RANDOM FOREST AND XGBOOST ALGORITHM FOR EARTHQUAKE ANALYSIS AND PREDICTION

NURFARAHIN BINTI AMIR HAMZAH

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF THESIS

Author's full name	:	NURFARAHIN BINTI AMIR HAMZAH			
Student's Matric No.	:	MCS241020	Academic Session	:	2024/2025
Date of Birth	:	16/03/1999	UTM Email	:	nurfarahin99@graduate.utm.my
Project Report Title	:	RANDOM FOREST AND XGBOOST ALGORITHM FOR EARTHQUAKE ANALYSIS AND PREDICTION			

I declare that this thesis is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the thesis bers to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this thesis for academic exchange.

Signature of Student:

Signature :

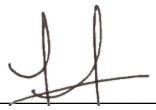
Full Name : Nurfarahin Binti Amir Hamzah
Date : 20 July 2025

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I ASSC.PROF. DR. HAZA NUZLY	Full Name of Supervisor II
Date : 24 July 2025	Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

"I hereby declare that I have read this dissertation and in my opinion this dissertation is sufficient in term of scope and quality for the award of the degree of Master of Data Science

Signature : 

Name of Supervisor I : ASSC. PROF. DR HAZA NUZLY

Date : 24 JULY 2025

Signature : _____

Name of Supervisor II :

Date :

Signature : _____

Name of Supervisor III :

Date :

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration Click or tap here to enter text. and Click or tap here to enter text.

Certified by:

Signature :
Name :
Position :
Official Stamp
Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

RANDOM FOREST AND XGBOOST ALGORITHM FOR EARTHQUAKE ANALYSIS AND PREDICTION

NURFARAHIN BINTI AMIR HAMZAH

A dissertation submitted in fulfilment of the
requirements for the award of the degree of
Master of Data Science

Faculty of Computing
Universiti Teknologi Malaysia

JULY 2025

DECLARATION

I declare that this dissertation entitled *"Random Forest and XGBoost Algorithm for Earthquake Analysis and Prediction"* is the result of my own research except as cited in the references. The dissertation has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : .....

Name : NURFARAHIN BINTI AMIR HAMZAH

Date : 20 JULY 2025

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Madya Dr. Haza Nuzly for encouragement, guidance, critics and friendship.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Master study. Librarians at UTM also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have helped at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

Earthquakes or seismic activity are devastating geological disasters that can cause massive destruction and loss of lives. Traditional geoscience methods in earthquake studies alone are not sufficient to accurately analyze and predict earthquake events. The advancement of technology nowadays in combining traditional geoscience methods and data science including machine learning such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) in analyzing and predicting earthquake events is very important as preparation and mitigation measures for future earthquake events to reduce the impact on humans' lives. Exploratory data analysis (EDA) shows that throughout 2015-2025, the most frequent earthquakes occurred in 2021 with more than 2000 cases. In addition, Indonesia experienced the most frequent earthquakes while Jamaica experienced the highest average frequency of earthquakes both located on the Pacific Ring of Fire. The Pacific Ring of Fire records the total frequency of earthquakes relative to active faults, which is also closely related to the distance of the earthquake epicenter to the fault. The Spearman correlation matrix shows that most of the feature measurements of the earthquakes are not strongly correlated. This study focusses on identifying the geological location of the earthquake in Ring of Fire and active fault zone. Further study will be conducted in master project 2 to analyze the use of Random Forest and XGBoost algorithms in improving accuracy of this research.

ABSTRAK

Gempa bumi atau aktiviti sismik merupakan sebuah bencana geologi yang dahsyat dan boleh mengakibatkan kemusnahan yang besar dan kehilangan nyawa. Penggunaan kaedah tradisional geosains dalam kajian gempa bumi sahaja tidak mencukupi untuk mengalisis dan meramal kejadian gempa bumi dengan tepat. Kepesatan dalam teknologi kini dalam menggabungkan kaedah tradisional geosains dan data sains seperti penggunaan pembelajaran mesin seperti Random Forest (RF) dan Extreme Gradient Boosting (XGBoost) dalam menganalis dan meramal kejadian gempa bumi sangatlah penting sebagai persiapan dan langkah mitigasi dalam menghadapi kejadian gempa bumi pada masa hadapan bagi mengurangkan impak kepada kehidupan manusia. Eksplorasi data analisis (EDA) menunjukkan bahawa sepanjang 2015-2025, gempa bumi paling kerap berlaku pada tahun 2021 dengan catatan lebih daripada 2000 kes. Selain itu, Indonesia mengalami gempa bumi yang paling kerap manakala Jamaica mengalami purata kekerapan gempa bumi yang paling tinggi kerana kedua-duanya berada di kawasan Lingkaran Api Pasifik. Lingkaran Api Pasifik merekodkan jumlah kekerapan kejadian gempa bumi berbanding sesar aktif yang mana ia juga berkait rapat dengan jarak epicenter gempa bumi ke sesar. . Korelasi Spearman menunjukkan hubungan yang lemah antara kebanyakan ciri-cirinya. Kajian ini memberi tumpuan mengenal pasti lokasi geologi gempa bumi di Ring of Fire dan zon sesar aktif. Oleh itu, kajian selajutnya akan dijalankan dalam projek master 2 bagi menganalisis penggunaan Random Forest dan XGBoost algorithm dalam meningkatkan ketepatan kajian ini.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACC	-	Accuracy
ANN	-	Artificial Neural Network
FN	-	False Negative
FP	-	False positive
FPR	-	False Negative Rate
ML	-	Machine Learning
RF	-	Random Forest
TN	-	True Negative
TP	-	True Positive
TPR	-	True Positive Rate
XGBoost	-	Extreme Gradient Boosting

LIST OF SYMBOLS

A_t	-	actual value
c	-	constant
$h_i(x)$	-	prediction of the *t*-th tree
p	-	AR model.
y_i	-	the expected value
$\hat{y}$	-	final prediction.
$\mathbf{Y_t}$	-	predicted value at time
$l(y_i, \hat{y}_i)$	-	the loss function
ϵ_t	-	error term
ϕ_1	-	autoregressive coefficients
$\Omega(f_k)$ $\gamma T + \frac{1}{2}\lambda\sum\omega_j^2$	= -	regularization term that penalizes complexity
ω_j^2	-	leaf weight

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Earthquakes are natural geological disasters that caused the highly damage and threatened the lives and infrastructure. The occurrence of these events is often unpredictable and this becomes the main challenge for seismologists and scientists to predict these events. Sometimes traditional statistical methods are not suitable for earthquake prediction due to the complex and nonlinear behavior of earthquake that need the advanced methods including machine learning algorithms to generate more accurate earthquake data. This research aims to discover and discover the use of Random Forest and XGBoost to enhance the prediction accuracy of the earthquake events to support more effective disaster preparation and risk mitigation especially in areas with limited seismic history.

1.2 Problem Background

Earthquakes or seismic activity are catastrophic natural events that can threats to human life, infrastructure, and socioeconomic stability. The formation of this event due to the releasing energy of geological stress along fault created by tectonic processes including subduction, plate collision or lateral movement of plates that can cause the sudden shaking af lithosphere and can be felt by human (Lay et al., 2017; Stein & Okal, 2020).

As the stress exceeds the regional rock, seismic waves are generated and the ground shakes and often initiating secondary hazard such as tsunamis and landslides. Although earthquake occurrence is mainly focused in active tectonic zones as the Pacific Ring of Fire, their impact also can be felt in nonactive zone such as Malaysia

that affected from the seismic activity in Sumatra and to the south of Philippines that happened in 2004 and 2015.

The occurrences of earthquakes are difficult to predict by the traditional prediction methods of the past statistical modelling, so the development of new statistical models and system is necessary to improve the accuracy and timely forecasting systems. As the old methods are limited due to the complexity, unpredictability, and constant transformations of earthquakes, recent researchers have used advanced techniques such as time series and machine learning to help more clearly distinguish underlying patterns in seismic data and improve forecast accuracy.

The research on earthquakes mostly relying on the prediction magnitude and the occurrences tie rather than spatial perspective of seismic hazards. Risk analysis requires outlining the exact geological areas prone to earthquakes such as areas located on the Pacific Ring of Fire and near to the major active fault. However, the current predictive models do not display spatial generalization and most are built on incomplete datasets which fail to consider the geospatial or tectonic context. The advanced classification techniques of machine learning such as Random Forest and XGBoost can be used to classify and point out high-hazard areas by using historical trends, seismic frequency, and geological characteristics that rarely used in geological research.

Additionally, the existing solutions to earthquake analysis in machine learning involve concentrating on numerical and time series information without explicitly incorporating other geological factors such as fault types, tectonic zone, and plate movement boundaries which have decisive influence on seismic activity. This inconsistency between domain and algorithm modelling weakens the interpretability and limits the utility of earthquake forecasts especially areas relating to creating disaster plans. The proposed research thus incorporates geospatial and geological parameters in a Random Forest and XGBoost-based model with a two-fold idea of improving predictive performance and the higher interpretability of the areas of high seismic hazard.

The collected seismic data usually has missing values, noise, and class quantity, where the low magnitude seismic event exceeds high-impact seismic event

tremendously, especially when collected from global repositories like United State Geological Survey (USGS) database. These inconsistencies can prevent machine learning models from learning generalizable patterns especially when involving predicting or classifying unusual and disastrous rare events. As a result, appropriate pre-processing is important in scaling features, outlier removal and class balancing before training. Failure to put these considerations into account may enable models such as Random Forest or XGBoost to overfit or to generalize insufficiently across verticals and periods.

1.3 Problem Statement

The occurrence of earthquakes is still unpredictable and affect the large losses of human lives, infrastructure damage and long-term economic problems. Despite significant advances in seismology and geoscience over the years, the ability to predict the timing, location, depth, impact score and magnitude of earthquakes with higher accuracy are still major issues in science. The underlying complexity, non-linearity and unpredictable of tectonic processes make it challenging for traditional geophysical and statistical models to offer the accurate prediction.

At the same time, the increasing availability of seismic data along with advances in data science and ML opens an array of opportunities to improve earthquake prediction. The application of time series analysis and ML algorithms especially in capturing complex, multidimensional temporal relationships can reveal subtle indicators and correlations in earthquake data that may obscure signs of large-scale seismic activity and better insights.

Therefore, the development of an interdisciplinary structured approach combined with geoscientific knowledge and high analytical power from data science is critical to resolve those challenges. The combination of machine learning techniques and time series analysis of large seismic and geological dataset can enhance the precision, interpretability, and reliability of earthquake prediction models.

Modern studies tend to treat modern earthquake prediction, impact analysis and risk visualisation approaches as separate task that generates incomplete workflows that are difficult to implement or scale. It consequently requires the development of a more coherent architecture that incorporates exploratory data analysis, predictive modelling, spatial classification and the visualization in a single consistent pipeline.

This paper will present a comprehensive pipeline consistent with data collection, data pre-processing, model development and deployment of a dashboard into a holistic solution in predicting earthquakes and the evaluation of the hazards and take advantage of the advanced machine learning such Linear Regression, Random Forest, and XGBoost algorithms.

1.4 Research Questions

Earthquakes are random events that cause significant damage to the area which becomes a significant barrier to geoscientists and data analysts. Seismic data traditionally employed conventional statistical methods, but these often neglect the complicated, non-linear, and dynamic nature of earthquakes. Due to these limitations, attention is being focused on advanced analytical tools that include time series analysis and machine learning (ML) with a view to discovering hidden patterns and greater accuracy in forecasts. However, priority areas for the improvement are the enhancement of the understanding of regional earthquake patterns, the selection of the best machine learning techniques and proving their accuracy performances in real life earthquake prediction systems.

This research aims to address these gaps through the following key questions:

1. How geosciences and data science knowledge can be integrated in earthquakes analysis and prediction?

2. How Random Forest and XGBoost algorithm determine the zone classification of earthquake locations in active fault zone and Ring of Fire?

3. How Random Forest and XGBoost improve the accuracy and practical applicability of earthquake forecasting analysis compared to traditional statistical methods?

1.5 Research Objectives

(a) To develop exploratory data analysis, feature engineering and modelling for earthquake analysis and prediction.

(b) To predict the geological location of earthquake occurrences in Ring of Fire using classification algorithms of Random Forest and XGBoost.

(c) To develop an interactive visualization dashboard for presenting the earthquakes historical trends, modelled forecast and area of risk zone for seismic hazard analysis.

1.6 Research Goal

The aim of this study is to design an interactive and insightful visualization model of Random Forest and XGBoost algorithm for analysing the global earthquake data and make predictions of earthquake occurrences to improve the understanding of earthquake patterns and facilitating disaster preparedness.

1.7 Scope

The scope of this research focuses on performing exploratory data analysis (EDA), feature engineering and building a dashboard for analyzing and predicting the the geological location of earthquakes occurrences on Ring of Fire. Algorithms like Random Forest and XGBoost will be used to derive knowledge from the dataset by emphasizing the value of the relationship patterns between key features such as magnitude, depth, geographic location and root means square time interval for the event with earthquake occurrences. However, this study is limited to the earthquake dataset from 2015 - 2025 and the its geological locations as geological indicator for seismics analysis.

1.8 Significance of Research

This study represents a significant step toward scientific discovery and has practical implications for efforts seeking to reduce disaster risks. The combination of machine learning and geoscience knowledge in this research proposes to improve earthquake prediction models by overturning the classical restrictions resulting from the limitation in understanding the complicated and nonlinear pattern of seismic data. Through improving prediction models, the efficiency of early warning systems in seismically sensitive and non-sensitive areas can be improved, making communities in those areas more vulnerable to disasters, with a greater time to prepare and reduce destruction. Simultaneously, this research further the understanding in the emerging field of geoinformatics, with a synergistic combination of geological sciences and data science. Since urbanization and climate changes entail the increased likelihood of earthquakes, the increased accuracy of prediction may substantially reduce human and economic cost in the face of these events. Further, the results of this study may contribute to the improvement of predictive systems applicable to various natural catastrophes, therefore extending the scope and impact of this future seismology study.

1.9 Chapter Summary

In summary, this chapter is the fundamental stage of research where it explains the challenges of forecasting earthquakes and the limitations of traditional methods due to the complicated nature of seismic activity. The application of machine learning and time series analysis highlighted in this study to improve earthquake forecasts by exposing hidden patterns in seismic data. The chapter also outlines the approach of researching and comparing machine learning models for disaster preparedness and risk reduction especially in locations with a limited of earthquakes history.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews the past study of earthquake events and the machine learning tools used in earthquake analysis and predictions. The strength and limitations of past study in various predictive approaches and highlighting their relevance to current development discussed in this chapter.

2.2 Formation of Earthquake

Seismic events can cause major catastrophic natural disaster that may lead to disastrously high magnitudes and risk of secondary disaster such as volcanic activity, landslides, fires, liquefaction and tsunamis that affect the destruction of the environment, economy, properties and loss of lives (Pwavodi et al., 2024). Earthquake happens when there is a sudden movement of a tectonic plate due to energy release along a fault zone, leading to a vibration in lithosphere of the earth and failure, slipping and shifting of the tectonic plate (Pwavodi et al., 2024).

According to Lay (2016), earthquakes started at subduction zones, the Atlantic ridges and transform fault zones. Subduction zone is a tectonic plate boundary where an oceanic plate is forced under another oceanic or continental plate, based on gravitational and slab pull force, producing major seismic activity which is due to accumulated strain energy. As the subducting plate sinks into mantle, stress accumulates at the megathrust interface due to friction and resistance at this interface until a point when a rapid rupture occurs and triggers an earthquake (Lu et al., 2017) shown in Figure 2.1.

Subduction zones are most tectonic active areas that induce tsunamis. It occurrence commonly along convergent boundaries with various collision of tectonic plates. Plate collision classified into three major types which are oceanic-continental plate, oceanic-oceanic plate and continental-continental plate. Subduction zone mostly occurs at active seismic activity known as “Pacific Ring of Fire” (Figure 2.1) that also has the active volcano activities and deep oceanic trenches.

The presence of active fault also related to the formation of earthquake due the continuous crustal movement that generated the frictional resistance between plates and accumulation of the tectonic stress. Over the time, when the stress exceeds the frictional resistance, the earth crust fractured and sudden slip formed then releasing energy in form of seismic wave. The larger and deeper the slip occur, the highest magnitude of earthquake generated. This process called as elastic rebound(Biswas et al., 2023). The active faults such as East Anatolian Fault and San Andreas Fault are prone to earthquake due to active fault interaction and accumulation of stress at faster rate (Güvercin et al., 2022).

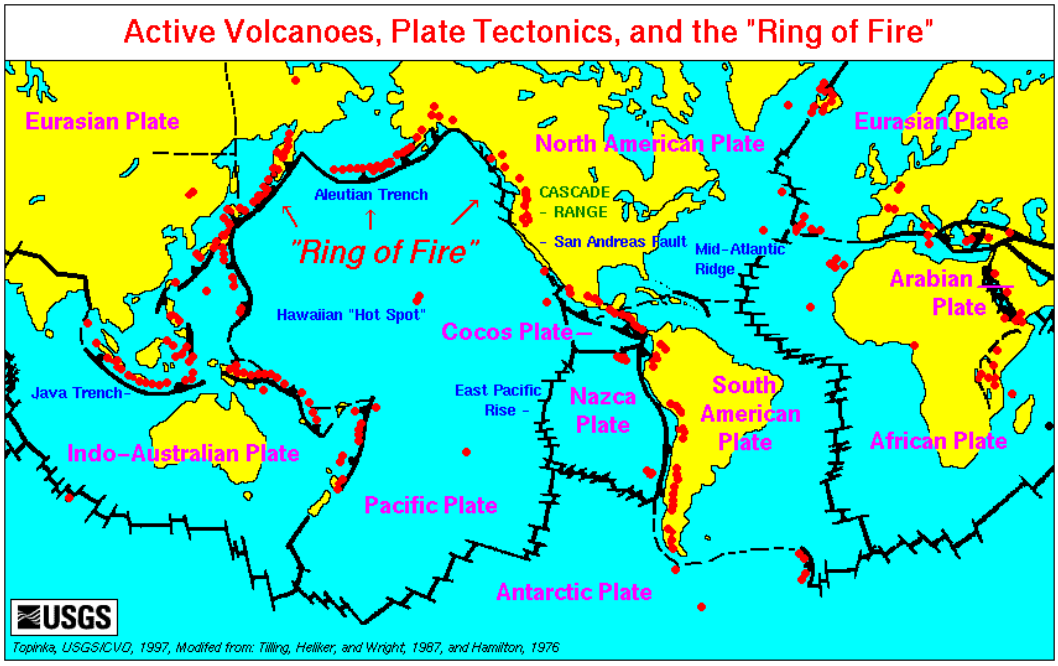


Figure 2.1 Location of active volcanoes, plate tectonics and the Ring of Fire (USGS,1997)

2.3 Types of Plate Boundary

2.3.1 Convergent Plates:

i. Continental-Continental Boundary

In continental-continental boundaries, two tectonic plates of continental crusts meet and collide into each other producing a large compression that deform the crust resulting in earthquakes of large magnitude. The best example is the collision of the Indian and Eurasian plate which has led to the formation of Himalayas ridge that uplift and produce large earthquake due to crustal shortening and the formation of fault (Bao et al., 2022). Earthquakes in these boundaries are pointed out at shallow depth and triggered the thrust faulting amongst complex fault systems that did not form a deep subduction since neither the plate can subduct easily enough to the mantle (Luo et al., 2022)

Another example is the Zagros Mountains that have been formed as a resulted from the Arabic-Eurasian continental collision have moderate-to-strong earthquakes. In this area significant seismic activity results from the sharp release of the accumulated tectonic stress via fault failure (Manaman & Shomali, 2010). It can be concluded that crustal compression and fault reactivation within thickened lithosphere is the major cause of earthquakes at continental-continental margins (James, 2021).

ii. **Oceanic – Continental boundary**:

Oceanic–continental convergence boundary occurs when a high density of oceanic plate subducts under a continental plate leading to intensive tension and pressure causing deep-focus earthquakes and inland volcanism (Duarte & Schellart, 2016). This mechanism creates deep-sea trenches, forearc basins, back arcs basin and volcanic arcs such as those found in the Andes and Cascadia regions which also generated the intermediate to high magnitude of deep. These large earthquakes in these zones capable to create the tsunami event due to the subduction of plates (Bilek & Lay, 2018).

