**ML Predictive Model for Earthquakes:**

Integrating Mass, Distance, Gravity, and Magnitude

Aadarsh Kushwaha

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

September 2024

Supervisor: Taufique Ahmed

**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Capstone Project |
| **Assessment Title:** | ML Predictive Model for Earthquakes: Integrating Mass, Distance, Gravity, and Magnitude |
| **Supervisor Name:** | Taufique Ahmed |
| **Student Full Name:** | Aadarsh Kushwaha |
| **Student Number:** | 2023398 |
| **Assessment Due Date:** | 27 September 2024 |
| **Date of Submission:** | 26 September 2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Declaration

I, Aadarsh Kushwaha, 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.

Acknowledgments

I s my deepest gratitude to my supervisor, Taufique Ahmed, for his valuable guidance, support, and encouragement throughout this research. His expertise and insights were instrumental in shaping this dissertation.

I am also grateful to CCT College Dublin for providing the resources and environment conducive to my research. Special thanks to the faculty members of the MSc in Data Analytics 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

This study explores using machine learning regression models to improve earthquake prediction by incorporating geophysical and astronomical factors such as Earth-Moon gravitational forces, their distance variations, and localized gravity fluctuations. Using an extensive dataset from 2011 to 2024, sourced from the US Geological Survey (USGS) and additional web scraping, the performance of several machine learning models was tested on four sample sizes: 25%, 50%, 80%, and 100% of the data. For each, 70% was allocated for training and 30% for testing. The XGBRegressor model surpassed all others, attaining an R² score of 0.8706 on the training data and 0.8632 on the test data. Additionally, it recorded a Mean Squared Error (MSE) of 0.1114 and a Mean Absolute Error (MAE) of 0.2473. Significant features were identified using methods like Information Gain, ANOVA, and Lasso, revealing strong correlations with seismic activity. However, hypothesis testing showed that Moon-Earth distance had no significant impact, though gravity variations did. Despite encouraging results, the moderate R² score suggests that the model cannot capture all the complexities of earthquake prediction. The study suggests future improvements by including more comprehensive geological metrics, real-time seismic data, and advanced machine learning techniques like convolutional and recurrent neural networks. It highlights the importance of conducting cross-regional research and obtaining further validation of established hypotheses by specialists.

Table of Contents

[Chapter 1 Introduction 1](#_Toc178197109)

[1.1 Research Context and Motivation 1](#_Toc178197110)

[1.2 Problem statement and research question 2](#_Toc178197111)

[1.3 Research Objectives 2](#_Toc178197112)

[1.4 Scope of the Research 3](#_Toc178197113)

[Chapter 2 Literature Review 4](#_Toc178197114)

[2.1 Earth’s Gravity 4](#_Toc178197115)

[2.2 Earthquakes 4](#_Toc178197116)

[2.3 Lunar Distance 5](#_Toc178197117)

[2.4 Machine learning approach 7](#_Toc178197118)

[2.5 Tidal Forces and Earth's Response 12](#_Toc178197119)

[2.6 Gravitational Variations on Earth 13](#_Toc178197120)

[2.7 Moon - Earth distance 15](#_Toc178197121)

[Chapter 3 Methodology 17](#_Toc178197122)

[3.1 Research Hypothesis 17](#_Toc178197123)

[3.2 Architecture of Machine learning Model 18](#_Toc178197124)

[3.3 Data collection 19](#_Toc178197125)

[3.4 Data Preprocessing 21](#_Toc178197126)

[3.4.1 Distance between moon and location on earth 21](#_Toc178197127)

[3.4.2 Gravitation acceleration(g) 23](#_Toc178197128)

[3.4.3 Gravitational Force calculation between Earth-Moon 24](#_Toc178197129)

[3.4.4 Exploratory Data Analysis (EDA) 25](#_Toc178197130)

[3.4.4.1 Analysis on numerical features 25](#_Toc178197131)

[3.4.4.2 Categorical Feature 25](#_Toc178197132)

[3.4.4.3 Univariant Analysis 26](#_Toc178197133)

[3.4.4.4 Bivariant Analysis 31](#_Toc178197134)

[3.4.4.5 Correlation between features 35](#_Toc178197135)

[3.4.5 Variation of Gravitational Acceleration (g) on world map 37](#_Toc178197136)

[3.5 Feature Engineering 38](#_Toc178197137)

[3.5.1 Ordinal Encoding 38](#_Toc178197138)

[3.5.2 Feature Scaling (for continuous feature) 38](#_Toc178197139)

[3.5.3 Cyclic Encoding 39](#_Toc178197140)

[3.6 Feature selection 40](#_Toc178197141)

[3.6.1 Correlation Coefficient 40](#_Toc178197142)

[3.6.2 ANOVA (Analysis of Variance) 40](#_Toc178197143)

[3.6.3 Information Gain for Regression 40](#_Toc178197144)

[3.6.3.1 SelectFromModel (Lasso) 41](#_Toc178197145)

[3.7 Sampling 41](#_Toc178197146)

[3.8 Machine Learning Models 42](#_Toc178197147)

[3.8.1 ML Model Selection and Training 42](#_Toc178197148)

[3.8.2 Random Forest Regressor 43](#_Toc178197149)

[3.8.3 Decision Tree Regressor 44](#_Toc178197150)

[3.8.4 Gradient Boosting Regressor 45](#_Toc178197151)

[3.8.5 Linear Regression 45](#_Toc178197152)

[3.8.6 XGBRegressor 46](#_Toc178197153)

[3.8.7 CatBoosting Regressor 46](#_Toc178197154)

[3.8.8 AdaBoost Regressor 46](#_Toc178197155)

[3.8.9 Artificial Neural Network 47](#_Toc178197156)

[3.8.10 Polynomial Regression 48](#_Toc178197157)

[3.9 Hyperparameter Tuning 48](#_Toc178197158)

[Chapter 4 Result 50](#_Toc178197159)

[4.1 Hyperparameter Tuning score 50](#_Toc178197160)

[4.2 Correlation score 51](#_Toc178197161)

[4.3 ANOVA test results 52](#_Toc178197162)

[4.4 Information Gain scores 53](#_Toc178197163)

[4.5 R-squared (R2 Score) 54](#_Toc178197164)

[4.6 Root Mean Squared Error (RMSE) 55](#_Toc178197165)

[4.7 Mean Squared Error (MSE) 55](#_Toc178197166)

[4.8 Mean Absolute Error (MAE) 56](#_Toc178197167)

[4.9 Mean absolute percentage error (MAPE) 57](#_Toc178197168)

[4.10 Validation curve for best model 58](#_Toc178197169)

[Chapter 5 Discussion 60](#_Toc178197170)

[Chapter 6 Conclusion and Future Work 62](#_Toc178197171)

[Chapter 7 Ethical Considerations: 63](#_Toc178197172)

[Chapter 8 References 64](#_Toc178197173)

List Of Figures

[Figure 1: Earth’s Gravity(mGal) (NASA Jet Propulsion Laboratory, July 30, 2003). 4](#_Toc178197174)

[Figure 2: The World’s Major Earthquakes from 1956‒2022 (Symington, Feb 2023). 5](#_Toc178197175)

[Figure 3: Distance of Moon from earth during Apogee and Perigee (Wibisono, 2018). 6](#_Toc178197176)

[Figure 4: The gravity expressed in mGal vs 30 December 2009 from Italy(Straser, 2010). 12](#_Toc178197177)

[Figure 5: Distance to Moon form Earth(Rösch, 1972). 15](#_Toc178197178)

[Figure 6: Flowchart of ML Predictive Model. 18](#_Toc178197179)

[Figure 7: Image from the National Center for Polar and Ice Research (NCPOR, 2022). 22](#_Toc178197180)

[Figure 8: Distribution before removing outliers. 28](#_Toc178197181)

[Figure 9: Distribution plot after removing outliers. 28](#_Toc178197182)

[Figure 10: Box plot after removing outliers. 30](#_Toc178197183)

[Figure 11: Cardinality of categorical feature before outliers. 31](#_Toc178197184)

[Figure 12: Cardinality of categorical feature after outliers. 31](#_Toc178197185)

[Figure 13: Relation between magnitude vs depth, distance, force, gravity and year. 32](#_Toc178197186)

[Figure 14: Line graph for Magnitude Vs depth, distance, gravity, force, and year. 34](#_Toc178197187)

[Figure 15: Correlation between features. 35](#_Toc178197188)

[Figure 16: Variation of Gravitational Acceleration (g) on world map. 37](#_Toc178197189)

[Figure 17: Feature selection Lasso. 41](#_Toc178197190)

[Figure 18: Score of Linear and Non-Linear Dataset. 43](#_Toc178197191)

[Figure 19: Working of ANN (Koech, 2022) 47](#_Toc178197192)

[Figure 20: Correlation coefficient for feature selection. 51](#_Toc178197193)

[Figure 21: ANOVA result for feature selection. 52](#_Toc178197194)

[Figure 22: Feature selection mutual info. 53](#_Toc178197195)

[Figure 23: R2 score for ML models with sample size 54](#_Toc178197196)

[Figure 24: RMSE score for ML models with sample size 55](#_Toc178197197)

[Figure 25: MSE score for ML models with sample size 56](#_Toc178197198)

[Figure 26: MAE score for ML models with sample size 56](#_Toc178197199)

[Figure 27: MAPE score for ML models with sample size 57](#_Toc178197200)

[Figure 28: Validation Cure for Training and Testing Data 58](#_Toc178197201)

List Of Tables

[Table 1: Sample size from dataset. 42](#_Toc178197202)

[Table 2: R2-Score of models Hyperparameter Tuning. 50](#_Toc178197203)

[Table 3: Top 4 models trained and tested with 80-20%, sample size 100% 59](#_Toc178197204)

List Of Abbreviations

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

# Introduction

## Research Context and Motivation

Earthquake prediction continues to be a significant challenge in seismology, with crucial implications for disaster management and risk reduction. As natural disasters that inflict substantial destruction and loss of life, earthquakes underscore the urgent need for accurate forecasting to lessen their impact. Traditional forecasting methods, which depend on historical data and seismic monitoring networks, often lack precision and reliability. Recently, there has been increasing interest in using machine learning to enhance earthquake forecasting capabilities. Machine learning excels at analysing large datasets and identifying complex patterns that traditional methods may overlook (The Geological Society of London, 2012). This research proposes a novel approach to earthquake forecasting by incorporating mass, distance, gravity, and magnitude data into a machine learning (ML) predictive model. The goal is to create a model that can accurately predict earthquake occurrences, thereby improving early warning systems and reducing potential damages. To accomplish this, the study expands the dataset used in earthquake prediction models by integrating additional features such as the Moon's gravitational forces and the Earth's non-uniform gravitational field. Feature engineering techniques will be employed to extract relevant features from seismic activity records, gravity variations, and earthquake magnitudes, followed by the development of an ML model using supervised learning algorithms.

