# **INFO6105 Data Science Engineering Tools and Methods**

# **Earthquake Analysis and Prediction**

# Team 6:

Mubin Modi NUID: 002822945

Dhriti Kanchan NUID: 002794620

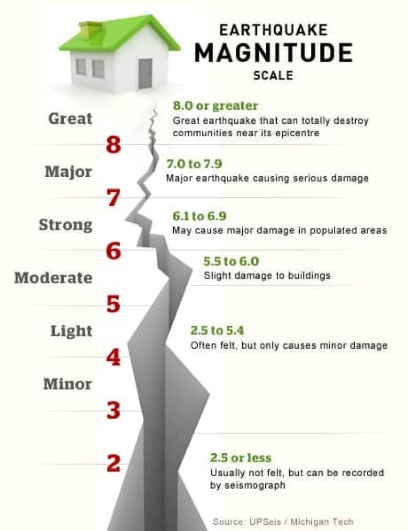
## **Introduction**

**Background:**

Earthquakes, sudden bursts of energy from the Earth's crust, can cause widespread damage and disrupt lives. While they're naturally occurring, predicting them could significantly reduce their impact. Traditionally, we only reacted to earthquakes after they happened, focusing on helping those affected. But now, advancements in earthquake science and data analysis offer a glimmer of hope: the possibility of forecasting seismic activity. By studying past earthquakes and identifying patterns, scientists can build more accurate models to predict future ones. This proactive approach empowers authorities to take preventive measures. They can issue warnings, design earthquake-resistant structures, and prepare for emergencies. Ultimately, predicting earthquakes could save lives and protect livelihoods from the devastating consequences of these natural disasters.

**Richter Scale:**

The Richter scale, developed by Charles F. Richter in 1935, is a logarithmic scale used to measure the magnitude of earthquakes. The magnitude quantifies the energy released at the earthquake's source. Each whole number increase on the Richter scale represents a tenfold increase in measured amplitude of seismic waves and approximately 31.6 times more energy release. For example, an earthquake with a magnitude of 6.0 is 10 times more powerful than one with a magnitude of 5.0. The Richter scale provides a standardized way to communicate the size and impact of seismic events, serving as a crucial tool for seismologists and emergency responders in assessing earthquake severity.



**Motivation:**

**Addressing the Pressing Need for Earthquake Prediction: A Project for Safer Communities**

The devastating consequences of earthquakes, with their potential for widespread destruction and loss of life, necessitate the urgent development of effective prediction models. This earthquake analysis and prediction project aims to address this critical need by leveraging the power of data and past seismic events to enhance our understanding and forecasting capabilities.

Through in-depth exploration of historical data and identification of recurring patterns, we strive to develop highly accurate models that can predict future earthquakes with greater certainty. This knowledge will empower us to transition from a reactive to a proactive approach, enabling communities to implement crucial measures for risk mitigation and preparedness.

Our project seeks to deliver tangible outcomes that contribute to a safer future:

* **Early warnings:** Timely predictions will allow for evacuation of at-risk populations and implementation of emergency protocols, potentially saving countless lives.
* **Resilient infrastructure:** By incorporating earthquake-resistant design principles into critical infrastructure, we can minimize damage and ensure faster recovery times.
* **Enhanced preparedness:** Informed by accurate forecasts, communities can develop comprehensive response plans, allocate resources effectively, and raise public awareness, leading to a more effective response in the event of an earthquake.

By proactively addressing the threat of earthquakes through improved prediction capabilities, we can mitigate their devastating impact and build a more resilient future for vulnerable communities. This project represents a significant step towards achieving this vital goal.

Top of Form

**Goal:**  
This project delves deep into earthquake data, dissecting the geological zones impacted by these tremors and their accompanying tsunamis. We aim to unravel the mysteries behind earthquakes, pinpointing the factors that influence their frequency and severity. This knowledge will serve as a cornerstone for preventive measures on local and global scales, allowing us to prioritize the construction of earthquake-resistant structures in vulnerable areas. By scrutinizing tectonic movements and uncovering long-term seismic patterns, we strive to enhance public awareness and preparedness. Ultimately, our objective is to predict high-risk seismic zones and develop accurate forecasts for future earthquakes, empowering communities to mitigate their devastating impact.