Moreover, these boundaries also generate characteristic seismic anomalies consist of a double seismic zone where the dehydration of the sinking slab increases the fluid pressures, weakens the fault strength, and causes intermediate earthquakes (Geersen et al., 2022). Some of major earthquakes occurred within the outer rise, seaward of the trench, or on plate-bending related faults below the megathrusts, while others are originated within the upper plate near to the zone of high slip. Even though the deeper megathrust ruptured are rarely happened, but some intraplate aftershock may occur due to the significant afterslip in the upper plate that resulting from the complicated faults and high activity aftershocks (Bilek & Lay, 2018; Lay et al., 2020).

Understanding mechanism and advancement of technology in the imaging techniques such as seismic tomography combined with the advanced statistical technique such as machine learning played a critical role in improving understanding of how plate subduction geometry affects seismicity activity in convergent margin for assessing earthquake hazards and mitigation.

iii. Oceanic – Oceanic boundary:

Oceanic-oceanic plate boundaries formed when two oceanic plate collides where one plate subduct under another plate produced the deep ocean trenches of large earthquakes due to the frictional locking and massive energy released throughout the subduction zone. The subduction initiate when the older dense oceanic lithosphere being forced beneath the younger and less dense oceanic lithosphere that basically facilitated by transform faults (Wu et al., 2019). For example, Tonga-Kermadec subduction zone is the example of megathrust earthquake caused by the accumulation of tectonic stress (Wang et al., 2023). Subduction at these boundaries also lead to volcanic eruptions and tsunamis often associated with the sudden rise of the seafloor (Buforn & Udías, 2010).

In these regions, seismic activity usually traces the geometry of the descending slab and exhibits constant subduction dynamics. Further, geophysical studies show that there is an internal deformation of the downgoing slab in ocean-ocean boundaries, which explains the complexity of the earthquake (Gurnis et al., 2000) . Seismic profiles and focal mechanisms recorded in these areas indicate that reverse and thrust faulting are the dominant patterns of rupture, consistent with the overall compressional characteristics of tectonic processes (Sartori et al., 1994).

2.3.2 Divergent Plate:

Divergent plate boundaries are zones occur when two tectonic plates move separately formed new seafloor and mid-ocean ridges or continental rift zones due to upwelling magma and induced shallow earthquakes from stretching crust (Duarte & Schellart, 2016). Earthquake in this plate less prone to high magnitude compared to convergent plate due to tensional stress that pulls the plate apart along rift zones and primarily occur at the normal fault where the hanging wall moves down relative to the footwall which generate small tension and seismic activity (Olive, 2023).

However, divergent plate also has possibility to experience the large seismic activity triggered by dynamic stresses transfer from distant seismic events, magmatic processes such as dike intrusion that spread plates apart and large tensional stresses and faulting movement (Hill & Prejean, 2015). As example, earthquake of 5.0 magnitude at Reykjanes Peninsula in Iceland on 2021 of the divergent Mid-Atlantic Ridge are associated to both intense seismic swarms and magma intrusion into formed the 9 km dyke beneath Fagradalsfjall area (Fischer et al., 2022).

Similarly, the earthquake near the East African Rift Zone with magnitude 5.9 affect disruption in North Tanzania demonstrated the impact of divergent zone quakes due to the magma intrusion of Fentale diking resulting the seismic activity and volcanic events (Macheyeki, 2024). These events highlight the need of rigorous monitoring of rift regions especially in areas of higher population growth.

2.3.3 Transform Plate

Transform plate boundaries formed when tectonic plates move and slide alongside each other horizontally produced large shear stress which often generate earthquakes (Duarte & Schellart, 2016). The characteristics of these boundaries include strike slip faults such as California's San Andreas Fault where the Pacific plate moves northward to the North American plate. If the motion of tectonic plates is hindered by friction, built-up stress may eventually be abruptly discharged through short earthquakes causing extensive damage (Liu & Buck, 2018). The East Anatolian Fault in Turkey is the example of major transform fault that triggered fatal earthquakes in 2023 with the loss of tens of thousands of individuals (Biswas et al., 2023).This case demonstrates the huge seismic risks involved in transform boundaries especially in areas where large cities are located close to fault zones.

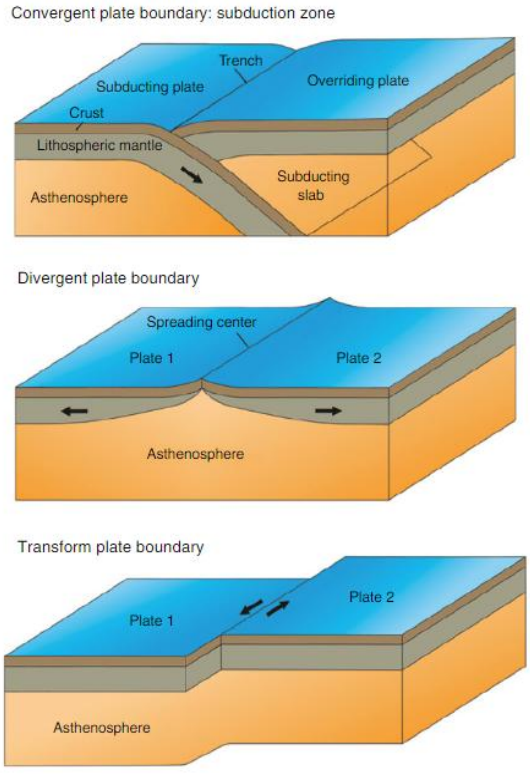


Figure 2.2 Schematic representation of three types of plate boundaries: convergent (top), divergent (centre) and transform (bottom) modelled from (Duarte & Schellart, 2016)

2.4 Pacific Ring of Fire

Approximate 90% of global earthquakes and 75% of dormant world's active volcano located at active seismic zone known as "Ring of Fire" mainly due to active interactions between the Pacific Plate and the nearby plates such as the Philippine, Cocos, and Nazca plates shown in Figure 2.1 and Figure 2.3. These zones have high seismicity, active volcanoes and large oceanic trenches resulting from the intense tectonic forces (Pwavodi et al., 2024). Most of these interactions take place at subduction zones where oceanic lithosphere subducts under continental or other oceanic plates causing slabs to be circulated into the mantle and the epirogenic of high heat and magma that feeds explosive volcanic arcs (Wu et al., 2019).

The Ring of Fire is bounded by subduction zones that have deep ocean trenches with volcanism occurred along the arc systems and high earthquake activities at convergence and subduction zones (Wang et al., 2018; Kerr et al., 2022). As example, Japan Trench and Peru-Chile Trench located at major subduction zones where the Pacific Plate subducted under Eurasian and South American Plates (Dragoni & Santini, 2022; Nakamura et al., 2023). These processes of slab pull, and mantle convection are important for developing the dynamics of tectonic motion and the ongoing transformation of the Ring of Fire.

However, 10% possibility of the earthquake's events occur outside this zone can also lead to the massive disaster due to the active movement of plate boundaries such 7.8 magnitude of earthquake with 17.9-km depth in Turkey and Syria 2023 which killed more than 45, 000 peoples (Biswas et al., 2023). The collision of Eurasian and Africa plates (Figure 2.3) triggered by several active faults including North Anatolian Fault, the East Anatolian Fault and the Dead Sea Transform Fault has triggered varying magnitude of earthquakes ranging from minor to major (Biswas et al., 2023).

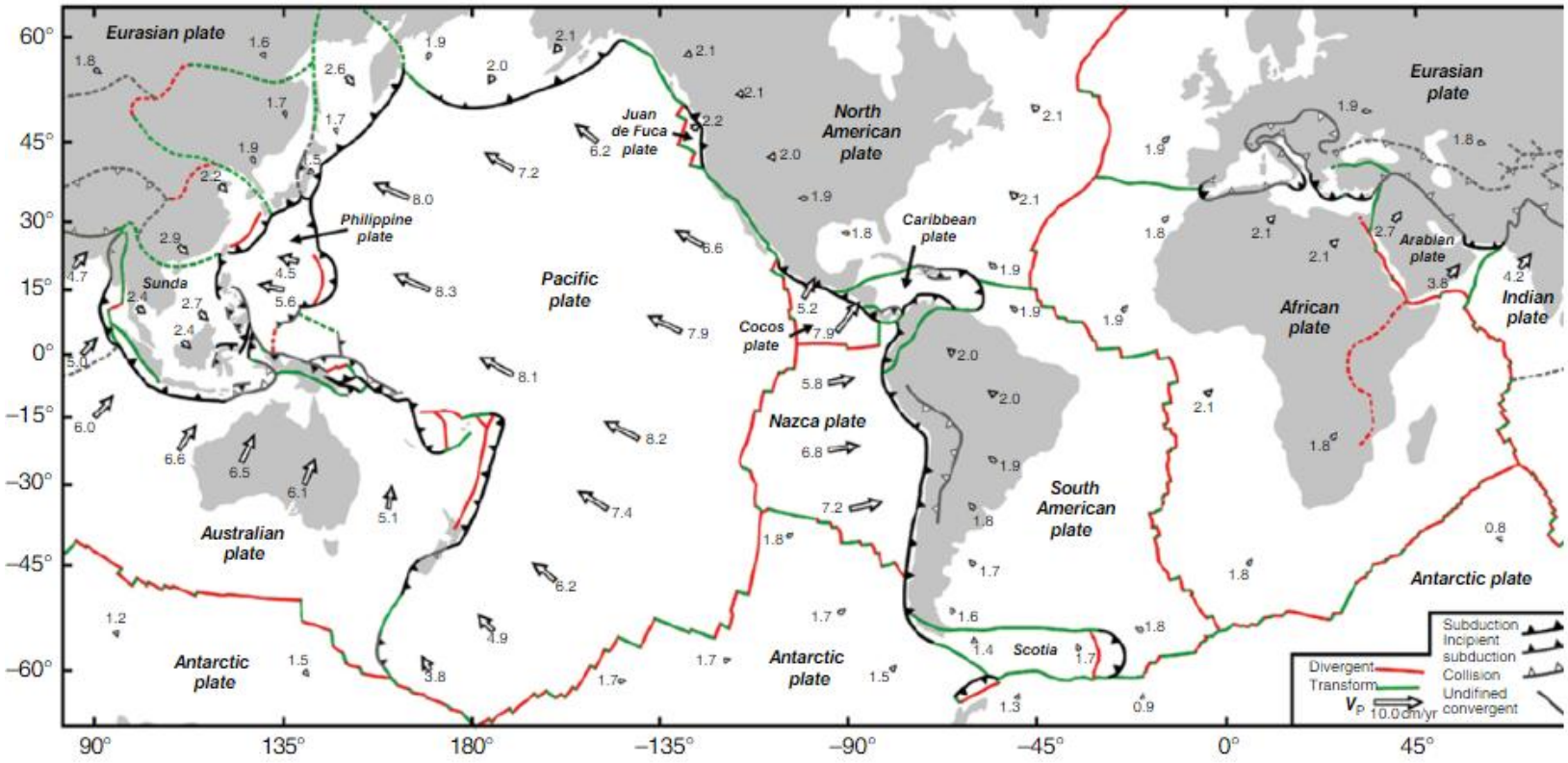


Figure 2.3 Global plate tectonic map illustrating of major tectonic plates, their velocities (in cm/yr), and the major plate boundaries: convergent plate boundaries (black and grey segments with triangles), divergent plate boundaries (grey segments), and transform plate boundaries (black segments) modified from (Duarte & Schellart, 2016)

2.5 Earthquakes due to active faults

A massive 7.8-magnitude earthquake occurred in southeast Turkey near the East Anatolian Fault (EAF) due to a significant strike-slip fault zone ascribing to the fact that the high tectonic activity is capable of being experienced outside the Pacific Ring of Fire (Yenidoğan, 2024). Geologically, the earthquake was triggered by the tectonic stresses that had been built up along complex fault systems due to the relative movement of Anatolian, Arabian, and African plates in an intraplate setting (Biswas et al., 2023).

Recently on March 2025, catastrophic earthquakes happened at Mandalay, Myanmar which located outside the Ring of Fire which indicates that this region can also experience massive seismic activity due to its location at the complex boundary between the Indian and Sunda plates (J. Wang et al., 2025). The Sagaing Fault that formed from a large right-lateral strike-slip fault merely acts as the main factor in providing the space for the northward movement of the Indian Plate with respect to the Sunda Plate interactions (Shahzada et al., 2025).

The fault that crossed the densely urban area can even be capable of causing even moderate earthquakes that can affect the massive disruption hazard. This 7.9 magnitude earthquake has highlighted to ongoing strain accumulation and revealed the of complexity subsurface fault structures existed through geophysical surveying (Cai et al., 2025). Through geophysical surveying of seismic reflection and crustal imaging techniques have proven that intraplate deformation is the main factor of Myanmar's seismicity influencing to sudden shaking depths and fault interactions (T. Wang et al., 2023). The history of highest magnitude earthquakes for 10 years from 2015 to early 2025 can be classified in Table 2.1.

Table 2.1 History of high magnitude earthquake in from 2015-2025

Year	Citation	Location	Characteristics of Earthquake	Risk Factor (Outcome)	Impact
2015	Quantifying The Benefit of Risk Mitigation Strategies on Present and Future Seismic Losses in Katmandu Valley, Nepal. (Mesta et al., 2023)	Nepal	M7.8 Thrust fault	Himalaya fold zone, old rock structure, poor infrastructure and high population density	Massive destruction in Kathmandu killed nearly 8,800, 23,000 injuries, extensive property damage.
2016	April 2016 Ecuador Earthquake of Moment Magnitude Mw7.8: Overview and Damage Report. (Mera et al., 2017)	Ecuador	M7.8, Subduction zone	Subduction of the Nazca Plate under the South American Plate caused the crustal earthquakes and volcanism on the Andes Mountains	670 killed, 28,000 injured and massive destruction
2018	Field Insights and Analysis of the 2018 Mw 7.5 Palu, Indonesia Earthquake, Tsunami and Landslides (Cilia et al., 2021)	Indonesia, Sulawesi	M7.5, Strike-slip and Tsunami	Palu-Koro left-lateral strike-slip fault triggered tsunami due to earthquake-induced soil liquefaction and landslides	Destroy Palu city with more than 2,245 people killed, thousands missing, 10,000 injured and 4000 severe injured, 75,000 displaced.

2023	February 6, 2023, Earthquakes and Preliminary Assessment of Building Damage Based on Field Surveys. (Yenidoğan, 2024)	Turkey-Syria, 2023	M7.8 + M7.6, Strike-slip	Extensive rupture length (~560 km total), multiple large magnitude events in close succession, and the presence of seismic gaps along the fault segments	Over 53,000 killed, 11 provinces affected and extensive property damage
2024	The 2024 Noto Peninsula Earthquake: Preliminary Observations and Lessons to Be Learned. (Suppasri et al., 2024)	Noto Peninsula, Japan	M7.5, Active fault + Cascading hazard	Cascading hazard including geological uplift, liquefaction, landslides, fires and tsunami	240 deaths, severe infrastructure damage
2025	In the wake of the March 28, 2025 Myanmar earthquake: A detailed examination. (Shahzada et al., 2025)	Myanmar-Thailand, 2025	M7.7, Strike-slip (Sagaing fault)	A major dextral strike-slip boundary between the Burma Microplate and Sunda Plate caused supershear rupture propagated over 460 and surface displacements exceeding 6 m and violent shaking in urban centers like Mandalay, Sagaing, and Naypyidaw.	More than 4390 killed, over ~ 4900 fatalities, ~ 6000 injuries, and widespread destruction of infrastructure.

2.6 Supervised Machine Learning for Earthquake Analysis and Prediction

Supervised machine learning is the labelled training dataset that implemented to make a prediction of input data with chosen output. The application of supervised ML models in earthquake studies such as simulate the movement and analysis of tsunami movement through neural network-based hazard predictions which compared outputs of the model and observed wave heights (Pham et al., 2021). These interlocking mechanisms has improved the insights of researchers and disaster management organizations to measure and manage tectonic events in complex geologies such as in Sulawesi.

Combination of machine learning and artificial intelligence (AI) for developing new models for immediate post-earthquake damage assessment, futures probabilistic aftershock monitoring that played decisive roles in the guidance of rescue efforts. Advanced machine learning such as convolutional neural networks and probabilistic methodologies made the precise classification of damage from satellite imagery possible, thereby possible to produce detailed urban devastation mapping (Kilic et al., 2023). Turkey is not located at the Ring of Fire zone but the complex geology of Turkey makes it susceptible for earthquake activity in futureand it revealed the need for ML and AI in allow early warning systems for populated and seismically active areas (Gülen et al., 2023).

