EARTHQUAKE PREDICTION USING MACHINE LEARNING MODEL

Submitted To:
Farhana Afrin Duty
Assistant Professor, Department of Statistics and Data Science
Jahangirnagar University

Submitted By:

Name: Roksana Parvin
Student ID: 20231177

Date: December 22, 2023
Department of Statistics, Jahangirnagar University

LIST OF CONTENTS

I. INTRODUCTION

In recent years, the frequency and impact of earthquakes have become increasingly significant, posing serious threats to human lives and infrastructure. These natural disasters are influenced by various factors, including increasing temperatures, the velocity of tectonic plates, and anthropogenic seismic activities (Khurram, 2021). To address this challenge, the use of machine learning algorithms has emerged as a promising approach for earthquake prediction. By leveraging the large amounts of training and ground-truth data available, machine learning techniques have been successfully applied to various aspects of seismology, including seismic waveform classification, event localization, earthquake prediction, and earthquake early warning (Bianco et al., 2019). One specific objective of this ML project paper is to predict the probability of earthquakes and tsunami in a particular country for a given year. To achieve this, we will utilize a range of data sources and apply machine learning methods logistic regression. The algorithms will be implemented in the python programming language. The overall aim of this project is to develop a predictive model that can accurately assess the probability of earthquakes occurring in a specific country during a given year.

By effectively analyzing and interpreting seismic data, the developed model will contribute to enhancing early warning systems and enabling proactive measures to mitigate the impact of earthquakes. The potential applications of this ML project in earthquake prediction are promising not only for improving emergency response and preparedness but also for minimizing the overall damage caused by earthquakes.

II. LITERATURE REVIEW

Earthquake activity is presumed as a spontaneous phenomenon that can damage huge number of lives and properties, and currently there is no any model exists that can predict the exact position, magnitude, frequency and time of an earthquake. Researchers have conducted several experiments on earthquake events and forecasts, leading to a variety of findings based on the factors considered. The well-known Gutenberg and Richter statistical model found a correlation between the magnitude of earthquake and frequency of earthquake. For structural design, this earthquake probability distribution model was used. In supervision of the California Geological

Survey, Petersen conducted research and suggested a model that is time-independent. This time independent model demonstrating that chances of occurrence of earthquake follow the Poisson's distribution model. Shen suggested a probabilistic earthquake forecasting model based on the strain studied between the behaviour of tectonic plates. Based on this model, higher measured strain results in a higher risk of earthquake. Ebel provided a long-term prediction model that allowed for the extrapolation of previous earthquakes with magnitudes greater than and up to 5.2 in order to forecast possible seismic events. There are various methods for predicting earthquakes using Artificial Neural Networks and seismic precursors are discussed in the literature. Negarestani used a Back Propagation Neural Network to identify irrational behaviour in concentration of radon due to occurrence of earthquake. The presence of radon gas in soil is constantly measured and researcher have founded that it varies constantly due to changes in environment. The concentration of soil radon also rises due to seismic activity. This radon can be differentiated from natural variations caused by the environment through neural networks. Since splitting the entire globe in four quadrants, the system devices establish logic and correlation principles based on the historical record of earthquakes. The expert method will forecast earthquakes in each quadrant of the world for a period of 24 hours. Panakkat and Adeli presented an enthralling approach to earthquake prediction based on mathematically determined seismic indicators derived from the spatial variation of historical seismic events for Southern California. The algorithm makes monthly predictions, and the parameters are modelled using various Artificial Neural Networks. The estimation of all those parameters required to make sufficient earthquake database. For this limited number of times, the events were executed to measure the parameters of seismic event before taking the month into account. After this study, Adeli and Panakkat used exactly same parameters of seismic in collaboration with Probabilistic Neural Network to forecast earthquakes. Morales-Esteban and Reyes suggested separate seismic criteria for earthquake prediction using mathematical calculations in Chile and Iberia for a time interval of 8–9 days, respectively. For modelling the relationship between earthquake events and parameters, these parameters are determined using Bath's law and Omori's law. Zamani proposes using a combination of neural networks and mathematical logic to forecast earthquakes in Iran. For a selected group of seismicity indices, this study includes information normalization and corresponding feature extraction accompanied by principal component analysis. Mirrashid provides another design for earthquake prediction in Iran, which incorporates symbolic logic,

fuzzy C-means, subtractive clustering, and grid partitioning. Through this model, we try to predict earthquakes by training various Machine Learning models on seismic and acoustic data from a laboratory micro earthquake simulation.

