Earthquake Forecasting Using Machine Learning Algorithm.

By

Ghanshyam Kumar Mahato

**Submitted in accordance with the requirements for the degree of**

MSc Computer Science with industrial placement.

York St John University

**School of science technology and health**

August 2024

`

Declaration

I, Ghanshyam Kumar Mahato, confirm that the work presented in this dissertation is my own and has not been submitted for any other degree or professional qualification. Where I have consulted the work of others, this is clearly stated.

I confirm that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Haroon Sihan, for his invaluable guidance, support, and encouragement throughout this research. His expertise and insights were instrumental in shaping this dissertation.

I am also grateful to York St John University for providing the resources and environment conducive to my research. Special thanks to the faculty members of the MSc Computer Science program for their academic support and mentorship.

I extend my appreciation to my colleagues and friends who have provided moral support and constructive feedback during this study. Your camaraderie and intellectual discussions have been immensely beneficial.

Lastly, I would like to thank my family for their unwavering support, patience, and understanding throughout this academic journey. Your belief in me has been a constant source of motivation.

Abstract

Earthquake prediction remains a critical yet challenging task in geophysics, with traditional methods often lacking precision. This study explores the potential of machine learning to enhance earthquake prediction by integrating geophysical and astronomical variables, such as the gravitational force between the Earth and the Moon, their distance, and localized gravity variations. Utilizing a comprehensive dataset spanning from 2011 to 2024, collected from the US Geological Survey (USGS) and augmented through web scraping for additional parameters, various machine learning models were evaluated. The Random Forest Regressor emerged as the most effective model, achieving an R² score of approximately 0.299, with a Mean Squared Error (MSE) of 0.126 and a Mean Absolute Error (MAE) of 0.2369. Significant predictive features were identified through Information Gain, ANOVA, and Lasso methods, indicating meaningful correlations with seismic activities. Despite the promising results, the moderate R² value suggests that additional factors and complexities are not fully captured by the current model. Future research should focus on integrating more diverse features, such as real-time seismic data, advanced geological metrics, and employing sophisticated machine learning techniques like convolutional and recurrent neural networks. Cross-regional studies are also necessary to validate model applicability across different seismic zones. This research underscores the potential of machine learning to improve earthquake prediction, advocating for continuous advancements in data integration and model sophistication.

Table of Contents

[Chapter 1 Introduction 1](#_Toc172801541)

[1.1 Context 1](#_Toc172801542)

[1.2 This Study 1](#_Toc172801543)

[1.3 The Value of Understanding 1](#_Toc172801544)

[1.3.1 Earth’s Gravity 1](#_Toc172801545)

[1.3.2 Earthquakes 2](#_Toc172801546)

[1.3.3 Lunar Distance 3](#_Toc172801547)

[1.4 Research Aim 3](#_Toc172801548)

[1.5 Research Objectives 3](#_Toc172801549)

[Chapter 2 Literature Review 4](#_Toc172801550)

[2.1 Introduction 4](#_Toc172801551)

[2.2 Machine learning approach: 5](#_Toc172801552)

[2.3 Tidal Forces and Earth's Response: 8](#_Toc172801553)

[2.4 Gravitational Variations on Earth: 9](#_Toc172801554)

[2.5 Moon - Earth distance: 10](#_Toc172801555)

[2.6 Conclusion: 10](#_Toc172801556)

[Chapter 3 Methodology 11](#_Toc172801557)

[3.1 Aim and Hypothesis 11](#_Toc172801558)

[3.2 Machine learning approach 11](#_Toc172801559)

[3.2.1 Data collection: 12](#_Toc172801560)

[3.2.1.1 Secondary data collection: 13](#_Toc172801561)

[3.2.1.2 Primary data collection: 14](#_Toc172801562)

[3.2.2 Exploratory Data Analysis (EDA): 17](#_Toc172801563)

[3.2.2.1 Shape of the dataset: 17](#_Toc172801564)

[3.2.2.2 Datatypes in the dataset: 17](#_Toc172801565)

[3.2.2.3 Features(columns) in the dataset: 18](#_Toc172801566)

[3.2.2.4 Adding features in the dataset: 18](#_Toc172801567)

[3.2.2.5 Missing values: 19](#_Toc172801568)

[3.2.2.6 Numerical Variables 21](#_Toc172801569)

[3.2.2.7 Use Log Transformation to form normal distribution 24](#_Toc172801570)

[3.2.2.8 Outliers: 25](#_Toc172801571)

[3.2.2.9 Categorical Features: 26](#_Toc172801572)

[3.2.2.10 Cardinality of Categorical Variable: 26](#_Toc172801573)

[3.2.2.11 Earthquake throughout the world by year: 28](#_Toc172801574)

[3.2.2.12 A kernel density estimates of magnitude, depth, distance by using pair plot and kind kde. 29](#_Toc172801575)

[3.2.2.13 Concentration of force, distance, gravity on higher magnitude 30](#_Toc172801576)

[3.2.2.14 Conclusion 32](#_Toc172801577)

[3.2.3 Feature engineering and feature selection: 32](#_Toc172801578)

[3.2.3.1 Handel Missing Value 32](#_Toc172801579)

[3.2.3.2 Removing outliers: 33](#_Toc172801580)

[3.2.3.3 Perform log transformation to make skewed data to form Gaussian distribution. 34](#_Toc172801581)

[3.2.3.4 Handling Categorical Feature into numerical Variable. 34](#_Toc172801582)

[3.2.3.5 Correlation: 35](#_Toc172801583)

[3.2.4 Model selection and Feature Selection 35](#_Toc172801584)

[3.2.4.1 Selected Features 36](#_Toc172801585)

[3.2.4.2 Model selection 37](#_Toc172801586)

[3.2.5 Model training and evaluation: 37](#_Toc172801587)

[3.2.6 Final Test: 38](#_Toc172801588)

[3.2.7 Conclusion: 39](#_Toc172801589)

[Chapter 4 Discussion and Future Work: 39](#_Toc172801590)

[Chapter 5 Conclusion 41](#_Toc172801591)

[Chapter 6 References 42](#_Toc172801592)

Table Of Figures

[Figure 1 Earth’s Gravity(mGal) (NASA Jet Propulsion Laboratory, July 30, 2003) 2](#_Toc172459028)

[Figure 2: The World’s Major Earthquakes from 1956‒2022. 2](#_Toc172459029)

[Figure 3 Distance of Moon from earth during Apogee and Perigee. 3](#_Toc172459030)

[Figure 4: The gravity expressed in mGal vs 30 December 2009 from Italy. 8](#_Toc172459031)

[Figure 5: Distance to Moon form Earth. 10](#_Toc172459032)

[Figure 6 Flowchart of approach 12](#_Toc172459033)

[Figure 7: Earth-Moon distance scraping method. 15](#_Toc172459034)

[Figure 8: Earth's gravity based on the location. 16](#_Toc172459035)

[Figure 9: shape of the dataset. 17](#_Toc172459036)

[Figure 10: Datatypes in the dataset. 18](#_Toc172459037)

[Figure 11: Final columns names. 18](#_Toc172459038)

[Figure 12: Time Date. 19](#_Toc172459039)

[Figure 13: Null Values across dataset in percentage. 19](#_Toc172459040)

[Figure 14: The missing value counts for various features. 20](#_Toc172459041)

[Figure 15: Numerical variables. 21](#_Toc172459042)

[Figure 16: The distribution of numerical features. 22](#_Toc172459043)

[Figure 17: The relationship between magnitude and various features. 24](#_Toc172459044)

[Figure 18: Outliers plots. 25](#_Toc172459045)

[Figure 19: 3.2.6 Cardinality of Categorical Variable. 27](#_Toc172459046)

[Figure 20: Earthquake by year. 28](#_Toc172459047)

[Figure 21: KDE pair plot. 29](#_Toc172459048)

[Figure 22: Concentration of magnitude towards center. 30](#_Toc172459049)

[Figure 23: The relationship between force, magnitude, and distance. 31](#_Toc172459050)

[Figure 24: Correlation heatmap. 31](#_Toc172459051)

[Figure 25: Outlier removing using IQR method 33](#_Toc172459052)

[Figure 26: log Transformation. 33](#_Toc172459053)

[Figure 27: One-Hot Encoding. 34](#_Toc172459054)

[Figure 28: Correlation. 34](#_Toc172459055)

List Of Tables

[Table 1 Feature Selection Results 48](#_Toc172186605)

[Table 2 Models and Results. 49](#_Toc172186606)

[Table 3 Final Test Result. 50](#_Toc172186607)

List Of Abbreviations

|  |  |
| --- | --- |
| **Acronym** | **Full Form** |
| USGS | United States Geological Survey |
| NEHRP | National Earthquake Hazards Reduction Program |
| CSV | Comma-Separated Values |
| SVM | Support Vector Machine |
| RMSE | Root Mean Square Error |
| MAE | Mean Absolute Error |
| SCEC | Southern California Earthquake Center |
| ANN | Artificial Neural Network |
| LSTM | Long Short-Term Memory |
| KDE | Kernel Density Estimate |
| R² | R-squared |
| MSE | Mean Squared Error |

# Introduction

## Context

Earthquakes are natural disasters that can cause significant destruction and loss of life, highlighting the importance of accurate forecasting to mitigate their impact. Traditional earthquake forecasting methods rely on historical data and seismic monitoring networks, but they often lack precision and reliability. In recent years, there has been growing interest in leveraging machine learning algorithms to improve earthquake forecasting capabilities (The Geological Society of London, 2012).

Machine learning offers the potential to analyze vast amounts of data and identify complex patterns that may be indicative of impending seismic activity. By integrating various factors such as geological features, seismic data, and environmental parameters, machine learning algorithms can enhance the accuracy and timeliness of earthquake forecasts (Environment, June 2021).

The current state of research in this field, discuss challenges and limitations of existing methods, and propose innovative approaches to improve earthquake prediction accuracy. Additionally, we will highlight the potential benefits of incorporating machine learning into existing forecasting frameworks and discuss the implications for disaster preparedness and response efforts. Through this research, we aim to contribute to the advancement of earthquake forecasting techniques and ultimately enhance our ability to mitigate the impact of seismic events on society and infrastructure.