Some earthquakes triggered the occurrence of tsunami resulting the massive catastrophic impact of that area such as in Japan, 2011. The diverse earthquakes analysis on earthquake dataset from past event using machine learning such as random forest and logistic regression to perform binary classification of tsunami events based on seismic, geopspatial features enhanced the accuracy and lead time of tsunami forecasting (Satish et al., 2025). The performance of ML model improves the predictive performance in both tsunami and earthquake magnitude forecasting enabling the extraction of complex temporal and spatial trend from high-dimensional seismic data (Kaftan, 2025).

2.6.1 Random Forest (RF) Algorithm

Random Forest (RF) algorithm widely used in predicting earthquake magnitude based on diverse seismic indicator and time series features with high accuracy from the historical pattern of seismic data (Kukartsev & Degtyareva, 2024; Novick & Last, 2023). Random Forest is the Furthermore, this model also developed to make the prediction of ground motion intensity measure which important for evaluating seismic hazards and improving building design (Long et al., 2024; Nandu et al., 2025)

The performance of this model also shown high accuracy in predicting earthquake parameters. The accuracy performance achieve until 98.8% in predicting earthquake intensity using Indonesia's earthquake dataset and 94.49% for magnitude estimation using strong-motion data (Handayani et al., 2024; Wijaya et al., 2022).

RF also used as the comparative performance models with other machine learning model such as Support Vector Machines (SVM), Gradient Boosting Machines (GBM) and other models that shows this model is higher generalization performance and lower error rates(Gupta et al., 2024; Long et al., 2024; Nandu et al., 2025; Yavas et al., 2024).

Even though RF show the high performance accuracy, sometimes it can be overfitting when the model is too complex that tend to capturing noise data than the underlying trends in the data (Z.-N. Wu et al., 2024). This model also struggles with the highly nonlinear relationship of seismic data where sometimes the other machine learning like Gradient Boosting Machine (GBM) and Support Vector Machine (SVM) shown better performance in determining these nonlinearities.(Babu et al., 2024; Cornely & Wang, 2023).

The lack of interpretability of RF for decision making process due to complex decision tress where the earthquake analysis need the clear insights into factors influencing predictions are important for effective disaster mitigation management. (Alidadi & Pezeshk, 2025).

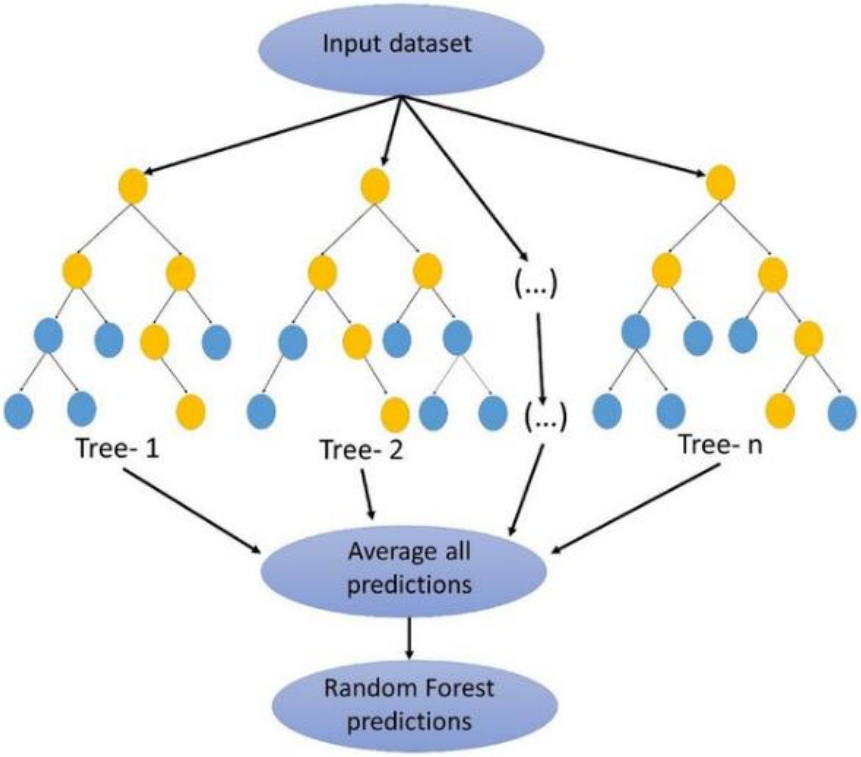


Figure 2.1 Schematic diagram of Random Forest algorithm

2.6.2 Extreme Gradient Boosting (XGBoost) Algorithm

Extreme Gradient Boosting algorithm is a powerful machines for earthquake analysis through statistical boosting methods and build classification ad regression trees by integrating multiple trees into a consensus prediction framework to improve the model precision (J. Wang et al., 2025).

This model demonstrates the high performance accuracy and efficiency for earthquake analysis, can handle large and complex dataset while prevent overfitting since the algorithm incorporates L1 (Lasso) and L2 (Ridge) regulation techniques (Babu et al., 2024). According to Wang et al., (2023), XGBoost well performed in classifying seismic event including earthquakes, explosion and mining induced earthquakes which shown the high accuracy of more than 90% compared to SVM algorithm that indicates the high possibility of seismic events discrimination.

XGBoost algorithm sometimes faces a few challenges when dealing with massive and complex datasets. As an example, this model can reduce the sensitivity to imbalance classes in differentiating the mining-induce earthquakes from explosion leading to skewed precision and recall that can cause the bias and affect the prediction accuracy especially for nonlinear structural responses (Babu et al., 2024; Wang et al., 2023).

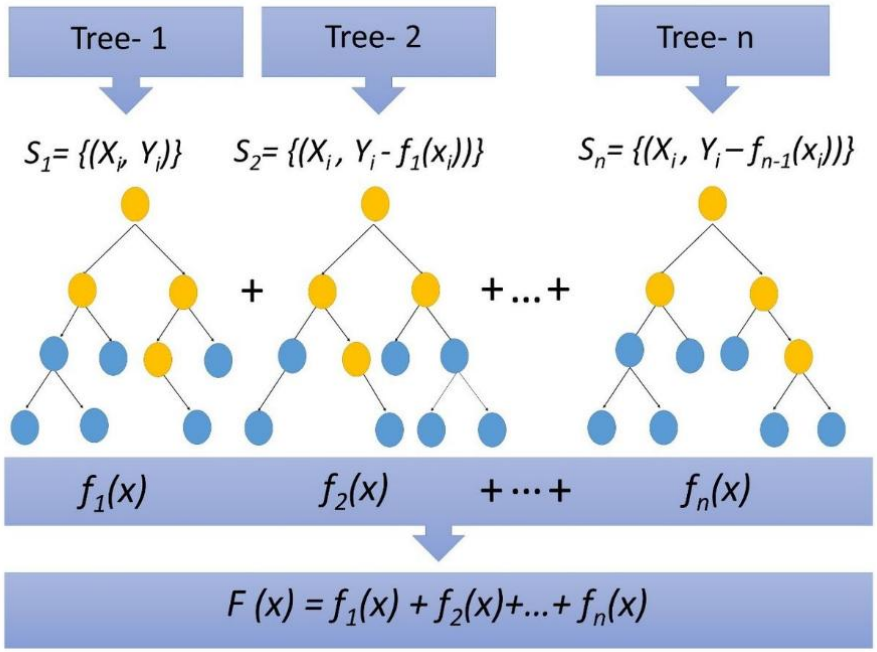


Figure 2.2 Schematic diagram of XGBoost algorithm

Table 2.2 Summary of Machine Learning Algorithms

Model	Tittle	Performance and Accuracy	Advantages	Limitation
Random Fores (RF)	Machine Learning Implementation for Estimation of Earthquake Magnitude using Strong Motion Data (Handayani et al., 2024)	High accuracy (>98.8% for intensity and >94.49% for magnitude estimation)	- Good in handling imbalanced data - High classification performance - Robust with noisy and incomplete data - Handle large data	- Overfitting with complex trees - Poor interpretability (black box)
XGBoost	Earthquake Prediction Model using Random Forest and Gradient Boosting Algorithm (Babu et al., 2024)	High accuracy >90% for seismic classification	- Reduce overfitting by regularization (L1/L2). - Good for large and complex datasets - Fast and scalable	- Sensitive to class imbalance - Difficult to interpret in critical disaster decisions

2.7 Research Gap

The growth of machine learning for earthquake analysis and prediction become crucial for generate the accurate real-time earthquake prediction and mitigation strategies. However, applying models like Random Forest and XGBoost within a geological framework showing the significant gap in most research. Most studies only focusing on statistical or geophysical technique without combining geological variables such as fault lines, rock types, and tectonic stress regimes into their models. This gap limits the interpretability and reduces the effectiveness of relevance predictions. Even though Random Forest and XGBoost have works well in managing high-dimensional, imbalanced and noisy data, their application is not fully used in predicting geological earthquakes because there is not enough collaboration between experts in this field and in machine learning.

Random Forest and XGBoost widely used in data science field especially for classification and regression analysis as their robustness and ability to address complex nonlinear relationships. Yet, their application in earthquake prediction often focuses only on seismic signals or magnitude forecasting without exploring the hidden spatial-temporal patterns influenced by geological factors. Moreover, most collected datasets limited in size, are not fully labelled and lack reliable geographical geology information which makes the models less accurate and useful in practice. This missing information challenges the use of geological concepts like stress increase, rock bending and fault connections in predicting seismic events.

To address this gap, the more comprehensive approach required in combining geology with various data science algorithms. Integrating geological datasets such as fault activity history into machine learning models like Random Forest and XGBoost can improve the performance of prediction and geophysical insight. Future research should aim to develop more advanced hybrid model with the large and various geological dataset that account for both geological processes and machine learning optimization. This approach can increase the accuracy of earthquakes prediction and provide clear insight and decision making especially for dealing with disasters mitigation strategies, cities planning and evaluating risks.

2.8 Chapter Summary

This chapter highlights the comprehensive review of earthquake formation and tectonic settings that highlights the main mechanism at various earthquakes and the worldwide distribution of seismic activity especially at the Pacific Ring of Fire and active fault zone such as East Anatolian and Sagaing faults. This chapter also describe the supervise machine learning model applications used for earthquake analysis and prediction using Random Forest and XGBoost for predicting and classifying earthquakes. There is a clear gap in research for combining geological information and machine learning indicating the necessity of using several disciplines to enhance prediction and decision-making abilities.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter discussed the comprehensive framework applied in this research to analysing and predicting the patterns of earthquake occurrences using machine learning algorithm of Random Forest and XGBoost. This research is built on earthquake events dataset collected from United State Geological Survey (USGS) dataset a globally recognized authority for seismic activity. The research methodology presents the detailed research process from the problem identification and initial study of the topic to the evaluation of the developed model. The evaluation performance of the machine learning model used also discussed in this chapter.

3.2 Research Framework

The methodology is designed into seven stages to analyze and predict earthquake events including problem identification and initial study, data collection, data pre-processing, exploratory data analysis (EDA), feature engineering, model development and evaluation and data visualization. The stages of this research framework are illustrated in flow diagram Figure 3.1. The work schedule of this research framework was shown in Appendix A.

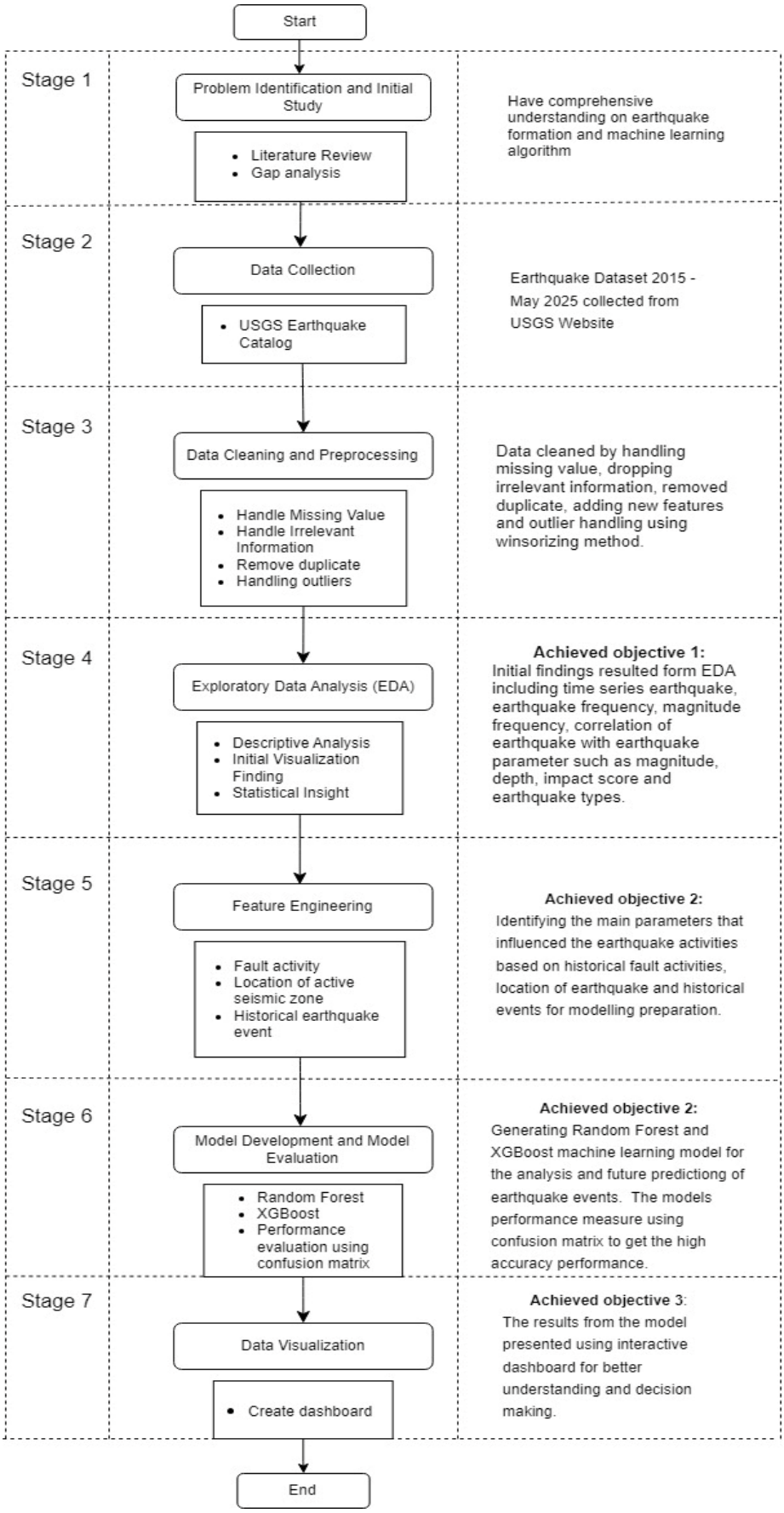


Figure 3.1 Flow diagram Framework

3.2.1 Stage 1: Problem identification and Initial study

This primary stage is a foundation of research study which is important before the research can be conducted. In this stage, the research problem was identified based on literature review of the related study of earthquakes and machine learning to gain information and understanding of the research field. The occurrences of earthquakes are unpredictable and become the main challenge for experts to minimize damage and casualties through data-driven predictions. This study aimed to predict the geological location of earthquake location on Ring of Fire.

3.2.2 Stage 2: Data Collection

The official USGS Earthquake Catalog from USGS website https://earthquake.usgs.gov/earthquakes/map is used in this research. This data has used by many researchers to train machine learning models for several algorithms such as Decision Tree, KNN, Random Forest, Gradient Boost, XG Boost, SVM, and Ridge Regression to predict earthquake magnitude (har et al., 2024)

The dataset extracted consist of recorded earthquake event from March 2015 until May 2025 consists of 17498 records and 22 features. The dataset extracted from 2015 until 2025 which is a decade to ensure comprehensiveness and effectiveness of predictions accuracy. The occurrence of earthquake is like a cycle in some active seismic zone and active fault.

The long-term history of earthquakes data is needed to predict the frequently affected areas, recurrence interval, location of epicentre and the historical of maximum magnitude to develop the reliable prediction model from machine learning. The good performance of machine learning will generate the accurate seismic hazard assessment and mitigation. However, the lack historical data will be led to the less accuracy of model performance and undermining preparedness and risk mitigation efforts. The

dataset extracted in CSV format to easier the process of handling and processing the dataset. Then, it exported into Google Colab for data processing and analysis.

3.2.3 Stage 3: Data Pre-processing

Data pre-processing is an important step to make sure the quality and reliability of the dataset before operating machine learning models. The cleaning process of both data includes deleting the unnecessary column, remove duplicate value and filled the missing rows with mean, median and mode. The missing values, outliers, inconsistent formats of the earthquake dataset must check to improve the accuracy of model. The outliers with unrealistic values such as negative or extreme need to remove to avoid skewed model training (Harirchian et al., 2021).

Data transformation used for converting raw data such as categorical data to numerical value to improve the performance of model. Label Encoding is a tools that used in this study to convert categorical data such as classification zone of ring of fire and active fault to numerical value that machine can learn and process it (Satish et al., 2025)

3.2.4 Stage 4: Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the early procedure of analyzing datasets to discover and grasp the key features through visualization and statistics before modelling. This process will uncover patterns, trends, visualize distribution, correlations and detect anomalies, missing value and outliers in the earthquake dataset then summarize it using descriptive statistics through plots such as plotting magnitude and depth distributions using histograms, box plots, and summary statistics (mean, median, skewness). This helps detecting imbalances data such as rare high-magnitude earthquakes, which may require resampling techniques (Senkaya et al., 2024).

The benefits of this process allow data scientists or seismologists to see the entire dataset clearly and handle the data cleaning, selecting the best features and the most effective modelling strategies. The application of EDA reduces the errors, enhances the accuracy and reduces time consuming in later analysis. Effectively, EDA will verify the fitness of the data for further use in the machine learning flow (Cui et al., 2024)

3.2.5 Stage 5: Feature Engineering

Feature engineering is the process of converting raw data into meaningful predictors before implementing it in Random Forest and XGBoost model. This step requires the adding or modifying features to expose the relationships and similarities that can increase the model's performance. This process involves extracting domain-specific insights from temporal, geospatial, and physical characteristics of seismic events (Senkaya et al., 2024).

