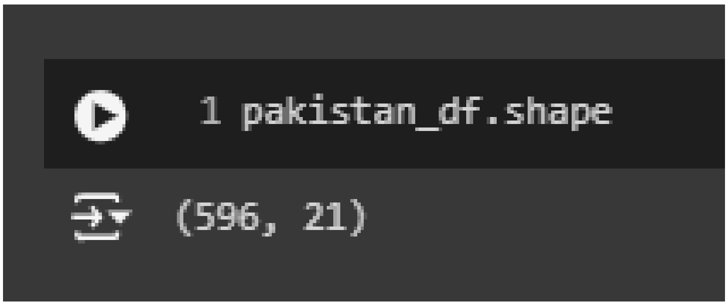


Figure 18: Number of events.

Occured Events In Pakistan Geospatial Visualization

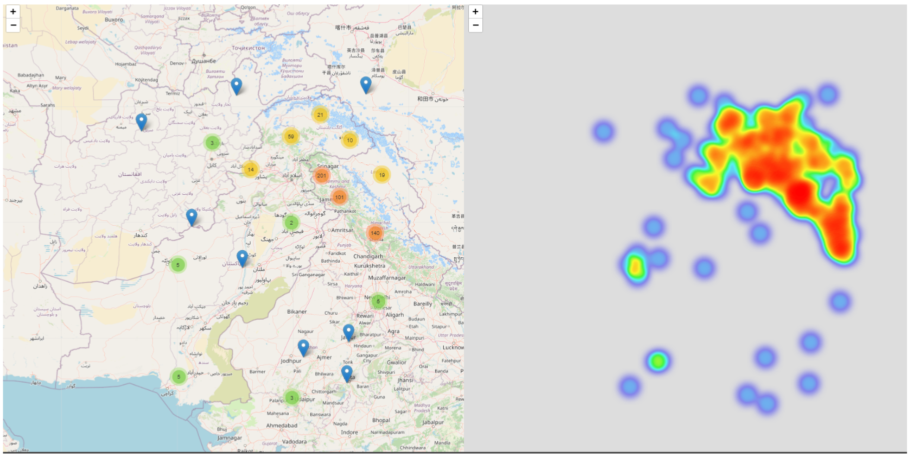


Figure 19: Geospatial Visualization

Acording To The Above Map, We can See That In Pakistan Most Of The Events Mountain Side Areas. Such As Gilgit Baltistan, KPK and Kashmir.

Time Series Of Occurred Events By Years In Pakistan

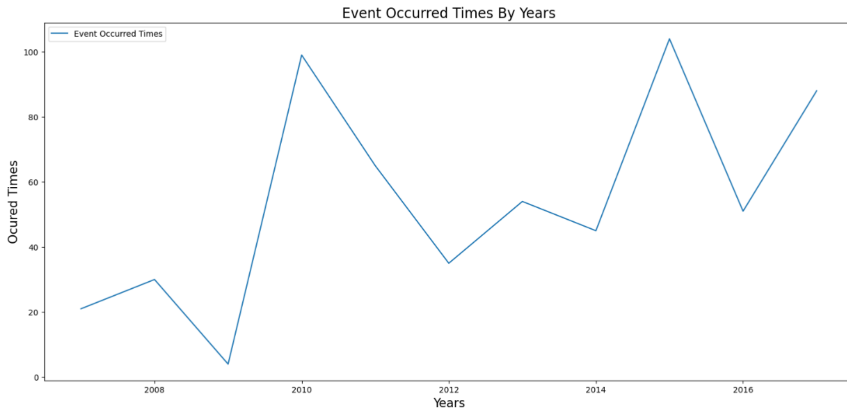


Figure 20: By years

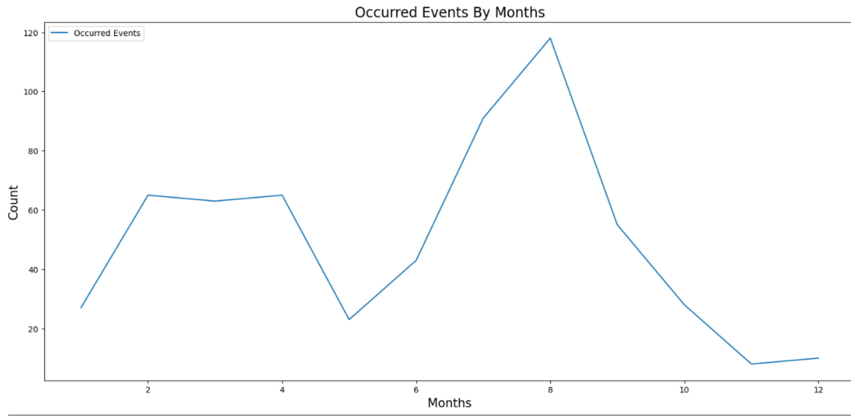


Figure 21: By month

We can see that most events occurred during 2010 and 2015 mostly during august.