## This Study

The variation in gravitational force experienced at different locations on the surface of the Earth. Gravity is the force of attraction between objects with mass, and its strength depends on the mass of the objects involved. Therefore, locations with more mass beneath them will experience a stronger gravitational force, while locations with less mass will experience a weaker gravitational force. This variation in gravitational force across different locations on Earth is due to differences in the distribution of mass beneath the Earth's surface. The prevailing technique involves analyzing historical earthquake data to identify patterns, relation between earth’s gravitational force between earth and moon, ununiform acceleration due to gravity. To enhance the efficiency of earthquake prediction models, we propose expanding the dataset by introducing additional features. These features would capture relevant factors such as the gravitational forces of the Moon, the non-uniform gravitational field of the Earth (geophysics, July 30, 2023 ).

## The Value of Understanding

### Earth’s Gravity

The observation of red and blue color variations in the Figure 1 indicates gravitational variations, with blue indicating lower gravity and red indicating greater gravity compared to the average gravitational acceleration of the Earth (approximately 9.8 m/s²).

This color variation likely represents a gravitational anomaly map, where regions with deviations from the average gravitational acceleration are highlighted. Blue areas would suggest areas where gravity is weaker than the average, while red areas indicate regions with stronger gravity than average.

Such gravitational anomaly maps are valuable in various fields such as geophysics, geology, and mineral exploration, as they provide insights into the underlying geological structures and density variations beneath the Earth's surface. These variations in gravity can be caused by factors such as variations in the density of rocks, mineral deposits, or tectonic features. (i.e.. 9.8m/s2) (NASA Jet Propulsion Laboratory, July 30, 2003).

A map of the world

Description automatically generated

Figure 1 Earth’s Gravity(mGal) (NASA Jet Propulsion Laboratory, July 30, 2003)

### Earthquakes

The Figure 2 displays dots representing earthquake events that have occurred at different locations. The density of dots indicates the frequency or concentration of earthquake events in those areas, with a higher density suggesting a greater occurrence of earthquakes in that region.

Additionally, the color scale ranging from blue to red represents the intensity or magnitude of the earthquakes, typically measured on the Richter scale. Blue hues indicate lower magnitudes, while red hues indicate higher magnitudes. Therefore, areas with a redder hue on the scale experienced earthquakes with greater magnitudes, potentially indicating regions of increased seismic activity or higher levels of tectonic stress.

In summary, the image provides a visual representation of both the frequency and magnitude of earthquake events, with denser clusters of dots and redder hue indicate areas of higher seismic activity and potentially greater earthquake magnitudes (Symington, Feb 2023).

A map of the world

Description automatically generated

Figure 2: The World’s Major Earthquakes from 1956‒2022 (Symington, Feb 2023).

### Lunar Distance

The Figure 3 illustrates the Moon's elliptical orbit around the Earth, with labels indicating the distances at apogee and perigee. Apogee refers to the point in the Moon's orbit when it is farthest from the Earth, while perigee is the point when it is closest to the Earth.

It's important to note that the Earth is not positioned exactly at the center of the Moon's orbit, as depicted in the image. Additionally, the eccentricity of the orbit has been exaggerated for illustrative purposes.

This image provides a visual representation of the Moon's orbit around the Earth, highlighting key points such as apogee and perigee, and demonstrating the elliptical nature of the Moon's path (Wibisono, 2018).

A diagram of the earth

Description automatically generated

Figure 3 Distance of Moon from earth during Apogee and Perigee (Wibisono, 2018).

## Research Aim

Earth's gravitational interaction with the Moon creates tidal forces that influence various geophysical processes on Earth, including tectonic activity and crustal deformation. By analyzing historical earthquake data in conjunction with gravitational dynamics between the Earth and the Moon, we seek to elucidate the relationship between mass distribution and seismic activity. Delving into the influence of lunar gravitational effects will provide a comprehensive understanding of their impact on earthquake occurrences, paving the way for more accurate seismic assessments. ML offers a powerful framework for analyzing complex datasets and identifying patterns that may not be apparent through conventional statistical approaches. By leveraging ML algorithms, we aim to enhance the accuracy and reliability of earthquake prediction models. Investigating the correlation between fluctuations in Earth's gravitational force and seismic events provides an opportunity to unravel the intricate connections between these factors, contributing to a deeper comprehension of earthquake dynamics and improving predictive models.

## Research Objectives

The objectives of this research are to investigate the impact of gravitational forces from both the Earth and the Moon on the occurrence of earthquakes. This involves studying variations in gravitational force influenced by the distribution of mass beneath the Earth's surface, gravitational interactions between the Earth and the Moon, non-uniform gravity on Earth's surface, and the distance between the Earth and Moon. By incorporating these factors, the research aims to enhance earthquake prediction models with new data, such as the gravitational forces of the Moon and the Earth's non-uniform gravitational field, seeking to develop more advanced and accurate prediction models.

Additionally, the study will identify significant predictive features by analyzing various potential elements within the dataset and assessing their relevance and impact on seismic forecasting. It will also evaluate the performance of different machine learning techniques in predicting earthquakes, aiming to identify the most effective algorithms for seismic prediction and compare their accuracy and efficiency.

An important aspect of the research is to explore the relationship between fluctuations in Earth's gravitational force and seismic events, along with the correlation between these fluctuations and variations in the Moon-Earth distance. This investigation will provide insights into the intricate connections between these factors and their impact on earthquake dynamics.

# Literature Review

## Introduction

This section explores several studies that employ hybrid machine learning models, seismic indicators, and historical earthquake data to forecast earthquake magnitudes and occurrences.

One notable study by Salam, Ibrahim, and Abdelminaam (2021) focuses on earthquake prediction models using hybrid machine learning techniques in Southern California. By utilizing seismic indicators and historical earthquake data from the Southern California Earthquake Center (SCEC), the researchers developed two models to predict earthquake magnitudes over fifteen-day periods. These models employ seven seismic indicators as inputs, which are processed using hybrid machine learning algorithms. Specifically, the Extreme Learning Machine (ELM) algorithm is optimized by Firefly Algorithm (FPA), and the Least Squares Support Vector Machine (LS-SVM) is optimized with FPA to enhance prediction accuracy. The performance of each model is evaluated using various metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and Percentage Mean Relative Error (PMRE). The results indicate that the FPA-LS-SVM model outperforms the FPA-ELM model across all evaluation criteria, demonstrating superior accuracy and reduced false alarm rates in earthquake prediction.

Furthermore, research surveys have highlighted the evolution of earthquake prediction methodologies, from early optimism in the 1970s to modern skepticism due to inconsistent results and the absence of statistically significant precursors (Geller, 1997; Galkina and Grafeeva, 2019). However, recent advancements in machine learning, such as the application of neural networks to analyze continuous seismic data, offer promising avenues for improved earthquake forecasting (Alexandridis et al., 2013). By uncovering previously unidentified signals within seismic data and understanding fault mechanics at a deeper level, machine learning techniques contribute to transformative progress in earthquake science (Zechar and Nadeau, 2012; Alves, 2006).

Moreover, studies in India have demonstrated the efficacy of supervised machine learning classifiers in earthquake type forecasting, with algorithms like Logistic Model Tree (LMT) and Simple Logistic exhibiting high accuracy rates (Kulkarni and Kulkarni, 2016; Debnath et al., 2021a). Additionally, the development of cross-region prediction models, such as SeisEML, represents a significant advancement in PGA prediction during earthquakes, surpassing conventional regression models in accuracy and reliability (Joshi et al., 2024; Debnath et al., 2021b).

The integration of machine learning techniques with seismic indicators and historical earthquake data holds immense potential for enhancing earthquake prediction accuracy and mitigating seismic hazards. Despite the inherent challenges and uncertainties in earthquake forecasting, ongoing research endeavors aim to leverage machine learning advancements to address one of the most complex challenges in geophysics.

## Machine learning approach:

Earthquake Prediction using Hybrid Machine Learning Techniques, researcher paper proposes two earthquake prediction models using seismic indicators and hybrid machine learning techniques in southern California. The Earthquake catalogue source was downloaded for free from the website (www.data.scec.org.). The historical earthquake data of Southern California between 1st January 1950 and 31st May 1978 is divided into 693 periods. Each period consists of fifteen days. Seven seismic indicators were calculated mathematically and statistically. These parameters are the inputs for the network to predict the future expected magnitude as the network output. Two hybrid machine-learning models are proposed to predict the earthquake magnitude for fifteen days. Then the ELM algorithm is optimized by FPA. FPA is used for the prediction of the occurrence of the Earthquake, and for optimizing LS-SVM with FPA to enhance the accuracy of earthquake magnitude prediction. The network architecture contains seven input indicators, which represent the seismic indicators, and the output shows the predicted magnitude during fifteen days. First, data is divided into 70% for training and 30% for testing. Then, data is divided into 80% for training and 20% for testing. At last, data is divided into 90% for training and 10% for testing. After processing data and introducing the indicators to the proposed models, the performance of each model is estimated using four performance evaluation criteria RMSE, MAE, SMAPE, and PMRE. The performance of the proposed FPA-LS-SVM outperformed the FPA-ELM model according to all compared criteria. Also, FPA-LS-SVM is the best in reducing the false alarm ratio in earthquake prediction (Salam, Ibrahim and Abdelminaam, 2021).

Research surveys suggest that there was once an optimistic outlook on the potential for early detection of earthquake hazards, which arose in the 1970s due to a few seemingly successful predictions. However, this optimism has since been supplanted by skepticism (Geller, 1997). This shift primarily occurred because of numerous instances of incorrect predictions, alongside the absence of statistically significant precursors (Galkina and Grafeeva, 2019). The study "Machine Learning Predicts Laboratory Earthquakes" suggests that predictions rely exclusively on the immediate physical properties of the acoustical signal, without considering its past. Machine learning discerns a previously disregarded signal emanating from the fault zone, initially perceived as low-amplitude noise. This newfound signal facilitates the forecasting of failure across the laboratory quake cycle. Additionally, it is inferred that the signal stems from ongoing grain motions within the fault gouge as fault blocks shift. Implementing this methodology with continuous seismic data could yield significant progress in uncovering signals presently unidentified, offering fresh insights into fault mechanics, and establishing limits on fault failure times.