Feature selection and dimensionality reduction methods are utilized to improve and optimize the feature set. Features like earthquake magnitude, depth, zone classification, types are choose to train the models. Correlation analysis, redundant features such as overlying temporal variables with high collinearity are identified and removed to increase model efficiency and decrease overfitting. Tree-based models of Random Forest and XGBoost provide strong feature importance scores allowing elimination of non-informative predictors to produce more reliable and accurate results by choosing the best splitting classes and geospatial trends (Kaftan, 2025)

3.2.6 Stage 6: Model Development and Evaluation

This stage involved the development and evaluation of Random Forest (RF) and XGBoost model. The cleaned dataset and the relevant feature engineering selected in will splitting into training and test set. The training and testing with range 80% training to ensures model has enough data to learn the pattern, relationships and trends within the dataset while the 20% test for evaluating the performance accuracy of the hidden data. (Senkaya et al., 2024). This range is enough for the model to perform well with sufficient data without having overfit or underfitting.

The training of Random Forest model involved the process of creating multiple decision trees from bootstrapped data and the prediction results taken from the average of the outcomes (Senkaya et al., 2024). Meanwhile, the XGBoost model built from the simple decision tree to more complex tree with the aim of fixing the errors of the previous gradient boosting. The hyperparameters of the number of trees, learning rate, and tree depth are modified either through cross-validation or an additional validation set to demonstrate the best-performing configuration. (Babu et al., 2024). Both models classifier will predict the classes zone of Ring of Fire into which the observation fell based on qualitative variables.

Afterward, RF and XGBoost model is evaluated using 20% test set to measure its generalization performance on different data. The confusion matric and key metrics of accuracy, precision, recall, and F1-score used for classification. Both models provide the importance features to identifying the best performance of the key metrics for the predictions.

3.2.7 Stage 7: Data Visualization

The final step for this research is data visualization and improvement. The results of the earthquake analysis and prediction from the modelling visualized using dashboard of Power BI. The interactive results allow the users to explore model outputs and key insights from the data for better understanding and enhance the accuracy of decision-making.

Graphical features such as confusion matrices, feature importance plots, and time-series plots assist users understand the results of the model easier and identify the factors influence the predictions. The spatial of earthquake data visualized into maps to demonstrate predicted risk areas or historical earthquake allocations, presenting the analysis more understandable and practical for decision-making.

3.3 Algorithm models

3.3.1 Random Forest (RF)

Random Forest is a group model that create bootstrap samples of training data to form each decision tree. Bagging (bootstrap aggregation) aims to minimize variance and overfitting which increases the model generalization abilities. It resulted the robust and dependable predictions about earthquakes by reducing bias and variance. The function is expressed in Equation 3.1.

$$\hat{y} = mode\ (h_1(x), h_2(x), \dots, h_n(x)) \qquad \text{Equation 3.1}$$

Where:

- n = number of trees,
- $h_i(x)$ = class prediction of the *i*-th decision tree
- $\hat{y}$ = final prediction.

3.3.2 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting or XGBoost model is generating from a simple decision tree to a more complex system tree that is designed to fixed the error from the original decision tree to create new tree. The outcomes from the trees will become predictions results. This model can measure based on Equation 3.2:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \qquad \text{Equation 3.2}$$

Where:

- $\widehat{y_i}$ = predicted value for instance *i*,
- f_k = function (a regression tree) added at the *k*-th step
- *K* = total number of trees.

The objective function that XGBoost minimizes is:

$$Obj = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{\kappa=1}^{\mathrm{K}} \Omega\,(f_k) \qquad \text{Equation 3.3}$$

Where:

- $l(y_i, \widehat{y_i})$ is the loss function (e.g., squared error for regression, log loss for classification),
- $\Omega\,(f_k) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum \omega_j^2$ is the regularization term that penalizes complexity (with T is number of leaves, ω_j^2 is leaf weight)

3.4 Performance Metrics:

3.4.1 Confusion Matrix

Confusion matric used to determine model performance for earthquakes events in binary classification for model assessment. In classification tasks, reliability of a certain model is traditionally evaluated using the measure of a few performance indicators. One of the major tools involved in these evaluations is the confusion matrix represented in Table 3.1 that indicates the true positives, stands for the false positives, represents false negatives, and is equivalent to true negatives.

Table 3.1 Confusion Matrix

Actual / Predicted	**Active Fault (Class 0)**	**Ring of Fire (Class 1)**
Active Fault (Class 0)	TN	FP
Ring of Fire (Class 1)	FN	TP

These measurements counting the precision, recall, accuracy and F1- Score were used to assess the efficacy of the specific model. The equations demonstrated in Equations 3.4 until Equation 3.7:

- **Accuracy:**

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Equation 3.4}$$

- **Precision**: Proportion of correctly identified positive cases.

$$Precision = \frac{TP}{TP + FP} \qquad \text{Equation 3.5}$$

- **Recall**: Sensitivity or True Positive Rate.

$$Recall = \frac{TP}{TP + FN}$$ Equation 3.6

- **F-1 Score :**

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$ Equation 3.7

Where:

- TP = true positive of predicted number
- TN = true negative of predicted number
- FP = false positive of predicted number
- FN = false negative of predicted number

3.5 Chapter Summary

This chapter outlines the methodology for analyzing and predicting earthquakes events using Random Forest and XGBoost model based on global earthquake dataset from USGS. The details of the steps discussed in this chapter include the primary study and problem identification, data collection, data pre-processing, exploration data analysis, feature engineering, model development and evaluation and visualization of results. The use of statistical and machine learning measuring performance matrices such as accuracy, precision, Recall and F1-score discussed in this chapter to validate the effectiveness in predicting seismic activities.

CHAPTER 4

EXPLORATORY DATA ANALYSIS (EDA)

4.1 Introduction

This chapter presents the initial findings from the exploratory data analysis (EDA) conducted on the Global Earthquake Dataset. The analysis aims to uncover patterns, relationships, and trends related to earthquake occurrences and their triggers, focusing on global patterns. The findings are presented on statistical summaries, visualizations, and machine learning techniques.

4.2 Overview of the Dataset

The earthquake dataset is obtained from global earthquake dataset from USGS websites and saved as .csv file with total of 17498 records and 22 features, which include date and time of the earthquakes, geographical location in terms of latitude and longitude on earth and the location of earthquakes to the epicentre with the factors that influence this activity such as magnitude, depth, dmin, rms, gap, nst, horizontalError, depthError magError, magNst and earthquake types.

Table 4.1 Key Earthquake Dataset

Attribute	Description	Data Type
time	Timestamp of the earthquake (UTC)	Object
latitude	Geographic latitude of the earthquake epicenter	Float64
longitude	Geographic longitude of the earthquake epicenter	Float64
depth	Depth of the earthquake in kilometers	Float64
mag	Magnitude of the earthquake	Float
magType	Type of magnitude measurement (e.g., ML, Mw, Mb)	Object

nst	Number of seismic stations used to determine the location	Float64
gap	Azimuthal gap (degrees) in station coverage around the epicenter	Float64
dmin	Minimum distance to the nearest station (in degrees)	Float64
rms	Root Mean Square value of the travel time residuals	Float64
net	Seismic network that provided the data	Object
id	Unique identifier for the event	Object
updated	Time the event was last updated	Object
place	Human-readable location description (e.g., 100km NW of Anchorage, Alaskaâ)	Object
type	Type of seismic event (e.g., earthquake, quarry blast)	Object
horizontalError	Horizontal location error (in km)	Float
depthError	Uncertainty of depth measurement (in km)	Float
magError	Uncertainty of the magnitude measurement	Float
magNst	Number of stations contributing to magnitude	Float
status	Review status of the data (e.g., reviewed, automatic)	Object
locationSource	Network or author who determined the location	Object
magSource	Network or author who determined the magnitude	Object

The dataset named as `df` for earthquake dataset and `fault` for faults dataset. These datasets read using prompt `.read_csv and`. The information of data presented in Figure 4.2 which used command `.info()`.

Earthquake dataset:

	time	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	...	updated	place	type	horizontalError	depthError	magError	magNst	status	locationSource	magSource
0	2025-05-05T21:14:02.950Z	64.6302	-17.4981	10.000	5.3	mww	91.0	42.0	1.650	0.36	...	2025-05-28T01:37:16.040Z	117 km WNW of Höfn, Iceland	earthquake	5.10	1.771	0.083	14.0	reviewed	us	us
1	2025-05-05T13:15:47.116Z	-28.2867	-176.5617	12.659	5.0	mb	61.0	78.0	1.534	0.59	...	2025-05-22T16:08:24.040Z	Kermadec Islands region	earthquake	10.68	3.974	0.074	58.0	reviewed	us	us
2	2025-05-05T10:53:27.776Z	23.9090	121.9671	27.000	5.6	mww	144.0	34.0	0.430	0.90	...	2025-06-09T13:07:21.412Z	37 km ESE of Hualien City, Taiwan	earthquake	4.76	1.822	0.053	34.0	reviewed	us	us
3	2025-05-05T10:09:59.032Z	23.9405	122.0201	10.000	5.3	mb	133.0	34.0	0.453	0.83	...	2025-05-29T10:20:13.704Z	42 km E of Hualien City, Taiwan	earthquake	3.65	1.787	0.033	315.0	reviewed	us	us
4	2025-05-05T09:46:46.882Z	-29.4675	-71.9242	22.933	5.1	mww	125.0	71.0	0.747	0.98	...	2025-06-09T13:12:30.040Z	78 km NW of Coquimbo, Chile	earthquake	4.03	3.521	0.083	14.0	reviewed	us	us

5 rows × 22 columns

Figure 4.1 Earthquake dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17498 entries, 0 to 17497
Data columns (total 22 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   time             17498 non-null  object
 1   latitude         17498 non-null  float64
 2   longitude        17498 non-null  float64
 3   depth            17498 non-null  float64
 4   mag              17498 non-null  float64
 5   magType          17498 non-null  object
 6   nst              5095 non-null   float64
 7   gap              17333 non-null  float64
 8   dmin             17324 non-null  float64
 9   rms              17497 non-null  float64
 10  net              17498 non-null  object
 11  id               17498 non-null  object
 12  updated          17498 non-null  object
 13  place            17498 non-null  object
 14  type             17498 non-null  object
 15  horizontalError  17376 non-null  float64
 16  depthError       17498 non-null  float64
 17  magError         16080 non-null  float64
 18  magNst           16182 non-null  float64
 19  status           17498 non-null  object
 20  locationSource   17498 non-null  object
 21  magSource        17498 non-null  object
dtypes: float64(12), object(10)
memory usage: 2.9+ MB
```

Figure 4.2 Earthquake info

4.3 Data Cleaning

Data cleaning is important to avoid hopeless inconsistency, error and misleading information in the data that will be analysed into machine learning algorithms. Incomplete and invalid data can lead to lack accuracy, biased models and poor performance affecting the credibility of the decision-making (Harirchian et al., 2021). The datasets of both earthquake and fault cleaned differently before both data merged and undergoes pre-processing step.

i. Check the Column

Prompt of `.columns` used to check the columns by printing the columns names of a data frame in Pandas which is a two-dimensional data structure. The DataFrame output of `df`, contains several columns relevant that are related to earthquake occurrences. This type of columns, for example `'time'`, `'latitude'`, `'longitude'`, `'depth'`, `'mag'`, `'magType'`, `'nst'`, `'gap'`, `'dmin'`, `'rms'`, `'net'`, `'id'`, `'update'`, `'place','type'`, `'horizontalError'`, `'depthError'`, `'magError'`, `'magNST','status'`, `'locationSource'`, `'magSource'`, that stores different types of data. The used of `dtype=object` indicate the columns may containing string, integer or date values such in Figure 4.3.

```
df.columns
```

```
Index(['time', 'latitude', 'longitude', 'depth', 'mag', 'magType', 'nst',
       'gap', 'dmin', 'rms', 'net', 'id', 'updated', 'place', 'type',
       'horizontalError', 'depthError', 'magError', 'magNst', 'status',
       'locationSource', 'magSource'],
      dtype='object')
```

Figure 4.3 Earthquake columns checked

ii. Dropping irrelevant column

Then, the unnecessary columns such as : `'nst','dmin'`, `'net'`, `'id'`, `'magNST','update`, `'status'`, `'locationSource'`, `'magSource'` dropped from the original data frame using `prompt .drop ()` to avoid from less accuracy of the model. This presented in Figure 4.4.

```
df = df.drop(['nst','dmin', 'magNst', 'id','updated','net','status','locationSource','magSource'], axis=1)
```

Figure 4.4 Remove irrelevant column of earthquakes

iii. **Change datetime format**

The `'time'` column changed to the datetime data type shown in Figure 4.5 The `pd.to_datetime()` function used as conversion function that is significant when performing date-based analysis and operations.

```
df['time'] = pd.to_datetime(df['time'])
```

Figure 4.5 Change datetime format of earthquake dataset

4.4 Pre-Processing

Pre-processing stage is important for maintaining the reliability and accuracy of modelling from the cleaned earthquake raw data. After cleaning process, this step will generated new features for providing additional information which can enhanced the analysis and model performance (Harirchian et al., 2021).

4.4.1 Earthquake Dataset

i. **Adding new column of Country**

To provide a geographical context to the earthquake data, a new column of country created using reverse geocoding enhancement of the same latitude and longitude data. The `reverse_geocoder` library used to enabled conversion of geographic coordinates to respective country codes. It was decided to write a custom function to retrieve the coordinates in every row and do the geocoding lookup and extract the two letters country code related to each event. This was then used over all the data set using the `.apply()` method and the output in the form of the country codes were saved under a new column of `'Country'`. Figure 4.6 indicates the codes and the output of the conversion based on longitude and latitude coordinate. This preprocessing activity allows analysing the data according to the location and makes the dataset easier to interpret.

```
#  generate new column of country based on longitude and latitude

!pip install reverse_geocoder

import reverse_geocoder as rg

def get_country(row):
  coordinates = (row['Latitude'], row['Longitude'])
  result = rg.search(coordinates)
  return result[0]['cc']

df['Country'] = df.apply(get_country, axis=1)
```

Figure 4.6 New column of country

ii. **Rename the country code**

Figure 4.7 shows the step of data preprocessing was replacing country codes in the column of the '`Country`' with full country names to make the data more readable and clearer in interpretation. A dictionary, henceforward referred to as `country_code_to_name`, was created containing each two-letter code country such as 'US', 'JP', 'ID' and the full designation are 'United States', 'Japan', 'Indonesia'. This replacement was implemented by using the column `Country` in DataFrame earthquakes and `.replace()` function and resulted in the dataset in which the full name of countries replaced the shorter codes of countries.

