A blue and white logo

Description automatically generated  
UNIVERSITY OF HERTFORDSHIRE  
School of Physics, Engineering and Computer Science

MSc Artificial Intelligence and Robotics  
**7COM1039-0501-2023 - Advanced Computer Science Masters Project**

Date:- 02/12/2024

EARTHQUAKE CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

Name: Mohammed Khaleeq  
Student ID: 21071212  
Supervisor: Dr. Efosa Osagie

Acknowledgement

I want to express my sincere gratitude to everyone who has assisted and motivated me while this work has been developed. To begin with, I would want to deeply thank  Dr. Efosa Osagie, my supervisor, whose insightful guidance and impactful feedback have been crucial in helping me complete this research successfully. Their contributions and experience have been an excellent source of support.

Additionally, I would like to thank the University of Hertfordshire for providing the infrastructure and vital resources required to conduct this study. I am appreciative of my fellow researchers for their thoughtful conversations, constant assistance, and interaction, which greatly aided in the progress of this research.

Additionally, I am indebted to the initial authors whose work combining seismology and machine learning served as a major inspiration for my study.

Lastly, I would like to convey my heartfelt thanks to my family and friends for their unwavering encouragement and support during the development of this project. Their belief in me has been a constant source of inspiration throughout this journey..

**Declaration**

I declare that this research is my original work and has not been submitted for any other degree or qualification. All sources and references have been acknowledged respectfully, and I have adhered to ethical standards throughout the research process.

Abstract

Evaluation of Earthquakes has a high impact on the property and Human life therefore stressing the importance of an effective classification system and early warning system. Conventional techniques of earthquake classification resolved in mechanical and statistical terms that do not let breaking down great volumes of seismic material produced by advanced seismological observation networks. Conventional data analysis provides a promising approach, although it cannot process big data and identify profound dependencies. This paper seeks to undertake a review of the existing literature in a bid to identify a range of ML techniques that can be employed to strengthen earthquake categorization, including the decision tree, logistic regression, and neural networks, among others. This work seeks to establish models for accurate tagging and predicting seismic activity with the view of enhancing disaster preparedness and management strategies. It shows how to overcome difficulties connected with ML application to seismology: data quality point, algorithm point, and computation effort. Altogether, this research contributes to the development of the subject via ML-based strategies for enhanced earthquake detection as an application for improving early earthquake warnings and reducing the devastating effects of earthquakes on society.

**Table of Contents**

[CHAPTER 1: INTRODUCTION 8](#_Toc183943756)

[1.1 Introduction 8](#_Toc183943757)

[1.2 Background 8](#_Toc183943758)

[1.3 Aim and Objectives 9](#_Toc183943759)

[Aim 9](#_Toc183943760)

[Objectives 9](#_Toc183943761)

[1.4 Research Questions 9](#_Toc183943762)

[1.5 Problem statement 10](#_Toc183943763)

[1.6 Rationale 10](#_Toc183943764)

[1.7 Scope of the Study 11](#_Toc183943765)

[1.8 Dissertation structure 12](#_Toc183943766)

[1.9 Summary 12](#_Toc183943767)

[CHAPTER 2: LITERATURE REVIEW 13](#_Toc183943768)

[2.1 Introduction 13](#_Toc183943769)

[2.2 Machine Learning in Seismology 13](#_Toc183943770)

[2.3 Application of Machine Learning in Earthquake Classification 15](#_Toc183943771)

[2.4 Performance Metrics for Evaluating Earthquake Classification Models 16](#_Toc183943772)

[2.5 Challenges and Future Directions 17](#_Toc183943773)

[2.6 Summary 19](#_Toc183943774)

[2.7 Literature Gap Table 20](#_Toc183943775)

[CHAPTER 3: METHODOLOGY 21](#_Toc183943776)

[3.1 Introduction 21](#_Toc183943777)

[3.2 Research Strategy 21](#_Toc183943778)

[3.2.1 Justification of Research Strategy 22](#_Toc183943779)

[3.3 Research Approach 22](#_Toc183943780)

[3.3.1 Justification of Research Approach 23](#_Toc183943781)

[3.4 Research design 23](#_Toc183943782)