III. OBJECTIVE

Main Objective

- To develop a predictive model for estimating earthquake probability in a country in a certain year

Specific Objectives:

- Identify key indicators and parameters for earthquake prediction
- Collect and analyze historical earthquake data for [Country]
- Develop a machine learning model for earthquake probability prediction
- Validate and optimize the model using relevant statistical methods

IV. TOOLS AND MODULS

- Python 3
- Jupyter notebook
- Important modules: Pandas, Numpy, sklearn, matplotlib, seaborn, plotly, geopy

V. METHODOLOGY

Data Collection

Dataset: The dataset is collected from Kaggle that has 1000 historical data of major earthquake from 1995 to 2023.

The data set contains 19 columns. The columns are explained in the following table.

Dataset Columns

Sl. No	Column Name	Description
1	title	title name given to the earthquake
2	magnitude	The magnitude of the earthquake
3	date_time	date and time
4	cdi	The maximum reported intensity for the event range
5	mmi	The maximum estimated instrumental intensity for the event
6	alert:	The alert level - “green”, “yellow”, “orange”, and “red”
7	tsunami:	"1" for events in oceanic regions and "0" otherwise
8	sig	A number describing how significant the event is. Larger the number, more significant the event. This value is determined on a number of factors, including: magnitude, maximum MMI, felt reports, and estimated impact
9	net	The ID of a data contributor. Identifies the network considered to be the preferred source of information for this event
10	nst	The total number of seismic stations used to determine earthquake location
11	dmin	Horizontal distance from the epicenter to the nearest station
12	gap	The largest azimuthal gap between azimuthally adjacent stations (in degrees). In general, the smaller this number, the more reliable is the calculated horizontal position of the earthquake. Earthquake locations in which the azimuthal gap exceeds 180 degrees typically have large location and depth uncertainties
13	magType	The method or algorithm used to calculate the preferred magnitude for the event
14	depth	The depth where the earthquake begins to rupture
15	latitude	coordinate system by means of which the position or location of any place on Earth's surface can be determined and described
16	longitude	coordinate system by means of which the position or location of any place on Earth's surface can be determined and described
17	location	location within the country
18	continent	continent of the earthquake hit country
19	country	affected country

Importing Libraries: After downloading data from Kaggle, the important libraries are installed and imported in Jupiter notebook. Imported the data using pandas.

Pre-Processing

1. ***Initial Data Exploration:***
    - Checked the first ten, last ten, and randomly selected ten rows:
    - Verified the dataset's structure and content to gain an initial understanding.

- Checked the data shape:
- Examined the dimensions of the dataset to understand the number of rows and columns.
- Checked the metadata information:
- Inspected metadata information, including data types and non-null counts, to identify potential issues.

2. ***Handling Null Values:***
    - Checked for null values: Identified null values in the dataset, revealing 716 null values in the 'continent' column, 449 null values in the 'country' column, and 6 null values in the 'location' column.
    - Null value handling: Recognizing that 'location,' 'country,' and 'continent' can be derived from 'latitude' and 'longitude,' we opted to address null values by removing these three columns. This decision was made to ensure the integrity of the dataset while maintaining essential geographical information.
3. ***Updated Dataset:***

    Removed 'location,' 'country,' and 'continent': Executed the removal of the identified columns, as they were deemed redundant and could be reconstructed from other available information.
4. ***Revised Dataset Check:***
    - Checked the first ten, last ten, and randomly selected ten rows:
    - Verified the dataset after the removal of redundant columns to ensure the integrity of the remaining data.
    - Checked the data shape:
    - Reassessed the data dimensions to confirm the impact of column removal.
    - Checked the metadata information:
    - Revisited metadata information to ensure proper handling of null values and column removal.