```
# Change country codes to country names
country_code_to_name = {
    'AF': 'Afghanistan', 'AX': 'Aland Islands', 'AL': 'Albania', 'DZ': 'Algeria', 'AS': 'American Samoa', 'AD': 'Andorra', 'AO': 'Angola', 'AI': 'Anguilla', 'AQ': 'Antarctica',
    'AG': 'Antigua and Barbuda','AR': 'Argentina', 'AM': 'Armenia', 'AW': 'Aruba', 'AU': 'Australia', 'AT': 'Austria', 'AZ': 'Azerbaijan','BS': 'Bahamas', 'BH': 'Bahrain',
    'BD': 'Bangladesh', 'BB': 'Barbados', 'BY': 'Belarus', 'BE': 'Belgium','BZ': 'Belize', 'BJ': 'Benin', 'BM': 'Bermuda', 'BT': 'Bhutan', 'BO': 'Bolivia',
    'BQ': 'Bonaire, Sint Eustatius and Saba','BA': 'Bosnia and Herzegovina', 'BW': 'Botswana', 'BV': 'Bouvet Island', 'BR': 'Brazil', 'IO': 'British Indian Ocean Territory',
    'BN': 'Brunei Darussalam', 'BG': 'Bulgaria', 'BF': 'Burkina Faso', 'BI': 'Burundi', 'CV': 'Cabo Verde','KH': 'Cambodia', 'CM': 'Cameroon', 'CA': 'Canada',
    'KY': 'Cayman Islands', 'CF': 'Central African Republic','TD': 'Chad', 'CL': 'Chile', 'CN': 'China', 'CX': 'Christmas Island', 'CC': 'Cocos (Keeling) Islands',
    'CO': 'Colombia', 'KM': 'Comoros', 'CG': 'Congo', 'CD': 'Congo, The Democratic Republic of the','CK': 'Cook Islands', 'CR': 'Costa Rica', 'CI': 'Cote d\'Ivoire',
    'HR': 'Croatia', 'CU': 'Cuba', 'CW': 'Curacao','CY': 'Cyprus', 'CZ': 'Czech Republic', 'DK': 'Denmark', 'DJ': 'Djibouti', 'DM': 'Dominica', 'DO': 'Dominican Republic',
    'EC': 'Ecuador', 'EG': 'Egypt', 'SV': 'El Salvador', 'GQ': 'Equatorial Guinea','ER': 'Eritrea', 'EE': 'Estonia', 'ET': 'Ethiopia', 'FK': 'Falkland Islands (Malvinas)',
    'FO': 'Faroe Islands','FJ': 'Fiji', 'FI': 'Finland', 'FR': 'France', 'GF': 'French Guiana', 'PF': 'French Polynesia', 'TF': 'French Southern Territories', 'GA': 'Gabon',
    'GM': 'Gambia', 'GE': 'Georgia', 'DE': 'Germany','GH': 'Ghana', 'GI': 'Gibraltar', 'GR': 'Greece', 'GL': 'Greenland', 'GD': 'Grenada', 'GP': 'Guadeloupe','GU': 'Guam',
    'GT': 'Guatemala', 'GG': 'Guernsey', 'GN': 'Guinea', 'GW': 'Guinea-Bissau', 'GY': 'Guyana','HT': 'Haiti', 'HM': 'Heard Island and McDonald Islands',
    'VA': 'Holy See (Vatican City State)', 'HN': 'Honduras','HK': 'Hong Kong', 'HU': 'Hungary', 'IS': 'Iceland', 'IN': 'India', 'ID': 'Indonesia', 'IR': 'Iran',
    'IQ': 'Iraq', 'IE': 'Ireland', 'IM': 'Isle of Man', 'IL': 'Israel', 'IT': 'Italy', 'JM': 'Jamaica', 'JP': 'Japan','JE': 'Jersey', 'JO': 'Jordan', 'KZ': 'Kazakhstan',
    'KE': 'Kenya', 'KI': 'Kiribati', 'KP': 'Korea, Democratic People\'s Republic of','KR': 'Korea, Republic of', 'KW': 'Kuwait', 'KG': 'Kyrgyzstan',
    'LA': 'Lao People\'s Democratic Republic','LV': 'Latvia', 'LB': 'Lebanon', 'LS': 'Lesotho', 'LR': 'Liberia', 'LY': 'Libya', 'LI': 'Liechtenstein','LT': 'Lithuania',
    'LU': 'Luxembourg', 'MO': 'Macao', 'MK': 'Macedonia','MG': 'Madagascar', 'MW': 'Malawi', 'MY': 'Malaysia', 'MV': 'Maldives', 'ML': 'Mali', 'MT': 'Malta',
    'MH': 'Marshall Islands', 'MQ': 'Martinique', 'MR': 'Mauritania', 'MU': 'Mauritius', 'YT': 'Mayotte','MX': 'Mexico', 'FM': 'Micronesia, Federated States of',
    'MD': 'Moldova, Republic of', 'MC': 'Monaco', 'MN': 'Mongolia', 'ME': 'Montenegro', 'MS': 'Montserrat', 'MA': 'Morocco', 'MZ': 'Mozambique','MM': 'Myanmar', 'NA': 'Namibia',
    'NR': 'Nauru', 'NP': 'Nepal', 'NL': 'Netherlands', 'NC': 'New Caledonia','NZ': 'New Zealand', 'NI': 'Nicaragua', 'NE': 'Niger', 'NG': 'Nigeria', 'NU': 'Niue',
    'NF': 'Norfolk Island','MP': 'Northern Mariana Islands', 'NO': 'Norway', 'OM': 'Oman', 'PK': 'Pakistan', 'PW': 'Palau','PS': 'Palestine, State of', 'PA': 'Panama',
    'PG': 'Papua New Guinea', 'PY': 'Paraguay', 'PE': 'Peru', 'PH': 'Philippines', 'PN': 'Pitcairn', 'PL': 'Poland', 'PT': 'Portugal', 'PR': 'Puerto Rico', 'QA': 'Qatar',
    'RE': 'Reunion', 'RO': 'Romania', 'RU': 'Russian Federation', 'RW': 'Rwanda', 'BL': 'Saint Barthelemy','SH': 'Saint Helena, Ascension and Tristan da Cunha',
    'KN': 'Saint Kitts and Nevis', 'LC': 'Saint Lucia','MF': 'Saint Martin (French part)', 'PM': 'Saint Pierre and Miquelon', 'VC': 'Saint Vincent and the Grenadines',
    'WS': 'Samoa', 'SM': 'San Marino', 'ST': 'Sao Tome and Principe', 'SA': 'Saudi Arabia', 'SN': 'Senegal', 'RS': 'Serbia', 'SC': 'Seychelles', 'SL': 'Sierra Leone',
    'SG': 'Singapore', 'SX': 'Sint Maarten (Dutch part)', 'SK': 'Slovakia', 'SI': 'Slovenia', 'SB': 'Solomon Islands', 'SO': 'Somalia', 'ZA': 'South Africa',
    'GS': 'South Georgia and the South Sandwich Islands', 'SS': 'South Sudan', 'ES': 'Spain', 'LK': 'Sri Lanka','SD': 'Sudan', 'SR': 'Suriname', 'SJ': 'Svalbard and Jan Mayen',
    'SZ': 'Swaziland', 'SE': 'Sweden','CH': 'Switzerland', 'SY': 'Syrian Arab Republic', 'TW': 'Taiwan, Province of China', 'TJ': 'Tajikistan','TZ': 'Tanzania', 'TH': 'Thailand',
    'TL': 'Timor-Leste', 'TG': 'Togo', 'TK': 'Tokelau','TO': 'Tonga', 'TT': 'Trinidad and Tobago', 'TN': 'Tunisia', 'TR': 'Turkey', 'TM': 'Turkmenistan',
    'TC': 'Turks and Caicos Islands', 'TV': 'Tuvalu', 'UG': 'Uganda', 'UA': 'Ukraine', 'AE': 'United Arab Emirates','GB': 'United Kingdom', 'US': 'United States',
    'UM': 'United States Minor Outlying Islands', 'UY': 'Uruguay','UZ': 'Uzbekistan', 'VU': 'Vanuatu', 'VE': 'Venezuela, Bolivarian Republic of', 'VN': 'Viet Nam',
    'VG': 'Virgin Islands, British', 'VI': 'Virgin Islands, U.S.', 'WF': 'Wallis and Futuna', 'EH': 'Western Sahara','YE': 'Yemen', 'ZM': 'Zambia', 'ZW': 'Zimbabwe'
}
```

Figure 4.7 Define Country Code

```
# Replace country codes with full names
df['Country'] = df['Country'].replace(country_code_to_name)
df.head()
```

	time	latitude	longitude	depth	mag	magType	gap	rms	place	type	horizontalError	depthError	magError	Country
0	2025-05-05 21:14:02.950000+00:00	64.6302	-17.4981	10.000	5.3	mww	42.0	0.36	117 km WNW of Höfn, Iceland	earthquake	5.10	1.771	0.083	Iceland
1	2025-05-05 13:15:47.116000+00:00	-28.2867	-176.5617	12.659	5.0	mb	78.0	0.59	Kermadec Islands region	earthquake	10.68	3.974	0.074	Tonga
2	2025-05-05 10:53:27.776000+00:00	23.9090	121.9671	27.000	5.6	mww	34.0	0.90	37 km ESE of Hualien City, Taiwan	earthquake	4.76	1.822	0.053	Taiwan
3	2025-05-05 10:09:59.032000+00:00	23.9405	122.0201	10.000	5.3	mb	34.0	0.83	42 km E of Hualien City, Taiwan	earthquake	3.65	1.787	0.033	Taiwan
4	2025-05-05 09:46:46.882000+00:00	-29.4675	-71.9242	22.933	5.1	mww	71.0	0.98	78 km NW of Coquimbo, Chile	earthquake	4.03	3.521	0.083	Chile

Figure 4.8 Rename Country Code to Full Name Country

iii. Adding a new column of Continent

New column of continent generated based on countries name. The `pycountry_convert` package is utilized to convert country names to their respective continents names making it easy to absorb common semantics flaws such the United States is renamed North America. Function `get_continent` used to convert the country names to ISA Alpha-2 codes, which are then mapped to continent codes and mapped to continent labels. This process provides a dictionary that can be used to assign individual countries to the proper continent and to create a new column of Continental into the DataFrame to allow geographical analysis (Figure 4.9).

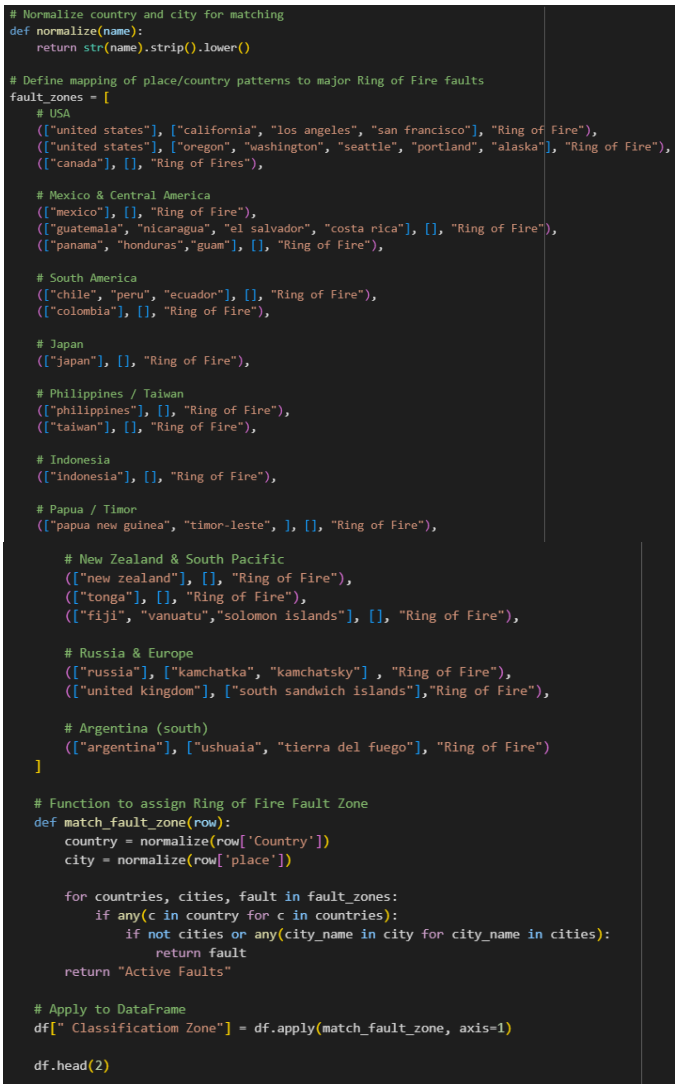
```
!pip install pycountry_convert
import pycountry_convert as pc

# Function to map country name to continent
def get_continent(country_name):
    try:
        # Handle name discrepancies manually
        rename_dict = {
            "Russian Federation": "Russia",
            "Iran, Islamic Republic of": "Iran",
            "Venezuela, Bolivarian Republic of": "Venezuela",
            "Korea, Republic of": "South Korea",
            "Korea, Democratic People's Republic of": "North Korea",
            "Syrian Arab Republic": "Syria",
            "Taiwan, Province of China": "Taiwan",
            "Viet Nam": "Vietnam",
            "United States": "United States of America",
            "Micronesia, Federated States of": "Micronesia",
            "Tanzania, United Republic of": "Tanzania",
            "Macedonia, The Former Yugoslav Republic of": "North Macedonia",
            "Congo, The Democratic Republic of": "Democratic Republic of the Congo"

        }
        if country_name in rename_dict:
            country_name = rename_dict[country_name]
        country_code = pc.country_name_to_country_alpha2(country_name)
        continent_code = pc.country_alpha2_to_continent_code(country_code)
        continent_name = pc.convert_continent_code_to_continent_name(continent_code)
        return continent_name
    except:
        return "Unknown"

# Map countries to continents
continent_mapping = {country: get_continent(country) for country in countries}
continent_mapping_sorted = dict(sorted(continent_mapping.items(), key=lambda x: x[1]))
df['Continental'] = df['Country'].map(continent_mapping_sorted)
```

Figure 4.9 New column of continental

iv. Adding a new column of Ring of Fire Zone

Figure 4.10 shows the country that related to the Ring of Fire classified based on the collision of two major tectonics (Figure 2.1) including Pacific Plate and Nazca Plate that form the San Andreas Fault and Motagua Fault in United State and Guatemala, Australian Plate and Eurasian Plate formed Sunda Trench and Papua New Guinea Trench in Indonesia, Timor – Leste and Papua New Guinea, Nazca Plate and South American Plate generate Peru-Chile (Nazca) Trench in Ecuador, Chile and Peru. The meeting of Cocos Plate with Pacific Plate form middle America Trench in Mexico and Pacific plate meet with Australia Plate in Tonga, New Zealand, Fiji and Vanuatu, Philippine Sea meet with Pacific Plate and Eurasian Plate in Philiphine, Taiwan, Japan and Guam, Scotia Plate with South America Plate in Argentina. A part of Russia caused by collision of Pacific Plate and Okhotsk Plate while Panama and Honduras caused by collision of Cocos Plate and Caribbean Plate (Wang et.al., 2018, Dragoni and Santini 2022, Nakamura et.al 2018, Biswas et.al 2023).

Prompt `.apply()` used to classify these fault name and generate new column of `Zone_Classification`. Subduction zone also classified as the Ring of Fire zone due to the active fault collision and the volcanic activities such as South Sandwich Island and Kuril Islands (Biswas et.al 2023).



```
# Normalize country and city for matching
def normalize(name):
    return str(name).strip().lower()

# Define mapping of place/country patterns to major Ring of Fire faults
fault_zones = [
    # USA
    (["united states"], ["california", "los angeles", "san francisco"], "Ring of Fire"),
    (["united states"], ["oregon", "washington", "seattle", "portland", "alaska"], "Ring of Fire"),
    (["canada"], [], "Ring of Fires"),

    # Mexico & Central America
    (["mexico"], [], "Ring of Fire"),
    (["guatemala", "nicaragua", "el salvador", "costa rica"], [], "Ring of Fire"),
    (["panama", "honduras","guam"], [], "Ring of Fire"),

    # South America
    (["chile", "peru", "ecuador"], [], "Ring of Fire"),
    (["colombia"], [], "Ring of Fire"),

    # Japan
    (["japan"], [], "Ring of Fire"),

    # Philippines / Taiwan
    (["philippines"], [], "Ring of Fire"),
    (["taiwan"], [], "Ring of Fire"),

    # Indonesia
    (["indonesia"], [], "Ring of Fire"),

    # Papua / Timor
    (["papua new guinea", "timor-leste", ], [], "Ring of Fire"),

    # New Zealand & South Pacific
    (["new zealand"], [], "Ring of Fire"),
    (["tonga"], [], "Ring of Fire"),
    (["fiji", "vanuatu","solomon islands"], [], "Ring of Fire"),

    # Russia & Europe
    (["russia"], ["kamchatka", "kamchatsky"] , "Ring of Fire"),
    (["united kingdom"], ["south sandwich islands"],"Ring of Fire"),

    # Argentina (south)
    (["argentina"], ["ushuaia", "tierra del fuego"], "Ring of Fire")
]

# Function to assign Ring of Fire Fault Zone
def match_fault_zone(row):
    country = normalize(row['Country'])
    city = normalize(row['place'])

    for countries, cities, fault in fault_zones:
        if any(c in country for c in countries):
            if not cities or any(city_name in city for city_name in cities):
                return fault
    return "Active Faults"

# Apply to DataFrame
df[" Classificatiom Zone"] = df.apply(match_fault_zone, axis=1)

df.head(2)
```

Figure 4.10 Adding Column Zone Classification

v. Dropping all the duplicated rows

Duplicate rows were deleted in a dataset (Figure 4.11) to maintain data quality and integrity by using the `` `drop_duplicates()` `` method of Pandas library. This method searches the whole DataFrame and drops the row that is a complete copy of another in all columns. With prompt of `` `inplace=True` ``, the process is done in place of changing the original DataFrame rather than returning a copy of it. The command `df.info()` was then entered to show a short summary of the augmented DataFrame with number of non-null values, data types of columns as well as memory consumption. The result showed that few duplicates were left (final number of 17,479 rows and 16 columns), which implied that the dataset was cleaned

and no duplicated data. The step is essential in avoidance of data distortions in further analyses and model training.

```
df.drop_duplicates(inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 17497 entries, 0 to 17497
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   time                  17497 non-null  datetime64[ns, UTC]
 1   latitude              17497 non-null  float64
 2   longitude             17497 non-null  float64
 3   depth                 17497 non-null  float64
 4   mag                   17497 non-null  float64
 5   magType               17497 non-null  object
 6   gap                   17497 non-null  float64
 7   rms                   17497 non-null  float64
 8   place                 17497 non-null  object
 9   type                  17497 non-null  object
 10  horizontalError       17497 non-null  float64
 11  depthError            17497 non-null  float64
 12  magError              17497 non-null  float64
 13  Country               17497 non-null  object
 14  Continental           17479 non-null  object
 15   Classificatiom Zone  17497 non-null  object
dtypes: datetime64[ns, UTC](1), float64(9), object(6)
memory usage: 2.3+ MB
```

Figure 4.11 Removed Duplicate

4.4.2 Discretization Task

i. Label Encoder

Categorical columns such as zone classification, continent, earthquake type and magnitude type was checked using `prompt .unique()` shown in Figure 4.12. Then, both categorical columns handled using label encoding technique that converted the unique integer into numerical values based on alphabetic ordering. The `LabelEncoder` was import from `sklearn.processing` library and `.fit_transform()` converted each unique categories to numerical value starting from 0. The results shown in Figure 4.13 indicate the continent, continent, earthquake type change from categorical to numerical value range from 0 until 9, while the zone classification change the string of active fault zone to 0 and ring of fire to 1.

```
print("Unique values for 'Continental':", df_new['Continental'].unique())
print("Unique values for 'Classification Zone':", df_new[' Classification Zone'].unique())
print("Unique values for 'type':", df_new['type'].unique())
print("Unique values for 'magType':", df_new['magType'].unique())
```

```
Unique values for 'Continental': ['Europe' 'Oceania' 'Asia' 'South America' 'North America' 'Africa'
 'Antarctica']
Unique values for 'Classification Zone': ['Active Faults' 'Ring of Fire']
Unique values for 'type': ['earthquake' 'volcanic eruption' 'nuclear explosion']
Unique values for 'magType': ['mww' 'mb' 'ml' 'mwr' 'mw' 'mwb' 'mwp' 'ml(texnet)' 'mwc' 'ms_20']
```

Figure 4.12 Identify unique value

```
#Label encoded fault proximity
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()

# Fit and transform the 'fault_proximity' column
df_new['Continent_encoded'] = label_encoder.fit_transform(df_new['Continental'])

df_new.drop('Continental', axis=1, inplace=True)
print("Label encoding mapping for 'Continent':")
for label, code in zip(label_encoder.classes_, label_encoder.transform(label_encoder.classes_)):
    print(f"{label} → {code}")
```

```
Label encoding mapping for 'Continent':
Africa → 0
Antarctica → 1
Asia → 2
Europe → 3
North America → 4
Oceania → 5
South America → 6
```

```
# Encode the zone classification
le = LabelEncoder()
df_new['Zone Classification'] = le.fit_transform(df[' Classification Zone'])
##print("Label encoding mapping for 'Zone Class':")
for label, code in zip(le.classes_, le.transform(le.classes_)):
    print(f"{label} → {code}")
```

```
Active Faults → 0
Ring of Fire → 1
```

```
] # Encode the zone classification
le = LabelEncoder()
df_new['Types of earthquake'] = le.fit_transform(df_new['type'])
df_new.drop('type', axis=1, inplace=True)
##print("Label encoding mapping for 'Zone Class':")
for label, code in zip(le.classes_, le.transform(le.classes_)):
    print(f"{label} → {code}")
```

```
earthquake → 0
nuclear explosion → 1
volcanic eruption → 2
```

```
# Encode the magtype classification
le = LabelEncoder()
df_new['Magnitude Type'] = le.fit_transform(df_new['magType'])
df_new.drop('magType', axis=1, inplace=True)
##print("Label encoding mapping for 'Zone Class':")
for label, code in zip(le.classes_, le.transform(le.classes_)):
    print(f"{label} → {code}")
```

```
mb → 0
ml → 1
ml(texnet) → 2
ms_20 → 3
mw → 4
mwb → 5
mwc → 6
mwp → 7
mwr → 8
mww → 9
```

Figure 4.13 Label encoding

4.5 Descriptive Statistics

Figure 4.14 consists of descriptive statistics which provide valuable information concerning the distributional characteristics and the scales of measurement of each of the numeric variables. The variability of depth of earthquakes is quite strong with the minimum of -1.01 km to the maximum of 670.81 km and the standard deviation being the highest of 103.5. This indicate strong disparity of seismic depth where the earthquake event can occur in various and unpredictable depth. Comparatively, magnitude (mag) has a considerably medium range of 5.0 to 8.3 and a mean magnitude of 5.33 suggesting that most of the events fall into the moderate to mildly intense category. The gap (azimuthal gap) in the station coverage, varies in the range of 8 and 340 degrees with the average of 59.87, thus, indicating the differences in the spatial limitation of seismic events.

In addition, rms and the three error measurements such as horizontalError, depthError, and magError have significant dispersion. DepthError specifically reaches a large number of 32.57, of which the impact on depth reliability can be significant. Some of the categorical variables such as Continent_encoded, Zone Classification, Types of Earthquake, and Magnitude Type are encoded in the numberic range where the Zone Classification variable is a binary code 0 or 1, and the variable Magnitude Type is in the range 0 to 9. These wide variabilities on these coded fields show that there is wide regional and the kind of event differences. This heterogeneity is desirable in developing general purpose machine-learning models. All these descriptive statistics show that the data is quite variant and thus requires close supervision of outliers.

