Earthquake Prediction

Using Clustering, Random Forest and Regression Analysis

**Team Members**

G. Bala Krishna – AP20110010033

**Abstract:**

The theme of earthquake predicting is one that I'd want to look into further. As a natural disaster, earthquakes continue to be damaging, not just to the economy but also to people's lives. This sparked the idea of developing an early warning system for seismic disasters to reduce fatalities. For some years, researchers have been developing earthquake predictions and seismic hazard ratings for specific locations by collecting the data. In this report, we attempt to predict earthquakes hotspots, tries to predict why the earthquakes are frequent in that region, and finally by classifying the clustering output to predict the precautions taken for the upcoming earthquake. It is probabilistic to determine the structural safety status of earthquake-damaged structures by using Clustering, and Random Forest prediction. The Clustering technique will forecast the earthquake-prone areas. The random forest technique can forecast the impact of the earthquake by taking latitude, longitude, magnitude, and depth as input. Here we used geopandas to visualize the predictions. In this study were correct between 95 to 99 percent.

**Introduction:**

In this system, we are going to predict the earthquakes by using the data collected by the researchers. This model can predict the scale and the precautions to taken for the earthquake by using Clustering, Random Forest, and Regression.

**Clustering**: This technique will forecast the earthquake-prone hotspots by clustering the points and calculating the centroid of these clusters.

**Classifiers (Random Forest, Decision tree, naive bayes)**: This technique can forecast the impact and scale of an earthquake by classifying latitude, longitude, magnitude, and depth.

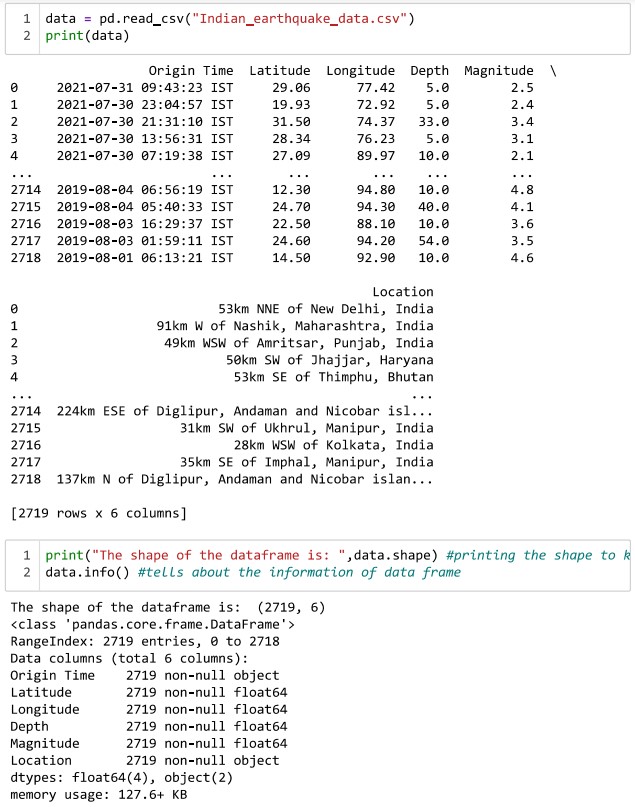
**Regression**: We use this technique to derive the relation between geographical position of India to earthquake positions.

This model has 95 to 99 percent accuracy in forecasting the exact location, magnitude, and depth of an earthquake.

**Dataset:**

The earthquake dataset was taken from Kaggle Website:

<https://www.kaggle.com/datasets/parulpandey/indian-earthquakes-dataset2018-onwards>



We use the dataset made by the National Center for Seismology. This is the nodal agency of the Government of India for monitoring earthquake activity in the country. NCS maintains the National Seismological Network of 115 stations each having state of art equipment and spreading all across the country. NCS monitors earthquake activity all across the country through its 24x7 round-the-clock monitoring center. NCS also monitors earthquake swarm and aftershock through deploying a temporary observatory close to the affected region. This dataset includes a record of the date, time, location, depth, magnitude, and source of every Indian earthquake since 2018.