Machine learning holds the promise of analysing extensive data and uncovering complex patterns that may indicate impending seismic activity. By integrating various factors such as geological features, seismic data, and environmental parameters, machine learning algorithms can improve the accuracy and timeliness of earthquake forecasts (Environment, June 2021). The variation in gravitational force at different locations on the Earth's surface is due to differences in mass distribution beneath it. Gravity, the force of attraction between objects with mass, varies in strength according to the mass of the objects involved. Thus, locations with more mass beneath them experience a stronger gravitational pull, while those with less mass feel a weaker force. This variation is attributed to the uneven distribution of mass below the Earth's surface. The prevailing technique involves analysing historical earthquake data to identify patterns, the relationship between the Earth’s gravitational force and that of the Moon, and the uneven 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 ).

## Problem statement and research question

This study addresses several research problems related to earthquake magnitude prediction, feature identification, machine learning algorithm effectiveness, and the integration of mass, distance, gravity, and magnitude data.

1. Integration of Mass, Distance, Gravity, and Depth: The research explores how incorporating factors like Earth and Moon mass, gravitational dynamics, and distance can enhance earthquake prediction models, as these factors are often overlooked in traditional methods.
2. Feature Identification and Prediction Methods: This study aims to identify relevant seismic features, such as gravity variations and mass distribution, while developing advanced prediction models that utilize these features to enhance the accuracy of earthquake predictions.
3. Evaluation of Machine Learning Algorithms: The research assesses the effectiveness of various machine learning algorithms for earthquake prediction, focusing on comparing their performance to identify which models yield the most accurate results.

**Research Question:** How can the integration of Earth-Moon mass, gravitational dynamics, and advanced feature identification methods enhance the accuracy of machine learning algorithms in predicting earthquakes?

## Research Objectives

1. Gravitational Influence on Earthquakes: Examine how gravitational forces from both the Earth and the Moon, including mass distribution beneath the Earth's surface and the Moon-Earth distance, affect earthquake occurrences.
2. Enhancement of Prediction Models Integrate new data, such as the Earth’s non-uniform gravitational field and Moon-Earth gravitational interactions, to refine earthquake prediction models.
3. Feature Identification: Identify significant predictive features within the dataset and evaluate their relevance and impact on seismic forecasting.
4. Machine Learning Evaluation: Assess and compare the performance of various machine learning algorithms for earthquake prediction to determine the most effective techniques.

## Scope of the Research

1. The primary objective of this study is to develop a machine learning model capable of forecasting earthquakes using data on mass, gravity, and magnitude.
2. The data is primarily sourced from the US Geological Survey (USGS), encompassing global seismic activity information collected over the last decade.
3. To train the model, the research will only use supervised learning techniques and Artificial Neural Network (ANN) from Deep Learning is used, other machine learning approaches will not be used.
4. The scope will be limited to technical prediction accuracy and model development; it will not address the social or economic effects of earthquakes.

# Literature Review

## Earth’s Gravity

The observation of red and blue color variations in 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). As per observation on current data the average gravitational acceleration on Earth is 9.79 m/s2 (~ 9.8m/s2) which is a universal constant. 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

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).

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 (Mustafa Abdul Salam1, 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 the 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 (Pankaj Chittora, 2022). However, recent advancements in machine learning, such as the application of neural networks to analyse continuous seismic data, offer promising avenues for improved earthquake forecasting (Alexandridis, et al., 2014). 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 (Bertrand Rouet-Leduc, 2017).

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 Logistics exhibiting high accuracy. 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, 2006). 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, the researcher’s 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 ([data2.scec.org](https://data2.scec.org/ugms-mcerGM-tool_v18.4/)). 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 Geological Society of London, 2012).

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 catalogue-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 neighbouring 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 analysed 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 (Joshi, 2006). Researchers have explored the application of ANN in modelling 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 behaviour 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 (Žalohar, 2018), 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 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 modelling 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 (Joshi, 2006).

## 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 centre 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 centre 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

Formula 1: Gravity formula (h, 2024)

Where,

* 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 5: Distance to Moon form Earth(Rösch, 1972).

The Moon, Earth's closest celestial neighbour, 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).

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 (S. Narayanakumar1\*, 2024). 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 modelling approaches, researchers work to increase the accuracy of earthquake predictions and improve disaster preparedness in potentially seismically active regions worldwide.

# Methodology

This study seeks to explore the connection between Earth's gravitational interactions with the Moon and seismic events by applying machine learning algorithms to analyse historical earthquake data and gravitational patterns, ultimately aiming to enhance the precision and reliability of earthquake prediction models.

## Research Hypothesis

**Hypothesis 1**: Although the Moon has a much smaller mass than Earth, its gravitational pull affects earthquake activity. Incorporating this factor into machine learning models will enhance the accuracy of seismic predictions.

* Null Hypothesis (H0): The Moon's gravitational pull does not influence earthquake occurrences, and its inclusion in prediction models does not improve their accuracy.
* Alternative Hypothesis (H1): The Moon's gravitational pull impacts earthquake occurrences, and incorporating it enhances the accuracy of seismic predictions.

**Hypothesis 2**: Differences in Earth's gravitational acceleration across various locations influence the probability and severity of seismic activity. Incorporating these gravitational variations into predictive models will result in more accurate earthquake forecasts.

* Null Hypothesis (H0): Variations in Earth's gravitational acceleration have no effect on the probability or intensity of seismic activities, and models that incorporate these fluctuations do not outperform those that exclude them.
* Alternative Hypothesis (H1): Earth's gravitational acceleration variations do influence the probability and intensity of seismic activities, resulting in more accurate predictive models.

**Hypothesis 3**: Variations in the Moon-Earth distance is associated with the frequency and magnitude of earthquakes. Machine learning algorithms that account for this distance will show improved predictive performance, reflected by higher accuracy (R²) and reduced error rates.

* Null Hypothesis (H0): Changes in the Moon-Earth distance is not related to earthquake frequency or magnitude, and models that incorporate this variable do not show enhanced performance.
* Alternative Hypothesis (H1): Variations in the Moon-Earth distance are linked to earthquake frequency and magnitude, resulting in improved performance in machine learning models that account for this variable.

## Architecture of Machine learning Model

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

A diagram of a process

Description automatically generated

Figure 6: Flowchart of ML Predictive Model.

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

Data collection is the process of obtaining and analysing accurate information from various sources to address research questions, identify trends and probabilities, and assess potential outcomes (Jain, 2024). To collect the necessary data, techniques like web scraping and mathematical computations were used. While earthquake events are available on many websites, the main challenge was obtaining a reliable dataset. For this, a dependable source managed by the [United States Geological Survey (USGS)](https://www.usgs.gov/programs/earthquake-hazards) was utilized. Furthermore, for important data do not present in the earthquake dataset, certain mathematical calculations were conducted, considering the timing and occurrence of seismic events based on research. 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. There are **357731 rows and 27 columns,** but later columns are modified as per research requirement.

Columns catalog:

* date\_time: Date and time when the earthquake event occurred.
* latitude: The geographical latitude (in degrees) where the earthquake occurred.
* longitude: The geographical longitude (in degrees) where the earthquake occurred.
* depth: The depth (in kilometres) at which the earthquake occurred below the Earth's surface.
* magnitude: Magnitude of the earthquake event.
* magnitude\_type: Type of magnitude measurement used (e.g., "ML" for local magnitude, "MW" for moment magnitude). Based on the range from where the earthquake occurred and may bias the ML model due to overlapping; later divided based on the Magnitude Scale.
* nst: Number of seismic stations that contributed to the earthquake magnitude determination.
* gap: The azimuthal gap (in degrees) of seismic stations used in determining earthquake magnitude.
* depth\_min: The minimum distance (in degrees) to the nearest station that reported the earthquake.
* rms: Root mean square (RMS) of the residual travel time (in seconds) between observed and predicted arrivals.
* net: The seismic network associated with the earthquake data.
* id: Unique identifier for the earthquake event.
* updated\_date: Time when the earthquake event data was last updated.
* place: Location description or name where the earthquake occurred.
* type: Type of seismic event (e.g., "earthquake," "explosion").
* horizontal\_error: Horizontal error (in kilometres) associated with the location of the earthquake.
* depth\_error: Depth error (in kilometres) associated with the depth of the earthquake.
* magnitude\_error: Magnitude error associated with the earthquake magnitude measurement.
* magnitude\_nst: Number of seismic stations used in determining the earthquake magnitude.
* status: Status of the earthquake event data (e.g., "reviewed," "automatic").
* location\_source: Source of location data for the earthquake event.
* magnitude\_source: Source of magnitude data for the earthquake event.
* distance: The distance between the centre of the Earth and Moon when an earthquake occurred.
* url: URL made by combining latitude and longitude to get geoid height.
* geoid\_height: The signed difference between the ellipsoid and geoid heights, used to calculate gravity.
* gravity: Gravity acceleration on Earth based on latitude and geoid height.
* force: The gravitational force acting between the Earth and Moon during an earthquake.
* year: Year of the event.
* month: Month of the event.
* hours: Hours of the event.
* minutes: Minutes of the event.
* day\_name: The names of the days.

The data not included in the earthquake dataset after secondary data collection but essential for research—such as the distance between the Moon and a specific location based on latitude and longitude on Earth, gravity at different locations on Earth, and the gravitational force between the Earth and Moon—was gathered through publicly available APIs. Let's explore how this data was collected.

## Data Preprocessing

Data preprocessing involves turning raw data into formats that are both understandable and usable through an iterative process. Raw datasets are typically defined by being incomplete, containing inconsistencies, lacking patterns, and trends, and having errors (Direct, 2022).