## **Methodology**

We used the following libraries:-

A screen shot of a computer

Description automatically generated

**Data preprocessing and cleaning:**

* We loaded the dataset into to access the earthquake data for further examination.
* We inspected column names and data types to gain a preliminary understanding of the dataset's structure and composition.
* We Scrutinized the dataset for values, aiming to detect and address any instances of duplicate entries that may compromise data integrity.
* We Converted the date column from an object type to a datetime format to facilitate temporal analysis and ensure consistency in date-related operations.
* Verified each column for unique values, providing insights into the diversity and distinctiveness of data across different features.
* Reevaluated values, particularly from newly created features, to ensure accuracy and consistency in the dataset.
* Cleanse columns, specifically addressing the 'status' and 'data type' columns, by removing unnecessary values and ensuring data consistency.

**Exploratory data analysis and Feature engineering:**

Our objective was to visually depict the geographical impact of earthquakes and tsunamis, identify areas in need of earthquake-resistant structures, and gain insights into the factors influencing their frequency and severity.

We utilized feature engineering to boost model interpretability and accuracy, strategically crafting and modifying features to extract relevant information. This process improved our understanding of data patterns, enhancing predictive model performance. Additionally, we employed a label encoder to convert categorical data into numerical format.

* Extracted relevant columns from the date column to create additional features, enriching the dataset for more in-depth analysis.
* Generated a comprehensive profile report to conduct a thorough exploration of dataset characteristics, distributions, and potential anomalies.
* Plotted graphs to visually explore various relationships between different variables, gaining insights into potential patterns or trends.
* Utilized external library data to establish geographical boundaries corresponding to tectonic plates, enhancing the dataset's contextual information.
* Employed label encoding and color-coding techniques to transform non-numerical data into a numerical format, facilitating quantitative analysis.
* Calculated correlation metrics, with a particular focus on the magnitude variable, to understand relationships between different features in the dataset.

**Model Selection and Evaluation:**

We used 40% of our dataset as our dataset has over 3 million rows.

1. Linear regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fit line that represents the overall trend in the data, allowing for predictions based on the input variables.

Step 1: We are taking random sample of the dataset which consist of 40% of the original dataset.

Step 2: Used train\_test\_split function to train and test the dataset with train size of 80% and random state 42.

Step 3: We have used gridsearchCV for hyperparameter tuning.

Step 4: Determined the test accuracy using y\_test and prediction of X\_test.

1. Ridge regression

Ridge regression is a regression technique that adds a penalty term to the traditional linear regression model. This penalty helps prevent overfitting by discouraging the model from assigning excessive importance to any feature, leading to a more robust and generalized prediction.

Step 1: We are taking random sample of the dataset which consist of 40% of the original dataset.

Step 2: Used train\_test\_split function to train and test the dataset with train size of 80% and random state 42.

Step 3: We have used gridsearchCV for hyperparameter tuning.

Step 4: Determined the test accuracy using y\_test and prediction of X\_test.

1. XGBoost regressor  
   XGBoost regressor is a machine learning algorithm that excels in predicting numeric outcomes. It combines the strengths of multiple decision trees to provide accurate and efficient predictions, making it widely used in various fields for tasks like forecasting and regression analysis.

Step 1: We are taking random sample of the dataset which consist of 40% of the original dataset.

Step 2: Used train\_test\_split function to train and test the dataset with train size of 80% and random state 42.

Step 3: We have used gridsearchCV for hyperparameter tuning.

Step 4: Determined the test accuracy using y\_test and prediction of X\_test.

1. Decision Tree:

A Decision Tree is a machine learning model that resembles a tree structure, where decisions are made based on conditions at each node. It recursively splits the dataset into subsets, assigning outcomes to each branch. Decision Trees are interpretable, making them valuable for understanding decision-making processes. However, they are prone to overfitting, and their simplicity may limit performance in complex datasets.

Step 1: We are taking random sample of the dataset which consist of 40% of the original dataset.

Step 2: Used train\_test\_split function to train and test the dataset with train size of 80% and random state 42.

Step 3: Determined the test accuracy using y\_test and prediction of X\_test.

1. Random Forest:

Random Forest is an ensemble learning method that leverages multiple decision trees during training. By aggregating the predictions of individual trees, it enhances overall accuracy and mitigates overfitting. Random Forest is versatile, capable of handling high-dimensional data and outliers effectively. Its robustness and ability to provide feature importance make it a popular choice in various machine learning applications.