The analysis of acoustic signals in laboratory settings, ML has demonstrated its capability to provide accurate forecasts of failure, independent of the slip cycle stage. Moreover, this approach has unveiled previously unidentified signals within the data, underscoring its potential for seismic signal analysis on a broader scale. This pioneering study marks the first instance of ML being applied to continuous acoustic/seismic data with the explicit aim of inferring failure times. By departing from traditional earthquake catalog-based analyses, which may overlook crucial factors, ML-based methodologies offer a promising avenue for comprehensive signal exploration. Notably, these approaches diminish human bias by autonomously exploring vast datasets, thereby enhancing the scope and depth of analysis (Alexandridis *et al.,* 2013). Transitioning from laboratory experiments to real-world applications presents a compelling challenge and opportunity. Analogous to fault patches exhibiting small repeating earthquakes, such as those observed near Parkfield on the San Andreas Fault, ML-based techniques may shed light on similar phenomena. Exploring the possibility of recording these signals through borehole and surface instruments presents an intriguing avenue for further investigation (Zechar and Nadeau, 2012). The broader implications of ML in seismic and geophysical research extend beyond earthquake prediction. By uncovering previously unknown signals and understanding fault physics at a deeper level, ML holds promise for a diverse range of applications, from industrial material failure prediction to natural hazard mitigation. At the nexus of advancements in instrumentation, ML algorithms, data handling capabilities, and computational power, seismic research stands poised for transformative progress. The convergence of these technologies sets the stage for significant breakthroughs in earthquake science, offering new insights, methodologies, and possibilities for addressing one of the most complex and pressing challenges in geophysics (Alves, 2006).

Analysis of Earthquake Forecasting in India Using Supervised Machine Learning Classifiers review the methodologies and findings of a study that employs ML algorithms to forecast earthquake types in various regions of India, focusing on datasets sourced from different states and neighboring countries. The study utilizes six distinct earthquake datasets sourced from a database, focusing on regions with high earthquake risk in India. Thirteen high-risk cities and fifteen medium-risk cities were identified for analysis. The datasets were categorized based on geographical regions, including Andaman & Nicobar, Gujarat, North India, Northeast India, Uttar Pradesh (UP), Bihar, and Nepal (Kulkarni and Kulkarni, 2016). The Weka tool, renowned for its capabilities in data mining and classification tasks, was employed to train and test ML models. Various performance metrics such as precision, recall, accuracy, F-measure, Matthews correlation coefficient (MCC), and confusion matrix were used to evaluate the models. For the Andaman & Nicobar dataset, the Simple Logistic and LMT methods exhibited the highest accuracy rate of 99.94%, followed by the Bayes Net, Random Forest, Random Tree, and Logistic Regression methods (Peng, Lee and Ingersoll, 2002). Conversely, the ZeroR method showed the lowest accuracy rate at 61.01%. The study aimed to forecast earthquake types to facilitate disaster management, with the Weka tool employed for this purpose. Seven supervised ML algorithms were compared, with accuracy rates serving as the primary evaluation metric. The study concludes that the Logistic Model Tree (LMT) and Simple Logistic algorithms are the most effective for forecasting earthquake impacts in India, based on the performance evaluation across different regions. These algorithms demonstrated high accuracy rates in predicting earthquake types, making them suitable candidates for earthquake forecasting systems. The findings underscore the potential of ML techniques in enhancing earthquake prediction and disaster preparedness efforts in seismic-prone regions (Debnath *et al.,* 2021a).

This research paper by (Joshi *et al.,* 2024) propose the development and application of a cross-region prediction model called SeisEML (Seismological Ensemble Machine Learning) for forecasting peak ground acceleration (PGA) during earthquakes. SeisEML integrates hybridized models, kernel-based algorithms, tree regression algorithms, and regression algorithms to enhance prediction accuracy. The review assesses SeisEML's performance, comparing it with conventional attenuation relations and ground motion prediction equations (GMPEs) across different seismic regions. SeisEML is trained and validated using datasets comprising 20,852 and 6,256 accelerograms from the Kyoshin Network in Japan (Si and Midorikawa, 2000). Various machine learning models and algorithms are evaluated based on statistical performance indicators, such as mean absolute error (MAE) and root mean square error (RMSE). The top-performing models are selected for integration into SeisEML. Additionally, the model's efficacy is tested on datasets from Iranian earthquakes to assess its cross-region predictive capability (Abrahamson and Litehiser, 1989). SeisEML demonstrates superior performance compared to conventional attenuation relations and GMPEs, yielding significantly lower MAE and RMSE values. Iso acceleration contour maps generated by SeisEML accurately depict PGA distributions during earthquakes, outperforming regional GMPEs. The model's effectiveness is further confirmed through cross-region predictions, particularly for earthquakes in similar tectonic environments. The comparisons highlight SeisEML's potential to enhance PGA prediction reliability across diverse seismic scenarios. SeisEML offers a robust machine learning approach for predicting PGA during earthquakes. By integrating various ML models and algorithms, SeisEML surpasses traditional regression models in accuracy and performance. The model's successful application in different seismic regions underscores its versatility and effectiveness in seismic hazard mapping. Ultimately, SeisEML represents a significant advancement in earthquake prediction, offering improved reliability and accuracy compared to conventional methods (Debnath *et al.,* 2021b).

Time series analysis methods have emerged as promising tools for earthquake prediction. This literature review focuses on predicting earthquake parameters in the Anatolian Peninsula using artificial neural network (ANN) methods. Specifically, an optimized Backpropagation Neural Network (BP-NN) model and a hyper-parameterized Long Short-Term Memory (LSTM) model are designed to predict earthquake magnitude, latitude, and longitude. The review evaluates the performance of these models against previous works using well-accepted metrics and explores the most significant contributing factors to earthquake occurrence (Emeç and Özcanhan, 2024). Earthquake data from Turkey spanning 1970 to 2019 is obtained from the United States Geological Survey. Feature extraction methods are employed to identify the most influential factors, with a focus on time, depth, and astronomical variables such as sun and moon distances to Earth. The data is then analyzed using the optimized BP-NN and LSTM models, with model configurations adjusted for optimal performance (Tan, Tapirdamaz and Yörük, 2008). Comparing the two models, the LSTM model demonstrates superior performance in earthquake magnitude prediction, achieving a mean squared error (MSE) of 0.062. However, while both models yield satisfactory results in latitude prediction, the BP-NN model outperforms the LSTM model with lower error rates. Notably, both models struggle with longitude prediction, indicating limitations in accurately pinpointing earthquake locations. Despite this, aligning predicted latitudes with fault lines provides valuable insights into potential earthquake locations (Yu *et al.,* 2019). The LSTM method offers improved accuracy in earthquake magnitude prediction compared to the BP-NN model. The BP-NN model performs better in predicting latitude, albeit with minor differences. Both models face challenges in accurately predicting earthquake locations, suggesting the need for further refinement. Future work will focus on enhancing latitude and longitude predictions by incorporating additional seismic inputs and geological precursors into the models. Overall, the study underscores the potential of time series analysis and ANN methods in earthquake prediction, highlighting avenues for future research and model improvement (Debnath *et al.,* 2021b).

Earthquake prediction research has focused on extracting input parameters from the temporal distribution of past seismic events to understand the frequency of seismic occurrences relative to their magnitudes. These parameters shed light on the underlying relationships between seismic activity and geophysical phenomena, such as seismic quiescence and foreshock frequency. Seismic quiescence, characterized by a decrease in seismic energy release from fault regions, may precede major earthquakes, with the accumulated energy often correlated with the magnitude of upcoming seismic events. Similarly, the frequency of foreshocks, earthquakes slightly higher in magnitude than background seismic activity, serves as a precursor to significant seismic events. The Gutenberg–Richter inverse power law provides insight into the relationship between earthquake magnitudes and the cumulative frequency of seismic events. Machine Learning (ML) and Artificial Neural Networks (ANN) have emerged as valuable tools in various fields, including computer vision, genetics, bioinformatics, and weather forecasting. Researchers have explored the application of ANN in modeling the nonlinear and complex relationships between geophysical factors and earthquakes, yielding meaningful results. This study aims to predict earthquakes with magnitudes of 5.5 and above in the Hindukush region monthly using ML approaches in conjunction with eight seismicity indicators. These indicators, derived from mathematical calculations based on past seismic events, provide insights into the seismic behavior of the region. ML techniques employed include Pattern Recognition Neural Network (PRNN), Recurrent Neural Network (RNN), random forest, and Linear Programming Boost (LPBoost) ensemble of decision trees. Comparative analysis of these techniques' prediction results provides valuable insights into earthquake forecasting in Hindukush. The eight seismic indicators used in this study represent the seismic state and potential of the ground. These indicators include parameters such as Time T, mean magnitude of past seismic events, square root of seismic energy release, Gutenberg–Richter b value, deviation from the Gutenberg–Richter inverse power law, and the difference between observed and expected earthquake magnitudes. Each indicator offers unique insights into seismic activity and earthquake potential. Neural networks, including PRNN and RNN, are employed to model the complex relationships between seismic indicators and earthquake occurrences. These networks utilize various transfer functions and training methodologies, such as Levenberg–Marquardt Backpropagation (LMBP), to optimize prediction accuracy. Decision trees, particularly in the form of random forest and LPBoost ensemble, are also utilized to enhance prediction performance through ensemble learning techniques. Evaluation of ML classifiers over unseen data reveals promising results, with each technique demonstrating varying levels of accuracy, sensitivity, specificity, and predictive value. While no single classifier emerges as superior in all aspects, each contributes valuable insights into earthquake prediction in Hindukush. These findings represent a significant step toward developing a robust earthquake prediction system, despite the inherently nonlinear and unpredictable nature of seismic phenomena. This study highlights the potential of ML techniques in earthquake prediction by leveraging mathematical seismic indicators and sophisticated modeling approaches. While further research is needed to refine prediction methodologies and address the inherent uncertainties of seismic activity, the results presented in this study offer promising insights into earthquake forecasting in the Hindukush region.

## Tidal Forces and Earth's Response:

The gravitational pull exerted by the moon and the Sun plays a crucial role in shaping Earth's tides, affecting not only the oceans but also the planet's crust and atmosphere. Recent studies, such as those by (Straser, 2010; Hagen and Azevedo, 2017) , found this result.

A graph showing the growth of the stock market

Description automatically generated

Figure 4: The gravity expressed in mGal vs 30 December 2009 from Italy(Straser, 2010).

The graph revels gravity monitoring data recorded on December 30, 2009. Gravity readings are in milligals (mGal) on the vertical axis, it portrays the time of oscillation This unique representation reverses the curve, aligning it with the fluctuations in water levels, which tend to rise concurrently with lower gravity levels. Along the horizontal axis, time is delineated from midnight, with intervals of approximately 28 minutes corresponding to 1000 oscillations of the oscillator. A line below signifies the average trend of these data points, while a line above illustrates the anticipated tidal pattern for the same date and location, specifically the Canal Porto of Venice Lido. The similarity between the observed and tidal patterns is evident. On December 30, 2009, the graph captures the measured variations in Earth's gravitational field, influenced by the same forces responsible for generating tides (Straser, 2010).

