Data Science Project Report:

Landslide Data of Bangladesh

**1. Introduction**

Landslides are a natural hazard that can result in significant loss of life and property, especially in mountainous regions. In Bangladesh, landslides occur mostly in the Chittagong Hill Tracts due to heavy rainfall, soil erosion, deforestation, and poor urban planning. This project aims to explore landslide data from Bangladesh to understand the causes, patterns, and trends in landslides, and to develop predictive models that could help mitigate their impact.

Objective:

Analyze the frequency, severity, and distribution of landslides.

Identify key factors contributing to landslides.

Build predictive models for early landslide prediction.

**2. Data Collection**

The landslide data used in this project was collected from various open sources, including:

Government disaster reports (Bangladesh Disaster Management Bureau).

Geospatial data from satellite images (e.g., MODIS, Sentinel-1).

Meteorological data from the Bangladesh Meteorological Department.

Topographical data from open sources such as OpenStreetMap (OSM) and digital elevation models (DEM).

The dataset includes the following variables:

Location: Latitude and longitude of the landslide event.

Date: Time of occurrence.

Magnitude: Severity of the landslide (e.g., small, medium, large).

Rainfall: Amount of rainfall in millimeters preceding the event.

Slope: Slope gradient of the area.

Soil Type: Classification of soil in the area.

Population Density: Number of people living in proximity to landslide-prone areas.

3. Data Preprocessing

Data preprocessing is an essential step to clean and prepare the data for analysis. The steps taken include:

Missing Value Treatment:

Imputed missing rainfall data using interpolation methods.

Removed rows with excessive missing values (e.g., those lacking both date and location).

Handling Categorical Data:

Converted categorical variables such as 'Soil Type' into numerical codes using label encoding.

Feature Engineering:

Created new features such as 'days\_since\_last\_rain' to better capture the relationship between rainfall and landslide occurrence.

Outlier Detection:

Identified and handled outliers in variables such as rainfall and slope using the Z-score method.

Data Scaling:

Standardized continuous features such as slope, population density, and rainfall using Min-Max scaling to ensure all features are on the same scale.

**4. Exploratory Data Analysis (EDA)**

EDA helps to understand the distribution of key variables and their relationships.

Landslide Frequency by Year and Location:

Plotted a time series to visualize the number of landslides over the years.

Mapped landslides geographically to identify regions most at risk.

Correlation Analysis:

Conducted a correlation analysis between variables such as rainfall, slope, soil type, and landslide occurrence to identify significant predictors.

Found that rainfall, slope, and population density have high correlation with landslides.

Rainfall and Landslide Severity:

Visualized rainfall patterns preceding landslide events and observed that landslides were more frequent during the monsoon season (June to September).

Grouped landslides by severity and found that heavy rainfall (>300 mm) was linked to more severe landslides.

Topographical Influence:

Analyzed the relationship between slope angle and landslide frequency. The findings showed that landslides were most common in areas with slopes greater than 30 degrees.

**5. Modeling**

We developed predictive models to forecast landslide events based on various features.

**Model Selection:**

Logistic Regression: Used for predicting whether a landslide will occur (binary classification).

Random Forest: Employed for feature importance analysis and to predict the likelihood of landslides.

XGBoost: Applied for enhanced performance and accuracy, given its success in handling tabular data.

Model Training and Validation: Split the data into training and testing sets (80-20 split).

Performed cross-validation using a 5-fold strategy.

Evaluated models using metrics such as accuracy, precision, recall, F1-score, and the ROC-AUC curve.

**6. Results**

Predictive Accuracy:

The Random Forest model achieved an accuracy of 85%, with precision and recall values of 0.82 and 0.88, respectively.

Key Factors for Landslide Occurrence:

Rainfall is the most important factor, with significant landslides occurring after rainfall above 300 mm.

Slope is the second most important factor, particularly in areas with a gradient above 30 degrees.

Areas with high population density are at greater risk due to the combination of natural factors and human settlements.

Geospatial Insights:

Landslides are most frequent in the Chittagong Hill Tracts and adjacent areas.

Seasonal patterns show a higher frequency of landslides during the monsoon season (June to September).

**7. Conclusion**

This study demonstrates the significant role of environmental and geographical factors in predicting landslides in Bangladesh. The key predictors identified—rainfall, slope, and population density—can be used to build an early warning system for high-risk areas. Future work can focus on incorporating real-time meteorological data and satellite imagery to improve prediction accuracy and provide better recommendations for mitigation strategies.

**8. Recommendations**

Early Warning System: Implement a monitoring system that collects real-time rainfall and weather data to trigger alerts when conditions are ripe for landslides.

Urban Planning: Focus on reducing population density in high-risk areas and implementing soil erosion control techniques.

Further Research: Extend the analysis to include more granular data from remote sensing satellites for real-time monitoring of landslide risks.