Step 1: We are taking random sample of the dataset which consist of 40% of the original dataset.

Step 2: Used train\_test\_split function to train and test the dataset with train size of 80% and random state 42.

Step 3: Determined the test accuracy using y\_test and prediction of X\_test.

**Description of the Dataset:[1]**

The earthquake dataset serves as an extensive repository capturing every recorded earthquake globally from 1990 to 2023. Comprising an impressive three million entries, each row corresponds to a distinct earthquake event, offering a comprehensive perspective. With a dataset size of 471MB, it provides valuable insights into seismic activities worldwide. The dataset includes the following columns:

1. Time: The timestamp of the earthquake event in milliseconds.

2. Place: Geographical location where the earthquake occurred.

3. Status: Represents the current state or condition of the event.

4. Tsunami: Boolean value indicating the presence of associated ocean waves.

5. Significance: Denotes the importance or impact level of the earthquake event.

6. Data\_Type: Specifies the type of referenced data.

7. Magnitude: Measurement of the earthquake's size or intensity.

8. State: Represents the administrative division or state where the event occurred.

9. Longitude: Coordinates of the earthquake's location.

10. Latitude: Coordinates the earthquake's location.

11. Depth: The depth of the earthquake's epicenter.

12. Date: Date when the earthquake occurred.

Data Source  
<https://www.kaggle.com/datasets/alessandrolobello/the-ultimate-earthquake-dataset-from-1990-2023/data>

<https://github.com/fraxen/tectonicplates>

## **Results and Analysis**

The methodology used for the project had the above-mentioned steps and their respective results are as follows:  
  
**Data Preparation and Exploratory Data Analysis:**

A graph showing the earthquake

Description automatically generated

Figure 1.1

The distribution of earthquake magnitudes is evident in the histogram, spanning from approximately 2 to 9. A notable peak is observed in the frequency of earthquakes with a magnitude around 5. However, earthquakes with a magnitude of 7 or higher are comparatively infrequent.

The histogram further illustrates a temporal trend, indicating an increase in earthquake magnitudes. This upward trajectory is likely attributed to advancements in scientific capabilities for earthquake detection and recording.

Examining the line graph, it becomes evident that the rise in earthquake magnitudes over time is not consistently linear. Certain years exhibit no recorded earthquakes, and the occurrence of earthquakes with a magnitude of 7 or higher remains relatively scarce.

A graph of different colored bars

Description automatically generated with medium confidence

Figure 1.2

The image shows a plot showing the magnitude of earthquakes per year from 1990 to 2023. The average magnitude of earthquakes has remained relatively constant over time. There has been a slight increase in the number of earthquakes in recent years, but this could be due to improved detection capabilities.

* Earthquake magnitudes have remained relatively constant over time.
* The number of earthquakes has increased slightly in recent years, but this may be due to improved detection capabilities.

A chart of different colored bars

Description automatically generated with medium confidence

Figure 1.3

This graph shows the range of earthquake sizes across different years. It gives a visual idea of how often earthquakes of different strengths happen over time. From the graph, we can see that most earthquakes are mild to moderate, with magnitudes between 2 and 6 on the Richter scale. However, there are some rare cases, outliers, where earthquakes are much stronger. Even though these stronger earthquakes are not as common, we need to consider them in our analysis because they can cause significant damage.

A graph with orange and black lines

Description automatically generated

Figure 1.4

This information pertains to the most powerful earthquakes documented within the period spanning from 1990 to 2023, as measured on the Richter Scale. The Richter Scale is commonly used to quantify the magnitude of seismic events, providing a numerical representation of their strength. The statement indicates a focus on the exceptionally strong earthquakes recorded during this specific timeframe, offering valuable insights into the most impactful seismic occurrences over the given duration. This data is significant for understanding and assessing the seismic activity's intensity and its potential implications during the specified period.

The strongest earthquakes plotted by years on the world map with year slicer.