```
analysis.describe()
```

	depth	mag	gap	rms	horizontalError	depthError	magError	Continent_encoded	Zone Classification	Types of earthquake	Magnitude Type
count	17497.000000	17497.000000	17497.000000	17497.000000	17497.000000	17497.00000	17497.000000	17497.000000	17497.000000	17497.000000	17497.000000
mean	52.992789	5.333634	59.866380	0.819244	7.605025	2.88626	0.065819	3.800480	0.660056	0.006458	5.082300
std	103.533028	0.404520	34.249738	0.231617	2.398491	1.84488	0.024940	1.731002	0.473703	0.112712	4.391958
min	-1.010000	5.000000	8.000000	0.060000	0.080000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	10.000000	5.100000	35.000000	0.660000	6.100000	1.80000	0.050000	2.000000	0.000000	0.000000	0.000000
50%	11.990000	5.200000	53.000000	0.800000	7.500000	1.90000	0.065000	5.000000	1.000000	0.000000	9.000000
75%	45.480000	5.500000	76.000000	0.960000	9.100000	4.00000	0.078000	5.000000	1.000000	0.000000	9.000000
max	670.810000	8.300000	340.000000	2.820000	21.100000	32.57000	0.563000	6.000000	1.000000	2.000000	9.000000

Figure 4.14 Descriptive Statistics

4.5.1 Identify the Outlier using Boxplot

The boxplots and descriptive statistics shown that most of the observations of rms, horizontalError, depthError, and magError are within a reasonable range, but there are a number of outlier especially in depthError that maximum is 32.57 and horizontalError of maximum value is 21. The presence of such outliers indicates that there is a group of earthquakes in which we found it necessary to use high uncertainty most probably due to poor sensor coverage or challenges with depth calibration. Nevertheless, the majority of data fall into limited interquartile ranges, such as magError range between 0.05 and 0.078. The insignificant difference can affect the results of classification that need to handling these outliers since they could indicate a rare phenomenon or data quality problems that can affect model performance.

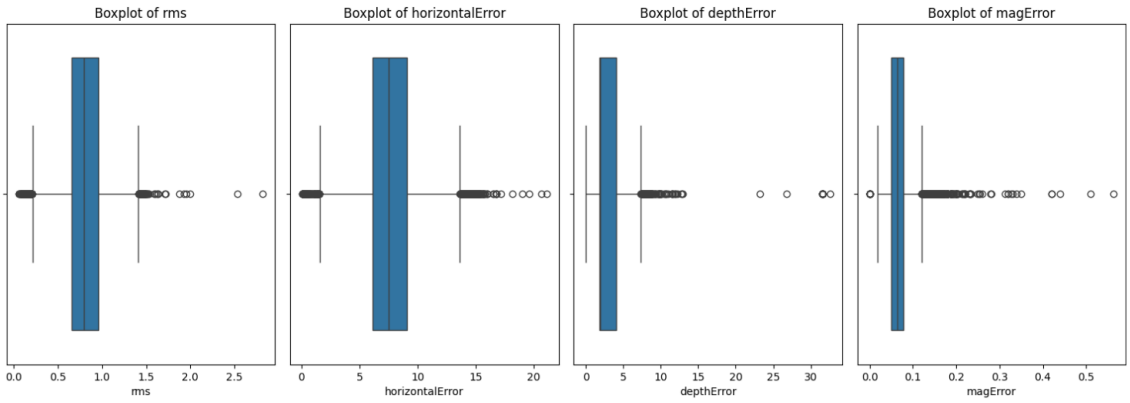


Figure 4.15 Box plot of outliers

The library function of `from scipy.stats.mstats import winsorize` used for winsorize the outliers while `import matplotlib.pyplot as plt` and `import seaborn as sns` used plotting the boxplot graph such in Figure 4.16. Then, the existing of this outlier handled using winsorization without removed the existed extreme values. The standardized data resolved by applying winsorization with function `mstats.winsorize` and `limit=[0.05, 0.05)`. The boxplot chart plotted to differentiate the results before and after the outliers handled.

```python
from scipy.stats.mstats import winsorize

# Winsorize the specified columns to handle outliers
columns_to_winsorize = ['rms', 'horizontalError', 'depthError', 'magError']
for col in columns_to_winsorize:
    df[col] = winsorize(df[col], limits=[0.05, 0.05]) # Winsorize at the 5th and 95th percentile

# Display descriptive statistics to see the effect of winsorization
print("\nDescriptive Statistics After Winsorizing:")
display(df[columns_to_winsorize].describe())

# Optional: Visualize the boxplots again to see the effect of winsorization
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(15, 5))
for i, col in enumerate(columns_to_winsorize):
    plt.subplot(1, 4, i + 1)
    sns.boxplot(x=df[col])
    plt.title(f'Boxplot of {col} After Winsorizing')
plt.tight_layout()
plt.show()
```

