Name: Lasta Maharjan

Project Name: Earthquake Prediction

# Introduction

My project for this semester’s Data Mining Class is “Earthquake Prediction”. As recently Turkey was hit by high magnitude earthquake and thousands of people have lost their lives, and family members. Forecasting earthquakes is one of the most important problems as the consequences of earthquakes are devastating. For this project I am using dataset from Kaggle, which comes from a well-known experimental set-up used to study earthquake physics. This white paper is intended to provide an overview of data mining for the Earthquake dataset, and also its data preparation, exploratory data analysis, modeling, and evaluation.

# Dataset:

The Earthquake dataset is available in several formats, including CSV & excel and can be downloaded from **Kaggle, or UCI Machine Learning Repository.**

# Data Preparation:

As the first step in data mining is to collect the data required for the project and then prepare the data for further analysis and modeling. The preparation of data involves cleaning and transforming the data to ensure that it is in a format to be easily analyzed. The next step is to check for missing values, outlier, and other inconsistencies, and handle those properly either by removing or replacing with appropriate values. If needed scaling or normalization of dataset can be done.

# Exploratory Data Analysis:

Once the outliers, missing values are treated, and data is prepared for further steps. The next step is to perform exploratory data analysis (EDA). During EDA we can summarize the descriptive dataset, build histograms, scatterplots, to see the patters of dataset, normality of dataset. We can also look for correlation, also build pairwise graph to deep dive into relationship between variables. This will help us understand the dataset better and also help in choosing the model.

# Modeling

The next step is to select a modeling technique. I will use three algorithms for the Earthquake dataset, including **RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory Networks), and Catboost**. The goal of modeling here is to predict when the earthquake is going to happen. The dataset contains 3 files.

1. **train.csv** - the training set contains the acoustic data and the tie failure to predict the earthquake.
2. **test.csv** - the test set contains multiple datasets to test the model with.
3. **sample\_submission.csv** – a sample submission file in the correct format.

The modeling process involves splitting the data into training, and testing sets, training the model on the training set, and then evaluating the model's performance on the testing set.

# Evaluation

Model evaluation is an important step in data mining on the earthquake dataset. For this prediction and models, I plan to use Root-Mean Squared Error (RMSE) or accuracy to compare which model is performing better. Higher accuracy indicates the better performing model.