### Distance between moon and location on earth

JPL's DE421 ephemeris data and the Python package Skyfield (rhodesmill.org, 2024) were used to calculate the distance between Earth and Moon now of an earthquake. To determine the Moon's position in relation to a specific location on Earth, ephemeris data must be loaded and the url for Api are [https://rhodesmill.org/skyfield/.](https://rhodesmill.org/skyfield/) The steps are summarized as follows:

* Data Preparation: Skyfield tools load ephemeris data with accurate celestial body information for space location and distance calculations.
* Location and Time: Earthquake events noted with date, time, latitude, longitude. Observer location defined by coordinates.
* Distance Calculation: Earth-Moon distance is calculated based on observer's location and ephemeris data, converting AU to 149,597,870. 7 km.

This method gives each event a specific Earth-Moon distance, which is essential for examining any possible links with seismic activity Ellipsoid height for gravitation acceleration(g) (Jenkins, 2023). Determining the geoid height is crucial for predicting gravitational acceleration (g) at different Earth locations. This involves finding the ellipsoid height, not directly available in the dataset. The geoid height, representing global mean sea level, differs from the mathematically defined reference ellipsoid. Knowing this height is key to accurately calculating gravity based on latitude and longitude, which are provided in the dataset. The ellipsoid height must be determined before gravity calculations can be made.

How to Measure an Ellipsoid?

* Orthometric height (H): This is the elevation above sea level at a given location, which can be obtained using a digital elevation model (DEM) or similar resources.
* Geoid Undulation (N): Using a geoid model like EGM96 or EGM2008, obtain the geoid undulation for the specified latitude and longitude, which represents the geoid’s height relative to the reference ellipsoid.
* Ellipsoid Height (h): The ellipsoid height is calculated by adding the orthometric height (H) to the geoid undulation (N).

Formula 2: Ellipsoidal Height (NCPOR, 2022)

where,

* **h** is calculated Ellipsoid height
* **H** is the orthometric height,
* and **N** is the geoid undulation

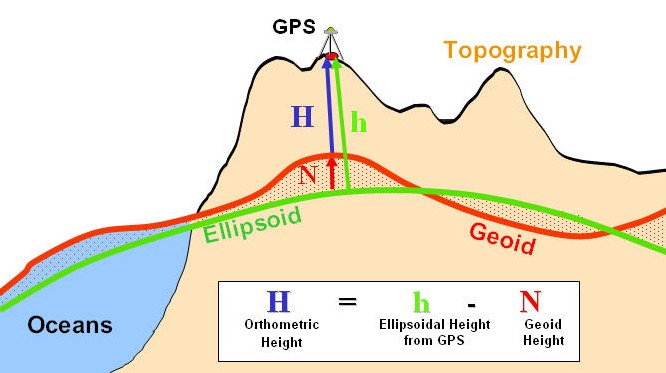


Figure 7: Image from the National Centre for Polar and Ice Research (NCPOR, 2022).

To find the geoid height, a URL like [https://geodesy.noaa.gov/api/geoid/ght?lat={lat}&lon={lon}&model=13](https://geodesy.noaa.gov/web_services/grav-d.shtml) is constructed using latitude and longitude which is dynamically changed and referred as lat and lon. The gravitational acceleration at a specific location is determined by calculating the ellipsoid height, which varies due to the Earth's gravity changes. To accurately determine gravity at the site of an earthquake, the National Geodetic Survey website is utilized to extract Earth's gravity based on location. Web scraping is used to obtain this information. The geoid height considers the Earth's surface irregularities, requiring precise calculation of gravitational acceleration. The ellipsoid serves as a reference model, while the geoid represents actual sea level, accounting for gravitational fluctuations. Accurate gravity measurement is crucial for seismic activity analysis, achieved through the difference between these two heights. The code aims to retrieve and update ellipsoid height data for specific locations, using parallel processing to enhance efficiency and saving progress to a CSV file to prevent data loss during interruptions.

### Gravitation acceleration(g)

The acceleration that an item experiences because of the gravity of the Earth is known as gravitational acceleration (g) (Jenkins, 2023). The Earth's shape, changing with latitude and altitude, affects gravitational acceleration. Understanding these variations is essential for predicting earthquakes.

* Seismic Activity Correlation: Changes in gravity could be linked to seismic events, indicating shifts in mass distribution on Earth affecting earthquake frequency.
* Precision of Data: Accurate gravitational data essential. Integration with latitude and longitude aids understanding spatial fluctuations impacting seismic forecasts.
* Model Accuracy: Considering other variables can enhance earthquake forecast accuracy using gravity acceleration.

Overall, calculating 𝑔 based on latitude and longitude helps refine earthquake prediction models by providing a more detailed understanding of how gravitational forces might interact with seismic processes (Tang, 2023). The WGS-84 ellipsoidal gravity formula calculates the gravitational acceleration (g) at a specific latitude and height above the ellipsoid:

Formula 3: Calculation of value of local Gravity (Jenkins, 2023).

Where,

* g(ϕ,h): Gravitational acceleration at latitude ϕ and height h (in meters).
* ϕ: Latitude in degrees (convert to radians for calculations).
* h: ellipsoid height in meters.
* 9.780327: Standard gravitational acceleration at the equator (in m/s²).
* 0.0053024: Coefficient for latitude-related adjustment to gravity.
* 0.0000058: Coefficient for adjustments due to the variation of gravity with latitude. 0.000003086: Coefficient for adjustments due to height above the ellipsoid.

The formula reflects how gravity varies with latitude and decreases with altitude, considering variations in height above the Earth's reference ellipsoid and variations in latitude.

### Gravitational Force calculation between Earth-Moon

The force of the object that pulls other objects towards its centre. 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. The gravitational force between the Earth and the Moon, can be calculated by using Newton's Law of Universal Gravitation:

Formula 4: Calculation of value of Gravitational Force (study.com, 2024).

Where,

* F is the gravitational force between two masses
* G is the gravitational constant (6.67430 × 10^−11m^3kg^−1s^−2).
* Me is the mass of the Earth (5.972\*10^24kg).
* Mm is the mass of the Moon (7.348\*10^22kg).
* r is the distance between the location on the Earth and the Moon in meters

Suppose the distance between the earth and moon is 380000km then the gravitational force bet­ween earth and moon will be 2.03\*1020 Newtons. In Similar way the formula is applied on each row of dataset to get Gravitational Force between location on earth and moon based on distance calculated above.

### Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an essential step in data analysis that focuses on examining and visualizing datasets to reveal patterns, detect anomalies, and gain insights before applying formal statistical techniques. EDA aids in understanding the data's structure, guiding feature selection, and identifying issues like missing values or outliers. It plays a vital role in helping researchers and analysts form hypotheses, make informed decisions, and improve the quality of data-driven models by ensuring the data is well-understood and properly prepared for further analysis (IBM, 2024). Knowledge of descriptive statistics, data visualization, and Python libraries such as NumPy, Matplotlib, Plotly, and Seaborn is crucial for machine learning. NumPy manages large arrays and math functions, while Matplotlib and Plotly help create various visualizations. Seaborn, based on Matplotlib, simplifies making statistical graphics. These tools support data exploration, summarization, and visualization for advanced machine learning tasks. The dataset contains 357,731 rows and originally had 27 columns, but this may change during feature engineering. It currently consists of 26 columns, with 16 float and 11 object data types, totalling about 302.2 MB in memory. Column names were standardized for clarity and usability, including renaming 'time' to 'date\_time' and 'mag' to 'magnitude'. New features were derived from the 'date\_time' column using Python's 'datetime' package to analyse the relationship between temporal elements and earthquake activity. Certain columns were removed, such as date\_time and geoid\_height, to improve focus and reduce distractions as explained in the catalogue. One missing value in the magnitude\_type column was removed, resulting in a dataset size of (357,730, 13).

#### Analysis on numerical features

Latitude, longitude, depth, magnitude, distance, gravity, force, year, month, hour, and minutes are the eleven numerical properties in total. Four of these are related to time features and are called temporal variables (year, month, hour, and minute). There are two categories for the numerical features: continuous and discrete. None of the features are distinct, though. Thus, including the four temporal features, seven of the 11 features—latitude, longitude, depth, magnitude, distance, gravity, and force—are continuous.

#### Categorical Feature

There are 2 categorical features in total: magnitude\_type and day\_name. Cardinality of a categorical variable refers to the number of unique categories it contains.

* magnitude\_type feature has 4 categories: least damage, moderate damage, strong damage, and major damage.
* day\_name feature has 7 categories: Tuesday, Monday, Sunday, Saturday, Friday, Thursday, and Wednesday.

Classify the magnitude type according to the magnitude scale and its destructive nature.

Four categories of earthquake severity are identified by the function categorize\_magnitude: 'least damage' for magnitudes less than 5.0,'moderate damage' for magnitudes between 5.0 and 6.0,'strong damage' for magnitudes between 6.0 and 7.0, and 'major damage' for magnitudes of 7.0 or higher (Tech, 2024). Using the earthquake\_df DataFrame, this function is used to the 'magnitude' column to create a new column called 'magnitude\_type'. This classification makes it easier to distinguish between various damage levels and reduces data ambiguities by minimizing the overlap in magnitude data, which can be impacted by factors like the angle and distance of the seismic event (USGS, 2023).

#### Univariant Analysis

Examining a single variable to determine its distribution, central tendency, and spread is known as univariate analysis. It aids in enumerating and outlining the primary attributes of that variable, including its variation, range, mean, median, and mode. It is used for following reason:

* to obtain understanding of a variable's fundamental characteristics.
* to recognize anomalies, trends, or patterns.
* to comprehend the distribution of data to prepare it for additional study.

##### Descriptive Analysis

The given's descriptive analysis on **Numerical feature** of DataFrame offers the following observations.

