**Lasta Maharjan**

**Earthquake Magnitude Prediction Using Machine Learning**

**Introduction:**

Earthquakes can be described sudden shaking of the ground caused by seismic activity that happens between the movement of tectonic plates, and when these movements accumulate energy in the form of rock stress, then it is suddenly released. 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. The main goal of this project is to predict the magnitude of the earthquake that will take place.

**Dataset:**

For this project I am using dataset from United States Geological Survey (USGS), which is updated every 15 minutes with real time data of earthquake and their details. My data set is 2.27 MB, and in .csv format which can be downloaded from [USGS](https://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php) for any time frame as needed. My data consists of 22 attributes, regarding date, time of earthquake, its magnitude, location, depth, depth error, magnitude error, etc.

1. time: Date & Time when earthquake exactly took place
2. latitude: The latitude is the number of degrees north (N) or south (S) of the equator and varies from 0 at the equator to 90 at the poles.
3. longitude: The longitude is the number of degrees east (E) or west (W) of the prime meridian which runs through Greenwich, England.
4. depth: The depth where the earthquake begins to rupture.
5. Mag: Earthquake magnitude is a measure of the size of an earthquake at its source. It is a logarithmic measure.
6. magType: The method or algorithm used to calculate the preferred magnitude for the event.
7. Nst: Number of seismic stations which reported P- and S-arrival times for this earthquake.
8. Gap: The largest azimuthal gap between azimuthally adjacent stations (in degrees).
9. Dmin: Horizontal distance from the epicenter to the nearest station (in degrees).
10. Rms: The root-mean-square (RMS) travel time residual, in sec, using all weights.
11. Net: The ID of a data contributor. Identifies the network considered to be the preferred source of information for this event.
12. Id: A unique identifier for the event.
13. Updated: Time when the event was most recently updated.
14. Place: Textual description of named geographic region near to the event.
15. Type: Type of seismic event.
16. locationSource: The network that originally authored the reported location of this event.
17. magSource: Network that originally authored the reported magnitude for this event.
18. horizontalError: The horizontal location error, in km, defined as the length of the largest projection of the three principal errors on a horizontal plane.
19. depthError: The depth error, in km, defined as the largest projection of the three principal errors on a vertical line.
20. magError: Uncertainty of reported magnitude of the event. The estimated standard error of the magnitude.
21. magNst: The total number of seismic stations used to calculate the magnitude for this earthquake.
22. status: Indicates whether the event has been reviewed by a human. Status is either automatic or reviewed. Automatic events are directly posted by automatic processing systems and have not been verified or altered by a human.

**Data Preprocessing:**

After data was collected, the major focus was on to cleaning the dataset and preparing for the further analysis and modeling. The preparation of data involved around cleaning and transforming the data to ensure that it is in a format to be analyzed and modelled. For the first step, I looked at the data types, what my data are and attributes and their meaning. I further checked for null values, I had 6 features which had missing values, hence by percentage analysis of what percent of that particular feature I am missing values on, one of the feature “dmin” had about 46% of the values missing, upon checking it would not have been optimum to replace the data with some other value, hence I removed this column for my further analysis. And, for the other 5 attributes, I was missing values about 20-30% and those attributes were very important for the further modeling and analysis, hence I replaced the null values for those by their mean values.

**Exploratory Data Analysis:**

For further analysis, I checked the statistical features such as mean, median, standard deviation, maximum value, minimum value, quartiles, etc. for the numerical feature. There is a column “updated”, which is a time stamp when the event was updated in the USGS website, hence I removed that column as it was not an important feature for my further analysis. For further analysis I checked their pair plots to see the relationship of the featured with one another. Upon analysis I could see that magnitude had a strong relationship with gap azimuthally adjacent stations (in degrees). I did find correlation between MagnitudeNst and Nst of 0.61, but the other features did not have much correlation, meaning not much strong relationship. For further, for categorical variables, such as Earthquake type, magnitude type, magnitude sources, source of earthquake, I used Label Encoder to label unique numbers to each category for those features. Upon looking at the boxplots and histogram I could see that many of my features were skewed either left or right, for longitude and latitude mostly because the location we were selecting for earthquake prediction. I used MinMaxScaler to normalize my data to get the optimum results for the further modeling.

**Modeling and Evaluation:**

After handling all outliers, skewness of data, I further decided to choose Support vector Machine, Random Forest, and K-Nearest Neighbors (KNN) algorithm models to predict the magnitude of the earthquake in the future. I decided to go with these models as my number of features were not very huge, and support vector works well with both linear and non-linear datasets. Random Forest would be helpful to predict the earthquake magnitude based on the input features such as the location, time, etc. I used mean square error(MSE) as evaluation metrics to see which model worked best for the prediction. I had following as results for MSE with corresponding models:

Chart, bar chart

Description automatically generated

As, we can see from the above comparison that RandomForest model has least Mean Square error compared to KNeighbors and then followed by Support Vector Machine. This helps us decide that RandomForestRegressor model worked better between these three models.

**Conclusion and Future Work:**

After comparing the predicted magnitude values Vs. actual earthquake magnitudes, and as we used 3 models, however, within the time contrain I could not try Artificial Neural Network this time but would definitely would like to try and see in the future. And, in addition to predicting the magnitude of earthquake in future I really do think that also trying to predict the time stamp, would be my next few steps for this project.