Time Series Of Fatalities and Injuries Comparing To Events Occurred Times In Pakistan

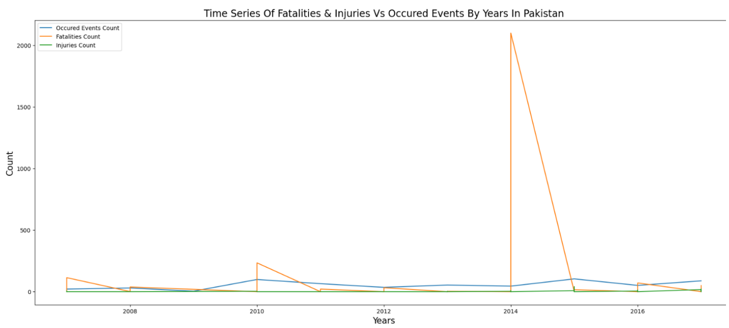


Figure 22: Fatility and Injury count through years

We can observe that the highest fatility count reported were during 2014-2015 which was previously marked as highest events reported year, with over 2k fatility count.

Biggest Landslide Disaster In Pakistan

In 2014, Pakistan experienced one of the biggest landslide disasters in its history. The details of this event can be found in the *Annual Flood Report 2014* provided by the Federal Flood Commission (FFC) of Pakistan. The report can be accessed at the following URL: `https://mowr.gov.pk/SiteImage/Misc/files/2014%20Annual%20Flood%20Report%20of%20FFC.pdf`.

Figure 23: Event Details

Distribution Of Categories In Occurred Events In Pakistan

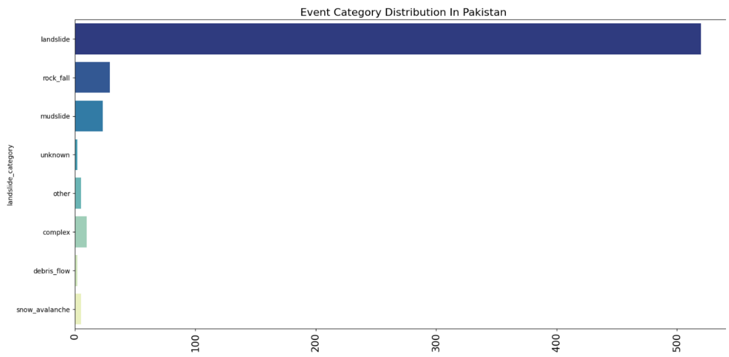


Figure 24: Categories

We can see that majorly Landslides have occurred in Pakistan following rockfall.

Distribution Of Triger Type In Occurred Events In Pakistan

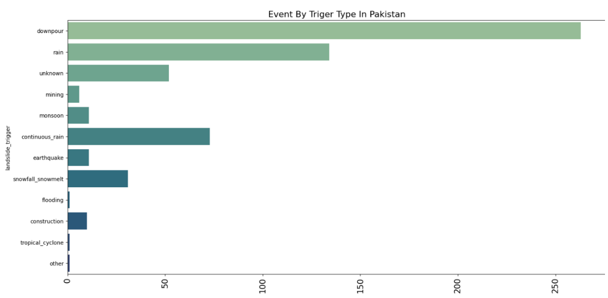


Figure 25: Trigger Types

Downpour and rain has triggered landslides more often.