* **Latitude**: The data has a mean of 18.477 and a range of -84.422 to 87.386. The standard deviation of 30.491 indicates a significant fluctuation, indicating a broad geographic dispersion of earthquakes.
* **Longitude**: The values have a mean of -29.762 and range from -179.9997 to 179.9999. The large standard deviation (126.050) suggests a wide longitudinal distribution of earthquake locations.
* **Depth**: The mean earthquake depth is 58.452, ranging from -3.74 to 697.36 km. The significant variation in earthquake depths is reflected by the high standard deviation of 106.952.
* **Magnitude**: Earthquake magnitudes have an average of 3.826 and a standard deviation of 0.886, ranging from 0.7 to 9.1. The data reveals a broad spectrum of earthquake intensities, with most events falling between magnitudes 2.9 and 4.5.
* **Distance**: The average Moon-Earth distance during earthquakes is 384,940.614 km, ranging from 350,275.685 km to 413,063.916 km. A low standard deviation of 15,540.309 suggests that distance measurements are precise.
* **Gravity**: The average gravity value is 9.795943 m/s², with a range of 9.7802 to 9.8321 m/s². The small standard deviation of 0.013217 indicates highly consistent gravity measurements with minimal variation.
* **Force**: The average force is 1.986326e+20, with values ranging from 1.716565e+20 to 2.387122e+20. The standard deviation of 1.620298e+19 shows that the force measurements are stable, with only slight variations in the recorded values
* **Year**: The data has a mean year of 2017.561 and covers the period from 2011 to 2024. The moderate spread around the mean year is indicated by the standard deviation (3.756).
* **Month**: The dataset, covering all months from 1 to 12, has a mean of 6.368 and a standard deviation of 3.405, indicating a relatively even distribution throughout the year.
* **Hour**: Earthquake events occur at all hours of the day, with a mean of 11.465 and a range from 0 to 23. The standard deviation of 6.939 reflects variability in the times of occurrence.
* **Minutes**: The data on minutes reveals variation in the exact timing of earthquakes, ranging from 0 to 59, with a mean of 29.456 and a standard deviation of 17.319.

The given's descriptive analysis on **Categorical feature** of DataFrame offers the following observations.

* **Magnitude Type**: There are 4 distinct categories and 357,731 entries. With 333,876 appearances, the category "least damage" is the most prevalent.
* **Day Name**: There are 7 distinct categories among the 357,731 entries. With 52,148 occurrences, "Sunday" is the most common day.

##### Distribution Plot

A group of red and white graphs

Description automatically generated

Figure 8: Distribution before removing outliers.

A group of red and white graphs

Description automatically generated

Figure 9: Distribution plot after removing outliers.

* **Latitude**: Data clustered around -50 to 50 degrees latitude, likely from frequent measurements. After outliers removed, distribution focused on mid-latitudes, suggesting concentration of events in populated or geologically active regions.
* **Longitude**: Initially, longitude data exhibited wide distribution with peaks at -150 to 150 degrees. Post outlier removal, distribution cantered around -100 to 100 degrees, hinting at concentrated events near landmasses or geographic zones.
* **Depth**: Data initially showed a peak at shallower depths, with a tail at deeper levels. Outliers were removed, emphasizing shallow measurements and suggesting most events occur near the surface. This simplification underscores the importance of analysing shallow depths in seismological data, such as earthquakes.
* **Magnitude**: Data points originally ranged from 2 to 5, with outliers reaching up to 8, showing many moderate events and few extreme occurrences. Removal of outliers’ shifts focus to moderate events between 2 and 5, making analysis easier.
* **Distance**: The data initially showed two event clusters between 350,000 and 390,000 units. After removing outliers, two main event groups at varying distances remained, indicating data from separate regions or contexts.
* **Gravity**: Initial gravity values cantered around 9. 79 to 9. 80, with outliers flagged as errors or rare conditions. Removed outliers highlighted noise.
* **Force**: Distribution peaks at 1.8 to 1.9e20 before outlier removal, indicating high forces. After removal, range is 1.8 to 2.0e20 with fewer extreme values.
* Outliers don't affect distribution of events by year, month, hours, and minutes over time.

##### Handle Outliers

An outlier is an abnormal observation in a random sample, determined by the analyst or consensus process, and requires characterization of normal observations (NIST, 2024). Box plots offer a swift visual representation of the dispersion of values within a dataset. They display key statistics such as the median, upper and lower quartile, minimum and maximum value, and any value way from maximum and minimum are considered as outliers which are not good for further analysis because they provide incorrect analysis (Insights, 2023). Here in the figure below, we can see that some points are far away from the maximum point, and I removed them by using the Interquartile Range (IQR) technique which is nothing but a formula to remove outliers from data sets.

IQR = Q3-Q1

lower limit = Q1 - 1.5\*IQR

upper limit = Q3 + 1.5\*IQR

Q1= first quartile,

Q3 = third quartile,

IQR = Interquartile Range (Anon., 2023)

Remove all the value that is less than the lower limit and more than that upper limit.

A group of red boxes with numbers

Description automatically generated

Figure 10: Box plot after removing outliers.

Most data points cluster around middle latitudes, covering a wide range of longitudes, with exceptions possibly near the poles. Occurrences are common in distinct geographic regions. Depth is mainly under 50 units, with rare deep events exceeding 100 units. The size distribution is balanced, with moderate intensity events between 3 to 5. The distance range spans 350,000 to 410,000 units evenly. Gravity readings consistently fall between 9. 79 to 9. 80, showing reliability. Force values range from 1. 8e20 to 2. 1e20, with moderate variability. Data collection spans from 2012 to 2024, evenly spread throughout all months and times of day. Dataset dimension reduced to (314351,13) after preprocessing.

##### Cardinality of categorical feature

A comparison of a graph

Description automatically generated with medium confidence

Figure 11: Cardinality of categorical feature before outliers.

A comparison of a graph

Description automatically generated with medium confidence

Figure 12: Cardinality of categorical feature after outliers.

Prior to removing outliers, the category "least damage" had the highest count (333,876), followed by "moderate damage," "strong damage," and "major damage. " After removing outliers, the count for "least damage" decreased to 292,971, with decreases in other categories as well. This suggests outliers were present in all categories. The distribution of earthquakes by day remained consistent, with Sunday having the highest count (52,148) and Wednesday the lowest (49,623) before outliers were removed. After removing outliers, counts for each day decreased uniformly, with Saturday having the highest count (45,814) and Wednesday the lowest (43,675). Eliminating outliers had a greater impact on total counts than on distribution patterns.

#### Bivariant Analysis

Bivariant analysis looks at how two variables are related to one another. It examines how changes in one variable relate to those in another (Columbia, 2024). It is used when to ascertain whether two variables are associated or have a relationship. To evaluate the relationship's direction and strength. Investigate causality, the idea that one variable depends on another. Distance, depth, gravity, force and year are bivariant features.

##### Relation between magnitude vs depth, distance, force, gravity and year

A screenshot of a graph

Description automatically generated

Figure 13: Relation between magnitude vs depth, distance, force, gravity and year.

* **Depth vs. Magnitude:** Earthquake depth varies with magnitude, with most earthquakes occurring between 0 and 60 km. Higher magnitude earthquakes tend to be shallower, while those causing strong damage to occur across all depths. Abnormalities can be seen at negative and extreme depths.
* **Distance vs. Magnitude:** This graphic shows no significant correlation between earthquake magnitude and distance. Most earthquakes occur between 340,000 and 415,000 kilometres from a certain spot. Anomalies or outliers appear at smaller distances.
* **Gravity vs.** **Magnitude**: This analysis examines the link between earthquake magnitude and gravitational force, which typically ranges from 9. 77 to 9. 84 m/s². Earthquakes causing least to moderate damage fall within 9. 82 to 9. 83 m/s². Most "least damage" quakes cluster in the middle range, while stronger (high damage) quakes occur across all gravity values, implying an unclear linear relationship between gravitational changes and quake intensity. Some exceptions involve lower gravity levels with varying magnitudes.
* **Force vs. Magnitude**: Plot shows the relationship between earthquake magnitude and force, which could be tectonic or tidal. Force levels range from 1. 6 to 2. 4e20, with higher forces typically linked to higher magnitudes. Most "least damage" earthquakes occur across the force range, while "strong damage" ones have a similar distribution. Fewer outliers compared to previous plots.
* **Year vs. Magnitude**: From 2010 to 2024, earthquake frequency and magnitude remain consistent with no clear trend in strength or distribution across categories. No pattern of increasing or decreasing earthquake strength is evident.

##### Line graph for Magnitude Vs depth, distance, gravity, force, and year

A group of graphs showing different types of data

Description automatically generated with medium confidence

Figure 14: Line graph for Magnitude Vs depth, distance, gravity, force, and year.

* **Depth vs Magnitude:** Deeper earthquakes tend to occur with higher magnitudes, especially between magnitudes 2 and 5. Variability in depth is more significant around magnitude 2, while depths stabilize at a generally constant level for earthquakes above magnitude 5 around 30-40KM.
* **Distance vs Magnitude:** For earthquakes under magnitude 2, distance from a reference point decreases significantly, indicating localized events. Magnitudes 3-6 show greater distance variations, suggesting moderate earthquakes cover a wider range. Magnitude 6+ earthquakes have less distance variation, indicating they occur within a consistent range between 380000-400000KM.
* **Gravity vs Magnitude:** Lower gravity levels observed with earthquakes of magnitudes 2 to 4, followed by increasing gravity values. The relationship between gravity levels and earthquake magnitudes varies, suggesting a potential correlation.
* **Force vs Magnitude:** Magnitude 4 shows an increase in force followed by stabilization. Magnitudes 2 to 4 display varied force levels, indicating a range of forces at play. A saturation point may exist where more force doesn't significantly raise earthquake magnitude, seen in force stabilization at higher magnitudes.
* **Year vs Magnitude:** The decline in low-magnitude earthquakes may be due to improved detection technology or reporting trends. Magnitudes 3 to 5 show consistent distribution over time, with a slight increase around 2024. This uptick could be from data anomalies or future predictions.

Correlation patterns between earthquake magnitude, depth, and force show variability, especially in magnitude ranges 2 to 4. Above magnitude 5, some features stabilize, suggesting a ceiling effect limiting the impact of changes on magnitude.

#### Correlation between features

A colorful squares with numbers

Description automatically generated

Figure 15: Correlation between features.

The Pearson correlation coefficients between several earthquake-related characteristics, including latitude, longitude, depth, magnitude, distance, gravity, and force, are shown in this correlation table. The correlation values range from -1 to 1 (Direct, 2022) A correlation of 1 indicates a perfect positive relationship, where an increase in one variable corresponds with an increase in the other. Conversely, a correlation of -1 represents a perfect negative relationship, meaning that as one variable increases, the other decreases. A correlation of 0 indicates no relationship between the two variables

1. **Latitude and Gravity:** There is a moderate positive correlation between latitude and gravity (0.53), indicating that as latitude increases, gravity tends to increase as well.
2. **Magnitude and Longitude:** A moderate positive correlation (0.60) exists between magnitude and longitude, suggesting that certain longitudes are associated with higher magnitudes.
3. **Latitude and Magnitude:** Latitude has a moderate negative correlation with magnitude (-0.54), meaning that higher latitudes tend to be associated with lower magnitudes.
4. **Distance and Force:** Distance and force have a very strong negative correlation (-0.999), suggesting that they are almost perfectly inversely related. As one increases, the other decreases.
5. **Other Weak Correlations:** Most other correlations are quite weak or close to zero, implying that the variables involved (like latitude and force, or depth and longitude) are not strongly related to each other.