5. ***Feature Selection:***

    The following features have been identified as crucial for the earthquake prediction model:

- **Magnitude:** Represents the size of the earthquake.
- **date_time:** Records the timestamp of earthquake occurrences.
- **cdi (Community Decimal Intensities):** Quantifies earthquake effects over an area.
- **mmi (Modified Mercalli Intensity):** Estimates shaking intensity at a specific location, considering its effects on people, objects, and buildings.
- **Tsunami:** Indicates the presence of a larger wave resulting from the earthquake.
- **Depth:** Measures the depth of earthquake effects, ranging from 0 to 700 KM.
- **latitude:** Specifies the north-south position of the earthquake.
- **longitude:** Specifies the east-west position of the earthquake.

Rationale for Selection

- **Magnitude:** The size of the earthquake is a fundamental parameter influencing its impact and potential consequences.
- **date_time:** Temporal information is crucial for understanding patterns and trends in earthquake occurrences over time.
- **cdi and mmi**: These intensity measures provide valuable insights into the seismic effects on both communities and specific locations, aiding in risk assessment.
- **Tsunami:** The occurrence of a tsunami is a significant indicator of the earthquake's potential for widespread impact. Depth: Understanding the depth of earthquake effects contributes to assessing the potential severity of the event.
- **latitude and longitude:** Geographical coordinates are essential for spatial analysis and mapping the distribution of earthquakes.

6. ***Feature Engineering:***

It involves either Feature Selection or Feature Extraction and Feature Scaling. A data set contains numerous of features which are random and may not be useful in prediction. Feature Engineering deals with reduction of random features under consideration and obtaining a set of minimum features which contribute to accurate prediction. Many algorithms are provided by ML for feature selection/extraction. Feature scaling is strategy used to standardize or normalize the range of features in the data-set. Feature Engineering

is useful as it compresses the data, reduces the storage space, computation time and removes redundant features. Used feature Engineering technique to derived year from the "date_time" columns

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical phase to understand the characteristics of the earthquake dataset. This report focuses on analyzing key features, including 'magnitude,' 'date_time,' 'cdi,' 'mmi,' 'tsunami,' 'depth,' 'latitude,' and 'longitude,' aiming to gain insights into seismic events and their patterns.

	magnitude	cdi	mmi	tsunami	depth
count	1000.000000	1000.000000	1000.00000	1000.000000	1000.000000
mean	6.940150	3.605000	6.02700	0.325000	74.612541
std	0.438148	3.328972	1.43399	0.468609	130.812590
min	6.500000	0.000000	1.00000	0.000000	2.700000
25%	6.600000	0.000000	5.00000	0.000000	16.000000
50%	6.800000	4.000000	6.00000	0.000000	29.000000
75%	7.100000	7.000000	7.00000	1.000000	55.000000
max	9.100000	9.000000	10.00000	1.000000	670.810000