Distribution Of Event Size In Occurred Events In Pakistan

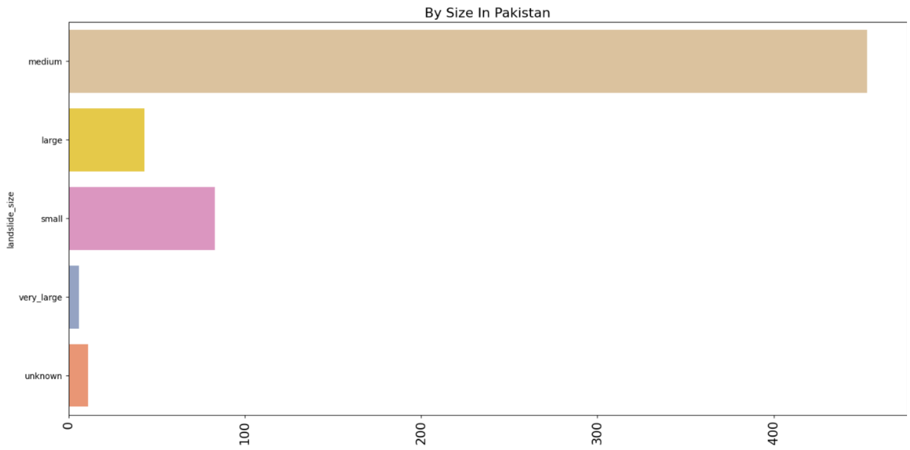


Figure 26: Landslide Size.

Pakistan had alot of medium sized landslides rather than large scale or small.

Event Reported Sources About In Pakistan

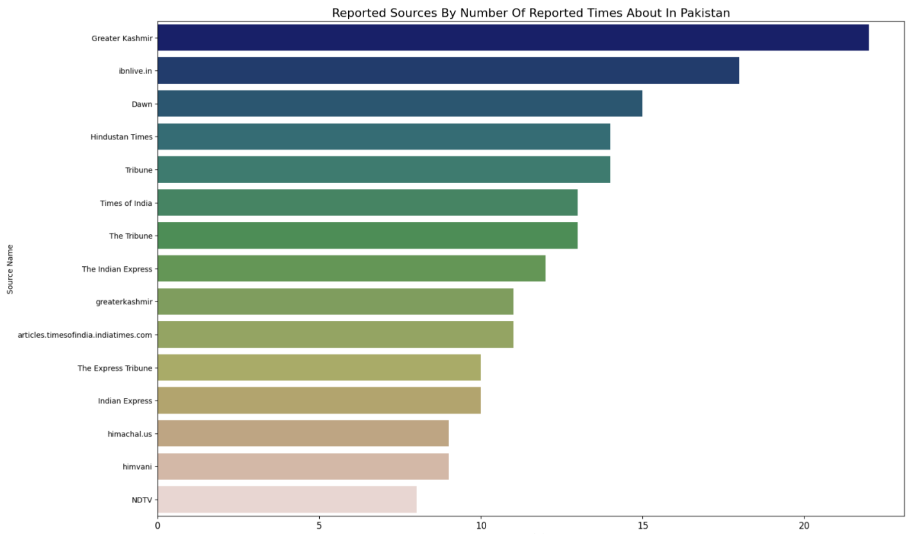


Figure 27: Event Reported Sources About In Pakistan

Greater Kashmir Mirror Has Did Good Job When Reporting Events.

Insights gained

The analysis of Pakistan's landslide data from 1988 to 2017 reveals concentrated occurrences in mountainous regions like Gilgit Baltistan, KPK, and Kashmir. Temporally, peaks in events during 2010 and 2015, particularly in August, are notable. Correlations between event frequencies and consequences, such as fatalities, are evident. Effective communication channels, like Greater Kashmir Mirror, play a vital role. These insights inform targeted risk management strategies and proactive disaster preparedness efforts in Pakistan.

Machine Learning Algorithms Applied

SVM

SVM is ffective for classification tasks when there is a clear margin of separation between classes. SVMs can handle both linear and non-linear relationships between features, making them suitable for predicting landslide categories.

```
svm_model = SVC()
svm_model.fit(X_svm_encoded, y_encoded)
y_svm_pred = svm_model.predict(X_svm_encoded)
```

Evaluation Metrics

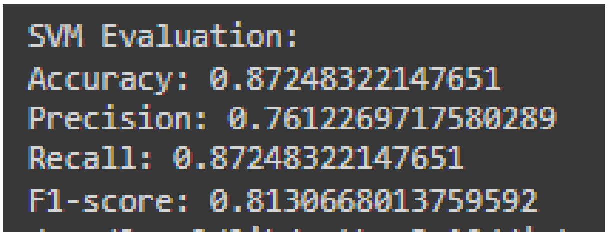


Figure 28: Svm evaluation.

Bagging

Bagging (Bootstrap Aggregating) is an ensemble method that builds multiple base models (in this case, decision trees) on different subsets of the dataset and then combines their predictions. The features chosen for the Bagging Classifier include geographical coordinates ('longitude' and 'latitude'), 'year', and 'month', which are likely to have a significant impact on the occurrence of landslides. These features are chosen because they provide information about the location and time of landslide events, which can be crucial for predictive modeling.