The gravitational pulsations occur with the moon's position. When the moon is at the Perigee, Earth intensifies gravitational forces. Peak force occurs during New (mostly) and Full Moons, while minimum force is observed during the First or Third Quarter. The most pronounced effects coincide with the highest tides, particularly impacting subduction zones along shorelines. It's noteworthy that the largest earthquakes tend to occur during these specific moon phases, suggesting a correlation between lunar gravitational influence and seismic activity (Hagen and Azevedo, 2017).

The interaction between the Moon and Earth results in gravitational oscillations, closely tied to tidal patterns and partially influencing geohazard events at subduction zones. Conversely, the interaction between the Sun and Earth involves electromagnetic variations, with the Sun possessing a magnetic field twice as intense as Earth's. Statistical analysis spanning from 1996 to 2016 yielded inconclusive findings regarding a direct correlation between sunspot maxima and heightened earthquake activity. Nevertheless, our future research will explore a notable increase in earthquakes during strong geomagnetic storms and delve into the occurrence of deeper seismic events (Hagen and Azevedo, 2017).

## Gravitational Variations on Earth:

The Earth's shape has long been a subject of profound significance for humanity, encompassing philosophical, religious, and practical dimensions, particularly for early seafarers. Today, advancements in technology enable us to determine the Earth's figure with remarkable precision, utilizing gravity measurements obtained from both space-based and terrestrial sources.

These measurements offer insights not only into the Earth's overall shape but also into local and temporal variations in gravity across the planet. Such variations provide valuable clues about the density and structure of the Earth's deep subsurface, as well as the dynamics of geodynamic processes occurring within it. Thus, our understanding of the Earth's shape continues to evolve, driven by advancements in scientific instrumentation and analytical techniques (Mölg and Kaser, 2021)

While Earth possesses an average gravity of 9.8, gravity on earth surface vary from 9.78 to 9.83 across different locations on the planet due to variations in mass. Gravity, a fundamental force of attraction between objects, is weaker for objects with smaller mass and stronger for those with larger mass. The reason you are bound to Earth's surface is its substantial gravitational force, although you also exert a gravitational force on Earth, albeit significantly smaller due to your lesser mass compared to Earth's. Gravity is quantified by the rate of acceleration between objects, with Earth's average gravitational pull measuring at 9.8 meters per second squared (m/s²). Earth's composition, comprising substances like air, rock, and water, contributes to this variation in gravitational forces owing to differences in density; for instance, rock boasts a higher density than air.

When celestial bodies like the Earth, the Moon, the Sun, and other planets draw closer together, the gravitational forces among them intensify. Consequently, plates heavier than others experience stronger gravitational effects. Despite the proximity of other planets, significant earthquakes occasionally fail to materialize, often attributed to the stable positioning of tectonic plates. However, earthquakes typically occur during such celestial alignments (Mahmud, 2019). The Moon orbits around the Earth, while both the Earth and other planets revolve around the Sun. Occasionally, as the other planets near Earth along their orbits around the Sun, the combined gravitational forces exerted on Earth intensify. This heightened gravitational influence can trigger increased motion among tectonic plates, leading to earthquakes.

Despite the close approach of other planets, there are instances when major earthquakes do not occur. This could be attributed to the stable positioning of Earth's tectonic plates. However, earthquakes typically do occur during such celestial alignments, with most earthquakes happening in regions where tectonic plates can readily shift, such as Indonesia, Peru, and Japan.

Earthquakes often strike suddenly and last only for a few seconds, making early warnings challenging. However, by monitoring the positions of planets, it may be possible to anticipate periods when larger earthquakes are more likely to occur, although predicting their exact locations remains elusive. Through ongoing research, we strive to gain a deeper understanding of the precise factors that contribute to earthquakes (Mahmud, 2019).

Firstly, Earth's shape is not perfectly spherical—it's slightly flattened at the poles and bulges out near the equator, causing points near the equator to be farther from the center of mass. Consequently, the gravitational force on an object is weaker at the equator compared to the poles, resulting in a reduction in gravitational acceleration of about 0.18%.

Secondly, Earth's rotation induces an apparent centrifugal force that opposes the gravitational force, particularly pronounced at the equator where it points directly opposite to gravity. This centrifugal effect, combined with the difference in center of mass distance, results in a further reduction in g of approximately 0.53% at the equator compared to the poles (h, 2024).

To obtain accurate gravity by any latitude we can use this formula:

A black text on a white background

Description automatically generated

Figure 5 Gravity formula (h, 2024)

here,

gpoles = 9.832 m/s2

g45 = 9.806 m/s2

gequator = 9.780 m/s2

ϕ = latitude, between −90° and +90°

This formula is derived from Newton's law of universal gravitation (Rosenfeld, 1965).

## Moon - Earth distance:

The moon's orbit around Earth isn't a perfect circle; it's slightly elliptical. This means that the distance between the Earth and the Moon varies throughout its orbit. When the Moon is closest to Earth, it's at the perigee (363,396 km), and when it's farthest away, it's at the apogee (405,504 km) (Rösch, 1972). This variation in distance can affect things like the apparent size of the Moon in the sky and the strength of tides (Agrawal, 2015).

A close up of a graph

Description automatically generated

Figure 6: Distance to Moon form Earth(Rösch, 1972).

The Moon, Earth's closest celestial neighbor, boasts a diameter of 3476 km and a mass of approximately 0.7 × 10^23 kg. Positioned at about 400,000 km from Earth, it possesses distinct characteristics such as low gravity, substantial temperature fluctuations, and a sparse atmosphere.

Regarding gravity, the Moon's total mass and radius dictate its gravitational acceleration. Consequently, gravitational acceleration on the Moon measures approximately 1.622 m/s², which is merely one-sixth of Earth's gravitational force (Höber, Taschner and Fimbinger, 2021).

## Conclusion:

Literature has shown that celestial bodies have a multifaceted influence on Earth, causing tides, earthquakes, moonquakes, and gravitational variations. Although great advances have been made toward understanding these intricate interactions, much further research is needed to fully dissect the mechanisms responsible. By explaining the relationship between Earth and the cosmos, the dynamics in geophysical processes of our planet and their place within our universe are brought closer to the scientists. We present significant improvements and potential of machine learning (ML) techniques in seismic hazard mitigation and earthquake prediction. From hybrid machine learning models to supervised classifiers and cross-region prediction models, researchers are using ML algorithms to find seismic indicators with remarkable levels of accuracy and reliability of historical earthquake data. The effectiveness of the prediction has been duly considered by the ML approaches noted for forecasting earthquake magnitudes, types, and peak ground acceleration across diverse seismic regions. The integration of several ML models and algorithms like Extreme Learning Machine (ELM), Least Squares Support Vector Machine (LS-SVM), and Logistic Model Tree (LMT) has further improved the prediction accuracy and reduced false alarm rates. The application of ML techniques for continuous seismic data analysis allows insight into the signals and fault mechanics that have never been observed before. This fact paves the way for a landmark progression in earthquake science. These advances not only better earthquake prediction but also help in the development of an improved understanding of seismic phenomena and fault dynamics. Despite the inherent problems and uncertainties in earthquake forecasting, this convergence of ML advancements, seismic indicators, and historical earthquake data offers almost unlimited potential to address one of the most critical problems in geophysics. Thus, by continuously refining the ML methodologies, integrating newer seismic inputs, and the exploration of novel modeling approaches, researchers work to increase the accuracy of earthquake predictions and improve disaster preparedness in potentially seismically active regions worldwide.

# Methodology

## Aim and Hypothesis

The Earth is 81 times heavier than the Moon, and their distances are, on average, 384,400 km; 31 times diameters of moon can be fitted in the diameter of the Earth. It thus goes without saying that such big differences between them regarding weight and size create gravitational forces between them, because of which various phenomena take place. One such phenomenon is the tidal effect: the Moon's gravitational pull gives rise to oceanic water bulging towards it, which leads to tides. On the other hand, latest studies have shown that the Earth's gravitational force can even stir up moonquakes. The idea is to investigate if the Earth and Moon masses play a role in causing earthquakes. In this regard, machine learning can be applied by analyzing various forms of data, such as seismic data with respect to earthquakes and data regarding the location and distance of the Earth and Moon to identify patterns or correlations existing within the given datasets for acquiring information regarding seismic activities and gravitational interaction. Models shall be developed for predicting probabilities of future seismic occurrence from historical seismic data, with variations in location, magnitude, and time. Identification and forecasting of earthquakes to be faced in the future have been tried by many scholars in these recent times from earthquake databases. Still, a lot of shortcomings shall have to be worked out by future research or maybe are unknown. Identify what exact factor to be determined which is critical in predicting earthquakes. Out of thousands of algorithms, many researchers use the two most common of these linear regression and classification. Researchers use these methods in determining factors like the earthquake intensity, category (natural or artificial), and even detecting any seismic waves considered due to human activity. To discover the relationship between earthquake magnitude and different factors that have a bearing on the actuality of earthquakes. Seismic data, geological characteristics, and other relevant variables will be analyzed to identify factors affecting the incidence of earthquakes and their magnitudes.

## Machine learning approach

This research is based on machine learning methods which include data collection, data visualization, feature engineering, model selection, training & testing and model evaluation.

Figure 7 Flowchart of approach

Start

Model Selection

And Training

Model Evaluation

Prediction, Results And

Refinement

End

Feature Engineering

Feature Engineering

Feature Selection

Feature

Selection

Data Collection

(Open Source, USGS)

Data Preprocessing

yes

no

no

yes

This flowchart provides a step-by-step visualization of the methodology, detailing each phase from initial data collection to final evaluation and conclusion.

### Data collection:

For collection data I have used several methods like web scraping, mathematical calculation.. etc. the data like earthquake events are easily available on many websites, the main challenge was to access earthquake dataset from trusted source. For the earthquake dataset I’ve used one of the trusted sources which is maintained by [USGS](https://www.usgs.gov/programs/earthquake-hazards) science for a changing world.

And data which were not available with the earthquake dataset for that I have done some mathematical calculation based on the time and data on which seismic events occurred.