1. ***Summary Statistics***

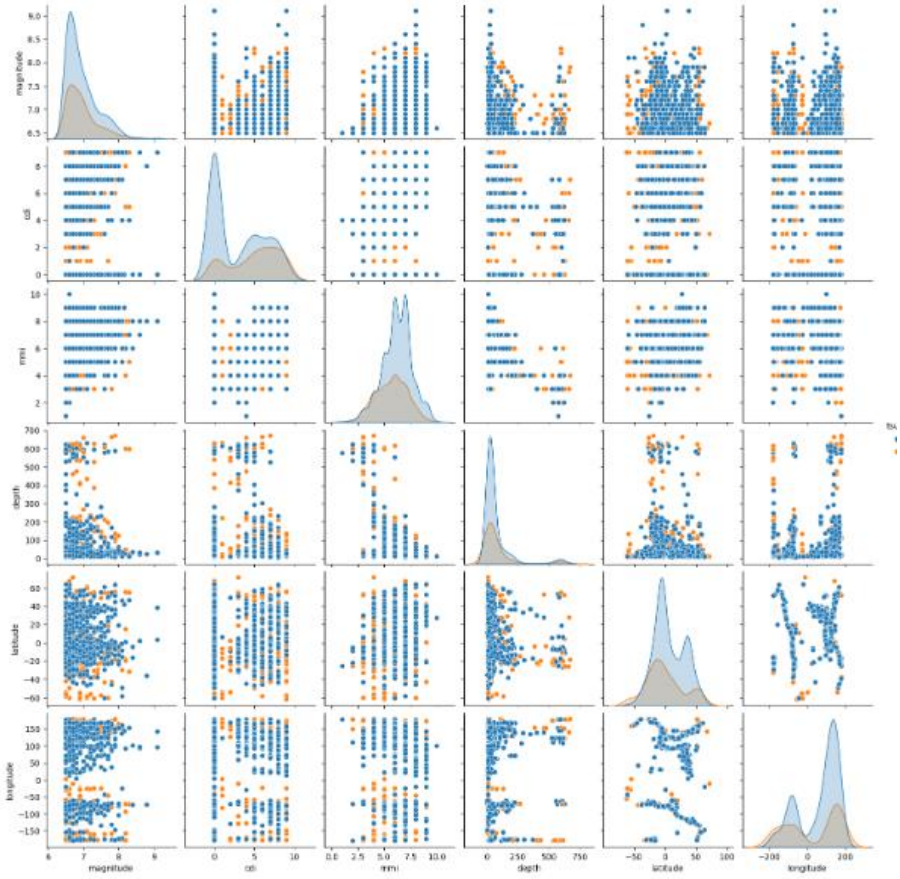
In summary statistics we have explored the *total count, mean, standard deviation, minimum value, maximum value, 25th percentile, 50th percentile or median, 75th percentile, and maximum value of* magnitude, cdi, mmi, tsunami, and depth column.

2. ***Visualization:***

Pair Plot:

Objective: The pair plot is designed to visualize relationships between pairs of variables in the dataset.

Implementation: A pair plot was generated for key variables such as 'magnitude,' 'cdi,' 'mmi,' 'depth,' 'latitude,' and 'longitude.' Scatterplots are included on the diagonal to visualize the univariate distribution of each variable. Off-diagonal plots provide scatterplots for each pair of variables, revealing potential patterns and correlations.



Heatmap:

Objective: The heatmap is employed to visualize the correlation matrix between numerical variables in the dataset.

Implementation: A heatmap was generated using the Seaborn library, displaying a color-coded matrix of correlation coefficients. High correlation is depicted by warmer colors, while cooler colors represent lower correlation. This visualization aids in identifying relationships between variables, guiding feature selection and model development.

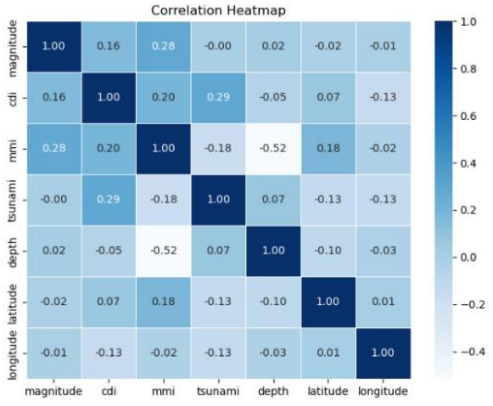
From the heatmap we can see that there is no multicollinearity among the features.

Boxplot:

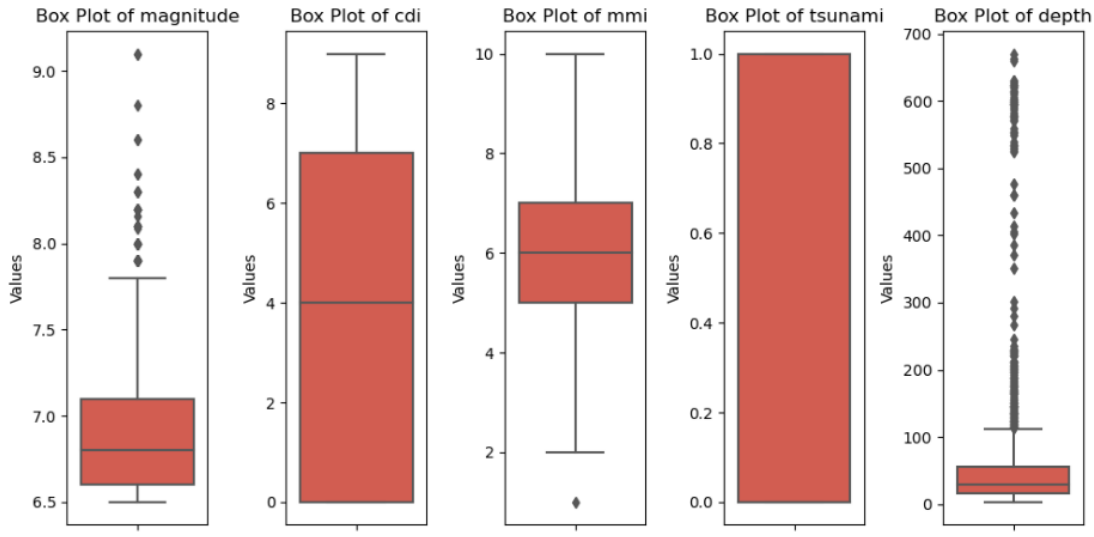
Objective: Boxplots are utilized to visualize the distribution and spread of specific variables, providing insights into central tendency and variability.

Implementation: Boxplots were created for 'magnitude' and 'depth' to understand their distributions. The central box represents the interquartile range (IQR), while the whiskers extend to the minimum and maximum values within 1.5 times the IQR. Outliers are identified as individual points beyond the whiskers, aiding in the detection of extreme values

Despite the presence of outliers observed in the boxplot, it is crucial to note that these data points, which may appear as outliers, actually signify instances of elevated seismic activity or high magnitude earthquakes. The variability beyond the typical range in the boxplot reflects the inherent diversity in earthquake magnitudes within the dataset. Therefore, these apparent outliers hold significance in capturing extreme seismic events, contributing valuable information to



our understanding of the earthquake data.



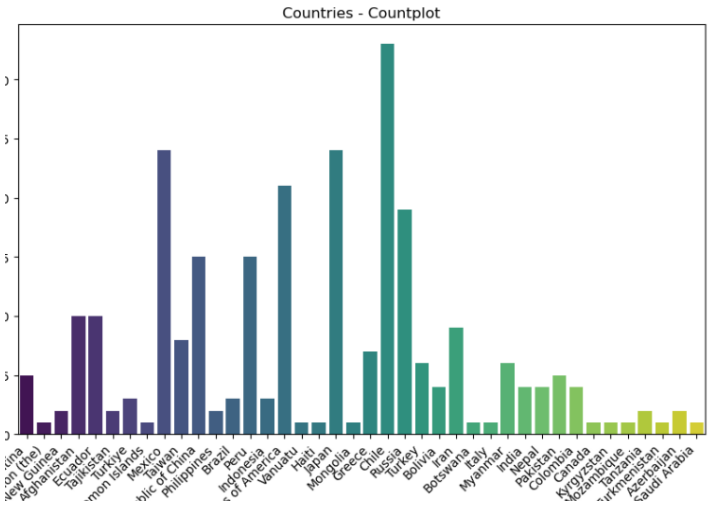
Feature Engineering

In this feature engineering step, geo py was employed to derive 'country' and 'continent' names from latitude and longitude coordinates, amplifying the dataset with geographical context. Utilizing geo py, we reverse-geocoded coordinates, extracting 'country' and 'continent' details for each earthquake record. This enhancement enables more granular data exploration, offering insights into earthquake patterns based on countries and continents. The inclusion of geographical features may enhance the predictive capacity of the model, considering location-specific nuances in earthquake occurrences.

Count plot:

Objective: Illustrate earthquake frequencies per country and continent from 1995 to 2023. **Implementation:** Utilized a countplot to visually represent the occurrences of earthquakes over the specified time frame.