### Variation of Gravitational Acceleration (g) on world map

A map of the world

Description automatically generated

Figure 16: Variation of Gravitational Acceleration (g) on world map.

This image illustrates the worldwide distribution of gravitational acceleration (g) on Earth, showcasing varying gravitational forces at different locations. The color-coded scale represents different levels of gravitational force: yellow indicating higher values (~9. 83 m/s²) and purple indicating lower values (~9. 79 m/s²). Earth's oblate shape results in increased gravitational forces in regions marked in yellow and green, particularly in polar areas and parts of the northern hemisphere. Conversely, areas of darker purple, such as those near the equator, experience lower gravity due to the Earth's equatorial bulge increasing the distance from the centre and reduced gravity from centrifugal forces due to rotation. Variations in Earth's mass distribution led to lower gravity in oceanic regions, like those near mid-ocean ridges such as the Atlantic Ridge, where gravity is weaker due to less dense materials. The Himalayas and mountainous regions exhibit lower gravity due to the inverse relationship between gravity and altitude. Additionally, distinctions in gravitational forces between landmasses and oceanic regions suggest that local gravity is influenced by the density and geological composition of continental and oceanic crust.

## Feature Engineering

Feature engineering involves transforming raw data into meaningful features that can be utilized in supervised learning. It involves training new machine learning features to effectively handle different tasks. Features can be any measurable input, like color or voice. It includes converting raw data into meaningful features using statistical or machine learning techniques. Ordinal encoding and standard scaling are used on specific columns in the dataset to enhance the features for predictive models’ techniques (Patel, 2024).

### Ordinal Encoding

Ordinal encoding is a preprocessing technique that converts categorical data into numerical values while preserving the natural order of categories, making it suitable for machine learning models like neural networks (Wojtek Fulmyk, 2023). It has two key benefits: enabling algorithms to process categorical data and maintaining the ordinal relationships between categories, unlike one-hot encoding. For example, magnitude types like least damage, moderate damage, and strong damage are encoded as 0.0, 1.0, and 2.0. In the dataset, categorical features like magnitude\_type and day\_name are ordinally encoded, which helps the model capture the underlying order or ranking of these variables. This is especially useful when the model benefits from understanding the progression or relationship between categories.

### Feature Scaling (for continuous feature)

Feature scaling in machine learning is crucial for ensuring numerical features are on the same scale, improving algorithm performance and convergence. This preprocessing step aligns values with a mean of 0 and standard deviation of 1, preventing distortion in estimates caused by features with wider numerical ranges like force (BL, 2024). Continuous features such as latitude, longitude, depth, distance, gravity, and force are typically scaled using standard scaling instead of normalization. Standard scaling is preferred because it maintains the relative distribution of the data, centres values around 0, and avoids unnecessarily compressing the range. This scaling method is more suitable for models that can handle outliers, like linear regression, XGBRegressor, and Artificial Neural Network, as it adjusts feature values based on the mean and variance of the data. Standard scaling is particularly beneficial when feature ranges vary significantly, as it is not dependent on specific value ranges, unlike normalization.

Formula 5: Standard Scaling (Z-score normalization) (BL, 2024)

Where,

* χ is the original feature value
* μ is the mean of the feature
* σ is the standard deviation of the feature

### Cyclic Encoding

The method of representing cyclical properties, such time-related variables (like month, hour, and minute), in a way that maintains their cyclical nature is called cyclic encoding. These variables are converted into a continuous form by this encoding, which explains their innate periodicity. There is a natural cyclic order to time-related aspects, such as the months in a year, the hours in a day, and the minutes in an hour. For instance, hour 23 (11 PM) follows hour 0 (midnight), or hour 12 (December) follows hour 1 (January). This cyclical relationship may be misrepresented if these qualities are treated as linear. Problem with Linear Encoding: whereas we numerically represent hours or months (for example, 0-23 for hours), machine learning models may see 0 and 23 as being far apart whereas they are next to each other on the clock. As a result, false patterns are produced. Cyclic encoding solves this by transforming the values into sine and cosine components:

Formula 6: Cyclic Encoding (Chawla, 2024)

The model can "understand" that December is close to January and that hour 23 is close to hour 0. The continuous sine and cosine values, which show the cyclical proximity between values, maintain the cyclical nature of the features. Cyclic encoding, using sine and cosine dimensions, accurately represents cyclical values like hours 23 and 0, unlike one-hot encoding that increases complexity and sparsity. This approach preserves the cyclical structure, helping models capture temporal trends for improved accuracy in time series forecasting and seasonal analysis. Cyclic encoding, using sine and cosine dimensions, accurately represents cyclical values like hours 23 and 0, unlike one-hot encoding that increases complexity and sparsity. This approach preserves the cyclical structure, helping models capture temporal trends for improved accuracy in time series forecasting and seasonal analysis (Chawla, 2024) .

## Feature selection

The act of choosing the most crucial features to include in machine learning algorithms is known as feature selection, and it is one of the key elements of feature engineering. By removing redundant or unnecessary features and condensing the set of features to those that are most pertinent to the machine learning model, feature selection approaches are used to decrease the number of input variables (Heavy.ai, 2024). There are other feature selection techniques; I prefer to use the Lasso Correlation Coefficient, Information Gain for Regression, ANOVA (Analysis of Variance), and SelectFromModel based on my dataset.

### Correlation Coefficient

The process calculates the correlation matrix for numerical columns in a dataset using the Pandas library. It first selects only the numerical features, excluding non-numeric types like strings, and then computes the Pearson correlation coefficients between each pair of these features. This matrix helps identify the relationships and dependencies between numerical variables in the dataset (Fernando, 2024).

### **ANOVA (Analysis of Variance)**

ANOVA is very helpful when comparing group means to assess features in categorical environments. It is perfect for choosing features with significant differences between categories since it aids in discovering features that have a major impact on the variance in the target variable (gajawada, 2019).

This process utilizes the SelectKBest class from the Scikit-learn library to perform feature selection based on the f\_regression scoring function. It selects the top 10 features most relevant to the target variable by evaluating their F-scores. After applying this selection method to the dataset, the F-scores and p-values of each feature are calculated and combined with the feature names in a DataFrame. The resulting table is sorted to display the top 10 features based on their F-scores, indicating their significance in predicting the target variable.

### Information Gain for Regression

knowledge Gain, which is particularly helpful in regression scenarios, quantifies the amount of knowledge a feature offers about the target variable. It aids in the selection of characteristics that lower target uncertainty or variance, enhancing model accuracy (Maddukuri, 2024). The process involves using the mutual\_info\_regression function from the scikit-learn library to calculate the mutual information between each feature in the dataset and the target variable. Mutual information measures the dependency between variables, indicating how much information about the target variable is gained by knowing a specific feature. The calculated mutual information values are stored and associated with the corresponding feature names. These values are then sorted in descending order to identify the features with the highest predictive power, helping to prioritize important features for regression tasks

#### SelectFromModel (Lasso)

Lasso regression selects features by utilizing L1 regularization to reduce some coefficients to zero, effectively eliminating less significant features from high-dimensional datasets. Out of fifteen initial features, nine were chosen as most pertinent for forecasting the target variable based on their significance in the Lasso model. This technique enhances the model's interpretability and prevents overfitting by incorporating only the features with the greatest predictive power.

A screenshot of a graph

Description automatically generated

Figure 17: Feature selection Lasso.

## Sampling

Sampling is the process of selecting a subset of data points from a larger dataset. It is essential to ensure that the important features of the entire population are reflected in this subgroup. In machine learning, sampling is used to train models on smaller data sets, reducing computing complexity while maintaining the validity of the results. Random sampling is recommended for earthquake datasets as it provides each data point with an equal chance of being selected, ensuring representativeness without biases. Balancing components like location, magnitude, and depth, random sampling helps in developing models that can generalize to new data, making them suitable for complex datasets. While other sampling methods may be useful in certain situations, random sampling is typically adequate for most analyses, ensuring representativeness and variety (Qualtrics, 2024). Earthquake datasets are large and diverse, covering various locations, times, magnitudes, and depths. Random sampling ensures that subsets (e.g., 25%, 50%) are representative of the full dataset, reducing biases from focusing on specific regions or magnitudes. Sampling also minimizes computational load, allowing models to be trained and tested on smaller, manageable subsets without losing significant trends or patterns. These subsets maintain the original dataset's statistical characteristics, ensuring the development of reliable machine learning models. Furthermore, setting random\_state=42 guarantees that random samples are selected consistently, which improves reproducibility.

Table 1: Sample size from dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample Size | 25% | 50% | 80% | 100% |
| Rows | 78,587 | 157,175 | 251,480 | 314,351 |
| Column | 11 | 11 | 11 | 11 |

## Machine Learning Models

Machine learning is a branch of artificial intelligence that employs data and algorithms to mimic human learning processes. It emphasizes making predictions and classifications based on input data. An error function gauges the accuracy of the model's predictions by comparing them to known examples. Model optimization entails adjusting weights to enhance the model's alignment with training data (IBM, 2024). Model evaluation involves applying various metrics to assess a machine learning model's performance, crucial in determining its effectiveness during research and ongoing monitoring (Domino, 2024). The evaluation involves the use of regression metrics to evaluate the models, where models are trained with 80% and tested with 20% of each sample size data.

### ML Model Selection and Training

The dataset was analysed to determine whether a linear or non-linear relationship better represented the data for building a machine learning model. Splits of 80% training and 20% testing data were used, with a random state of 42. Various proportions of data samples were evaluated (25%, 50%, 80%, 100%) using Linear Regression and Polynomial Regression (degree 5), assessing their performance with R-squared (R²) and Root Mean Squared Error (RMSE). The linear model had an R² of 0. 6130 and RMSE of 0. 5613, showing moderate fit and difficulty minimizing errors. In contrast, the polynomial model (degree 5) had an R² of 0. 8012 and RMSE of 0. 4024, indicating a better fit and lower errors. In general, the polynomial regression model demonstrated superior accuracy and fit compared to the linear model, indicating a non-linear trend in the data.