The two main data collection methods I have used they are:

#### Secondary data collection:

The earthquake dataset is collected from [USGS](https://www.usgs.gov/programs/earthquake-hazards) science for a changing world (Program, 2024). The USGS takes on the responsibility of monitoring and reporting earthquakes, evaluating their impacts and associated hazards, and conducting focused research into the causes and consequences of seismic events. These endeavors are integral components of the broader National Earthquake Hazards Reduction Program ([NEHRP](https://www.nehrp.gov/)), a collaborative initiative established by Congress involving four agencies. From the website data downloaded from 2011 to 2024 Feb (link: <https://earthquake.usgs.gov/earthquakes/search/>). By entering basic parameters such as date, magnitude upper and lower limits, and area, 20,000 rows in CSV format can be downloaded per query. To gather all the necessary data, dates from 2011 to 2024 were repeatedly entered. After downloading the entire required dataset, all CSV files were merged into one and sorted by date. The resulting dataset includes earthquakes with magnitudes greater than 2.5 and covers the years 2011 to 2024. Additionally, it contains information about seismic events from around the world.

Columns catalog:

* date\_time: The "date\_time" denotes the date and time when the earthquake event occurred.
* latitude: The "latitude" represents the geographical latitude, measured in degrees, where the earthquake occurred.
* longitude: The "longitude" denotes the geographical longitude, measured in degrees, where the earthquake occurred..
* depth: The "depth" indicates the depth at which the earthquake occurred beneath the Earth's surface, measured in kilometers..
* magnitude: The "magnitude" represents the magnitude of the earthquake event.
* magnitude\_type: The "magnitude\_type" specifies the type of magnitude measurement employed, such as "ML" for local magnitude or "MW" for moment magnitude.
* nst: The "nst" denotes the number of seismic stations that contributed to the determination of the earthquake magnitude.
* gap: The "gap" indicates the azimuthal gap, measured in degrees, of the seismic stations used in determining the earthquake magnitude.
* depth\_min: The "depth\_min" represents the minimum distance, measured in degrees, to the nearest station that reported the earthquake.
* rms: The "rms" refers to the root mean square (RMS) of the residual travel time, measured in seconds, between observed and predicted arrivals of seismic waves.
* net: The "net" indicates the seismic network associated with the earthquake data.
* id: The "id" serves as a unique identifier for the earthquake event.
* updated\_date: The "updated\_date" denotes the time when the earthquake event data was last updated.
* place: The "place" provides a description or name of the location where the earthquake occurred.
* type: The "type" specifies the type of seismic event, such as "earthquake" or "explosion."
* horizontal\_error: The "horizontal\_error" represents the horizontal error associated with the location of the earthquake, measured in kilometers.
* depth\_error: The "depth\_error" denotes the depth error associated with the depth of the earthquake, measured in kilometers.
* magnitude\_error: The "magnitude\_error" indicates the magnitude error associated with the earthquake magnitude measurement.
* magnitude\_nst: The "magnitude\_nst" specifies the number of seismic stations utilized in determining the earthquake magnitude.
* status: The "status" denotes the status of the earthquake event data, such as "reviewed" or "automatic."
* location\_source: The "location\_source" indicates the source of location data for the earthquake event..
* magnitude\_source: The "magnitude\_source" specifies the source of magnitude data for the earthquake event.
* distance: The "distance" typically represents the distance between the center of the Earth and the Moon at the time when an earthquake occurred.
* force: The "force" refers to the gravitational force acting between the Earth and the Moon during an earthquake event.
* year: The "year" denotes the year in which the earthquake event occurred.
* month: The "month" indicates the month in which the earthquake event occurred.
* day: The "day" specifies the day of the month on which the earthquake event occurred.
* hours: The "hours" denotes the hour of the day at which the earthquake event occurred.
* minutes: The "minutes" specifies the minute of the hour at which the earthquake event occurred.
* day\_name: The "day\_name" provides the name of the day corresponding to the date of the earthquake event.

#### Primary data collection:

The information which are not available with the earthquake dataset is collected by different methods like web scraping. The data like distance, force, gravity are not available earthquake dataset. So let’s have a look how I’ve collected it.

##### web scraping:

Web scraping is a method with the help of which we can extract contents from the website. In this method I had to create a bot which was able to extract data and save in the list of arrays. Later it was merged with the main dataset.

The distance and gravity information web scraped based on time, date and location form other well-known and trusted website which I managed by [national geodetic survey](https://geodesy.noaa.gov/web_services/grav-d.shtml) (survey, 2024).

##### Scraping Distance between earth-moon:

I have web scraped the distance from the ‘https://www.nasa.gov/’ a bot which I created using python library like requests, concurrent. Futures. By passing out the date in the html link I was able to get distance to moon form earth. This distance is measured from center of both earth and moon in kilometers(km). And this was done for every label which is in the dataset. So, the distance which I have extracted is the distance between moon and earth when every earthquake events occurs.

A screenshot of a computer program

Description automatically generated

Figure 8: Earth-Moon distance scraping method.

The figure 6 illustrate how web scraping was done. Form the figure, the get\_distance() function is called using concurrent ThreadPoolExecutor(). Which calls the URL which is constructed using the https link and the date time of the earthquake event. When request is sent to the server it returned the distance between moon and earth.

##### Scraping Gravity of the earth based on location:

The [gravity of the earth](https://en.wikipedia.org/wiki/Gravity_of_Earth) varies based on the location. So, to get the actual gravity by the location where earthquake occurred, I have selected [National Geodetic Survey](https://geodesy.noaa.gov/) website to extract gravity of earth based on location. Like distance I have used web scraping method to fetch gravity. The below image illustrates how it was done.A screenshot of a computer program

Description automatically generated

Figure 9: Earth's gravity based on the location.

##### Gravitational Force between earth-moon calculation:

The force of the object that pulls other objects towards its center. Gravitational Force between two object is calculated multiplying mass of the objects with universal [gravitational constant](https://en.wikipedia.org/wiki/Gravitational_constant) and dividing the result by square of the distance. The result is measured in newton.

Mass of moon is (mm) = 7.35\*1022 kg

Mass of earth is (Me) = 5.9722\*1024 kg

Gravitational constant (G) = 6.67\*10-11 N.m2.kg-2

Force between them is calculated by using formula (Quinn T, 2014 Oct 13).

F = G\*(Me\*mm)/d2

Suppose the distance between the earth and moon is 380000 then the gravitational force action between earth and moon will be.

F = ((7.35 \* 1022) \* (5.977 \* 1024)\*( 6.67 \* 10-11))/ (380000 \*1000)2

F = 2.04 \* 1019 Newtons

A screenshot of a computer

Description automatically generatedForce for each label is calculated by shown method in the figure above.

### Exploratory Data Analysis (EDA):

Descriptive statistics and data visualization are crucial steps in machine learning research, providing valuable insights into the data that facilitate effective data preparation and feature engineering. Utilizing Python libraries such as NumPy, Matplotlib, and Seaborn enhances these processes significantly. NumPy supports large multi-dimensional arrays and offers a comprehensive collection of mathematical functions, making it indispensable for numerical computing. Matplotlib is a versatile plotting library that creates a wide range of static, animated, and interactive visualizations. Built on Matplotlib, Seaborn simplifies the creation of attractive and informative statistical graphics. Together, these libraries enable thorough data exploration, statistical summarization, and visualization, forming a solid foundation for advanced machine learning tasks.

#### Shape of the dataset:

The dataset consists of 18059rows and 28 columns (features). The number of columns may increase during the feature engineering and data preparation stages. Currently, the dataset's dimensions are (18059, 28).



Figure 10: shape of the dataset.

#### Datatypes in the dataset:

The dataset contains 28 columns, with 17 of them being float types and 11 being object types. The total memory occupied by the dataset is approximately 16 MB.

A screenshot of a computer

Description automatically generated

Figure 11: Datatypes in the dataset.

#### Features(columns) in the dataset:

For convenience in data visualization and selection, I have changed the column names to lowercase and corrected any spelling errors. This standardization enhances clarity and usability in subsequent data analysis tasks.

A screenshot of a computer

Description automatically generated

Figure 12: Final columns names.

#### Adding features in the dataset:

To better understand the relationship between time factors (year, month, day, hour, minutes) and earthquakes, I have created additional features by extracting these components using the `datetime` library in Python. This allows for a more detailed analysis of how these temporal elements correlate with seismic events.

A computer screen shot of a program

Description automatically generated

Figure 13: Time Date.

#### Missing values:

The provided code snippet and output illustrate an analysis of missing values in the earthquake dataset. The process includes two steps: identifying features with missing values and then printing the number and percentage of missing values for each feature.

A screenshot of a computer code

Description automatically generated

A graph of a bar graph

Description automatically generated with medium confidence

Figure 14: Null Values across dataset in percentage.

The plot illustrates the percentage of null values across various features in the earthquake dataset. Notably, the 'magnitude\_type' feature has the highest percentage of null values at 61.9%, followed by 'place' at 30.0%, 'depth\_error' at 32.9%, 'magnitude\_error' at 12.2%, and 'magnitude\_net' at 25.9%. Other features with significant null values include 'nst' at 23.5% and 'gap' at 1.3%. Several features, such as 'date\_time', 'latitude', 'longitude', 'depth', 'magnitude', and various time-related features, have 0% null values, indicating complete data in these columns.

A group of graphs with numbers

Description automatically generated with medium confidence

Figure 15: The missing value counts for various features.

The series of bar plots provide a visual representation of the missing value counts for various features in the earthquake dataset. Each plot displays the count of missing (indicated by '1') and non-missing (indicated by '0') values for a specific feature, with the y-axis representing the magnitude of the counts.

* nst: The plot shows a significant number of missing values, with over 11,000 instances missing. The count of non-missing values is lower, highlighting a substantial portion of the data is missing for this feature.
* gap: This feature has a relatively small number of missing values compared to non-missing values. The plot indicates that only a few hundred instances are missing.
* depth\_min: Like 'nst', this feature has a notable number of missing values, with over 4,000 instances missing. Most of the data for this feature is present.
* rms: The plot for 'rms' shows a low number of missing values, comparable to 'gap'. The majority of the data is non-missing.
* horizontal\_error: There is a significant number of missing values for this feature, with over 5,000 instances missing. The non-missing count is still higher but indicates considerable missing data.
* depth\_error: The missing values for 'depth\_error' are also notable, with over 2,000 instances missing. The plot indicates that while most data is present, a substantial portion is missing.
* magnitude\_error: This feature has a high count of missing values, with nearly 6,000 instances missing. The plot shows a significant disparity between missing and non-missing values.
* magnitude\_nst: Like 'magnitude\_error', this feature has a high count of missing values, with over 4,000 instances missing. The non-missing data count is higher but still indicates a considerable amount of missing data.

These plots collectively highlight the extent of missing data for critical features in the dataset, underscoring the need for careful handling of these missing values during the data preprocessing and analysis stages.

#### Numerical Variables

A screenshot of a computer

Description automatically generated

Figure 16: Numerical variables.

The total number of numerical variables identified is 21.