A map of the world

Description automatically generated

Figure 1.5

A map of the world

Description automatically generated

Figure 1.6

A map of the world

Description automatically generated

Figure 1.7

A graph showing the number of data

Description automatically generated with medium confidence

Figure 1.8

A graph showing the state of state

Description automatically generated

Figure 1.9

A graph showing the number of earthquakes

Description automatically generated

Figure 2.0

A graph showing the number of earthquakes

Description automatically generated

Figure 2.1

A graph showing the growth of earthquakes

Description automatically generated

Figure 2.2

A graph showing the number of earthquakes

Description automatically generated

Figure 2.3

**Graphs made using tectonic data [2]**

A map of the world with red dots

Description automatically generated

Figure 2.4

Mapping earthquake coordinates onto an image depicting tectonic plates and country boundaries provides a spatial understanding of seismic activity in relation to the Earth's structure. This integrated approach offers a valuable tool for researchers to study the correlation between earthquakes, tectonic plate interactions, and proximity to country borders, aiding in a holistic interpretation of seismic patterns and regional impacts. From this graph we can see the distribution of all earthquakes from 1990-2023 over the tectonic plates beneath the earth surface. However we need to filter this data to get better idea of more devastating earthquakes

A graph showing the formation of a volcano

Description automatically generated with medium confidence

Figure 2.5

This visual representation illustrates the correlation among earthquake magnitude, significance, and the depth of the epicenter, offering insights into the interplay between these factors. The graph will help identify threshold values for magnitude, , significance, and the depth.

A map of the earth showing the natural disasters

Description automatically generated

Figure 2.6

By applying filters for earthquakes with a magnitude exceeding 8, significance surpassing 500, and depth greater than 100, our analysis suggests that high-impact seismic events tend to cluster around the faults or boundaries of tectonic plates. We can make the following conclusions from the graph in fig 2.6:-

- Tectonic Plate Correlation:

- Earthquake occurrences align closely with tectonic plate boundaries.

- Indicates higher earthquake frequency in areas where tectonic plates meet.

- High Seismic Activity Zones:

- Red dots on the map, representing earthquakes, densely cluster along plate boundaries.

- Active regions include the circum-Pacific belt ("Ring of Fire") and boundary lines in Asia and the Himalayas.

- Potential Risk Assessment:

- Countries along tectonic plate boundaries face a higher earthquake risk.

- Implications for infrastructure development and emergency preparedness planning.

- Data Distribution Patterns:

- Earthquakes are not uniformly distributed along plate boundaries.

- Some sections exhibit clusters of activity, indicating more active seismic zones.

- Mid-Ocean Ridges and Continental Boundaries:

- Seismic activity along mid-ocean ridges (e.g., Mid-Atlantic Ridge) and continental boundaries.

- Illustrates different types of plate boundaries: divergent, convergent, and transform.

A map of the world with red lines

Description automatically generated

Figure 2.7

We attempted to visualize earthquake locations where the tsunami value is equal to 1 (indicating a potential tsunami occurrence). However, during this process, we encountered inconsistencies in the dataset. Tsunamis occur near coasts because they are triggered by undersea earthquakes along tectonic plate boundaries. During these earthquakes, the seafloor's vertical displacement displaces a large volume of water, generating tsunami waves. While the tsunami may not be immediately noticeable in deep-sea regions, it becomes highly destructive near coastlines due to the amplification of wave height as it approaches shallow coastal areas. So we have to filter our data with respect to depth of the epicenter.

A graph showing the number of particles

Description automatically generated with medium confidence

Figure 2.8

In this analysis, we aimed to explore the extreme values of depth, significance, and magnitude, seeking a better understanding of their distribution and potential impact.

A diagram of earthquakes

Description automatically generated with medium confidence

Figure 2.9

To address the inconsistency observed in the previous Figure 2.7, we aimed to determine the optimal criteria for filtering tsunami data based on magnitude and significance. Our goal was to identify earthquakes that potentially trigger tsunamis by comparing the magnitudes and significance levels of the seismic events.

A map of the world

Description automatically generated

Figure 3.0

In this visual representation, we focus on filtered regions that specifically showcase instances of tsunamis and their proximity to coastlines. The highlighted areas underscore the geographical locations where tsunamis take place in connection with coastal boundaries. Following our analysis, we noted that the depth values for deep-sea earthquakes were consistently zero, a logical observation given the challenge of determining the epicenter depth for such events.