**Model:**

To help with training the model, we use Clustering, Random Forest, and Regression analysis.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. In this process, combination of multiple classifiers is used to solve a complex problem and to improve the performance of the model.

Decision Tree is a Supervised learning technique, It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

Naive Bayes use a similar approach to forecast the likelihood of various classes based on various attributes. This approach is commonly used in text categorization and when dealing with problems that have several classes.

Polynomial Regression is a regression algorithm that models the relationship between a dependent and independent variable as nth degree polynomial. It is a linear model with some modification in order to increase the accuracy.

Clustering is a machine learning technique, which groups the unlabeled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

With the help of the above models, we can learn and implement along with that we can predict the impact of earthquake in scale.

**Methodology:**

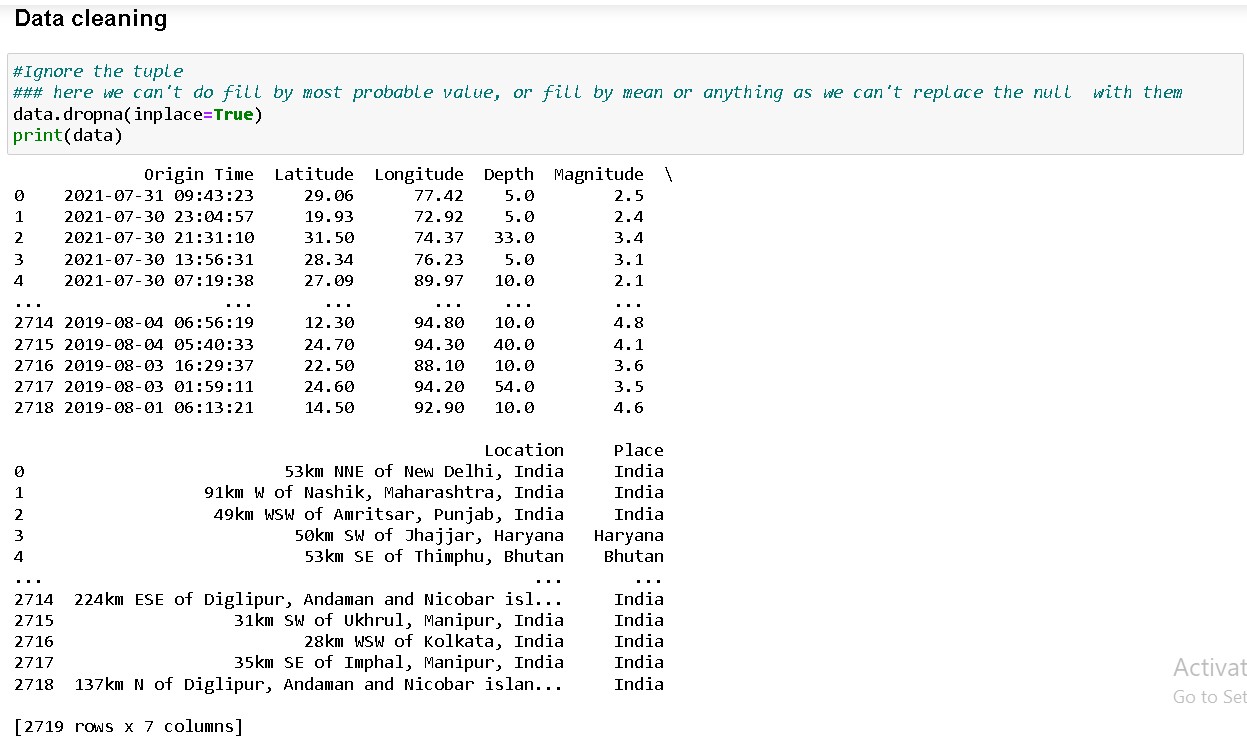
In methodology we use the dataset which is provided information about the earthquake prediction, firstly, the data must be in the form of error free, and usable format to get into that we need to pre-process the data.

**Data - Cleaning**:

The dataset which was acquired from the Kaggle, we have to check on the correctness instead of duplicates, so we need to firstly implement the data - cleaning process.