A group of red and orange columns

Description automatically generated with medium confidence

Figure 17: The distribution of numerical features.

The provided plot displays the distribution of various numerical features in the earthquake dataset using histograms.

* date\_time: The count of earthquake occurrences over time, showing a relatively even distribution with slight variations across different years.
* latitude: The distribution of earthquake events based on latitude, with a concentration around certain latitudes and a noticeable decline toward the extremes.
* longitude: The distribution based on longitude, showing multiple peaks, indicating more frequent occurrences at specific longitudes.
* depth: The depth (in kilometers) at which the earthquake occurred below the Earth's surface, with most occurrences at shallow depths and a steep decline as depth increases.
* magnitude: The distribution of earthquake magnitudes, showing a higher frequency of lower magnitudes and fewer occurrences of higher magnitudes.
* nst: The number of seismic stations that contributed to the earthquake magnitude determination, with a peak at lower values and a gradual decline.
* gap: The azimuthal gap (in degrees) of seismic stations used in determining earthquake magnitude, indicating more frequent occurrences at lower gap values.
* depth\_min: The minimum distance (in degrees) to the nearest station that reported the earthquake, mostly concentrated at lower values.
* rms: Root mean square (RMS) of the residual travel time (in seconds) between observed and predicted arrivals, showing a peak at lower values.
* horizontal\_error: Horizontal error (in kilometers) associated with the location of the earthquake, showing a symmetrical shape with most values around the center.
* depth\_error: Depth error (in kilometers) associated with the depth of the earthquake, concentrated at lower values.
* magnitude\_error: Magnitude error associated with the earthquake magnitude measurement, mostly concentrated at very low values.
* magnitude\_nst: The number of seismic stations used in determining the earthquake magnitude, showing a peak at lower values and a gradual decline.
* distance: The distance between the center of the Earth and Moon when an earthquake occurred, with several peaks indicating more frequent occurrences at specific distances.
* gravity: Gravity of Earth based on latitude and longitude, with a high concentration at certain values and fewer occurrences as the values deviate.
* force: The gravitational force acting between the Earth and Moon during an earthquake, showing multiple peaks indicating variations in the recorded force values.
* day: The day of the month distribution, showing a relatively even spread across all days.
* hours: Distribution of earthquake events based on the hour of occurrence, showing a relatively even spread.
* minutes: Distribution of earthquake events based on the minute of occurrence, showing a relatively even spread.

These histograms provide a visual understanding of how each numerical feature is distributed within the dataset, revealing insights into the characteristics and patterns of earthquake occurrences.

#### Use Log Transformation to form normal distribution

A screenshot of a computer program

Description automatically generated

The provided code snippet defines a function to filter and return continuous features from a DataFrame that are numeric, not datetime, and have only positive values.

A group of red and white graphs

Description automatically generated

Figure 18: The relationship between magnitude and various features.

The relationship between the 'magnitude' feature and various other features within the dataset. The features plotted against 'magnitude' include 'nst', 'gap', 'rms', 'horizontal\_error', 'distance', 'gravity', 'force', 'year', 'month', and 'day'.

* Magnitude vs. Magnitude: This plot serves as a reference, showing a perfect linear relationship, as expected.
* Magnitude vs. Nst: Displays a scattered distribution without a clear trend, indicating no obvious relationship.
* Magnitude vs. Gap: Shows a downward trend, suggesting a possible inverse relationship between 'gap' and 'magnitude'.
* Magnitude vs. RMS: Exhibits a clustered pattern with no strong correlation.
* Magnitude vs. Horizontal Error: Like 'rms', this plot does not show a clear trend, indicating no significant correlation.
* Magnitude vs. Distance: Presents a dispersed pattern, indicating a weak or no direct relationship.
* Magnitude vs. Gravity: Displays a densely packed scatter with little variation in 'gravity', suggesting minimal influence on 'magnitude'.
* Magnitude vs. Force: Shows scattered points without a clear trend, suggesting no strong relationship.
* Magnitude vs. Year: This plot shows vertical clusters corresponding to each year, indicating discrete values with no obvious trend over time.
* Magnitude vs. Month: Exhibits a scattered distribution, showing no clear relationship.
* Magnitude vs. Day: Shows points scattered across the days, with no apparent trend.

Overall, most scatter plots indicate a lack of strong linear relationships between 'magnitude' and the other features, except for a slight inverse trend observed in the 'gap' feature.

#### Outliers:

A group of boxes with numbers

Description automatically generated

Figure 19: Outliers plots.

The distribution of various features in the dataset. The features included are 'magnitude', 'nst', 'gap', 'rms', 'horizontal\_error', 'distance', 'gravity', 'force', 'year', 'month', and 'day’. Magnitude: The box plot shows that most of the magnitudes are clustered between 5.5 and 6.5, with a few outliers reaching up to around 8.5.

* Nst: This plot shows a wide range of values for 'nst', with many outliers beyond 200, indicating a highly variable dataset.
* Gap: Most 'gap' values are between 50 and 100, with outliers extending up to 300.
* Rms: The 'rms' values are mostly below 1, with a significant number of outliers reaching up to 2.5.
* Horizontal Error: This feature has values mostly between 7 and 10, with outliers extending beyond 15.
* Distance: The 'distance' values are distributed over a wide range with no apparent outliers, indicating a more uniform distribution.
* Gravity: The 'gravity' values are tightly clustered around 9.79, with a few outliers on the higher side.
* Force: Most 'force' values are between 1.8e20 and 2.1e20, with outliers extending beyond 2.3e20.
* Year: The values for 'year' range from 2012 to 2024, showing a uniform distribution without any outliers.
* Month: This feature shows a uniform distribution across all months.
* Day: The 'day' values are uniformly distributed across the days of the month.

Overall, these box plots indicate that some features like 'nst', 'gap', 'rms', 'horizontal\_error', and 'force' have significant outliers, while others like 'distance', 'year', 'month', and 'day' show more uniform distributions.

#### Categorical Features:

These features are likely to contain string values or other non-numeric data that represent different categories or identifiers within the dataset.

A screenshot of a computer

Description automatically generated

#### Cardinality of Categorical Variable:

From the code, the features that have between 2 and 24 unique categories are added to the categorical\_feature list. These features are:

* magnitude\_type (15 categories)
* net (14 categories)
* magnitude\_source (24 categories)
* day\_name (7 categories)

A screenshot of a computer

Description automatically generated

A group of graphs with numbers

Description automatically generated with medium confidence

Figure 20: 3.2.6 Cardinality of Categorical Variable.

The bar plots for four categorical features ('magnitude\_type', 'net', 'magnitude\_source', and 'day\_name') showing the distribution of the number of occurrences for each category in the dataset.

* Magnitude Type: The categories are relatively balanced, with each type having a similar count.
* Net: There is a significant variation in the number of occurrences across different networks, with some networks having a much higher count than others.
* Magnitude Source: Like the 'net' feature, there is variation in the counts, but it appears to be more balanced compared to the 'net' feature.
* Day Name: The counts are evenly distributed across the days, indicating no particular day has significantly more or fewer occurrences.

These bar plots help visualize the frequency distribution of the categories within each of these features, providing insights into the data's structure and helping identify any potential biases or patterns.

#### Earthquake throughout the world by year:

Here the global map displays earthquake occurrences from 2011 to 2024. Earthquakes are represented by dots color-coded according to their magnitude, with a scale ranging from blue for lower magnitudes to red for higher magnitudes (2 to 9). The map highlights significant seismic activity along tectonic plate boundaries, especially around the Pacific Ring of Fire, where dense clusters of blue dots indicate frequent smaller earthquakes. Higher magnitude events, shown in red and yellow, appear less frequently but are prominently marked. An interactive timeline at the bottom allows for the visualization of earthquake activity over the specified years, offering a dynamic view of the temporal and geographical distribution of these events. This visualization effectively illustrates global seismic patterns and trends over time..

A screenshot of a computer screen

Description automatically generated

Figure 21: Earthquake by year.

#### A kernel density estimates of magnitude, depth, distance by using pair plot and kind kde.

A screenshot of a graph

Description automatically generated

Figure 22: KDE pair plot.

The pair plot shows a comprehensive visualization of the relationships between multiple numerical features in the earthquake dataset.

Key observations include that most earthquakes occur at shallow depths, as indicated by the KDE plot for `depth`. The `distance` feature shows a broad distribution, indicating variability in the distances at which earthquakes are measured. The `force` and `gravity` features display narrower distributions. Pairwise relationships, such as between `magnitude` and `depth` or `distance`, reveal how these features interact, with denser contours highlighting areas with a higher concentration of data points.

#### Concentration of force, distance, gravity on higher magnitude

A line graph with red dots

Description automatically generated

Figure 23: Concentration of magnitude towards center.

Figure 22 illustrates the concentration of earthquake magnitudes in relation to force, distance, and gravity. Each plot shows magnitude on the x-axis and one of the features on the y-axis, with the line representing the average trend and the shaded area indicating variability, likely reflecting confidence intervals.

In the first plot, force exhibits significant fluctuations at lower magnitudes (around 5 to 6), with a slight decrease in variability as magnitude increases. Notably, there is a spike in force around magnitude 8. The second plot shows the relationship between magnitude and distance, revealing considerable variability at lower magnitudes. As magnitude increases, the distance stabilizes but still shows some fluctuations, with a notable increase around magnitude 8. The third plot, depicting magnitude versus gravity, indicates less fluctuation compared to the other two features. The trend remains relatively stable, though there is a dip in gravity around magnitude 8, followed by an increase. We can clearly see that after magnitude 6 and above, there is a concentration of magnitudes in the center, indicating that force, gravity, and distance significantly affect earthquake magnitude..

A group of red and pink dots

Description automatically generated

Figure 24: The relationship between force, magnitude, and distance.

The relationships between gravity and three other features: force, distance, and magnitude. In the first plot, gravity values are plotted against force, showing that gravity remains tightly clustered around 9.78 to 9.84 across the range of force values, indicating no strong relationship between these variables. The second plot displays gravity against distance, again demonstrating a consistent clustering of gravity values within the same range, regardless of the distance, suggesting that distance does not significantly affect gravity. The third plot shows gravity versus magnitude, where gravity values remain stable across different magnitudes, though there is a slight trend where lower magnitudes show more spread in gravity values, while higher magnitudes exhibit a more compact distribution. Overall, these scatter plots indicate that gravity values are stable and do not have strong relationships with force, distance, or magnitude, remaining consistently clustered between 9.78 and 9.84 across these features.