[3.5 Data Collection Method 25](#_Toc183943783)

[3.6 Data analysis 26](#_Toc183943784)

[3.7 Ethical Consideration 28](#_Toc183943785)

[3.8 Summary 29](#_Toc183943786)

[CHAPTER 4: DATA ANALYSIS AND MODEL PERFORMANCE 30](#_Toc183943787)

[4.1 Introduction 30](#_Toc183943788)

[4.2 Data Exploration 30](#_Toc183943789)

[4.3 Exploratory Data Analysis (EDA) 33](#_Toc183943790)

[4.4 Data Preprocessing 35](#_Toc183943791)

[4.5 Model Training and Evaluation 36](#_Toc183943792)

[4.6 Model Comparison 39](#_Toc183943793)

[4.7 Conclusion 40](#_Toc183943794)

[CHAPTER 5: CONCLUSION AND RECOMMENDATIONS 42](#_Toc183943795)

[5.1 Conclusion 42](#_Toc183943796)

[5.2 Recommendations 43](#_Toc183943797)

[5.3 Linking with Objectives 43](#_Toc183943798)

[5.4 Future Work 44](#_Toc183943799)

[References 46](#_Toc183943800)

[Gantt Chart 52](#_Toc183943801)

[Appendix 53](#_Toc183943802)

**List of Figure**

[Figure 1.8.1: Structure of the earthquake classification report 12](#_Toc183802859)

[Figure 2.2.1: Machine learning process for seismic analysis 14](#_Toc183802860)

[Figure 3.4.1: Flow Chart showing the experimental pipeline 24](#_Toc183802861)

[Figure 3.5.1: Dataset based on earthquake alert 25](#_Toc183802862)

[Figure 3.6.1.1: Naïve Bayes (Source: Nyungen *et al.* 2024) 27](#_Toc183802863)

[Figure 3.6.1.2: Random Forest Algorithms (Source: Feng *et al.* 2024) 27](#_Toc183802864)

[Figure 3.6.1.3: Gradient Boosting Working (Source: Jain *et al.* 2023) 28](#_Toc183802865)

[Figure 4.2.1: Histogram of Earthquake Magnitudes 30](#_Toc183802866)

[Figure 4.2.2 shows the missing values 31](#_Toc183802867)

[Figure 4.2.3: Bar Chart of Alert Level 32](#_Toc183802868)

[Figure 4.3.1: Scatter Plot of Magnitude and Significance 34](#_Toc183802869)

[Figure 4.4.1: Correlation Heatmap 35](#_Toc183802870)

[Figure 4.5.1: Naïve Byes Classification Report 37](#_Toc183802871)

[Figure 4.5.2: Random Forest Classification Report 37](#_Toc183802872)

[Figure 4.5.3 Gradient Boosting Classification Report 38](#_Toc183802873)

[Figure 4.6.2: Hyperparameters tuning for best model 39](#_Toc183802874)

[Figure 4.6.1: Model Accuracy Comparison 40](#_Toc183802875)

# CHAPTER 1: INTRODUCTION

## 1.1 Introduction

Earthquakes are natural calamities that may lead to the loss of property and even human lives since they affect buildings and other facilities. Classification of the earthquake is critically important as the monitoring and early warning systems are developed based on the suitable classifications. The classification of earthquakes based on traditional approaches of data analysis and statistical modelling encounters problems, particularly with large datasets and data intensity. Machine learning can process large amounts of data and distinguish complex relationships, it becomes a more suitable approach. This paper focuses on the application of different techniques in machine learning to improve the task of earthquake classification (Rao *et al.* 2023). Therefore, using conceptual calculation methods, the study aims to enhance the recognition of seismic events and, thus, develop stronger predictive analysis techniques.

## 1.2 Background

Seismic taxonomy or the way earthquakes are categorised is a vital part of seismology, helping in the comprehension of earthquakes and perhaps their prediction. Classically, seismology is analysed earthquake information by mechanical means as well as by statistical tests. Some of these are the Richter scale of magnitude and the Mercalli intensity scale used to quantify the extent of the damage. Although helpful to an extent these approaches can become inefficient when analysing large and highly detailed data sets that are produced by today’s seismic

Current research has revealed machine learning’s efficiency in earthquake identification and categorization. For example, CNNs have been applied in classifying seismic signals and SVMs as well as decision trees in classifying different types of seismic events. These developments indicate that through the application of machine learning, how the occurrence of earthquakes and ways of responding to the conditions might dramatically transform.