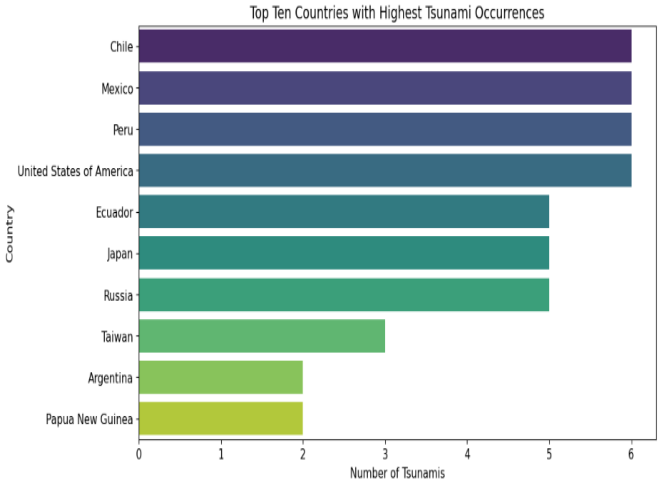
From the below count plot it is visible that within last 28 years earthquake occurs 33 times in Chile.



Barplot:

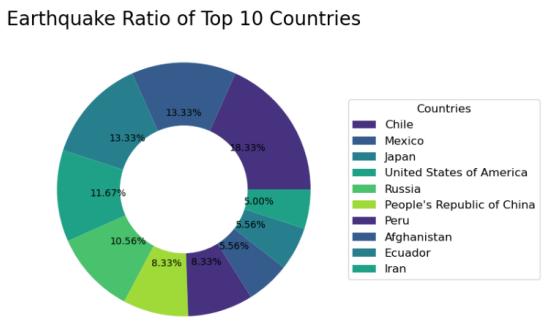
Objective: Visualize the frequency of tsunamis in the top 10 countries.

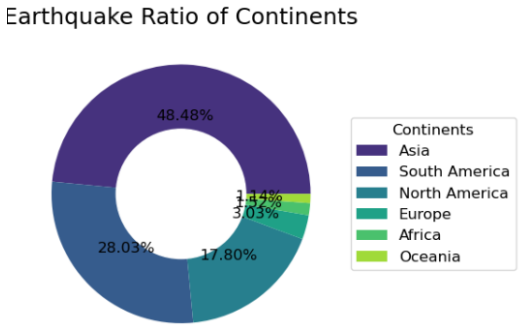
Implementation: Employed a barplot to display the occurrence of tsunamis in the selected countries.



Pieplot:

Objective: Demonstrate the ratio of earthquakes in the top ten countries and continents. **Implementation:** Utilized a pie plot to represent the distribution of earthquakes among the specified countries and continents.

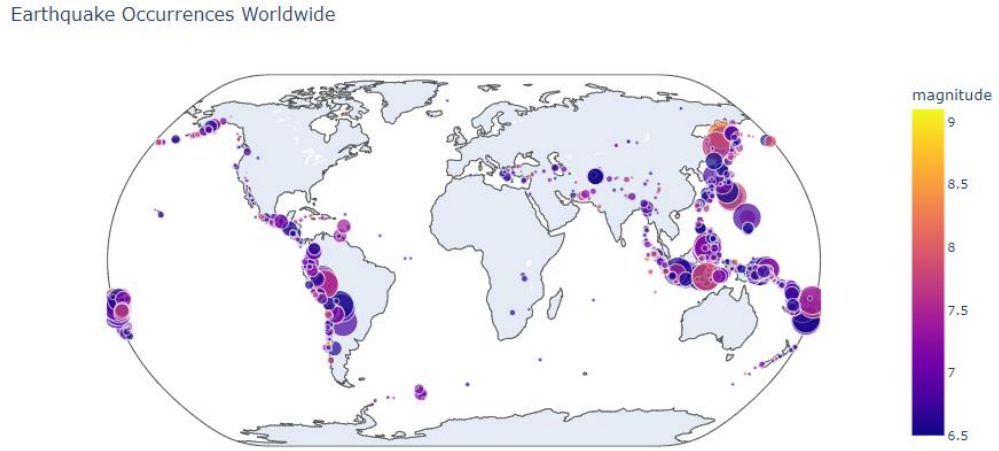




Geo Map:

Objective: Visualize global earthquake occurrences, with color indicating magnitude, size representing depth, and interactive features.