A screenshot of a graph

Description automatically generated

Figure 25: Correlation heatmap.

The heatmap provides correlations between various features in the earthquake dataset. Each cell displays the Pearson correlation coefficient between two features, with the color intensity indicating the strength and direction of the correlation. Positive correlations are represented by shades of purple and pink, while negative correlations are shown in shades of orange and red.

Key observations include that magnitude is positively correlated with rms (0.23), indicating that higher magnitudes tend to have higher rms values, and negatively correlated with gap (-0.11) and distance (-0.12), suggesting that these values tend to decrease as magnitude increases. Depth is negatively correlated with latitude (-0.22) and depth\_error (-0.18), but positively correlated with magnitude\_error (0.14). Latitude and longitude are positively correlated with each other (0.22) and latitude is negatively correlated with depth\_error (-0.4), indicating that higher latitudes are associated with lower depth errors. Nst shows a positive correlation with gap (0.19) and a slight positive correlation with magnitude (0.13). Gap is negatively correlated with depth (-0.21) and magnitude\_error (-0.13), suggesting that larger gaps are associated with shallower depths and lower magnitude errors. RMS has a positive correlation with magnitude (0.23) and a negative correlation with depth\_min (-0.25), indicating higher RMS values are associated with higher magnitudes and lower minimum depths. Horizontal\_error and depth\_error are strongly correlated (0.53), indicating that higher horizontal errors are associated with higher depth errors. Gravity shows minimal correlation with most other features, indicating its values remain relatively stable regardless of changes in other features.

#### Conclusion

The exploratory data analysis provides valuable insights into the structure and characteristics of the earthquake dataset. The dataset is extensive, with a mix of numerical and categorical features, some of which have significant missing values that need to be addressed during preprocessing. The correlation analysis highlights important relationships between features, which can inform feature selection and engineering steps for modeling. Visualizations reveal patterns and outliers that can impact the performance of machine learning models.

### Feature engineering and feature selection:

Feature engineering is an important step in machine learning

#### Handel Missing Value

Main aim features of this research project are “magnitude, distance between earth and moon during earthquake, gravitational force between earth and moon, gravity so I’m keening only these features. Other features are not the factor they are the methods and results.

Dealing with null entries in the dataset:

There are 9 features

A screenshot of a computer

Description automatically generated

#### Removing outliers:

Removing the outliers in the 'magnitude' column by first calculating the IQR and then determining the thresholds for outliers. It then prints these thresholds and filters the Data Frame to remove values outside a specific range.

A screenshot of a computer program

Description automatically generated

Figure 26: Outlier removing using IQR method

#### Perform log transformation to make skewed data to form Gaussian distribution.

A screenshot of a computer

Description automatically generated

Figure 27: log Transformation.

Log transformation is used to transform skewed data into a more Gaussian-like (normal) distribution. This is beneficial for many statistical techniques and machine learning algorithms that assume normality of the data.

By using np.log (), we ensure that all numerical features, even those with zero or very small values, are transformed without causing any mathematical errors, resulting in a more normal distribution which is suitable for further analysis or modeling.

#### Handling Categorical Feature into numerical Variable.

A screenshot of a computer code

Description automatically generated

Figure 28: One-Hot Encoding.

pd.get\_dummies is used to one-hot encode the magnitude type column. The original magnitude type column is dropped, and the resulting dummy variables are concatenated back to the Data Frame.



And later these data set is saved after saving 20 % of data set for final model evaluation.

#### Correlation:

A graph of numbers and a number

Description automatically generated with medium confidence

Figure 29: Correlation.

The pairwise relationships of the different attributes in the earthquake dataset. Each cell in a matrix holds the correlation coefficient for two variables. The coefficient of correlation is a value that can either be -1 or 1. If the coefficient of correlation is 1, then it is a positive correlation that indicates both variables would rise and fall together. On the other hand, a coefficient of -1 shows a perfect negative correlation: as one increases, the other decreases. The value is 0 in case there is no linear relation between the variables.

### Model selection and Feature Selection

Feature selection is crucial in building efficient and effective machine learning models. Different methods can be used to select the most important features, and here we'll discuss three methods: Information Gain, ANOVA (F-Test), and Lasso (using SelectFromModel). The table below summarizes the results from these methods and highlights the selected features.

Table 1 Feature Selection Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Information Gain | ANOVA  (F-Score) | ANOVA  (P-Value) | SelectFromModel  (Lasso) |
| mb | 0.136325 | 2832.808002 | 0 | Yes |
| mww | 0.109843 | 3246.872322 | 0 | Yes |
| depth | 0.094063 | 70.707019 | 0 | Yes |
| longitude | 0.030101 | 0.958198 | 0.3277 | Yes |
| latitude | 0.029206 | 0 | 0 | Yes |
| force | 0.02421 | 1.125785 | 0.2887 | No |
| distance | 0.023901 | 1.125785 | 0.2887 | No |
| mwc | 0.01318 | 71.856381 | 0 | Yes |
| mwb | 0.009684 | 0 | 0 | No |
| hour | 0.006936 | 0 | 0 | No |
| gravity | 0.006159 | 1.876483 | 0.1708 | No |
| month | 0.005488 | 0 | 0 | Yes |
| minutes | 0.005356 | 0 | 0 | Yes |
| year | 0.003155 | 0 | 0 | Yes |
| rare\_mt | 0.000802 | 1.830741 | 0.1761 | No |
| day | 0.000079 | 0 | 0 | Yes |
| mwr | 0 | 41.283737 | 0 | No |

Information gain measures the reduction in entropy or uncertainty from transforming a dataset in some way. In the context of regression, it indicates the importance of each feature by how much they reduce the uncertainty in predicting the target variable.

In this dataset, features like `mb`, `mww`, and `depth` have the highest information gain, indicating they are the most informative.

The ANOVA F-test measures the linear dependency between the feature and the target variable. It computes the F-score for each feature, which represents the ratio of explained variance to unexplained variance.

Features with higher F-scores and lower p-values (below a threshold like 0.05) are considered more significant. Here, features such as `mww`, `mb`, `mwc`, `depth`, and `mwr` are highly significant.

The SelectFromModel method uses a model (in this case, Lasso regression) to select features. Lasso (Least Absolute Shrinkage and Selection Operator) adds a penalty equal to the absolute value of the magnitude of coefficients to the regression model. It can shrink some coefficients to zero, effectively selecting a subset of features.

Features selected by Lasso include `latitude`, `longitude`, `depth`, `year`, `month`, `day`, `minutes`, `mb`, `mwc`, and `mww`.

#### Selected Features

* latitude and longitude: Consistently selected by SelectFromModel (Lasso), capturing spatial information.
* depth: Highly rated across Information Gain, ANOVA, and SelectFromModel, indicating its critical role.
* distance, gravity, and force: Included as main features due to their importance in the research context, even though their statistical significance varies.
* mb, mwb, mwc, mwr, mww: Selected due to their consistent importance across different methods, highlighting their relevance in earthquake data.
* rare\_mt: Included due to its categorical nature and potential impact as indicated by one of the methods.

#### Model selection

The target variable 'magnitude' is continuous, regression methods will be applied to predict earthquake magnitudes. The popular and effective machine learning regression models, including Random Forest, Gradient Boosting Regressor, and Artificial Neural Networks are the most suitable model for these types of datasets.

##### Random Forest:

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mean prediction for regression tasks. It is known for its robustness and ability to handle many features. The hyperparameters tuned for Random Forest include the number of trees (estimators) and the maximum depth of each tree.

##### Gradient Boosting Regressor:

Gradient Boosting is another ensemble technique that builds trees sequentially, each attempting to correct the errors of the previous one. It optimizes a loss function over function space using gradient descent. The primary hyperparameters tuned for this model are the learning rate, number of estimators, and the maximum depth of trees.

##### Support Vector Regressor (SVR):

It tries to fit the error within a certain threshold, making it robust to outliers. Key hyperparameters for SVR include the kernel type, regularization parameter (C), and epsilon, which defines the margin of tolerance.

##### Artificial Neural Networks (ANN):

ANNs are computing systems inspired by the biological neural networks that constitute animal brains. They can capture complex patterns in data through layers of interconnected neurons. For ANN, the hyperparameters tuned include the number of layers, number of neurons per layer, activation functions, and learning rate.

### Model training and evaluation:

The process of model training and evaluation begins with importing all the necessary libraries. Next, various models are evaluated based on hyperparameter tuning, including Random Forest, Gradient Boosting Regressor, SVR, and ANN. After evaluation, the best parameters are selected for each model. The model with the optimal parameters is then trained using the training dataset. Subsequently, the trained model is tested with a separate test dataset, which was kept aside from the training set to ensure unbiased evaluation. The performance of the model is assessed by comparing the best parameters, MSE, MAE, RMSE and R² values. Finally, the results are concluded based on these evaluations.

Table 2 Models and Results.

|  |  |  |  |
| --- | --- | --- | --- |
| Models | R2 score | Mean Square Error | Mean Absolute Error |
| Random Forest | 0.273 | 0.128 | 0.241 |
| Gradient Boosting | 0.269 | 0.129 | 0.242 |
| Support Vector Regressor | 0.176 | 0.145 | 0.224 |
| Artificial Neural Networks | 0.181 | 0.144 | 0.267 |

The Random Forest model demonstrates the best overall performance, with the highest R² score of 0.273, indicating it explains the largest proportion of variance in the dependent variable. Additionally, it has the lowest MSE of 0.128, reflecting the smallest average squared errors among the models. Although its MAE of 0.241 is slightly higher than that of the Support Vector Regressor, the Random Forest's superior R² score and MSE make it the most effective model in this comparison.

The Gradient Boosting model follows closely with an R² score of 0.269 and an MSE of 0.129. Its MAE is 0.242, very similar to that of the Random Forest. While Gradient Boosting's metrics are competitive, it falls just short of the Random Forest in terms of explaining variance and minimizing squared errors.

The Support Vector Regressor has a lower R² score of 0.176 and a higher MSE of 0.145 compared to Random Forest and Gradient Boosting. However, it achieves the lowest MAE of 0.224, indicating it performs better in terms of average absolute errors. Despite this, its overall performance is less favorable due to the lower R² score and higher MSE.

The Artificial Neural Network model shows an R² score of 0.181 and an MSE of 0.144, both of which are lower than the top-performing models. Additionally, it has the highest MAE of 0.267, indicating larger average absolute errors. These metrics suggest that the Artificial Neural Network is the least effective model among the ones compared.