Evaluation Metrics

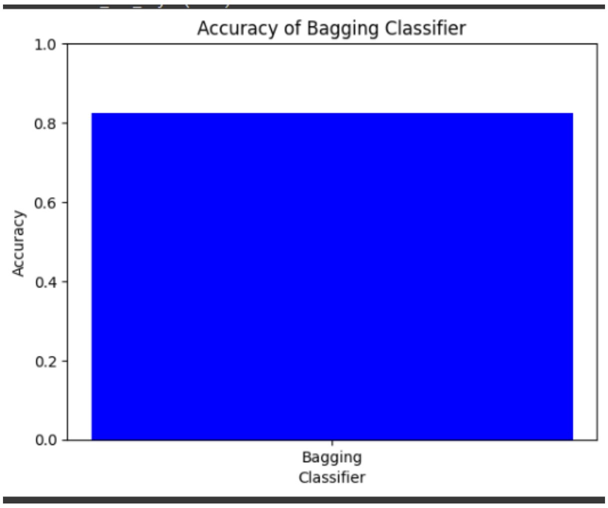


Figure 29: Bagging Results.

Gradient Boosting Classifier

Gradient Boosting is a powerful ensemble method that builds decision trees sequentially, focusing on minimizing the errors of the previous trees. The selected features cover various aspects related to landslide events, such as the triggering factors landslide trigger, the size of the landslide landslide size, the setting landslide setting, information about storms storm name, and the population of administrative divisions admin division population. These features are chosen to capture different dimensions of landslide occurrences and their potential causes.

Evaluation Metrics

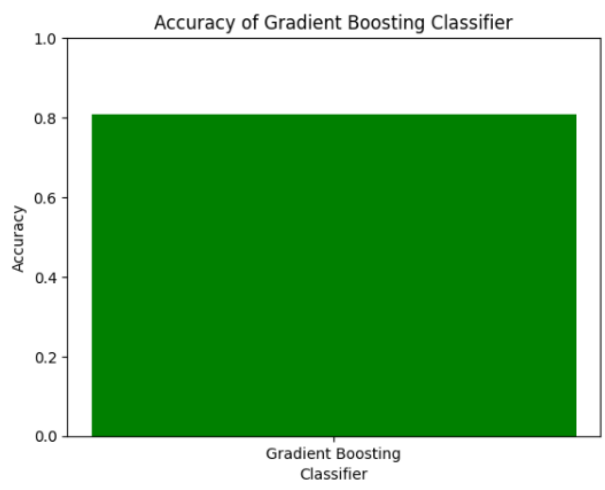


Figure 30: GB Results.

AdaBoost Classifier

Features Used: longitude, latitude, landslide size, admin division population

Reasoning: AdaBoost is an adaptive boosting algorithm that sequentially fits weak learners to the dataset, with a focus on instances that were misclassified by previous learners.

```
# Define and train the AdaBoost classifier
adaboost_classifier = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier',
    AdaBoostClassifier(base_estimator=DecisionTreeClassifier(),
    n_estimators=10, random_state=42))
])
```

The selected features include geographical coordinates ('longitude' and 'latitude'), 'landslide size', and 'admin division population'. These features are chosen to provide information about the location and magnitude of landslide events, as well as the population density of administrative divisions, which could influence the occurrence and impact of landslides.

Evaluation Metrics

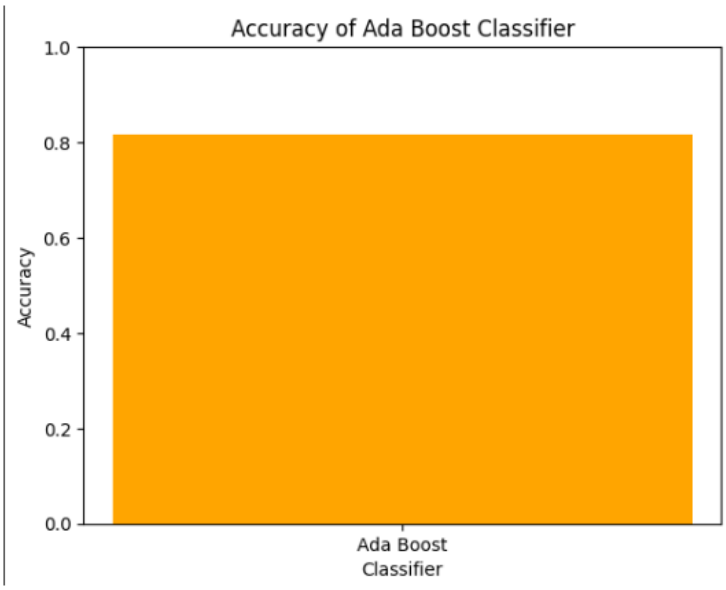


Figure 31: AdaBoosting Results

Kmeans

We used Kmeans to identify natural groupings or clusters of landslides based on their geographical coordinates or other features. K-means clustering helps in understanding spatial patterns in landslide occurrences.