A graph of different colored squares

Description automatically generated

Figure 18: Score of Linear and Non-Linear Dataset.

The polynomial regression model of degree 5 was selected for its ability to effectively capture non-linear patterns within the dataset. The models chosen for analysis include Random Forest, Decision Tree, Gradient Boosting, XGBoost, CatBoost, AdaBoost, Artificial Neural Network, and Polynomial Regression.

### Random Forest Regressor

The Random Forest Regressor (Towards Data Science Philip Wilkinson, 2022) is a powerful regression model capable of handling linear and non-linear relationships, performing effectively on large datasets. It excels at making precise predictions, especially when certain features like depth, gravity, and force are more significant than others. By leveraging multiple decision trees and averaging their predictions, it remains resistant to overfitting. This ensemble learning technique constructs numerous decision trees during training, utilizing random subsets of data to produce more reliable and accurate predictions. Ultimately, the Random Forest minimizes the risk of overfitting that individual decision trees may encounter.

Mathematically, for a set of T decision trees, the final prediction is:

Formula 7: Random Forest Regressor (Towards Data Science Philip Wilkinson, 2022)

Where is the prediction from the tree, and is the input data

Hyperparameter tuning involves modifying a model's learning parameters that are not derived from the data itself. Key hyperparameters in a Random Forest model include the number of trees (n\_estimators), the maximum depth of the trees (max\_depth), and the minimum number of samples required to split a node (min\_samples\_split). The goal is to find the optimal values to enhance model performance. For instance, tuning the number of trees can increase stability and reduce overfitting by averaging predictions. However, adding more trees can lead to higher computational costs and diminishing returns in performance improvements.

### Decision Tree Regressor

Decision tree regression is a simple method for predicting earthquake magnitude by dividing data based on key variables like depth and magnitude type. It can capture both linear and non-linear relationships to uncover patterns in the data. By creating a tree structure with decision nodes and leaves, it assigns predicted values. It selects features and thresholds to minimize loss at each split for accurate forecasting. (Towards Data Science Philip Wilkinson, 2022). Typical standards for reducing error consist of:

* The squared error, or the difference between the expected and actual values, is a measurement of the average of the squares of the mistakes.

Formula 8: Squared Error (Practicus, 2019)

* Friedman Mean Squared Error: A variation of squared error, optimized for gradient boosting.
* Absolute Error: Uses the absolute differences, focusing on the magnitude of errors.

Formula 9: Absolute Error (Practicus, 2019)

* Poisson Loss: Modelled using the Poisson distribution, this is helpful for count data.

Key hyperparameters for Decision Trees are criterion, affecting split quality, and splitter, determining split strategy ("best" or "random"). Tuning these is crucial for model effectiveness and generalization to new data. Techniques like grid search help identify optimal configurations by exploring various hyperparameter combinations.

### Gradient Boosting Regressor

Gradient Boosting (learn, 2024) is a predictive modelling method that aggregates weak learners, usually decision trees, to form a robust model. It is particularly useful for datasets with non-linear relationships, such as those involving earthquake variables and magnitude. The process begins with an initial prediction, usually the average of the target variable, and then uses a step-by-step learning method. During each iteration, the algorithm calculates errors from the previous model, trains a new weak learner on these errors, and adjusts the prediction by incorporating the new model's predictions. Hyperparameter tuning, including parameters such as the learning rate and number of estimators, plays a vital role in optimizing model performance. By experimenting with different hyperparameters, a balance can be struck between bias and variance, leading to improved predictive accuracy and generalization on new data (Kaladevi.A.C1, 2024).

### Linear Regression

Assuming a linear relationship between the independent variables and the goal, Linear Regression (learn, 2024) is the most basic type of regression procedure. It is helpful when the size of an earthquake is directly correlated (or almost directly correlated) with variables like depth, force, and gravity (Y, 2024). It offers a strong foundation for more intricate models. By fitting a linear equation to observed data, the statistical technique known as linear regression is used to model the connection between a dependent variable and one or more independent variables. A basic linear regression model's formula is as follows:

Formula 10: Simple Linear Regression (Y, 2024)

Where,

* is predicted output
* is y-intercept (the value of y when)
* is Coefficient for the independent variable (how much y changes for a one-unit change in )
* is independent variable (the input feature)
* is error term (the difference between the predicted value and actual value)

### XGBRegressor

The XGBRegressor (Boudreau, 2019) is an implementation of gradient boosting specifically designed for speed and performance. It works by building an ensemble of decision trees in a sequential manner, where each tree corrects the errors of the previous one. Mathematically, it optimizes a loss function 𝐿 ( 𝑦 , 𝑦 ^ ) by minimizing the error between the predicted values y ^ ​ and the actual values y. The model uses the following formula to update predictions:

Formula 11: XGBRegressor

where the new tree added to the model is denoted by f m ​(x) and the learning rate is represented by η, which is controlled by the hyperparameter learning\_rate. Optimizing the model's hyperparameters is essential to improving its performance. We can modify the model's complexity and generalization by varying parameters like learning\_rate and n\_estimators, which signifies the quantity of trees in the model. More trees might be needed for convergence at a lower learning rate, but if the number of trees is too high, overfitting could result from an improperly balanced population. To achieve the best possible model performance, careful adjustment of these hyperparameters is therefore necessary.

### CatBoosting Regressor

The CatBoost Regressor (Sanni, 2023) is a specialized gradient boosting algorithm designed to handle categorical features and reduce overfitting. It builds a series of decision trees to correct errors from previous trees, effectively capturing complex data patterns with high performance. Optimizing the model's performance relies on hyperparameter tuning, such as adjusting depth, learning rate, and iterations. Depth controls tree complexity, learning rate influences accuracy and convergence speed, and iterations determine the balance between underfitting and overfitting. By systematically testing various hyperparameter combinations, we can enhance the model's precision and reliability in handling new data effectively.

### AdaBoost Regressor

AdaBoost Regressor (Sanni, 2023) is an ensemble learning technique that combines weak learners, like decision trees, to create a strong predictive model. It works by fitting several weak models in succession, correcting errors from previous models. The technique assigns higher weights to mis predicted data points in each iteration, making subsequent models focus more on these challenging cases. To enhance the model's accuracy and robustness, hyperparameter optimization is crucial. Key hyperparameters like n\_estimators (number of weak models) and learning rate (contribution of each learner) play a significant role. Adjusting these parameters can prevent overfitting, improve prediction, and enhance the model's ability to generalize to new data. Careful optimization of these hyperparameters is essential in maximizing model performance in machine learning tasks.

### Artificial Neural Network

Using layers of connected nodes (neurones), the Artificial Neural Network (ANN) model (S. Narayanakumar1\*, 2024) simulates how the human brain processes information. In this configuration, the model's performance and convergence speed are enhanced by standardising the input features using StandardScaler to make sure they are on the same scale. The network can learn from the training data by configuring the MLPRegressor (Multi-layer Perceptron Regressor) with a maximum of 2000 iterations. Various hyperparameters, such as hidden\_layer\_sizes and activation function (e. g. ReLU or tanh), play a crucial role in determining the architecture and performance of a model. The solver settings, like 'Adam' and 'sgd', influence weight changes during training, with the alpha parameter aiding in preventing overfitting. Additionally, the learning\_rate impacts how quickly the model learns from data. Tuning these hyperparameters is essential for optimizing artificial neural network models. By systematically exploring different combinations, we can enhance accuracy, efficiency, and general performance on unseen data. Through rigorous experimentation and analysis of performance metrics, a more robust model capable of accurately capturing patterns within the dataset can be achieved.

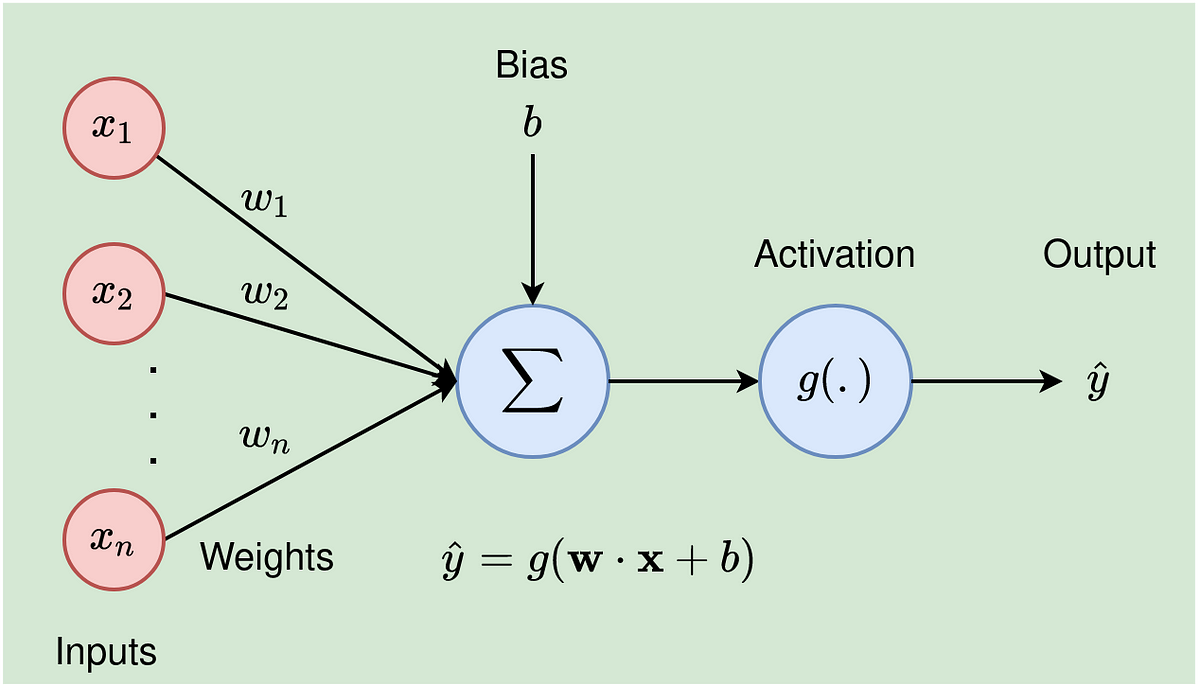


Figure 19: Working of ANN (Koech, 2022)

* Neuron: This is the fundamental building block of a neural network (NN). It takes weighted inputs, performs mathematical computations, and produces an output. It is also referred to as a unit, node, or perceptron.
* Input: These are the data or values provided to the neurons.
* Deep Neural Network (DNN): This term describes an artificial neural network (ANN) that consists of several hidden layers positioned between the input layer and the output layer.
* Weights: These values indicate the strength or importance of the connection between two neurons.
* Bias: This is a constant value added to the sum of the products of input values and their corresponding weights. It helps to adjust the activation of a given node, either speeding it up or slowing it down.
* Activation Function: This function introduces non-linearity into the neural network, enabling the model to learn more complex patterns (Koech, 2022).