Nevertheless, several difficulties can be encountered when applying machine learning in seismology. The following challenges have to be met; these include; data quality, algorithm comprehension and the requirement of computations. Therefore, this study seeks to establish these challenges and further evaluate the applicability of different machine-learning algorithms in earthquake classification (Torres and Dungca, 2024). Thus, it aims to help advance knowledge in the field of seismic monitoring and improve contemporary approaches toward earthquake prevention and response.

## 1.3 Aim and Objectives

### Aim

To develop and evaluate machine learning models such as Naïve bayes, Random forest and Gradient boosting for the effective classification of earthquakes, and to enhance early warning systems, and risk management strategies. This project is to design a multi-data machine learning model using seismic data for better earthquake classification. Given that feature engineering and ensemble methods enhance model performance, the study aims to increase the predictive capability, fine-tune the alert levels, and fill the existing research voids observed in real-time seismic event identification and early warning systems.

### Objectives

* Ensuring the dataset is clean and appropriate for machine learning applications which involves identifying and preprocessing important variables that affect earthquake classification, such as magnitude, depth, and significance.
* To increase the effectiveness and accuracy of earthquake identification and categorization techniques.
* To enhance disaster readiness and reaction programs, incorporate the developed model into current early warning systems.
* To evaluate the earthquake classification data with the help of machine learning and then measured the models with various metrics such as ***“accuracy, precision, recall and others”.***

## 1.4 Research Questions

1. What approaches and strategies can be provided to optimize seismic data features like magnitude, depth, and alert levels and assess the results of machine learning models such as Naïve Bayes, Random Forest, Gradient Boosting for earthquake classification?
2. What are machine learning algorithms' earthquake detection and classification accuracy, precision, recall, and F1 score?
3. How does hyperparameter optimization affect earthquake detection in machine learning?

## 1.5 Problem statement

Earthquakes remain one of the most difficult events to classify in seismology because of late complexity and large time series data. Generally, the static analysis of results and other conventional classification techniques do not suit well when faced with large amounts of data typically obtained from new-generation seismic monitoring instruments of earthquakes to lessen their impact.

Despite advancements in machine learning for earthquake detection, existing research faces significant challenges in data cleaning, comprehensive model evaluation, and integration into early warning systems. Many studies overlook the impact of preprocessing techniques on classification accuracy and rely on limited metrics, reducing their practical applicability. This study addresses these gaps by systematically comparing multiple machine learning models using diverse metrics, evaluating the effects of preprocessing, and contributing to the development of reliable disaster management solutions.