**Correlation Matrix w.r.t Magnitude:**

A close-up of a chart

Description automatically generated

Figure 3.1

Analyzing this correlation matrix allows us to pinpoint optimal parameters for our prediction algorithms by assessing their correlation with the target variable, which, in this case, is magnitude. This exploration helps identify key features that strongly influence our predictive models.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Model Evaluation:**

Evaluating various algorithms to determine the most effective model for the given task and dataset. This process involves assessing their performance metrics and overall suitability to identify the optimal algorithm for the specific context:

A screenshot of a computer code

Description automatically generated

A graph with a blue line

Description automatically generated

A screenshot of a computer code

Description automatically generated

A graph showing a red line and blue dots

Description automatically generated

A screenshot of a computer code

Description automatically generated

A graph showing the difference between earthquake and earthquake

Description automatically generated

A screenshot of a computer program

Description automatically generated

A graph with a red line and blue dots

Description automatically generated

A screenshot of a computer

Description automatically generated

Mean Squared Error on Test Set (Random Forest): 0.017138338539015818

Mean Absolute Error on Test Set (Random Forest): 0.01562594026571847

R-squared on Test Set (Random Forest): 0.9898979234163428

A graph with a red line and blue dots

Description automatically generated

## **Conclusion**

## Plate Boundaries and Tension: The concentration of earthquakes and tsunamis near tectonic plate boundaries is predominantly attributed to the tension along these boundaries. Tectonic plates, immense pieces of the Earth's crust, interact at these boundaries, creating stress and tension that can result in seismic activities.

## Identifying High-Impact Areas: Understanding this connection allows for the confident identification of high-impact areas prone to significant seismic events. This knowledge underscores the critical need for constructing resilient structures in regions located near tectonic plate boundaries, where the risk of earthquakes and tsunamis is notably higher.

## Predictive Model Success: The success in forecasting earthquake magnitudes using advanced models like Linear Regression, Ridge Regression, XGBoost Regressor, Decision Tree and Random Forest highlights the effectiveness of machine learning techniques in predicting seismic events. These models utilize historical data to provide valuable insights into potential future occurrences.

## Performance Metrics:

## Linear Regression: Mean Squared Error (MSE): 0.18299 || Mean Absolute Error (MAE): 0.29188 || R-squared: 0.89214

## Ridge Regression:Mean Squared Error (MSE): 0.18299 || Mean Absolute Error (MAE): 0.29188 || R-squared: 0.89214

## XGBoost Regressor: Mean Squared Error (MSE): 0.01768 || Mean Absolute Error (MAE): 0.01990 || R-squared: 0.98958

## Decision Tree: || Mean Squared Error (MSE): 0.03146 || Mean Absolute Error (MAE): 0.01931 || R-squared: 0.98145

## Random Forest: Mean Squared Error (MSE): 0.01714 || Mean Absolute Error (MAE): 0.01563 || R-squared: 0.98990

## Among the linear models (Linear Regression, Ridge Regression), both perform similarly with high R-squared values.

## XGBoost Regressor outperforms linear models with significantly lower MSE and MAE, and a higher R-squared value.

## Decision Tree and Random Forest models show strong performance, with Random Forest slightly edging out XGBoost in terms of MSE and MAE.

## In general, the tree-based models (XGBoost, Decision Tree, Random Forest) demonstrate superior predictive performance compared to the linear models in this scenario. Random Forest and XGBoost, in particular, showcase excellent performance with the lowest MSE, MAE, and highest R-squared values among all models

## Project Impact: The project's comprehensive exploration of earthquake data contributes significantly to our geological understanding. It untangles the complexities surrounding seismic events and identifies key factors influencing their occurrence and intensity. These findings are paramount for implementing preventive measures and prioritizing the construction of earthquake-resistant structures, particularly in regions vulnerable to seismic activities. The project's impact extends beyond data analysis, emphasizing the practical implications for enhancing resilience and preparedness in earthquake-prone areas.

1. The project's comprehensive exploration of earthquake data contributes significantly to our geological understanding. It untangles the complexities surrounding seismic events and identifies key factors influencing their occurrence and intensity. These findings are paramount for implementing preventive measures and prioritizing the construction of earthquake-resistant structures, particularly in regions vulnerable to seismic activities. The project's impact extends beyond data analysis, emphasizing the practical implications for enhancing resilience and preparedness in earthquake-prone areas.Top of Form

## **References**

[1] <https://www.kaggle.com/datasets/alessandrolobello/the-ultimate-earthquake-dataset-from-1990-2023/data>

[2] <https://github.com/fraxen/tectonicplates>