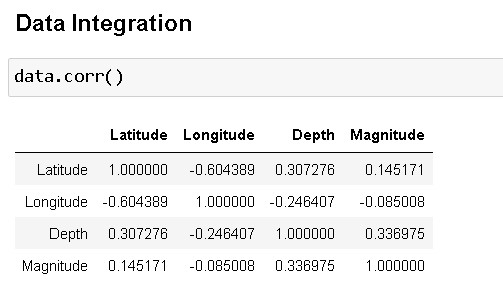
Firstly, data cleaning reduces errors and improves data quality, Data cleaning is the process of removing faulty data, organizing the raw data, and filling in missing information in order to prepare raw data for analysis. Cleaning data prepares it for data mining, which extracts the most useful information from the data collection.

Quality data management is primarily concerned with data quality. Data quality issues can arise anywhere in an information system. Data cleaning can help with these issues.



**Data - Integration**:

Data mining involves data integration, where multiple data sources are merged together and also checks for redundancy.

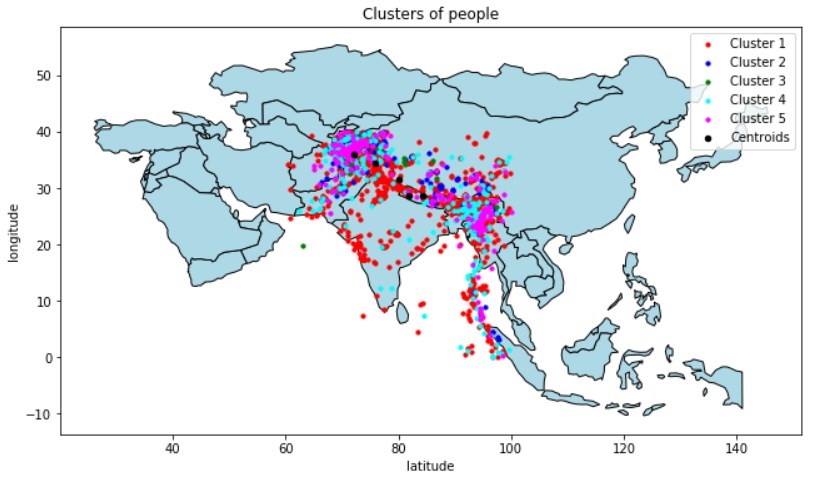


Here we can say that the values are not dependent to each other from correlation matrix

**Data - Transformation**:

The data needs to transform into unique appropriate format.

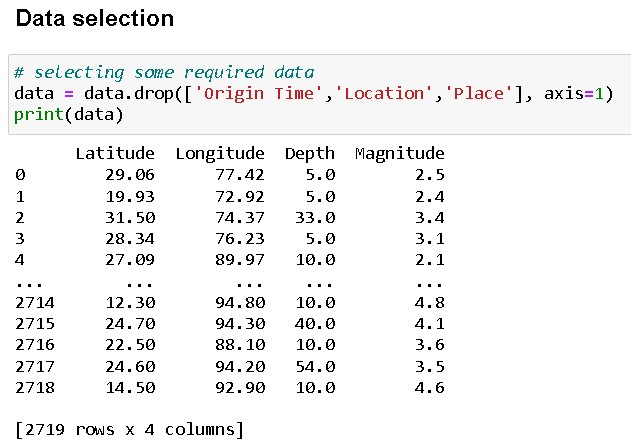




By reference to the decision tree, the multi-label distribution is calculated, and the labels are assigned.

**Data - Selection**:

The process of choosing the proper data kind and source, as well as appropriate data collection tools. The activity of data gathering is preceded by the selection of data. It helps to select over required data.