## 1.6 Rationale

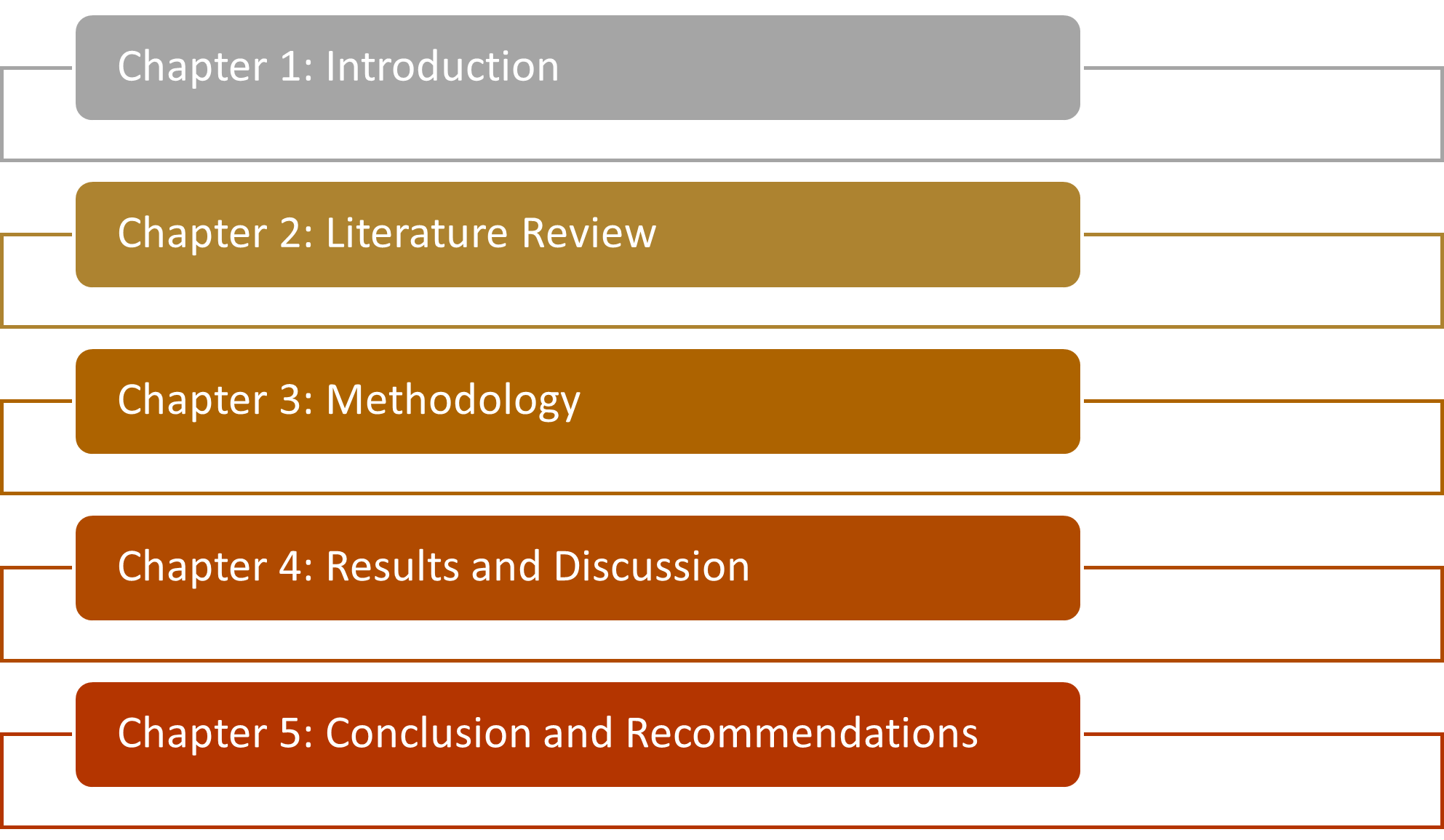
The impetus for conducting this research arises from the fact that the current classification of earthquake faults still has prepossessing errors and is time-consuming. Thus, the traditional migration methods, although serving as a starting point, are rather weak and cannot cope with the quantity and complexity of data that are generated by contemporary seismic monitoring systems. Machine learning is thus a revolutionary model as it utilizes efficient computational methods to interpret increased volumes of seismographic data coupled with improved accuracy in classifying intricate patterns. This research aims to expand on the methodological aspect, as the current approaches are rather basic, and machine learning can increase the possibilities for seismologists (Wen *et al.* 2023). It is by thus building, testing, and comparing a wide range of models that this study seeks to enhance the reliability of earthquake classification and, by extension, the means of making timely preparations and responses that are endeared to minimising the death tolls that could result from such unfortunate disasters. The intent of this work: The application of machine learning in seismology, especially to study earthquakes, has the potential of completely overhauling the way these natural events are monitored and prevented, thus making this investigation timely and imperative.

Consequently, the primary focus of this research is to provide a detailed analysis of different machine learning models that might help in improving earthquake classification and therefore contribute to more efficient preparation and response to disasters. The application of machine learning to seismology is a topic that has the potential to overhaul the current approaches for earthquake monitoring and prevention; this makes the book’s analysis timely and important (White *et al.* 2023). Besides, it also tries to reduce the gap between theoretical knowledge and practical Implementation in the field of earthquake risk management.