Implementation: Employed plotly's geomap for a dynamic map showcasing earthquake details upon mouse hover.

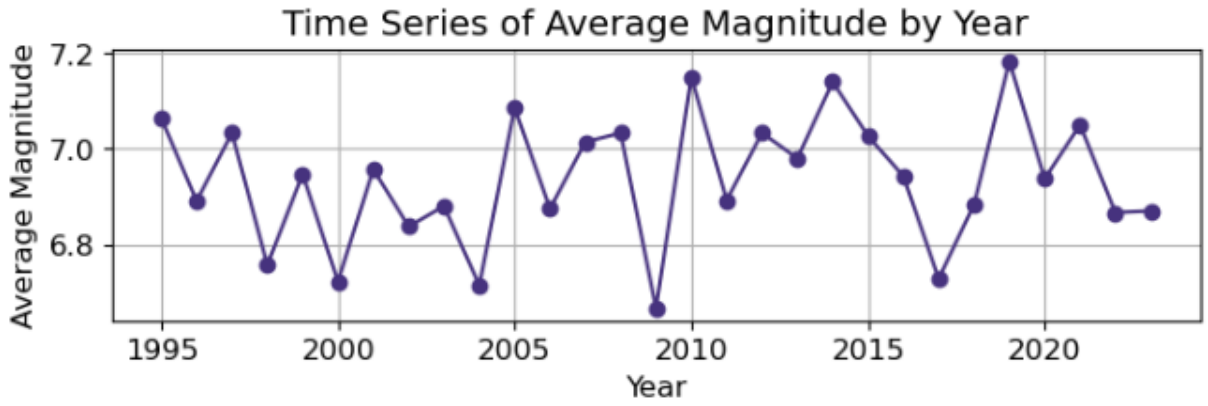


Time Series:

Occurrence of Earthquakes (1995-2023):

Objective: Illustrate the overall trend in earthquake occurrences over time.

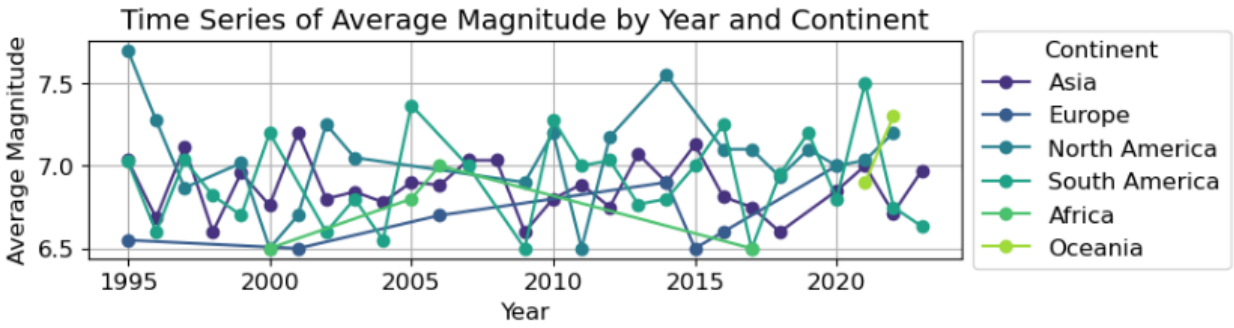
Implementation: Utilized a time series graph to visualize the frequency of earthquakes from 1995 to 2023.



Occurrence of Earthquakes in Top Ten Countries (1995-2023):

Objective: Showcase the temporal pattern of earthquakes in the top ten countries.