The Random Forest model stands out as the best performing model due to its highest R² score and lowest MSE, making it the most effective at explaining variance and minimizing prediction errors. GBR is a close second, while the SVR and ANN lag in overall performance.

### Final Test:

The final evaluation of the Random Forest Regressor model, trained on 90%, 80%, 70% of the data and tested on the remaining 10, 20,30 respectively.

Table 3 Final Test Result.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | MSE | MAE | R² |
| 70 training - 30 testing | 0.126 | 0.235 | 0.299 |
| 80 training - 20 testing | 0.126 | 0.245 | 0.299 |
| 90 training - 10 testing | 0.126 | 0.235 | 0.299 |

The R² score of 0.2991 indicates that the model explains approximately 29.91% of the variance in the dependent variable. While this score is not exceedingly high, it demonstrates a moderate ability of the model to capture the underlying patterns in the data. The MSE of 0.1261 reflects the average of the squared differences between the predicted and actual values, suggesting that the model's predictions are relatively close to the true values, with a lower MSE indicating better performance. MAE of 0.2369 represents the average magnitude of the prediction errors, showing that the model's predictions deviate from the actual values by about 0.2369 units on average.

Overall, the Random Forest Regressor, optimized with a maximum depth of 20, 300 estimators, and other fine-tuned hyperparameters, demonstrates a balanced performance with moderate explanatory power and low prediction errors. These results underscore the model's effectiveness and robustness in capturing the relationships within the dataset, making it a suitable choice for this regression task. The combination of these performance metrics highlights the model's ability to provide reasonably accurate and reliable predictions, making it a valuable tool for earthquake prediction.

### Conclusion:

The study successfully identifies the most significant features influencing earthquake occurrences and determined that the Random Forest model is the best-performing regression model for predicting earthquakes. These findings contribute to the understanding of seismic activities and offer a reliable tool for future earthquake prediction, thereby aiding in disaster preparedness and mitigation efforts. The integration of various feature selection methods and comprehensive model evaluation ensures a robust and effective approach to earthquake prediction using machine learning techniques.

# Discussion and Future Work:

The application of machine learning in earthquake prediction is determining key factors that could be influencing the development of seismic activity and in the predictive modeling. The Random Forest Regressor model turned out to perform well, making it more robust and reliable for complex datasets in capturing latent patterns. The model explains approximately 29.91% variance in the dependent variable. Having MSE 0.126 and MAE 0.2369, it still points some moderate explanatory power with quite low prediction error. Some of the added features are gravitational force, the distance between Earth and Moon, and gravity according to earthquake epicenter. They shade some valuable insight into the relation of these factors with earthquake occurrence. This underscores the importance of integrating different geophysical and astronomical variables into prediction models for earthquakes. The results concur the effect of gravitational forces and celestial alignments on the activities that lead to seismic events, hence accentuating that there may indeed be a need for integration into predictive frameworks. This process really makes feature selection relevant, using enhancement with the performance of the model through a set of techniques: Information Gain, ANOVA, and Lasso. A thorough exploratory data analysis creates a solid base from which to understand the structure and features of data, to guide further steps in modeling. Though the results are positive, the relatively low R² value indicates that there might be more variables and issues not captured by the model specified here.

The integration of more diverse and holistic features such as real-time seismic data, geospatial information, and advanced geological metrics into the model can be very helpful in improving the model accuracy and comprehensive understanding of earthquake dynamics. Use of advanced machine learning techniques like convolutional neural networks and recurrent neural networks will be useful to capture complex and intricate patterns in temporal dependencies inside seismic data. This would further improve the ability to apply earthquake forecasting systems in practice by developing real-time prediction models and constantly updating input on seismic and geophysical data. It presupposes robust integration of data as well as real-time processing to receive timely and accurate predictions.

Cross-regional studies that ensure the validation of applicability within different seismic zones and geographical contexts will bound to help foster generalization of the findings and assure the reliability of the model in diverse scenarios. Seismologists, researchers, and governmental agencies shall produce datasets and predictive models with better comprehensiveness and accuracy, a result of the sharing of data and insights. This can pave the road toward the standard methodology of earthquake prediction.

Adding more geological and environmental variables, including soil composition and fault line characteristics, to say weather conditions, will contribute to a more descriptive understanding of leading conditions in the occurrence of an earthquake. Such a holistic approach will improve predictability in earthquake forecasting models.

# Conclusion

The application of machine learning to earthquake prediction has demonstrated promising results in understanding and forecasting seismic events. By integrating various features such as the gravitational force between the Earth and the Moon, the distance between them, and localized gravity variations, this research has highlighted the potential of machine learning algorithms to enhance earthquake prediction models.

The study identified the Random Forest Regressor as the most effective model, achieving an R² score of approximately 0.299, with a MSE of 0.126 and a MAE of 0.2369. These metrics indicate a moderate explanatory power and low prediction errors, making the Random Forest model a reliable tool for predicting earthquake magnitudes.

Significant features influencing earthquake occurrences were identified through various feature selection methods, including Information Gain, ANOVA, and Lasso. These features were found to have meaningful correlations with seismic activities, underscoring the importance of incorporating geophysical and astronomical variables into earthquake prediction models.

Despite the positive outcomes, the relatively low R² value suggests that there are additional factors and complexities not captured by the current model. Future research should focus on integrating more diverse and holistic features, such as real-time seismic data, geospatial information, and advanced geological metrics. The use of more sophisticated machine learning techniques, like convolutional and recurrent neural networks, could further enhance the ability to capture complex temporal dependencies in seismic data.

Furthermore, cross-regional studies are necessary to validate the applicability of these models in different seismic zones and geographical contexts. Collaboration between seismologists, researchers, and governmental agencies in sharing data and insights can lead to the development of standardized methodologies for earthquake prediction.

In conclusion, while this research has made significant strides in leveraging machine learning for earthquake forecasting, continuous improvements and expansions in data integration and model sophistication are essential for achieving more accurate and comprehensive earthquake prediction systems.

# References

Abrahamson, N.A. and Litehiser, J.J. (1989) 'Attenuation of vertical peak acceleration', Bulletin of the Seismological Society of America, 79(3), pp. 549-580.

Agrawal, D.C. (2015) 'Micro moon versus macro moon: Brightness and size', arXiv Preprint arXiv:1507.03578, .

Alexandridis, A., Chondrodima, E., Efthimiou, E., Papadakis, G., Vallianatos, F. and Triantis, D. (2013) 'Large earthquake occurrence estimation based on radial basis function neural networks', IEEE Transactions on Geoscience and Remote Sensing, 52(9), pp. 5443-5453.

Alves, E.I. (2006) 'Earthquake forecasting using neural networks: Results and future work', Nonlinear Dynamics, 44, pp. 341-349.

Debnath, P., Chittora, P., Chakrabarti, T., Chakrabarti, P., Leonowicz, Z., Jasinski, M., Gono, R. and Jasińska, E. (2021a) 'Analysis of earthquake forecasting in india using supervised machine learning classifiers', Sustainability, 13(2), pp. 971.

Debnath, P., Chittora, P., Chakrabarti, T., Chakrabarti, P., Leonowicz, Z., Jasinski, M., Gono, R. and Jasińska, E. (2021b) 'Analysis of earthquake forecasting in india using supervised machine learning classifiers', Sustainability, 13(2), pp. 971.

Emeç, M. and Özcanhan, M.H. (2024) 'Application of artificial neural network methods to anatolian plate earthquake magnitude and location prediction', Journal of Engineering Technology and Applied Sciences, 9(2), pp. 47-62.

Galkina, A. and Grafeeva, N. (2019) 'Machine learning methods for earthquake prediction: A survey', Proceedings of the fourth conference on software engineering and information management (SEIM-2019), Saint Petersburg, Russia.pp. 25.

Geller, R.J. (1997) 'Earthquake prediction: A critical review', Geophysical Journal International, 131(3), pp. 425-450.

Hagen, M. and Azevedo, A. (2017) 'Sun-moon-earth interactions, external factors for earthquakes', Natural Science, 9(6), pp. 162-180.

Höber, D., Taschner, A. and Fimbinger, E. (2021) 'Excavation and conveying technologies for space applications', Berg Huettenmaenn.Mon, 166, pp. 95-103.

Joshi, A., Raman, B., Mohan, C.K. and Cenkeramaddi, L.R. (2024) 'Application of a new machine learning model to improve earthquake ground motion predictions', Natural Hazards, 120(1), pp. 729-753.

Kulkarni, E.G. and Kulkarni, R.B. (2016) 'Weka powerful tool in data mining', International Journal of Computer Applications, 975, pp. 8887.

Mahmud, S. (2019) 'The combined effects of the gravitational forces on the tectonic plates on earth‟ s surface exerted by the moon, the sun and the other planets are one of the main reasons of the earthquakes', Int.J.Adv.Res.Phys.Sci, 6, pp. 44-62.

Mölg, T. and Kaser, G. (2021) 'Springer textbooks in earth sciences, geography and environment. 483–495', .

Peng, C.J., Lee, K.L. and Ingersoll, G.M. (2002) 'An introduction to logistic regression analysis and reporting', The Journal of Educational Research, 96(1), pp. 3-14. Available at: <https://doi.org/10.1080/00220670209598786> .

Rösch, J. (1972) 'Laser measurements of earth-moon distances', The Moon, 3(4), pp. 448-455.

Rosenfeld, L. (1965) 'Newton and the law of gravitation', Archive for History of Exact Sciences, 2(5), pp. 365-386.

Salam, M.A., Ibrahim, L. and Abdelminaam, D.S. (2021) 'Earthquake prediction using hybrid machine learning techniques', International Journal of Advanced Computer Science and Applications, 12(5), pp. 654-6652021.

Si, H. and Midorikawa, S. (2000) 'New attenuation relations for peak ground acceleration and velocity considering effects of fault type and site condition', Proceedings of 12th World Conference on Earthquake Engineering.

Straser, V. (2010) 'Variations in gravitational field, tidal force, electromagnetic waves and earthquakes', New Concepts in Global Tectonics Newsletter, 57(2010), pp. 98-108.

Tan, O., Tapirdamaz, M.C. and Yörük, A. (2008) 'The earthquake catalogues for turkey', Turkish Journal of Earth Sciences, 17(2), pp. 405-418.

Yu, Y., Si, X., Hu, C. and Zhang, J. (2019) 'A review of recurrent neural networks: LSTM cells and network architectures', Neural Computation, 31(7), pp. 1235-1270.

Zechar, J.D. and Nadeau, R.M. (2012) 'Predictability of repeating earthquakes near parkfield, california', Geophysical Journal International, 190(1), pp. 457-462.