## 1.7 Scope of the Study

This research specifically deals with the use of machine learning algorithms for the analysis of seismological data towards the classification of earthquake occurrences. This process covers data gathering and cleaning of a wide range of seismic data and constructing as well as comparing and testing a list of machine learning algorithms including supervised algorithms like decision tree algorithms, and support vector machine algorithms as well as deep learning algorithms such as CNN. The effectiveness of these models is measured with the help of factors like accuracy, precision, recall and F1 score. Besides, it focuses on the technical characteristics of seismic data that affect the performance of these models and offers possible solutions for implementing such models into present seismological networks (Zhang *et al.* 2024). The work of the scope is limited to the classification of earthquakes only and does not include other aspects regarding loss estimation of earthquakes or ways of reducing such threats. The focused approach of this work is to offer feasible recommendations and hands-on methods to improve earthquake identification and categorization frameworks.

## 1.8 Dissertation structure



#### Figure 1.8.1: Structure of the earthquake classification report

## 1.9 Summary

This chapter provided an overview of the research on earthquake classification based on machine learning algorithms, explaining the importance of the research and its potential implications on the field. The paper described the purpose of the study, and its broad goal is to design and test machine learning models for identifying earthquake occurrences and described objectives and research questions that framed the investigation. Understanding these issues and the importance of this topic is provided in the background section: it describes typical methods and the new perspective of using machine learning in seismology, as well as the problem statement that highlights the current problems of this field and requirements for more effective solutions. The rationale explained the reason why machine learning was chosen as the subject of study by providing that it offered a means by which the classification of more accurate earthquakes could be achieved more quickly. The objectives of the study were formulated and the key point of the research related to the development and assessment of machine learning models was highlighted. In this first chapter, the reader is presented with the general context through which the subsequent chapters of the research will be developed.

# CHAPTER 2: LITERATURE REVIEW

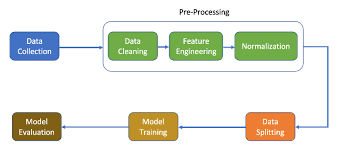
## 2.1 Introduction

Classification of earthquakes is a vital function in seismology as an aid in the analysis of seismic events and risks associated with them as well as the formulation of effective responses in case of occurrence of earthquakes. Earlier approaches that were used to classify earthquakes, including visual identification and probabilistic models based on some parameters derived from the seismic record, have helped to build the conceptual framework for earthquake study but cannot cope with modern sources and a wealth of information. These methods involve the use of professional judgement and can take a lot of time to complete hence delay in response efforts.

Over the past few years, ML has emerged as the go-to approach to modernise most classification methods and seismic data analysis. Supervised, unsupervised and deep learning methods of ML have been applied in the classification of earthquakes and it has been seen that they boost the probability of precise and improved classification (Abdalzaher *et al.* 2024). Decision trees and support vector machines (SVMs) are some of the supervised learning algorithms that can quickly learn features out of labelled data like seismic data for better classification of earthquake events while deep learning models work effectively for learning purposes.

## 2.2 Machine Learning in Seismology

Machine learning has gone a long way to improve the classification of earthquakes by providing effective techniques for analysing earthquake data. Among the various classes of supervised learning approaches used for earthquake classification, the most widely used is decision trees. The principle of decision trees is to split the dataset into subsets using the values of the characteristic defining a tree structure where each internal node is a test of a characteristic, each branch is the result of the test, and each terminal node is a class label or a decision. These trees are especially advantageous in the classification of earthquakes for their clear and easily understandable result where the seismologist can easily follow the steps or predominating factors that led to that classification (AlHamaydeh *et al.* 2024). Also, using decision trees can accommodate both numeric and alpha-numeric data, which may be common in seismic data including waveform features, geographical locations, and previous earthquakes among others.



#### Figure 2.2.1: Machine learning process for seismic analysis

Supervised learning also has popularity in earthquake classification and one of the highly used methods is random forests. Random forests can be defined as an ensemble method in the training process which builds multiple decision trees and, in the end, outputs a class that is the most frequent among the classes of the trees which were constructed. The trees in random forests work independently on a random sample of the training dataset as well as a random subset of the features to minimise overfitting. This Ensemble method helps to overcome the Biases and Errors of individual Decision trees and improve the Robustness of the Earthquake Classification models (Hasib *et al.* 2024). Random forests work best in noisy environments and learn the nonlinear transformations between seismic features to recognize different types of earthquake events.