```
Descriptive Statistics After Winsorizing:
```

	rms	horizontalError	depthError	magError
count	17497.000000	17497.000000	17497.000000	17497.000000
mean	0.818836	7.598919	2.839780	0.064627
std	0.209936	2.133354	1.460572	0.019115
min	0.480000	3.800000	1.600000	0.033000
25%	0.660000	6.100000	1.800000	0.050000
50%	0.800000	7.500000	1.900000	0.065000
75%	0.960000	9.100000	4.000000	0.078000
max	1.240000	11.700000	6.035000	0.102000

Figure 4.16 Handling Outliers

Figure 4.17 shows the results of the Winsorizing procedure which restricts extreme values and minimizes the role of outliers. A notable change occurs in the following in all four variables: there are no outliers, and the interquartile ranges are fully reduced, indicating that extreme values of the data may be successfully kept under control. The remaining distributions are mildly skewed to the right in an rms and depthError contrast, although horizontalError and magError are more symmetric and balanced. Winsorization creates a tighter unambiguous distribution with more readable pattern than those that were without, and thus a clearer more coherent pattern in subsequent analyses can be interpreted accurately.

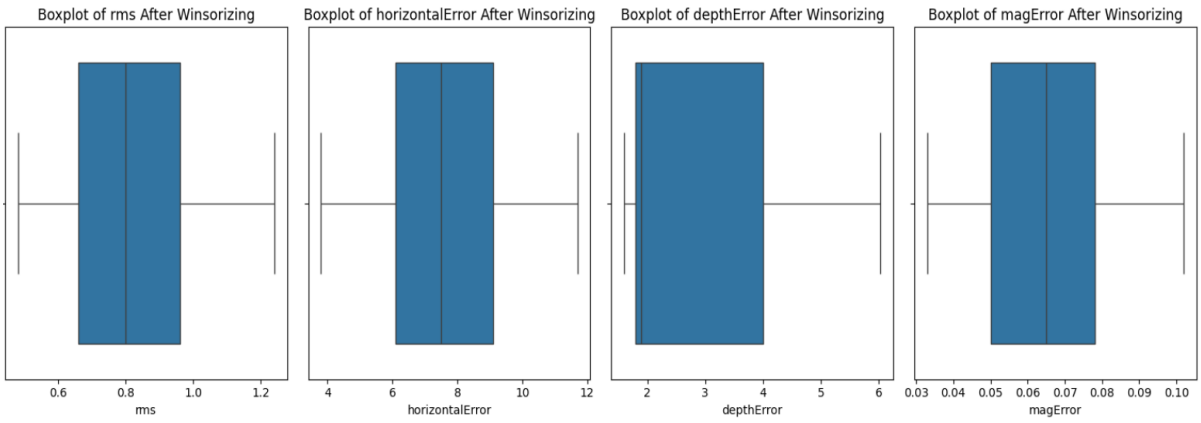


Figure 4.17 Boxplot after winsorizing

4.5.2 Depth and Magnitude Distribution

Figure 4.18 shows the histogram of key features distribution of earthquake activities including earthquake magnitude, depth and impact score. Earthquake Magnitude graph shows that most of the recorded earthquake falls in the moderate range which is between magnitudes 4.5 and 6.0. The distribution is highly right skewed indicates the high and low magnitude is the less frequent the events. This trend can be considered paralleling the overall seismology of the world, in which relatively moderate earth movements are more common than the devasting earthquakes. The maximum on magnitude 5.0 indicates that it is the most frequent magnitude interval recorded in the data.

The Depth (km) histogram shows most of the earthquakes are formed in shallow depths in the upper 100 kilometers of the earth crust. Events frequency decreases significantly with depth as shallower earthquakes have a higher rate and high chances to be detected and recorded. This distribution upholds geological facts that most earthquake-forming faults occur in upper lithosphere. This concentration is also very applicable to risk assessment and disaster preparation since shallow earthquakes are more destructive compared to deep earthquakes.

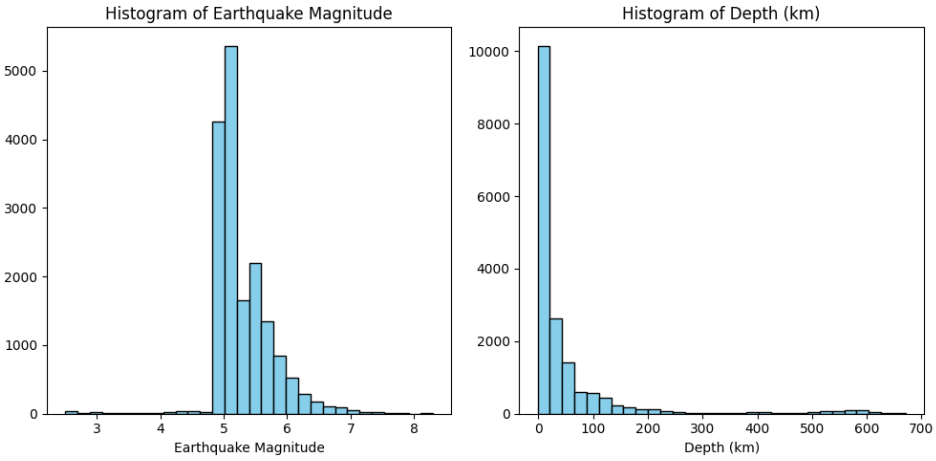


Figure 4.18 Earthquake magnitude and depth distribution

4.5.3 Location of Highest Magnitude

The strongest earthquakes recorded worldwide from 2015 to 2025 shown in Figure 4.19. Chile recorded the strongest magnitude of 8.3 earthquake in 2015 due to the subduction of the Nazca Plate underneath South America. Additionally, 8.2 magnitude earthquake recorded at United State Peninsula (likely Alaska), Fiji, and near Mexico's Chiapas region which are related to the active tectonic boundaries. Similarly, range magnitude of 7.9 to 8.1 recorded at the South Sandwich Islands, Kermadec Islands in New Zealand, Peru, the United State and Papua New Guinea are associated to tectonic plate collisions or subduction zones

Top 10 Highest Magnitude Earthquakes:

	Date	Country	City	Earthquake Magnitude
16503	16/9/2015	Chile	48 km W of Illapel	8.3
6605	29/7/2021	United States Peninsula	United States Peninsula	8.2
11762	19/8/2018	Fiji	267 km E of Levuka	8.2
13288	8/9/2017	Mexico	near the coast of Chiapas	8.2
6517	12/8/2021	South Sandwich Islands region	South Sandwich Islands region	8.1
7448	4/3/2021	New Zealand	Kermadec Islands	8.1
10328	26/5/2019	Peru	78 km NE of Navarro	8.0
12695	23/1/2018	United States	261 km SE of Chiniak	7.9
14451	17/12/2016	Papua New Guinea	140 km E of Kokopo	7.9
14216	22/1/2017	Papua New Guinea	35 km WNW of Panguna	7.9

Figure 4.19 The highest magnitude of earthquake

4.6 Initial Findings Visualization

The initial finding results visualized using matplotlib and seaborn library to generate various plots and graphs that help to understand the earthquake data visually. These visualizations provide extensive information on earthquake features in terms of magnitude and depth distribution, temporal dynamics, impacts based on the location and cross-correlation between different parameters.

4.6.1 Trend of Earthquake from 2015-2025

The bar chart in Figure 4.20 illustrates how the number of earthquakes changed each year from 2015 to 2025. For most years in this period, the earthquake counts generally stayed within a similar range between about 1,400 and 1,850 incidents. Notably, 2021 stands out with a significant spike in seismic activity, reaching the highest count of over 2,200 earthquakes. Other years, such as 2018, 2022, and 2023, also experienced relatively high numbers, while 2020 and 2024 saw moderate drops compared with surrounding years.

A dramatic shift occurs in 2025, with the number of earthquakes dropping steeply to slightly over 600, far below any other year on the chart. This sudden decline due to the recorded data only in half year of 2025. The chart, overall, highlights a decade marked by fluctuations, with mostly consistent or increasing earthquake activity until an abrupt and unusual decrease at the end of the observed period.

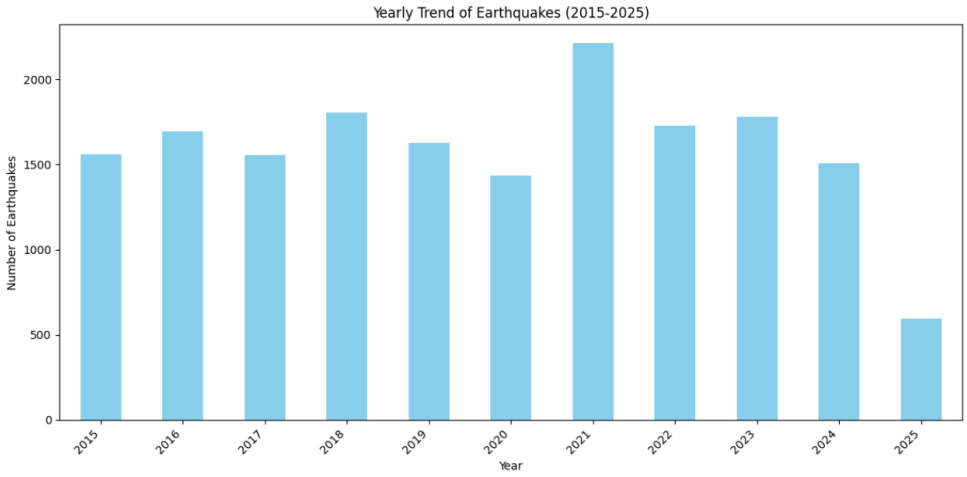


Figure 4.20 Yearly trend of earthquake

4.6.2 Most Recorded Earthquake From 2015-2025

Figure 4.21 presents the country with the highest earthquake record in a decade. Indonesia recorded the highest number of earthquakes with more than 1,750 events in a decade reflecting this country located on a major tectonic zone of active seismic activities. Papua New Guinea, Japan and the South Sandwich Islands recorded more than 1,000 quakes followed by Philippines, Tonga and the United States with less than 850 cases.

The island region such as South Sandwich Island, Kermadec Island and southeast of the Loyalty Islands also recorded with the highest earthquake events in a decade where these locations were located near the tectonic boundary and subduction zone. This data proved that the areas located near the active seismic zone and Ring of Fire have frequent seismic activity. This collected information allows authorities to identify the frequent seismic activity for the mitigation preparation and reduce the impact of earthquakes to the human population and infrastructures.

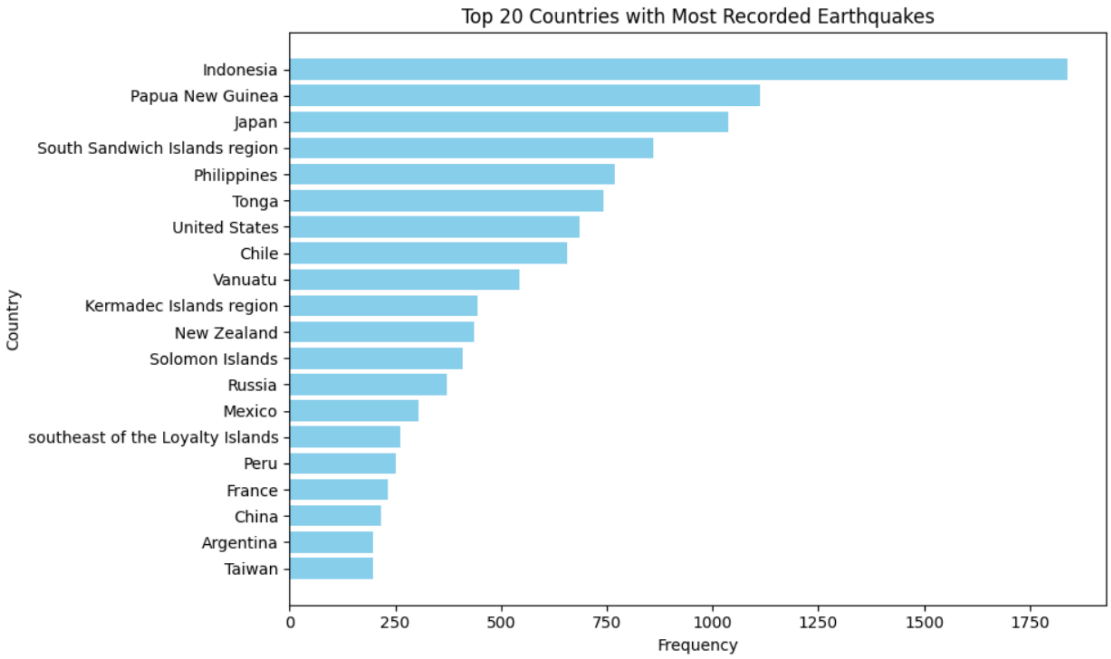


Figure 4.21 Countries with most recorded earthquake from 2015-2025

4.6.3 Most Frequent Locations of Earthquake

Figure 4.22 shows the two-bar chart of the locations land and sea that have frequent earthquakes between 2015-2025. The first chart demonstrates the land areas of highest frequency of earthquakes with Fiji region is the highest with over 120 recorded earthquakes. The location of Fiji located along the hotspot of tectonic activity influenced the high number of earthquake events.

Next, the most frequent land area affected with seismic activity are Pagan Region, Southern of Tonga and Vanuatu which all located in the Pacific basin which have the geological instability. However, the land regions like southern Africa and off the coast of Oregon show more than 30 earthquake events indicates that the presence of active tectonic movement in that area that caused the earthquake.

The sea-based region on the right bar chart shows the South Sandwich Islands region has the most frequent earthquake events with nearly 900 recorded events compared to Fiji region making it the most active seismic zone. This intense activity caused by subduction zones in the South Atlantic. Similarly, Kermadec Island and the area southeast

of the Loyalty Islands present the high earthquake event as they situated along convergent plate boundaries where the active zones of mid-ocean ridges like the Mid-Atlantic Ridge and Pacific-Antarctic Ridge formed new crust forms as plates move away from one another.

The huge different of land and sea earthquake data reveals an intense difference in activity levels where sea regions experienced much higher earthquakes than land areas. This comparison clearly shows that the sea-region greater tectonic instability under the oceans especially in areas with active subduction zones and spreading ridges.

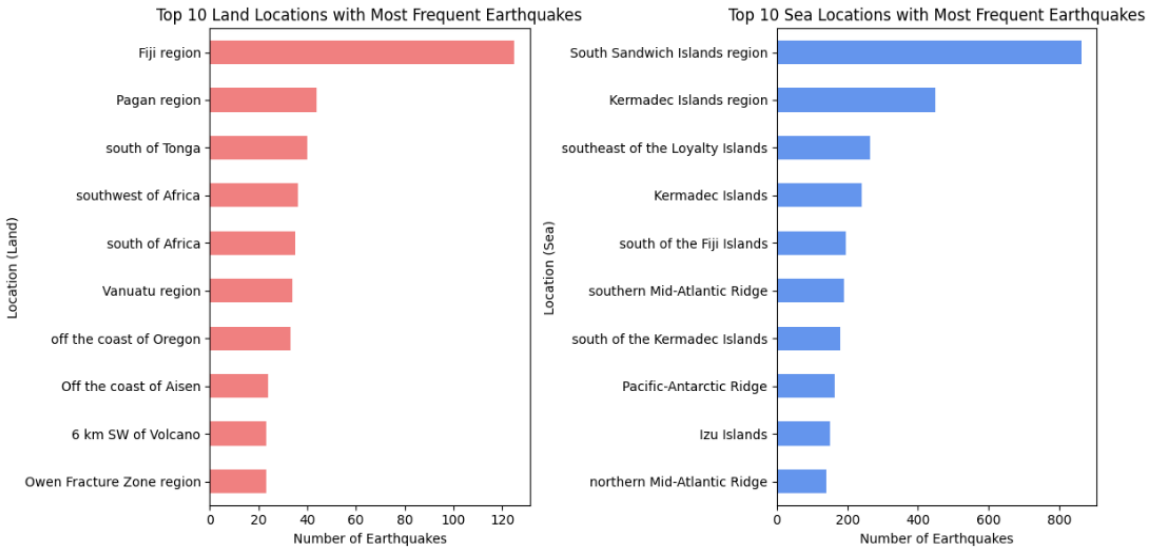


Figure 4.22 Location with most frequent earthquakes

4.6.4 Country of Highest Average Magnitude

Figure 4.23 demonstrates the countries with the highest average magnitude of magnitude in a decade. Jamaica recorded with the highest average magnitude compared to other countries with average magnitude more than 6.0. This data followed by the Croatia, Cayman Island, Thailand and Rwanda region with average magnitude more than 5.5.

Based on this data, the influenced of the highest average magnitude in all these countries caused by the active earth movement of seismic zone of Ring of Fire except Thailand because of Thailand located at the active movement of tectonic plate.

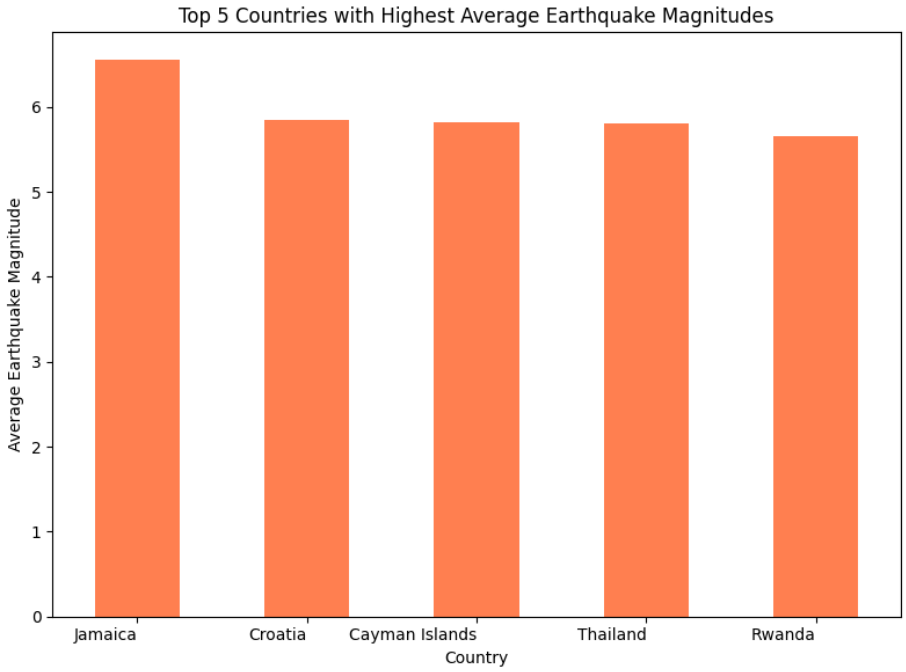


Figure 4.23 Countries with highest average earthquake magnitude

4.6.5 Earthquake Frequency based on geological location

Figure 4.24 illustrated pie chart of the distribution of earthquake zone based on its geological location of ring of fire (1) and active fault (0). Ring of fire dominated with high number of earthquakes with 66% while active fault zone recorded only 34%. The highest number of earthquake occurrences located on ring of fire confirming that this area has the most active tectonic activity and seismic events.

Comparative character of active fault regions and subtypes of a ring-of-fire point out to a remarkable similarity in frequency of earthquakes. The ring of fire, which is made of discontinuous plate margins, which are pushed to high levels of tectonic tension, may thus be predicted to manifest a significant difference in the rate of occurrence of earthquakes compared to the seismic activities on active faults at great distances. The

serious consequence that geological context and presence near fault lines can have in the formulation of risk assessment presents a strong case to be considered in future risk studies which is this information is crucial for disaster preparations and planning infrastructural details in risk-prone areas.

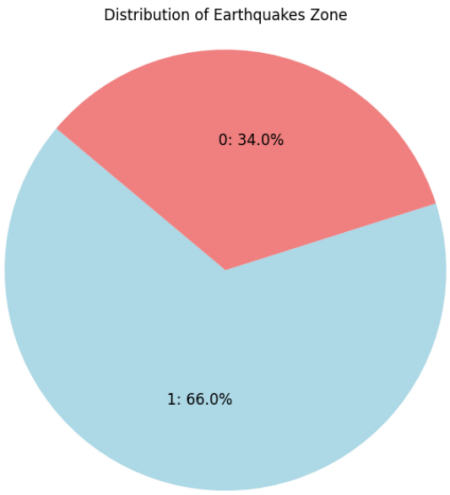


Figure 4.24 Pie chart of earthquake classification zone

4.6.6 Earthquake distribution based on Continent

The pie chart in Figure 4.25 demonstrates the distribution of earthquakes across different continents. The largest proportion of 34.8 % of global earthquakes is occupied by Oceania and is followed by Asia (30.2 %). South America has an equally large contribution of 15.6 % as well which is attributed to the high seismic activity in the region, a phenomenon attributed to the region lying in the boundaries of active tectonic plates.

Other continents have relatively lower percentages of the incidences of earthquakes. North America is 9 %, Europe 5 %, Africa 4.7 %, and the least is Antarctica where it is 0.7 %. These relative values imply that continents like Antarctica and Africa have less seismicity which could be ascribed to their location being far aside along tectonic plate boundaries or major fault lines. Overall, the chart highlights the asymmetrical worldwide seismic activity distribution by indicating that some areas of the world are much more prone to earthquakes as compared to others.

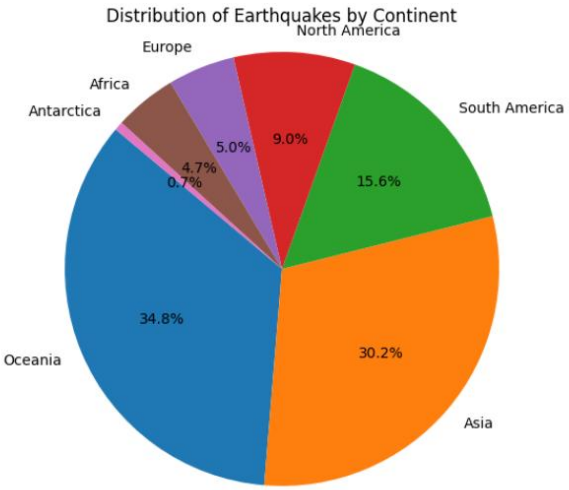


Figure 4.25 Earthquake distribution by continent

4.6.7 Correlation Matrix

The Spearman correlation matrix in Figure 4.26 explains the monotonically dependent nature of different attributes of earthquake that were scaled. As specifically, depth shows moderate positive relationship with depthError, where the coefficient is 0.62. This finding implies that the deeper the earthquake, the greater the chances of a large error in the determination of depth of the earthquake. A similar correlation of 0.55 can be discovered between mag (magnitude) and Magnitude Type that indicates an affinity between magnitude values and the assigned magnitude type (such as Mw, ML). These associations are probably also reflective of physical mechanisms in addition to methodological considerations of acquisition of earthquake data.

The current analysis exhibits the statistically significant negative correlation between magnitude (mag) and the azimuthal gap (gap) of -0.38. In particular, the earthquakes of larger magnitude exhibit fewer coverage gaps in azimuth, possibly to be attributed to better global coverage and the denser sensor networks of high-magnitude earthquakes. In comparison, the other variables showed small correlation lines with the azimuthal gap of magnitude error (magError), root-mean-square of the travel time residuals (rms) and horizontal error (horizontalError). The correlations of each were close to zero in this data set making them independent of the others. Lastly, the categorical features, Zone Classification, Types of Earthquake, and Continent encoded all showed

weak correlations hence as expected because these set of features are encoded or based on classification.

The matrix helps to identify the variables that provide duplicate or unique information during modelling activities. Overall, the low levels of strong multicollinearity between most variables are a good sign when using machine-learning models since it reduces the risk of overfitting due to feature redundancy. However, interrelation especially between depth and depthError and mag and Magnitude Type should be carefully examined during predictive modelling since they can reduced the quality of data and the credibility of the models, particularly the interpretation of variable significance or feature impact.

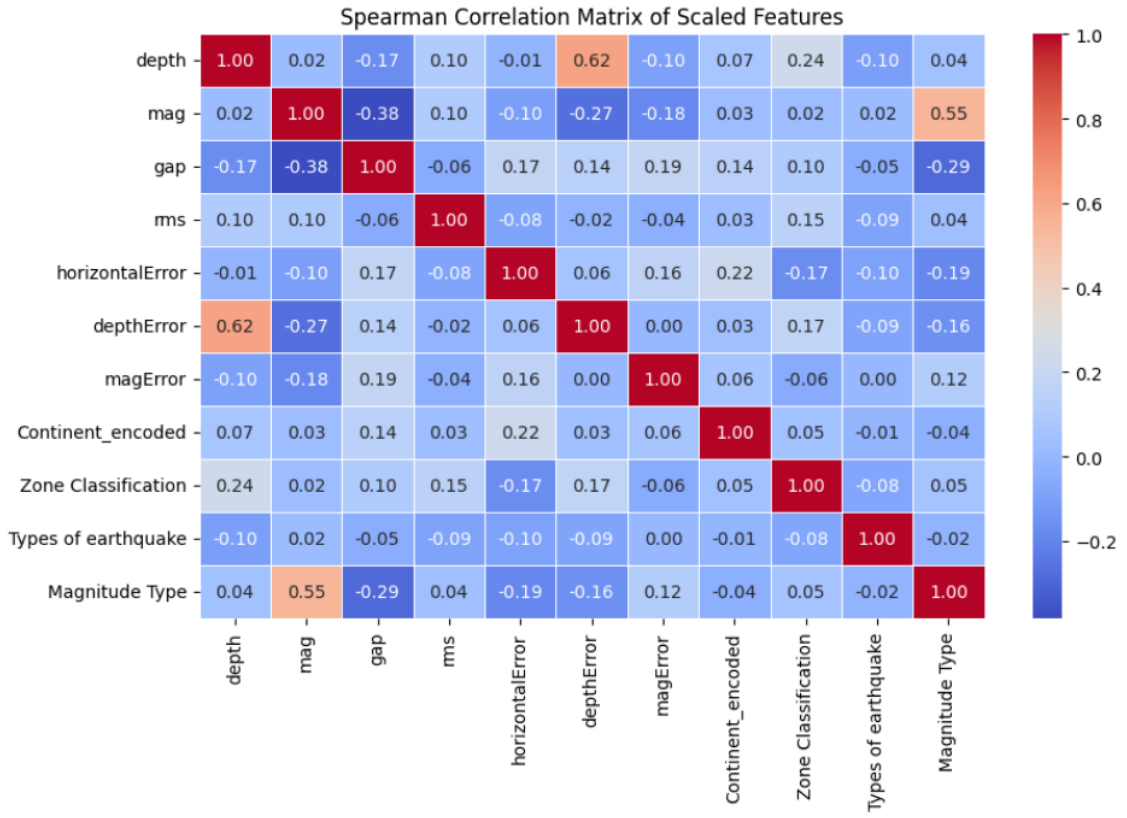


Figure 4.26 Spearman Correlation matrix of important features

4.7 Chapter Summary

In conclusion, this chapter discussed about the exploratory data analysis (EDA) steps started from data cleaning and pre-processing until the early findings of the earthquake from 2015 until 2025. The outlier existed in this data and solved using winosarisation methods without remove the original value. Analysis of the earthquake events indicate the most frequent earthquake occur at the active seismic zone of the ring of fire like Indonesia and Japan. Overall correlation matrix shows the low to moderate relationship between all features. Further analysis will perform in feature engineering to identify the nearest plate boundaries that influence the active seismic activities.

CHAPTER 5

CONCLUSION

5.1 Conclusion

The initial findings of exploratory data analysis (EDA) discover the important insights that provide the foundational understanding of global earthquake trends in a decade from 2015 until 2025. The analysis indicates most of the earthquakes happened at moderate magnitude at shallow. Indonesia recorded as the highest frequency of the earthquake events meanwhile Jamaica recorded with the highest average of the earthquake magnitude as both located on the active seismic zone of the Pacific Ring of Fire. As expected, Pacific Ring of Fire recorded the highest earthquake events compared to other active fault zones where these locations exhibit the most active fault movements with the volcano's activities.

The handling outliers and statistical summary demonstrated that large earthquakes usually result in a higher magnitude and the most earthquakes occur in subduction zone. Comparing the land-based area and sea-based area of the active tectonic plates, the sea-based area experience most frequent and intense seismic events. The Spearman correlation matrix shows most of the characteristics of earthquakes display narrow multicollinearity which implies little duplicity in the predictors. However, the moderate positive correlation between depth and depthError (0.62) and magnitude and MagnitudeType (0.55) would be good to pay attention to when modeling, as the two pairs would require treatment to avoid misleading interpretations.

In conclusion, this finding is important for the feature engineering, model development and evaluation phase for improving the model accuracy of earthquake predictions and earthquake mitigation strategies.

5.2 Future Work

This research only involved the halfway stage of Exploratory Data Analysis (EDA). The next future work will be focusing on advanced feature engineering, model development and evaluation and visualization using dashboards. Feature engineering will be applied to train and test both algorithms of Random Forest and XGBoost to compare the best performance of earthquake classifying earthquake magnitude, depth and impact score and location. The model will be evaluate using confusion matrix, Precision, Accuracy, F1 score and Recall. As a result, hyperparameters will be tuned and tested using cross-validation to produce the best performance model. Lastly, the results will be presented using dashboard for the effective earthquake model that can improve the earthquake analysis and prediction while creating the best insights and decision making for the disaster mitigation strategies.