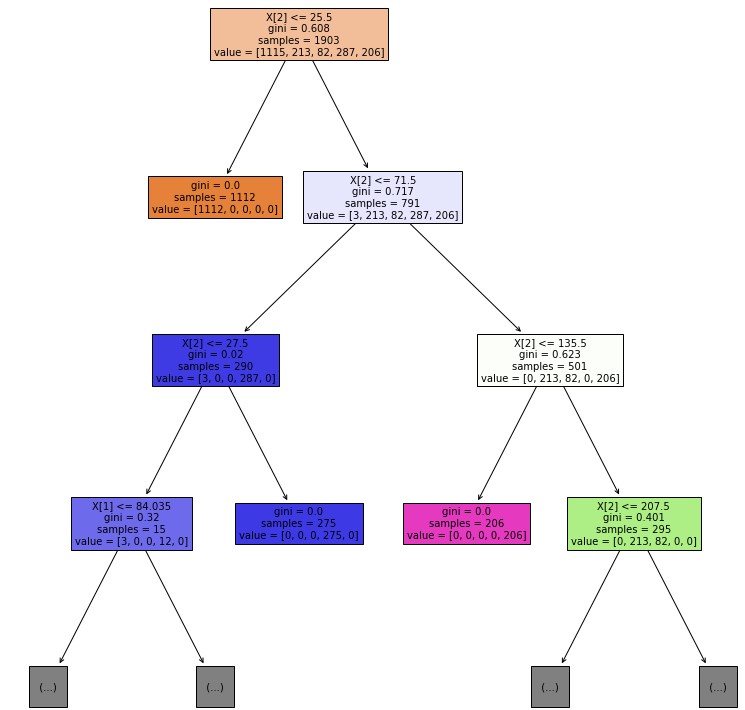
**Decision Tree**:

The supervised learning category includes the decision tree method. They can be used to address problems involving regression and classification. The problem is solved using the tree representation, in which each leaf node corresponds to a class label and characteristics are represented on the tree's interior node.

**Algorithm**:

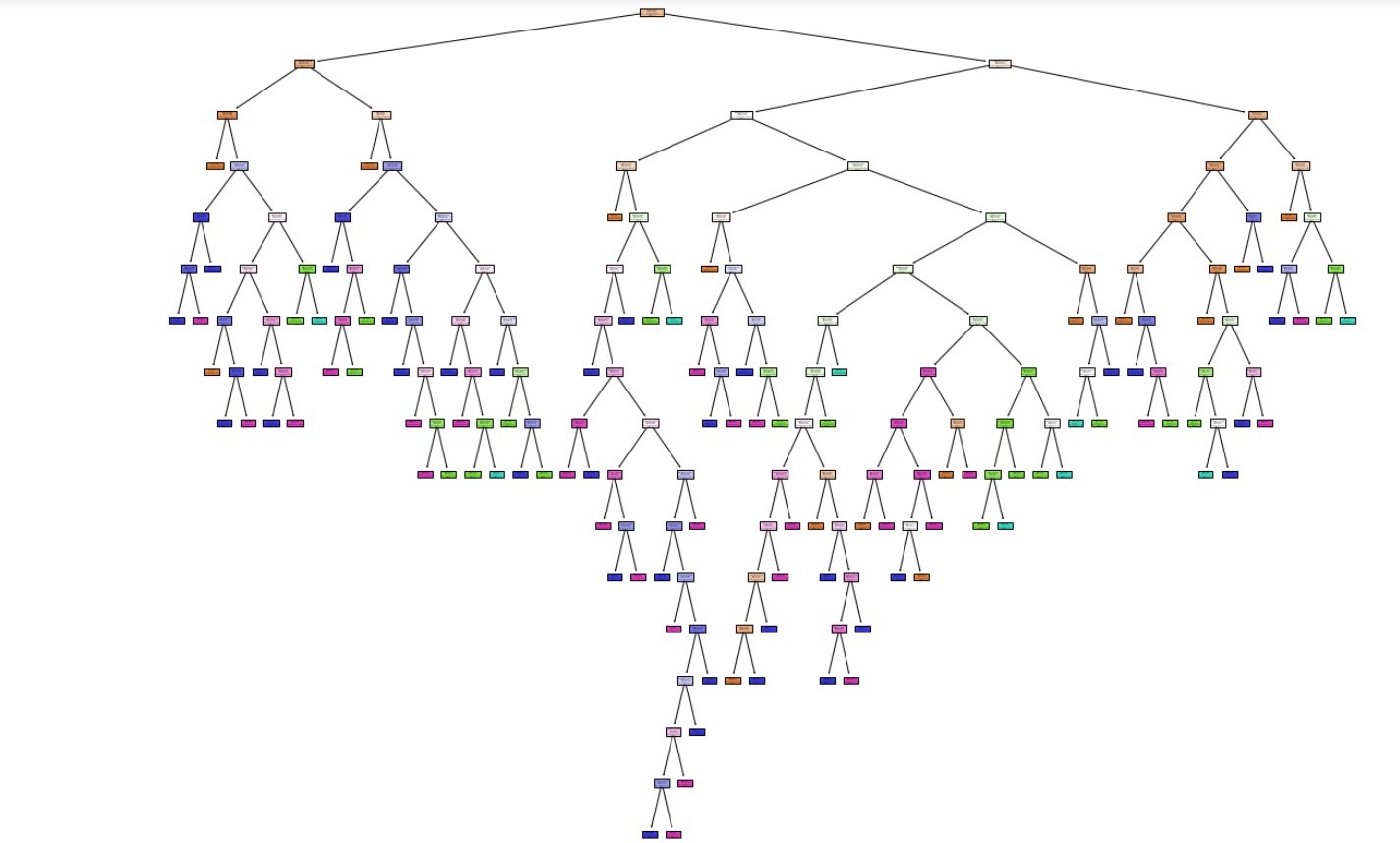
A fundamental challenge in the Decision Tree is determining the characteristic for the root node at each level. Attribute selection is the term for this procedure. There are two widely used attribute selection methods:

1. Obtaining Information
2. Gini coefficient



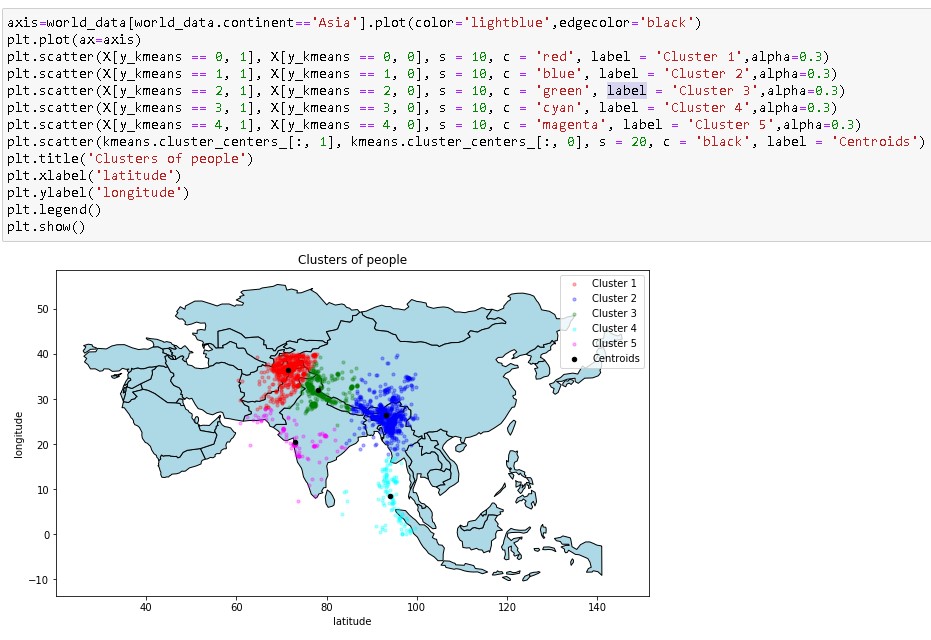
**Random Forest:**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. In this process, combination of multiple classifiers is used to solve a complex problem and to improve the performance of the model.