Decision trees and random forest classifiers have also been widely used in earthquake classification problems to improve the performance and robustness of the existing seismic surveillance systems. All these machine learning techniques can then help to minimise the reliance on manual analysis for the classification of earthquake events, thus the application of these techniques will help in speeding up the process of detection and classification of earthquake events (Hu *et al.* 2024). They play an essential role in enhancing the formulation of appropriate measures of disaster preparedness and response since they offer regular information on the occurrence of earthquakes. However, certain issues, for example, the quality of labelled datasets, interpretations of these models’ decisions, and the utmost computational resources are still crucial. Following research work will be conducted in the future where these techniques will be fine-tuned, optimised techniques of both methods will be studied, and different data preprocessing methods will be incorporated with the existing methods to make the application of machine learning in earthquake classification more accurate and precise.

## 2.3 Application of Machine Learning in Earthquake Classification

Machine learning (ML) has proved to be of more help in the classification of earthquakes in that it provides various means of analysing the data, increasing the accuracy of the results and aiding in automating processes such as the detection of earthquakes. Supervised learning techniques are one of the most common ways of using ML in earthquake classification since the technique uses labelled data to classify seismic events according to different features.

Two of the most frequently used categories of supervised learning algorithms used in earthquake classification include decision trees and random forests. Decision trees subdivide the data using feature thresholds and the solution turns into a tree structure in which the branches result in highly specific classification choices. These trees are better for seismological analysis because of the transparency of the algorithm used and can accommodate numerical data that is related to waveform characteristics, magnitude, and categorical data related to the type of events and location of the event (Huang *et al.* 2024). They are interpretable, thus, it becomes easy for seismologists to understand how some aspects support the categorization of earthquakes.

Frameworks like decision trees were improved by random forests which is a sub-type of ensemble learning’s use of many decision trees that are created with the help of various subsets of data and features. The advantage of this method is that it gives better decision margins since the mean of the values is computed and low risk of overfitting (Murti *et al.* 2022). Random forests in other classification algorithms such as the onset, sophistication of affiliation, how seismic characteristics interact, and the increased solidity of classification models. It is also valuable in separating one kind of seismic event from another and reducing the effects that noise has on the data.

Furthermore, in feature engineering and selection ML algorithms choose the features that have the highest contribution to an earthquake. This makes data preprocessing less strenuous and increases the work of the classification models by isolating important variables related to the seismic activities.

In addition to the aforementioned approaches, modern state-of-the-art methods based on deep learning algorithms such as CNNs are emerging in the realm of seismic signal processing. CNNs are effective in learning the hierarchical representations of the seismic waveforms to extract spatial and temporal features for the classification of earthquakes (Nguyen *et al.* 2024). While more advanced in terms of their capabilities and not described in detail in this paper, CNNs open up interesting directions for further increasing the accuracy of classification, which can be especially useful when studying the finer details in the available seismic data.

However, there are some issues still today; the availability of high-quality labelled data, dealing with influential computational requirements, and the fairly important problem of model interpretability. As for future work, research will proceed on refining the current algorithms, incorporating the higher level methods, and making the body of work as well as the classification systems themselves more scalable and robust in terms of ML approaches to classifying earthquakes (Peters *et al.* 2024). Based on the presented study concerning the development of these technologies, they will continue offering great prospects for increasing the exploration of seismic occurrences and the resistance against the risks related to earthquakes at the global level.

## 2.4 Performance Metrics for Evaluating Earthquake Classification Models

The algorithms used in this study are summarised below, followed by an evaluation of the classification performance of machine learning models in earthquake classification. This is because various metrics are often employed to evaluate those models so that they can properly classify the seismic events and supply the pertinent data for applications on the seismic observing and early indication systems.