We used Elbow method to determine number of cluster.

```
# Determine the number of clusters using the elbow method
wcss = []
max_clusters = 10  # Maximum number of clusters to consider
for i in range(1, max_clusters + 1):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
```

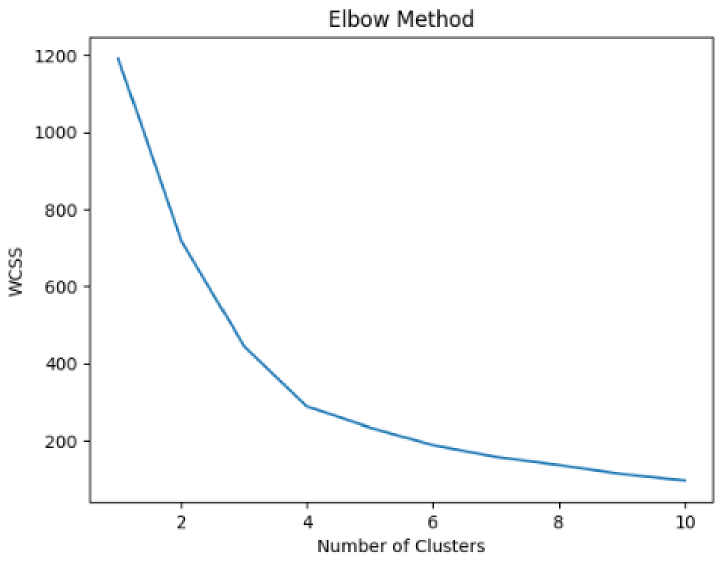


Figure 32: WCSS Plotting

Kmeans Graph

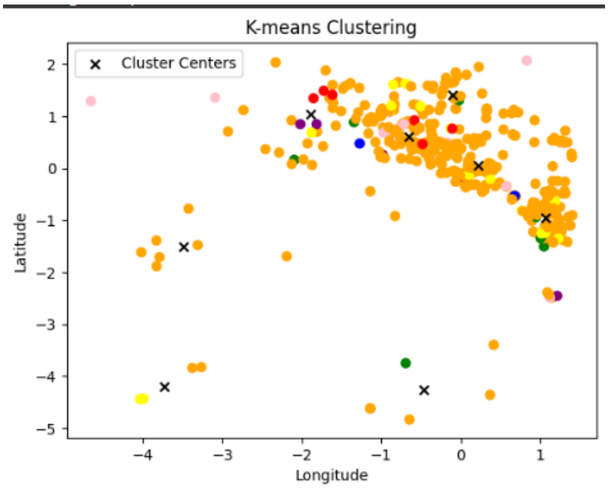


Figure 33: Kmeans Clusters

We can see that each category is plotted to a cluster wrt to longitude and latitude.

Logistic Regression

```
logistic_regression_model = LogisticRegression(C=4, penalty='l2')
logistic_regression_model.fit(X_train, y_train)
```

We applied L2 regularization to the logistic regression model by setting the penalty parameter to 'l2' and the regularization strength parameter (C) to 4. Because previously we noticed that without regularization out model was not performing well.

Evaluation Metrics

```
Accuracy: 0.8166666666666667
Precision: 0.6851694915254237
Recall: 0.8166666666666667
F1-score: 0.7451612903225807
```

Figure 34: Evaluation Matrices.

Decision Tree

We have used to predict the category of landslides based on various features like trigger, size, and setting. Decision trees provide interpret able rules for classification tasks.

```
# Decision Tree with regularization
decision_tree_model = DecisionTreeClassifier(max_depth=8)    ##channging depth
```

```
decision_tree_model.fit(X_dt_rf_encoded, y_encoded)
```

We have use used hyper-parameters and change the decision tree depth manually to get better results.