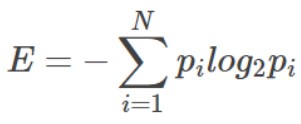


**Finding hotspots using Clustering**:

In this we take the longitude and latitude of the earthquake and plotted them using clustering, This forms the earthquakes into groups, each cluster has a different color and the cluster center is taken as hotspot. These hotspots are highlighted in black color.

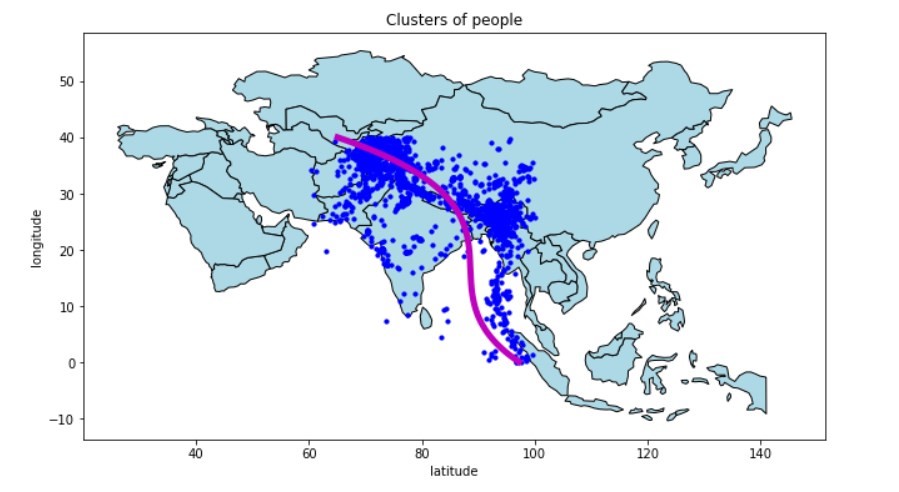


**Entropy Formula**:



**Why earthquakes are formed around India:**

First, we plot the earthquakes on the geographical map and we can see that the points are in curved manner so we used polynomial regression to get to find the relation between them and get a curved line. We get this curved line around the border of India because during the breakup of Pangea, the Indian subcontinent became isolated from the southern part of Pangea, called Gondwanaland and moved northwards and eventually collided with Eurasia to form the Himalayas. Due to this collision the tectonic plates move regularly which cause earthquakes around this line.



**Performance Analysis**:

Using the random forest classification, we can test the accuracy for the model is 100.0%

When decision tree is used, we received the accuracy of 99.01%

When Naive Bayes is used, we received the accuracy of 95.09%

**Conclusion and Future Work:**

Using K-MEANS clustering, and the Random Forest methodology, the project Earthquake prediction system was successfully completed. The data set was acquired from Kaggle and preprocessed according to attributes. After that, the data was used to develop a device model. When it comes to earthquake prediction, our algorithm performs best in terms of determining different levels of accuracy. This model can be used to assess how effective various sets of features are at predicting earthquakes. On the data set, we used k-means clustering to find the centroid which acts as hotspots and helps in dividing class labels, Classifier for classifying the the impact of earthquake by using the latitude, longitude, magnitude and depth. Regression to derive the relation between geographical position of India to earthquake positions. This model gives the accuracy range between 95% to 100%.

**References:**

1. K-Nearest Neighbor (KNN) algorithm for Machine Learning - Javatpoint. www.javatpoint.com. (n.d.). Retrieved April 4, 2022, from [https://www.javatpoint.com/k-nearest-neighbor-algorithmfor-machine-learning](https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning)

1. Regression analysis in machine learning - javatpoint. www.javatpoint.com. (n.d.). Retrieved

April 4, 2022, from <https://www.javatpoint.com/regression-analysis-in-machine-learning>

1. Machine learning random forest algorithm - javatpoint. www.javatpoint.com. (n.d.). Retrieved

April 4, 2022, from <https://www.javatpoint.com/machine-learning-random-forest-algorithm>

1. Journal of Soft Computing Exploration. SHM Publisher. (n.d.). Retrieved April 4, 2022, from <https://www.shmpublisher.com/index.php/joscex/article/view/51/43>