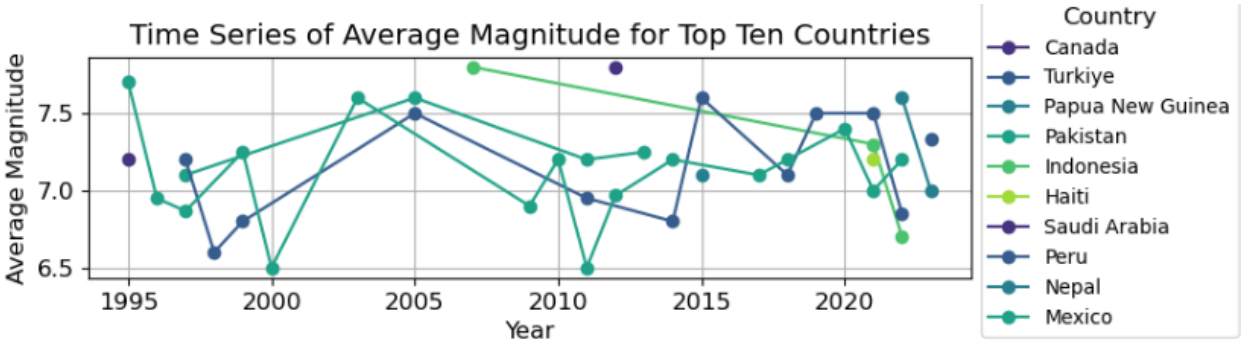
Implementation: Employed a time series graph to visualize earthquake occurrences in the selected countries over the specified period.



Occurrence of Earthquakes by Continents (1995-2023):

Objective: Demonstrate the temporal distribution of earthquakes across continents.

Implementation: Utilized a time series graph to represent the occurrence of earthquakes in continents from 1995 to 2023.



Model Development

Used logistic regression to predict the probability of earthquake and tsunami in a certain country in certain year. To fit the model, we selected the following columns as feature and target respectively.

Features = latitude, longitude, year, cdi, mmi, tsunami, depth

Target = magnitude

Then split the data in training and testing dataset.

Model Evaluation:

To evaluate the model we confusion matrix, accuracy, and classification report

Prediction

For prediction we input latitude and longitude, and date to predict the probability of earthquake and tsunami.

VI. RESULT AND DISCUSSION

The results obtained from the predictive model showcase its effectiveness in estimating earthquake probability in the input latitude and longitude The model's accuracy, precision, and recall rates demonstrate its potential for practical applications in earthquake risk assessment and preparedness. The identified key indicators and parameters shows that the model doesn't gives the result as expected. Therefore, we need to verify with other regression model for better performance.

VII. CHALLENGES

The development of a predictive model for earthquake probability posed several challenges, reflective of the complexities inherent in earthquake prediction and the limitations of available data. These challenges include:

Data Quality and Availability

The accuracy of the predictive model heavily relied on the quality and availability of historical earthquake data. Inconsistencies, gaps, or limited data coverage in certain regions posed challenges in building a comprehensive and representative dataset.

Model Generalization

Achieving a model that can generalize well to diverse geological and seismic conditions within presented difficulties. Ensuring the model's effectiveness across various terrains and geological features required careful consideration and validation.

Parameter Tuning

The optimization of model parameters to balance accuracy and overfitting proved to be a non-trivial task. Striking the right balance required extensive experimentation and iterative refinement.

Uncertainty in Earthquake Prediction

The inherent uncertainty associated with earthquake prediction posed a challenge in conveying the model's output accurately. Communicating the probabilities and uncertainties effectively to stakeholders, policymakers, and the public demanded a careful approach.

VIII. CONCLUSION

In conclusion, this research aimed to develop a predictive model for estimating earthquake probability in Country by identifying key indicators, collecting and analyzing historical earthquake data, and implementing a machine learning model. Despite the challenges encountered, the study yielded valuable insights and contributions to earthquake prediction research.

IX. REFERENCES

[1] Los Alamos National Laboratory, Geophysics Group: Builds on initial work from Paul Johnson. B. Rouet-Leduc Bertrand Rouet-Leduc, and Claudia Hulbert prepared the data for the research.

[2] PennState, Department of Geosciences: Data are from experiments performed by Paul Johnson, Prof. Chris Marone, Jacques Riviere, and Chas Bolton.

[3] Department of Energy, Geosciences and Biosciences Division, Office of Science, Chemical Sciences, Basic Energy Sciences: The Geosciences core research.

[4] Purdue University, Department of Physics and Astronomy: This stemmed from the DOE Council workshop “Information is in the Noise: Signatures of Evolving Fracture and Fracture Networks” held March 2018 that was organized by Prof. Laura J. Pyrak-Nolte.