DATA SCIENCE PROJECT CANVAS

Designed by: DSAI_127_121

Date: 06 August 2024

Title: Predict the flood and landslide with past data and data from weather/rainfall monitoring stations and river based sensors

1. Problem Statement

What problem are you trying to solve?
What larger issues do the problem address?

Due to the climate change there has been changes in climatic activity of rainfall and this year alone Nepal experienced its record time rainfall of ~>630mm of rainfall in 24hrs in Kanchanpur. Various cases of flood and landslides were experienced throughout the South Asian region and it has escalated this year alone.

This will address the issues related to loss of lives, economic and livelihood loss by providing an early detection of the probable disaster that might occur.

2. Outcomes/Predictions

What prediction(s) are you trying to make?
Identify applicable predictor (X) and/or target (y) variables.

Predictor:

- Real time rainfall and water level data from ground monitoring stations and sensors from Department of Hydrology and Meteorology.

Target:

- Chances of flooding and landslides in the next 24 hours with respect to the weather data.

3. Value Propositions

What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

With this information the concerned stakeholders can issue warnings at several levels from transportation to settlements to keep them safe or transfer them to a safer place.

4. Data Acquisition

Where are you sourcing your data from? Is there enough data? Can you work with it?

Series of real-time data collected by National Disaster Risk Reduction and Management Authority in Nepal with the help of DHM and DMG

6. Model Evaluation

How can you evaluate your model performance?

Flood Monitoring

- **Accuracy**: Percentage of correctly predicted instances.
- **Precision and Recall**: Measure the accuracy of positive predictions (flood occurrence).
- **F1 Score**: Harmonic mean of precision and recall.
- **ROC-AUC**: Area under the Receiver Operating Characteristic curve, indicating the trade-off between true positive rate and false positive rate.

Landslide Detection

- **Confusion Matrix**: True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN).
- **Precision-Recall Curve**: For imbalanced datasets, focus on precision and recall instead of ROC-AUC.
- **Geospatial Accuracy**: Overlay predictions on geographical maps and compare with known landslide locations.

5. Modeling

What models are appropriate to use given your outcomes?

Baseline Models: Start with simple models like logistic regression for classification or linear regression for prediction to establish a performance baseline.
Advanced Models: Implement and compare more complex models such as:

- **Machine Learning**: Random Forest, Gradient Boosting, Support Vector Machines.
- **Deep Learning**: Convolutional Neural Networks (CNNs) for spatial data, Recurrent Neural Networks (RNNs) for temporal data.

7. Data Preparation

What do you need to do to your data in order to run your model and achieve your outcomes?

- Use 10 years of past data and 24hrs to 2 days of real time data
- Cleaning the data for predictive precision
- Fill out the missing data caused by communication failure

Modified from Bill Schmarzo's Machine Learning Canvas and Jasmine Vasandani's Data Science Workflow Canvas for CP-DSAI @AIT