### Polynomial Regression

The Polynomial Regression (learn, 2024) model uses input features to create polynomial features of a set degree, allowing for the detection of non-linear relationships in data. These polynomial features are then used in conjunction with a Linear Regression model to better capture complex links between variables. The degree of the polynomial is a crucial hyperparameter that can impact model performance, with tuning necessary to find the right degree that balances accuracy and overfitting. By systematically varying the degree and evaluating metrics like RMSE or R², we can ensure the model generalizes well to new data while accurately capturing patterns in the training dataset.

## Hyperparameter Tuning

The process of modifying external configuration variables, or hyperparameters, for machine learning models, is known as hyperparameter tuning since every dataset and model has different requirements. To find out which combination of hyperparameters produces the greatest results entails putting the model through several experiments. Either a manual or automated approach can be used, but statistical analysis such as loss functions must be used to monitor the outcomes (AWS, 2024). Hyperparameters are manually set before training and impact a model's complexity, learning rates, and architecture. Choosing the right hyperparameters is crucial for a model's accuracy and performance, but there are no strict guidelines for selection. Adjusting hyperparameters can optimize a model's performance and structure. Automated and manual tuning methods are available to fine-tune hyperparameters. Grid search CV is a technique that evaluates various hyperparameter combinations to find the best setup, but it can be computationally demanding. The main grid search parameters include a list of hyperparameter values, a performance metric, and the use of cross-validation to assess the model across different data segments. Adequately tuning hyperparameters can lead to an increase in a model's R2 score.

# Result

Given that the data exhibits a non-linear, polynomial pattern, various machine learning models were developed and evaluated to predict earthquake magnitudes. As this is a regression task, different regression models were employed, and their performance was assessed using specified evaluation metrics. The results from all models support the alternative hypotheses for each proposed hypothesis, indicating that incorporating the Moon's gravitational effects, variations in Earth's gravity, and the distance between the Moon and Earth enhances the accuracy and reliability of earthquake prediction models. Integrating Earth-Moon mass dynamics and gravitational interactions with advanced feature identification methods can significantly improve the predictive performance of machine learning models for earthquake forecasting. The data shows that models like XGBRegressor, Random Forest, and Artificial Neural Networks already achieve high R2 scores above 0.85, even before tuning. After applying fine-tuning and possibly including additional geophysical features, most models show slight improvements in accuracy, highlighting the potential benefits of further optimization. Notably, the AdaBoost Regressor exhibits a substantial increase in R2 score from 0.5564 to 0.8137 after tuning, suggesting that certain models respond particularly well to feature engineering and tuning with complex geophysical data inputs. This demonstrates that integrating comprehensive Earth-Moon mass and gravitational data, along with advanced feature selection, can enable machine learning models to better capture the complex dynamics influencing seismic activities, leading to more accurate predictions.

## Hyperparameter Tuning score

Table 2: R2-Score of models Hyperparameter Tuning.

|  |  |  |
| --- | --- | --- |
| Models | R2 - Score Before Tuning | R2 - Score After Tuning |
| XGBRegressor | 0.8626 | 0.8632 |
| Random Forest | 0.8623 | 0.8629 |
| CatBoosting Regressor | 0.8635 | 0.8583 |
| Artificial Neural Networks | 0.8350 | 0.8572 |
| Gradient Boosting | 0.8435 | 0.8519 |
| AdaBoost Regressor | 0.5564 | 0.8137 |
| Polynomial Regression | 0.8012 | 0.8012 |
| Decision Tree | 0.7312 | 0.7335 |
| Linear Regression | 0.6130 | 0.6130 |

After hyperparameter tuning, most models showed improvement in their R2-Score, with significant gains observed in the AdaBoost Regressor, which jumped from 0.5564 to 0.8137, and Artificial Neural Networks, which increased from 0.8350 to 0.8572. Gradient Boosting also saw a moderate improvement from 0.8435 to 0.8519. XGBRegressor and Random Forest, already high-performing models, experienced slight increases, maintaining their strong performance. However, not all models benefited from tuning; the CatBoosting Regressor experienced a slight decrease in R2-Score from 0.8635 to 0.8583, and Polynomial Regression and Linear Regression saw no change. Overall, hyperparameter tuning positively impacted most models, enhancing predictive accuracy, especially for those initially underperforming.

## Correlation score

A colorful squares with numbers

Description automatically generated with medium confidence

Figure 20: Correlation coefficient for feature selection.

*Latitude, longitude, distance, force, and gravity* are critical factors chosen based on their correlation matrix for predictive modelling. They have moderate to high correlations, with latitude having a positive correlation with gravity and longitude having a negative correlation. To avoid multicollinearity, only force or distance should be selected due to their perfect inverse relationship. These features are important for models focusing on spatial or force-related factors, helping to improve model accuracy by reducing the risk of multicollinearity.

## ANOVA test results

A graph with different colored squares

Description automatically generated

Figure 21: ANOVA result for feature selection.

According to the ANOVA test results, longitude, latitude, magnitude\_type, and gravity are the primary and most significant features with the highest F-scores and p-values of 0. 0000. These features have a significant impact on the variance explanation of the target variable. While parameters like month\_cos, hour\_sin, and distance are also statistically significant but have lower F-scores, depth and year are somewhat important. Therefore, longitude, latitude, magnitude\_type, and gravity should be given priority in selection due to their high significance in the dataset.

## Information Gain scores

A graph with different colored bars

Description automatically generated

Figure 22: Feature selection mutual info.

The top five features with the highest Information Gain scores—*longitude, latitude, gravity, depth, and magnitude\_type*—are crucial for reducing uncertainty in the regression model. Latitude and longitude represent spatial dimensions essential for environmental or geographic models, while gravity can impact system characteristics. Depth and magnitude\_type capture measurement and magnitude qualities, enhancing predictive accuracy. Features like minutes\_cos and minutes\_sin with low or zero scores should be removed to improve model efficiency. In summary, prioritizing these five key features is recommended for effective model development.

After reviewing the results from the four feature selection techniques and considering the relevance of the features to this research, the following features have been chosen for model training and testing where magnitude\_type is target variable and reset of them are independent variable.

* longitude
* latitude
* gravity
* depth
* magnitude\_type
* year
* month\_cos
* month\_sin
* distance
* force

## R-squared (R2 Score)

R-squared is a statistical metric that indicates the proportion of variance in a dependent variable that is accounted for by the independent variables in a regression model. Higher R-squared values suggest a better fit of the model to the data, with values ranging from 0 to 1.

A graph of different colored bars

Description automatically generated with medium confidence

Figure 23: R2 score for ML models with sample size

As shown in Figure 23, the analysis of R² scores across various machine learning models and sample sizes indicated that Random Forest and XGBRegressor consistently achieved the highest performance, with R² values ranging from 0.855 to 0.863 across all sample sizes. These models showed little performance fluctuation with increasing sample size, making them robust options for the dataset. CatBoosting Regressor and Artificial Neural Network (ANN) also exhibited strong performance, with R² values averaging around 0. 857 for all sample sizes, like XGBRegressor and Random Forest. Conversely, linear regression had the lowest R² values and minimal improvement with increasing sample size. Polynomial regression performed moderately, falling behind tree-based models and neural networks. Gradient Boosting consistently achieved R² scores above 0.84, while Decision Tree had lower performance. AdaBoost Regressor performed moderately well. Increasing sample size often led to slightly improved R² scores, particularly for Random Forest and XGBRegressor, but Gradient Boosting and ANN showed good performance even with smaller datasets. Overall, XGBRegressor and Random Forest were deemed the best models for this dataset based on their R² scores.

## Root Mean Squared Error (RMSE)

RMSE is the square root of MSE, measuring residual spread in the model. It is in the same unit as the dependent variable. Lower RMSE values indicate a better fit of the model to the data

A graph of different colored bars

Description automatically generated

Figure 24: RMSE score for ML models with sample size

From figure 24, Analysis of RMSE scores across different machine learning models and sample sizes shows that XGBRegressor consistently performs the best, with RMSE values ranging from 0. 333 to 0. 339. Random Forest, Gradient Boosting, and CatBoosting Regressor also demonstrate strong performance, with RMSE values between 0. 334 and 0. 350. Artificial Neural Network (ANN) performs well with RMSE scores from 0. 341 to 0. 346, indicating good generalization across sample sizes. In contrast, Linear Regression struggles with RMSE values consistently above 0. 560, while Polynomial Regression shows higher RMSE scores with minimal improvement as sample size increases. Overall, larger sample sizes tend to result in better performance, with XGBRegressor, Random Forest, and CatBoosting Regressor standing out as the most accurate models.

## Mean Squared Error (MSE)

MSE calculates squared difference between actual and predicted values, emphasizing larger errors.

A graph of different colored bars

Description automatically generated

Figure 25: MSE score for ML models with sample size

From figure 25, Machine learning model evaluation using Mean Squared Error (MSE) shows a decrease in MSE as sample size increases (25%, 50%, 80%, 100%). XGBRegressor and Random Forest perform the best overall, with the lowest MSE across all sample sizes. XGBRegressor MSE decreases from 0. 1149 (25%) to 0. 1113 (100%), and Random Forest follows a similar trend (0. 1169 to 0. 1116). CatBoost and gradient boosting also perform well, with MSE stabilizing around 0. 1153 and 0. 1205 at 100% sample size. Linear Regression struggles with high MSE values, while AdaBoost and Artificial Neural Networks perform moderately well. Larger sample sizes result in lower MSE, highlighting the importance of using big datasets for improved model performance.

## Mean Absolute Error (MAE)

MAE calculates absolute difference between actual and predicted values, giving intuitive average error magnitude.