Evaluation Metrics

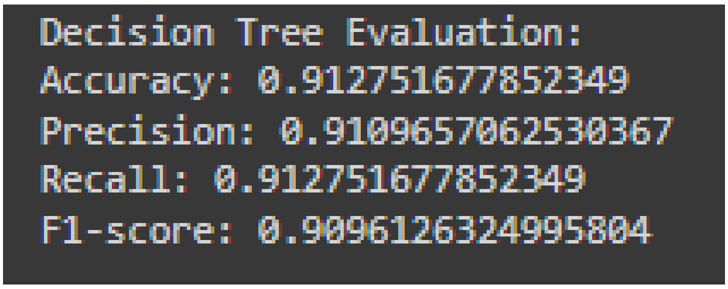


Figure 35: DT Results.

Random Forest

We have used to predict the category of landslides based on various features like trigger, size, and setting. Decision trees provide interpret able rules for classification tasks.

```
# Random Forest with regularization
random_forest_model = RandomForestClassifier(max_depth=7)
random_forest_model.fit(X_dt_rf_encoded, y_encoded)
```

We have use used hyper-parameters and change the random forest depth manually to get better results.

Evaluation Metrics

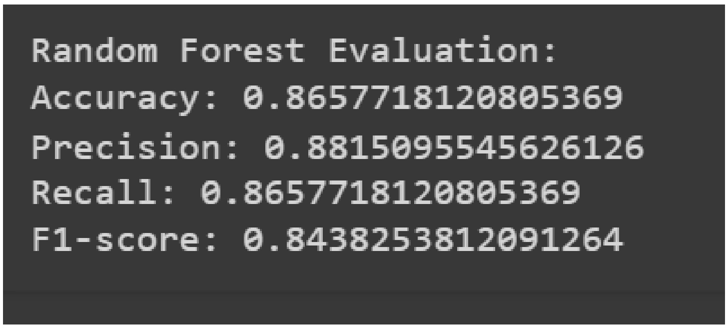


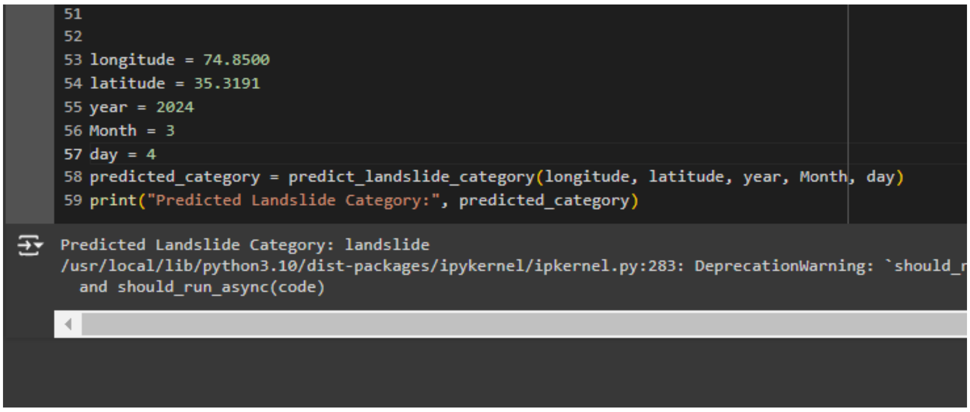
Figure 36: Random Forest Results.

BEST MODEL

So by using all the models we concluded that decision tree has better results than the other models so we will use this for our predictions.

LIVE PREDICTION

Lets test our decision tree on unseen data



```
longitude = 74.8500
latitude = 35.3191
year = 2024
Month = 3
day = 4
predicted_category = predict_landslide_category(longitude, latitude, year, Month, day)
print("Predicted Landslide Category:", predicted_category)
```

```
Predicted Landslide Category: landslide
/usr/local/lib/python3.10/dist-packages/ipykernel/ipykernel.py:283: DeprecationWarning: `should_r
  and should_run_async(code)
```

Figure 37: Input Values

Lets test out result through a quick google search and here is what we found out:

One killed as snow, landslide bring life to standstill in GB

Jamil Nagri | Published March 5, 2024

People walk along a snow laden street in Kalam on March 4, 2024. — AFP

Figure 38: Dawn News Article.

So our model predicted an in real life event. Showing how reliable it is!

You may check out the article: Click here

Conclusion

In conclusion, among the various machine learning algorithms applied, the decision tree stood out as the most effective for predicting landslide categories, achieving an impressive accuracy rate of 91. This high accuracy demonstrates the robustness and reliability of the decision tree model in classifying landslide events based on factors such as trigger type, size, and setting. With its interpretable rules and strong predictive power, the decision tree algorithm emerges as a valuable tool for landslide risk assessment and management