REFERENCES

Ahmed, F., Akter, S., Rahman, S. M., Harez, J. B., Mubasira, A., & Khan, R. (2024). Earthquake Magnitude Prediction Using Machine Learning Techniques. *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, *2*, 1–5. https://doi.org/10.1109/IATMSI60426.2024.10502770

Alidadi, N., & Pezeshk, S. (2025). State of the art: Application of machine learning in ground motion modeling. *Engineering Applications of Artificial Intelligence*, *149*, 110534. https://doi.org/10.1016/j.engappai.2025.110534

Babu, D. B., Revathi, M. L., Senthil, M., Parvathi, A. L., Sceenilai, B., & Sheema, Sk. (2024). Earthquake Prediction Model Using Random Forest and Gradient Boosting Algorithms. *2024 9th International Conference on Communication and Electronics Systems (ICCES)*, 1597–1607. https://doi.org/10.1109/ICCES63552.2024.10859534

Bao, H., Xu, L., Meng, L., Ampuero, J.-P., Gao, L., & Zhang, H. (2022). Global Frequency of Oceanic and Continental Supershear Earthquakes. *Nature Geoscience*, *15*(11), 942–949. https://doi.org/10.1038/s41561-022-01055-5

Bilek, S., & Lay, T. (2018). Subduction zone megathrust earthquakes. *Geosphere*, *14*. https://doi.org/10.1130/GES01608.1

Biswas, S., Kumar, D., & Bera, U. K. (2023). *Prediction Of Earthquake Magnitude and Seismic Vulnerability Mapping Using Artificial Intelligence Techniques: A Case Study Of Turkey*. In Review. https://doi.org/10.21203/rs.3.rs-2863887/v1

Buforn, E., & Udías, A. (2010). Chapter 3—Azores–Tunisia, A Tectonically Complex Plate Boundary. In R. Dmowska (Ed.), *Advances in Geophysics* (Vol. 52, pp. 139–182). Elsevier. https://doi.org/10.1016/S0065-2687(10)52003-X

Cai, J., Xi, N., Han, G., Deng, W., & Sun, L. (2025). Rapid report of the March 28, 2025 Mw 7.9 Myanmar Earthquake. *Earthquake Research Advances*, 100396. https://doi.org/10.1016/j.eqrea.2025.100396

Cilia, M. G., Mooney, W. D., & Nugroho, C. (2021). Field Insights and Analysis of the 2018 Mw 7.5 Palu, Indonesia Earthquake, Tsunami and Landslides. *Pure and Applied Geophysics*, *178*(12), 4891–4920. Scopus. https://doi.org/10.1007/s00024-021-02852-6

Cornely, P.-R., & Wang, J. (2023). Advancing Earthquake Prediction: A Comprehensive Review of Data Science Techniques. *2023 6th International Conference on Computing and Big Data (ICCBD)*, 9–16. https://doi.org/10.1109/ICCBD59843.2023.10607190

Cui, B., Guo, J., Han, G., & Liu, X. (2024). Earthquake Magnitude and Depth Prediction Based on Machine Learning and Multiple Linear Regression Models. *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, 1056–1060. https://doi.org/10.1109/ICSECE61636.2024.10729410

Dragoni, M., & Santini, S. (2022). Contribution of the 2010 Maule Megathrust Earthquake to the Heat Flow at the Peru-Chile Trench. *Energies*, *15*(6), Article 6. https://doi.org/10.3390/en15062253

Duarte, J., & Schellart, W. (2016). Introduction to Plate Boundaries and Natural Hazards. In *Plate Boundaries and Natural Hazards* (p. 352 pages). https://doi.org/10.1002/9781119054146.ch1

Fischer, T., Hrubcová, P., Salama, A., Doubravová, J., Ágústsdóttir, T., Gudnason, E. Á., Horálek, J., & Hersir, G. P. (2022). Swarm Seismicity Illuminates Stress Transfer Prior To The 2021 Fagradalsfjall Eruption In Iceland. *Earth and Planetary Science Letters*, *594*, 117685. https://doi.org/10.1016/j.epsl.2022.117685

Geersen, J., Sippl, C., & Harmon, N. (2022). Impact Of Bending-Related Faulting And Oceanic-Plate Topography On Slab Hydration And Intermediate-Depth Seismicity. *Geosphere*, *18*. https://doi.org/10.1130/GES02367.1

Gupta, A., Sharma, B., & Chingtham, P. (2024). Forecast of Earthquake Magnitude for North–West (NW) Indian Region Using Machine-Learning Techniques. *Lecture Notes in Electrical Engineering*, *1185*, 361–376. Scopus. https://doi.org/10.1007/978-981-97-1682-1_30

Gurnis, M., Zhong, S., & Toth, J. (2000). On The Competing Roles Of Fault Reactivation And Brittle Failure In Generating Plate Tectonics From Mantle Convection. *Geophysical Monograph Series*, *121*, 73–94. https://doi.org/10.1029/GM121p0073

Güvercin, S. E., Karabulut, H., Konca, A. Ö., Doğan, U., & Ergintav, S. (2022). Active seismotectonics of the East Anatolian Fault. *Geophysical Journal International*, *230*(1), 50–69. https://doi.org/10.1093/gji/ggac045

Handayani, T., Wijayanto, Wijaya, A., Himantara, L., Saputro, A. H., & Djuhana, D. (2024). Machine Learning Implementation for Estimation of Earthquake Magnitude Using Strong-Motion Data. *2024 4th International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, 351–355. https://doi.org/10.1109/RAAI64504.2024.10949525

Harirchian, E., Kumari, V., Jadhav, K., Rasulzade, S., Lahmer, T., & Raj Das, R. (2021). A Synthesized Study Based on Machine Learning Approaches for Rapid Classifying Earthquake Damage Grades to RC Buildings. *Applied Sciences*, *11*(16), 7540. https://doi.org/10.3390/app11167540

Hill, D. P., & Prejean, S. G. (2015). Dynamic Triggering. In *Treatise on Geophysics: Second Edition* (Vol. 4, pp. 273–304). Scopus. https://doi.org/10.1016/B978-0-444-53802-4.00078-6

Hoque, A., Raj, J., Saha, A., & Bhattacharya, D. (2020). *Earthquake Magnitude Prediction Using Machine Learning Technique*.

James, D. E. (2021). Lithosphere, Continental. In H. K. Gupta (Ed.), *Encyclopedia of Solid Earth Geophysics* (pp. 866–872). Springer International Publishing. https://doi.org/10.1007/978-3-030-58631-7_32

Joshi, A., Singh, P., & Raman, B. (2025). Isomapgen: Framework For Early Prediction Of Peak Ground Acceleration Using Tripartite Feature Extraction And Gated Attention Model. *Computers & Geosciences*, *196*, 105849. https://doi.org/10.1016/j.cageo.2024.105849

Kaftan, I. (2025). *Machine Learning Applications for Earthquake Magnitude Prediction in Western Türkiye* (SSRN Scholarly Paper No. 5234889). Social Science Research Network. https://doi.org/10.2139/ssrn.5234889

Kazbekova, G., Aben, A., Amanov, A., Zhunissov, N., & Abibullayeva, A. (2025). Effectiveness Of Machine Learning Methods In Determining Earthquake Probable Areas: Example Of Kazakhstan. *Scientific Journal of Astana IT University*. https://doi.org/10.37943/21KUXZ6354

Kukartsev, V., & Degtyareva, K. (2024). Forecasting seismic activity using machine learning algorithms. *E3S Web of Conferences*, *592*, 05002. https://doi.org/10.1051/e3sconf/202459205002

Lay, T. (2016). Great Earthquakes on Plate Boundaries. *Oxford Research Encyclopedia of Natural Hazard Science*. https://oxfordre.com/naturalhazardscience/display/10.1093/acrefore/9780199389407.001.0001/acrefore-9780199389407-e-32

Lay, T., Ye, L., Wu, Z., & Kanamori, H. (2020). Macrofracturing of Oceanic Lithosphere in Complex Large Earthquake Sequences. *Journal of Geophysical Research: Solid Earth*, *125*(10). Scopus. https://doi.org/10.1029/2020JB020137

Liu, Z., & Buck, W. R. (2018). Magmatic controls on axial relief and faulting at mid-ocean ridges. *Earth and Planetary Science Letters*, *491*, 226–237. https://doi.org/10.1016/j.epsl.2018.03.045

Long, X., Lu, C., Gu, X., Ma, Y., & Li, Z. (2024). Selection Of The Structural Severest Design Ground Motions Based On Big Data And Random Forest. *Engineering Applications of Artificial Intelligence*, *133*, 108238. https://doi.org/10.1016/j.engappai.2024.108238

Lu, R., Xu, X., He, D., John, S., Liu, B., Wang, F., Tan, X., & Li, Y. (2017). Seismotectonics of the 2013 Lushan Mw 6.7 earthquake: Inversion tectonics in the eastern margin of the Tibetan Plateau. *Geophysical Research Letters*, *44*(16), 8236–8243. https://doi.org/10.1002/2017GL074296

Luo, S., Yao, H., Zhang, Z., & Bem, T. S. (2022). High-Resolution Crustal and Upper Mantle Shear-Wave Velocity Structure Beneath the Central-Southern Tanlu

Fault: Implications For Its Initiation And Evolution. *Earth and Planetary Science Letters*, *595*, 117763. https://doi.org/10.1016/j.epsl.2022.117763

Macheyeki, A. S. (2024). Present-day Fault Kinematics and their Reactivation Likelihood within and South of the North Tanzania Divergence (NTD), East African Rift System: Implication for Geo-hazards Assessment. *Journal of the Geological Society of India*, *100*(1), 127–138. Scopus. https://doi.org/10.17491/jgsi/2024/172989

Mahmoud, A., Alrusaini, O., Shafie, E., Aboalndr, A., & S.Elbelkasy, M. (2025). Machine Learning-Based Earthquake Prediction: Feature Engineering and Model Performance Using Synthetic Seismic Data. *Applied Mathematics & Information Sciences*, *19*(3), 695–702. https://doi.org/10.18576/amis/190317

Manaman, N. S., & Shomali, H. (2010). Upper Mantle S-Velocity Structure And Moho Depth Variations Across Zagros Belt, Arabian–Eurasian Plate Boundary. *Physics of the Earth and Planetary Interiors*, *180*(1), 92–103. https://doi.org/10.1016/j.pepi.2010.01.011

Mera, W., Vera, X., Antonio, L. T., & Ponce, G. (2017). April 2016 Ecuador Earthquake Of Moment Magnitude Mw7.8: Overview And Damage Report. *Key Engineering Materials*, *747 KEM*, 662–669. Scopus. https://doi.org/10.4028/www.scientific.net/KEM.747.662

Mesta, C., Kerschbaum, D., Cremen, G., & Galasso, C. (2023). Quantifying The Potential Benefits of Risk-Mitigation Strategies on Present and Future Seismic Losses In Kathmandu Valley, Nepal. *Earthquake Spectra*, *39*(1), 377–401. Scopus. https://doi.org/10.1177/87552930221134950

Nakamura, Y., Kodaira, S., Fujie, G., Yamashita, M., Obana, K., & Miura, S. (2023). Incoming plate structure at the Japan Trench subduction zone revealed in

densely spaced reflection seismic profiles. *Progress in Earth and Planetary Science*, *10*(1), 45. https://doi.org/10.1186/s40645-023-00579-7

Nandu, P. P. H., Madhu, B. E., Reddy, K. S. K. K., & Adinarayana, B. (2025). AI-driven Development and Utilization Of 2024 Noto Earthquake Seismic Data for Prediction of Earthquake Intensity Measures for Japan Using Deep Machine Learning Models. *Innovative Infrastructure Solutions*, *10*(4). Scopus. https://doi.org/10.1007/s41062-025-01937-8

Navarro, A., Castro-Artola, O., García-Guerrero, E., Aguirre-Castro, O., Tamayo Pérez, U., López-Mercado, C., & Inzunza Gonzalez, E. (2025). Recent Advances in Early Earthquake Magnitude Estimation by Using Machine Learning Algorithms: A Systematic Review. *Applied Sciences*, *15*, 1. https://doi.org/10.3390/app15073492

Nayak, K., Romero-Andrade, R., Sharma, G., Lopez, C., Trejo Soto, M., & Vidal-Vega, A. (2024). Evaluating Ionospheric Total Electron Content (TEC) Variations as Precursors to Seismic Activity: Insights from the 2024 Noto Peninsula and Nichinan Earthquakes of Japan. *Atmosphere*, *15*, 1492. https://doi.org/10.3390/atmos15121492

Novick, D., & Last, M. (2023). Using Machine Learning Models for Earthquake Magnitude Prediction in California, Japan, and Israel. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13914 LNCS*, 151–169. Scopus. https://doi.org/10.1007/978-3-031-34671-2_11

Nugroho, H. A., Subiantoro, A., & Kusumoputro, B. (2024). A Model Fusion of 1D CNN with NARX for Accurate Earthquake Time Series Prediction. *2024 4th*

International Conference on Robotics, Automation and Artificial Intelligence (RAAI), 340–344. https://doi.org/10.1109/RAAI64504.2024.10949569

Olive, J.-A. (2023). Mid-Ocean Ridges: Geodynamics Written in the Seafloor. In *Dynamics of Plate Tectonics and Mantle Convection* (pp. 483–510). Scopus. https://doi.org/10.1016/B978-0-323-85733-8.00018-4

Pisarenko, V. F., & Pisarenko, D. V. (2022). A Modified k-Nearest-Neighbors Method and Its Application to Estimation of Seismic Intensity. *Pure and Applied Geophysics*, *179*(11), 4025–4036. Scopus. https://doi.org/10.1007/s00024-021-02717-y

Pwavodi, J., Ibrahim, A. U., Pwavodi, P. C., Al-Turjman, F., & Mohand-Said, A. (2024). The Role of Artificial Intelligence and Iot in Prediction of Earthquakes: Review. *Artificial Intelligence in Geosciences*, *5*, 100075. https://doi.org/10.1016/j.aiig.2024.100075

Sartori, R., Torelli, L., Zitellini, N., Peis, D., & Lodolo, E. (1994). Eastern Segment of The Azores-Gibraltar Line (Central-Eastern Atlantic): An Oceanic Plate Boundary with Diffuse Compressional Deformation. *Geology*, *22*(6), 555–558. https://doi.org/10.1130/0091-7613(1994)022<0555:ESOTAG>2.3.CO;2

Satish, S., Gonaygunta, H., Yadulla, A. R., Kumar, D., Maturi, M. H., Meduri, K., Cruz, E. D. L., Nadella, G. S., & Sajja, G. S. (2025). Forecasting the Unseen: Enhancing Tsunami Occurrence Predictions with Machine-Learning-Driven Analytics. *Computers 2025*, *14*(5), 175. https://doi.org/10.3390/computers14050175

Satish, S., Gonaygunta, H., Yudalla, A. R., Kumar, D., Mohan, H. M., Meduri, K., & Cruz, E. D. L. (2025). Forecasting the Unseen: Enhancing Tsunami Occurrence

Predictions with Machine-Learning-Driven Analytics. *Computers 2025*, *14*(5), 175. https://doi.org/10.3390/computers14050175

Shahzada, K., Noor, U. A., & Xu, Z.-D. (2025). In the wake of the March 28, 2025 Myanmar earthquake: A detailed examination. *Journal of Dynamic Disasters*, *1*(2), 100017. https://doi.org/10.1016/j.jdd.2025.100017

Suppasri, A., Kitamura, M., Alexander, D., Seto, S., & Imamura, F. (2024). The 2024 Noto Peninsula : Preliminary *International Journal of Disaster Risk Reduction*, *110*. Scopus. https://doi.org/10.1016/j.ijdrr.2024.104611

Styron, R., & Pagani, M. (2020). The GEM Global Active Faults Database. *Earthquake Spectra*, *36*(1_suppl), 160–180. https://doi.org/10.1177/8755293020944182

Truong, V.-H., Tangaramvong, S., Nguyen, M.-C., & Pham, H.-A. (2025). Machine Learning-Based Safety Assessment of Steel Frames Under Seismic Loadings Using Nonlinear Time-History Analysis. *Steel and Composite Structures*, *54*(4), 295–312. Scopus. https://doi.org/10.12989/scs.2025.54.4.295

Wang, J., Shahani, N. M., Zheng, X., Hongwei, J., & Wei, X. (2025). Machine Learning-Based Analyzing Earthquake-Induced Slope Displacement. *PLOS ONE*, *20*(2), e0314977. https://doi.org/10.1371/journal.pone.0314977

Wang, T., Bian, Y., Zhang, Y., & Hou, X. (2023). Classification Of Earthquakes, Explosions and Mining-Induced Earthquakes Based on Xgboost Algorithm. *Computers & Geosciences*, *170*, 105242. https://doi.org/10.1016/j.cageo.2022.105242

Wang, X., Cao, L., Zhao, M., Cheng, J., & He, X. (2023). What Conditions Promote Atypical Subduction: Insights from The Mussau Trench, The Hjort Trench,

And the Gagua Ridge. *Gondwana Research*, *120*, 207–218. Scopus. https://doi.org/10.1016/j.gr.2022.10.014

Wijaya, U., Kusrini, & Muhammad, A. H. (2022). Indonesian Seismic Mitigation using Earthquake Predicted Artificial Intelligence Model. 349–354. Scopus. https://doi.org/10.1109/ICOIACT55506.2022.9972091

Wu, F. Y., Wang, J. G., Liu, C. Z., Liu, T., Zhang, C., & Ji, W. Q. (2019). Intra-Oceanic Arc: Its Formation and Evolution. *Yanshi Xuebao/Acta Petrologica Sinica*, *35*(1), 1–15. Scopus. https://doi.org/10.18654/1000-0569/2019.01.01

Wu, Z.-N., Li, Z.-Q., Dong, Y., Han, X.-L., Zhang, G., Feng, R., & Zhu, K. (2024). Seismic Intensity Measure Selection Incorporating Interaction Effects for Damage Assessment Across Different Structural Sensitive Regions. *Structures*, *67*, 106917. https://doi.org/10.1016/j.istruc.2024.106917

Yavas, C. E., Chen, L., Kadlec, C., & Ji, Y. (2024). Improving Earthquake Prediction Accuracy In Los Angeles With Machine Learning. *Scientific Reports*, *14*(1). Scopus. https://doi.org/10.1038/s41598-024-76483-x

Yenidoğan, C. (2024). February 6, 2023, Earthquakes and Preliminary Assessment of Building Damage Based on Field Surveys. *Turkish Journal of Civil Engineering*, *35*(5), 75–113. Scopus. https://doi.org/10.18400/tjce.1335742

Appendix A : Gantt Chart

Task	Month								
	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep
1. Desk Study									
2. Proposal Development									
2.1 Research Question and Objective	X								
2.2 Literature Review	X								
2.3 Proposal Submission and Approval		X							
Project 1 :									
3. Data Collection									
3.1 USGS Global Earthquake Data Catalog	X	X							
4. Data Pre-processing									
4.1 Data Cleaning			X	X					
5. Exploratory Data Analysis (EDA)					X				
6. Thesis Writing	X	X	X	X	X	X			
7. Thesis Submission					X	X			
8. Thesis Presentation						X			
9. Thesis Revision and Final Submission							X		
Project 2:									
1. Feature Engineering					X	X	X		
2. Model Development and Evaluation									
2.1 Random Forest and XGBoost						X	X	X	
2.2 Training Machine Learning Model							X	X	X
2.3 Model Validation and Testing							X	X	X
3. **Results Analysis**									
3.1 Performance Metrics Analysis							X		
4.2 Visualization of Results								X	
4. Thesis Writing	X	X	X	X	X	X	X	X	X
5. Thesis Submission									X
6. Thesis Presentation									X
7. Thesis Revision and Final Submission									X