A graph of different colored bars

Description automatically generated

Figure 26: MAE score for ML models with sample size

When comparing machine learning models using Mean Absolute Error (MAE) across various sample sizes, XGBRegressor stands out with the lowest MAE of 0. 247 at full sample size, indicating its superior accuracy. Random Forest and CatBoost Regressor also demonstrate strong performance with competitive MAE values. As the sample size increases, models like Random Forest, XGBRegressor, and CatBoost Regressor exhibit improved accuracy. Artificial Neural Network (ANN) and Polynomial Regression yield satisfactory results, while Linear Regression consistently underperforms. Overall, increasing sample size enhances model accuracy, with XGBRegressor consistently leading in performance across all sample sizes.

## Mean absolute percentage error (MAPE)

Mean absolute percentage error is widely used as a loss function in regression and model evaluation due to its simplicity for understanding relative error.

A graph with different colored bars

Description automatically generated

Figure 27: MAPE score for ML models with sample size

XGBRegressor consistently achieves the highest prediction accuracy with the best MAPE scores across all sample sizes. Random Forest and CatBoosting Regressor also perform well with comparable MAPE scores. Artificial Neural Networks improve with larger sample sizes. Most models show lower MAPE scores with larger sample sizes, but XGBRegressor and Random Forest perform well across all sizes. Polynomial and Linear regressions show worse predictive accuracy with larger MAPE scores as the sample size grows. XGBRegressor consistently outperforms other models across all metrics. It excels in R², RMSE, MSE, MAE, and MAPE as sample size increases from 25% to 100%. High R² scores show strong predictive power, while RMSE, MSE, and MAE decrease with larger sample sizes, reducing error rates. MAPE remains consistently low, confirming accurate percentage error minimization. XGBRegressor shows reliable, minimal variation performance, improving marginally with larger sample sizes.

## Validation curve for best model

A graph with green and red lines

Description automatically generated

Figure 28: Validation Cure for Training and Testing Data

If the cross-validation score plateaus or drops slightly while the training score remains high, it suggests potential overfitting with more estimators. Overfitting happens when a model excels on training data but struggles with new or test data. Notable differences between training and cross-validation scores indicate this issue. The XGBRegressor model demonstrates minimal overfitting with close R2 values between training and testing (0. 8706 vs. 0. 8632). The model also shows strong predictive accuracy on both training and test sets, with low error values (MSE, RMSE, MAE, MAPE) confirming its effectiveness on unknown data while maintaining balance and generalizability. Overfitting (AWS, 2024) refers to a model becoming overly complex, capturing noise and irrelevant features along with underlying patterns, leading to high performance on training but poor performance on new data. Underfitting, on the other hand, occurs when a model is too simplistic to understand the underlying patterns in the data, resulting in low training and validation scores due to an inability to effectively learn from the data.

Table 3: Top 4 models trained and tested with 80-20%, sample size 100%

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | R2 score | Root Mean Square Error (RMSE) | Mean Square Error (MSE) | Mean Absolute Error (MAE) | Mean Absolute Percentage Error (MAPE) |
| XGBRegressor | 0.863236 | 0.333712 | 0.111364 | 0.247332 | 0.070678 |
| Random Forest | 0.862888 | 0.334136 | 0.111647 | 0.247364 | 0.071030 |
| CatBoosting Regressor | 0.858331 | 0.339644 | 0.115358 | 0.252918 | 0.072303 |
| Artificial Neural Networks | 0.857207 | 0.340989 | 0.116273 | 0.251290 | 0.071355 |

The study identifies key features influencing earthquake occurrences, with XGBRegressor as the top-performing model, slightly ahead of the Random Forest model in terms of R² score. However, when considering RMSE, Random Forest emerges as the best model by a small margin of 0.000424%. Both XGBRegressor and Random Forest are deemed as the best-performing models, offering valuable insights into seismic activities and aiding earthquake prediction for disaster preparedness and mitigation efforts. The study's utilization of various feature selection methods and comprehensive model evaluation ensures a robust and effective approach to earthquake prediction using machine learning.

# Discussion

The model developed in this research is highly generalized because it integrates multiple variables—mass, gravity, distance, and magnitude—which are not traditionally used in earthquake prediction models. By including these factors, the model can account for a wider range of seismic events across different regions, enhancing its adaptability. Generalization also stems from the inclusion of diverse machine learning techniques, such as XGBRegressor, Random Forest, and Neural Networks, which allow for handling both linear and non-linear relationships between features. The XGBRegressor, for example, demonstrated robust performance across various test sizes, achieving an R² score of 0.87 during training and 0.86 during testing with score of the variance in the dependent variable, with MSE of 0.111364, MAE of 0.247332, and MAPE of 0.070678, indicating moderate explanatory power and low prediction error. By combining geophysical and astronomical data, such as Earth-Moon gravitational forces, earthquake prediction has shown improvement. Traditional seismic data remains crucial, but adding gravitational data provides new insights. A model discovered important links between gravitational variations and seismic activity. Despite the Moon-Earth distance showing no direct influence on earthquakes, gravity fluctuations are strongly correlated with seismic events. This multi-factor approach enhances early warning systems and boosts disaster management effectiveness.

This study offers multiple benefits like

* Data Enrichment: Incorporating factors like gravity and Earth-Moon dynamics improves earthquake dataset comprehensiveness compared to traditional methods.
* Machine Learning Efficiency: Machine learning excels in identifying complex patterns from large datasets, ensuring relevant predictive features are captured.
* Error Reduction: The research achieved reduced error rates, with a Mean Squared Error (MSE) of 0. 1114 and a Mean Absolute Error (MAE) of 0. 2473, underscoring the reliability and precision of the developed predictive models.

The limitations of the research include data quality and availability, such as relying on historical data that may allowed to use real-time seismic activities because of research topic is to observe effect of gravity, distance and force, potentially leading to performance limitations in predicting near-term or real-time earthquakes. Additionally, gravitational data based on calculations instead of direct measurements could introduce inaccuracies, especially since important localized gravitational anomalies may not be included. High-magnitude earthquake events being relatively rare may result in data sparsity and limit accurate predictions. The simplification of seismic dynamics in the model overlooks critical factors such as crustal composition and deep fault interactions. The study's focus on the Moon-Earth gravitational relationship may have limited explanatory power on earthquakes. Geographical biases and limited cross-regional validation may reduce the model's effectiveness across regions with unique seismic behaviours. Despite using advanced machine learning models, the research's predictive accuracy remains moderate, with unexplained variances in earthquake occurrences. Potential overfitting of the model could be due to a small sample size of significant earthquake events or a complex relationship between features and seismic activities. The lack of real-time forecasting and temporal granularity in the dataset may limit the models' ability to provide dynamic or short-term predictions.

# Conclusion and Future Work

The study explores the potential of merging geophysical and astronomical data into machine learning models to forecast earthquakes. Traditional seismic models mainly rely on historical earthquake data, but by including factors like Earth's gravity, Moon-Earth distance, and gravitational variations, this study advanced the predictive capabilities. Various machine learning methods were tested, with the XGBRegressor model performing the best in accuracy. However, while the model successfully identified important features such as gravitational changes, it still struggles to fully grasp the complexity of earthquake occurrences. This indicates the need for further refinement to improve prediction accuracy. Despite confirming the null hypothesis for factors like Moon-Earth distance, the study suggests that enhancing prediction models by adding new elements and refining machine learning techniques can lead to better results. Additionally, while gravitational influences play a role, earthquakes are influenced by a range of geophysical and external forces that warrant further investigation.

Future research should explore incorporating real-time seismic data to enhance predictive capabilities. Advanced Machine Learning Techniques like CNNs and RNNs could capture intricate patterns in data. Integrating more comprehensive geological metrics, such as soil composition, study for movement of tectonic plates and its location and fault line stresses, can improve model accuracy. Cross-regional studies should test earthquake prediction models across different regions. Investigating the impact of lunar and solar gravitational forces on seismic activities, especially during lunar perigee, is recommended. Continuous validation of hypotheses with expert consensus and real-world data can establish relationships between gravitational forces and seismic events. These suggestions build on the promising results of your current research, which demonstrated the value of integrating various natural forces into earthquake prediction models. Also, use of advance hyperparameter tuning like **Optuna.** Hence, this study improves earthquake prediction by integrating geophysical factors and machine learning, paving the way for future advancements.

# Ethical Considerations:

Data privacy and confidentiality are crucial when collecting and analysing earthquake data, which may include sensitive information such as personal data and geographic coordinates. It is essential to adhere to data privacy regulations, obtain necessary permissions, and anonymize or aggregate data to protect individuals' privacy. Ensuring data quality and reliability is vital as historical earthquake data and ancillary datasets may contain errors or biases that could impact analysis accuracy. Implementing rigorous data validation, quality control procedures, and cross-validation with independent datasets can help identify and correct errors. Model uncertainty and prediction accuracy are significant considerations as machine learning models for earthquake prediction may have limitations. Evaluating model performance, quantifying uncertainty, and transparently communicating limitations to stakeholders are essential. Guarding against bias and unfair outcomes in model development is critical to prevent discriminatory results. Conducting bias and fairness assessments, ensuring representativeness in training data, and considering social implications can help mitigate biases. Researchers should prioritize the benefits of their work for society while considering potential risks. It is crucial to evaluate the repercussions of incorrect predictions, use predictive models responsibly and ethically, and prioritize beneficence and non-maleficence (Bristol, 2024).

The ethical use of predictive models in earthquake forecasting presents several challenges, including the risk of causing panic, spreading misinformation, and misallocating resources. To mitigate these issues, clear communication and guidance are essential for stakeholders and the public to understand the limitations and capabilities of these models. Additionally, the use of data in predictive modelling can lead to unintended environmental and societal changes, making environmental impact assessments and stakeholder consultations critical for sustainable decision-making. Data security is also a concern, necessitating robust cybersecurity measures, compliance with standards, and regular audits to protect seismic data from breaches. Engaging with communities is vital to build trust and foster collaboration, as a lack of communication can result in resistance. Moreover, adherence to data protection laws, such as CCPA and GDPR, is essential for responsibly managing personal data (Bristol, 2024). By addressing these ethical considerations proactively, researchers can conduct responsible earthquake prediction studies that minimize negative impacts while advancing knowledge in seismology, promoting transparency, and ensuring social responsibility.

# 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.