***Accuracy***

Accuracy calculates the ratio of correctly classified samples out of the total samples in the dataset. Although it gives the general notion of a model’s performance, it might not accurately represent the model for datasets that contain a highly imbalanced number of samples of different classes, for instance, the number of samples of non-earthquake events is much larger than that of events caused by earthquakes (Zhou and Lok, 2024).

***Precision and Recall***

Precision and recall have great importance especially while dealing with unbalanced datasets. Precision is the measure of the number of actual positive values relative to the number of positive values predicted by the algorithm. The term shows what proportion of the forecasted earthquake events were, indeed, accurate. Recall, also known as sensitivity, is the ratio of correctly identified positive predictions to accurately predicted positives; it measures the model’s capacity to identify all relevant earthquake events (Zhu *et al.* 2024). High recall often points to finding most actual earthquakes while high precision means most of the detected cases are quakes.

***F1 Score***

F1 score is calculated as the average of precision and recall in the same way as their combination in the denominator, which balances between false positives and false negatives. It is suitable when the value of precision and recall is almost equal since it provides a single measure of the model’s performance.

***Confusion Matrix***

The confusion matrix gives the results of classification to show true positive, true negative, false positive and false negative of the outcomes. It gives the outline of specific types of classification errors that can be used to make modifications to the model (Apriani *et al.* 2021).

Based on these measures, researchers and practitioners can have a detailed perception of the precise effectiveness of machine learning models in earthquake classification and identification. Appropriate testing guarantees the models are accurate and give long-lasting solutions that will assure the efficiency and effectiveness of the Seismic monitoring systems, as well as increase disaster preparedness and mitigation.

## 2.5 Challenges and Future Directions

In earthquake classification, several ML techniques have been shown to exhibit promising results; however, they have limitations that would have to be resolved to maximise their potential. Such as the quality and availability of the labelled seismic data is a major issue. Creating accurate ML models demands large-quality annotated datasets, but obtaining such datasets is challenging because of differences in seismic activity across different regions, data privacy laws, and the high costs of maintaining seismic networks (Ayele and Ray, 2023). There are several ways to address this challenge which include improving the collection methods for data, encouraging data sharing as well as data anonymization procedures and proper protocols added to the data annotation process.

Many studies in earthquake detection overlook the importance of preprocessing and its impact on model performance, resulting in inconsistent data handling and suboptimal outcomes (Ridzwan and Yusoff,2017).To address this, the report systematically preprocesses data, handling missing values and encoding categorical features like to ensure standardization. Additionally, existing research often lacks comprehensive model comparisons and relies on limited metrics like accuracy, failing to identify the most effective algorithms or evaluate performance holistically(Avinash Bhandiya,2024).

Another important topic is the explainability of the developed ML model. Although deep learning-based methods such as CNN deliver competitive results, the mechanism of the decision-making is opaque hence challenging to validate by seismologists. It seems significant to propose explainable AI techniques like feature importance analysis and model visualisation for improving the models’ interpretability and increasing their acceptance among the stakeholders.

The memory and computation demands needed for feeding, training, and running of ML models, especially deep ones, represent another difficulty. These models require a lot of processing power, during both training and day-to-day data processing, which remains a challenge in the current working environment (Blomeier *et al.* 2024). There are potential solutions such as optimising current algorithms, utilising parallel processing and building on cloud computing to address these computationally intensive demands and to scale up current ML-based earthquake classifications.

The usage of ML in the practical, working seismic observatory and early warning systems also poses technical and operational difficulties. Such systems require better and real-time models which should be able to give better and timely alerts. Preventing such losses requires the close cooperation of ML researchers, seismologists, and disaster management authorities to produce models that are fit for purpose and can provide actionable insights during earthquakes.

Further, the benefits and ramifications of applying ML in categorising earthquakes are not well elucidated in terms of ethics and society. Other factors include data privacy, the ability of patients to afford the latest monitoring devices, and the socioeconomic effects of receiving fake alarms or failing to detect an alert. These concerns show that it is critical to work on the frameworks that would help regulate how AI application is used in seismic activity monitoring.

## 2.6 Summary

This chapter described the use of ML in earthquake classification with a focus placed on the possibilities it brought and the difficulties arising from its implementation into practice.

Earthquake categorisation is now much better due to machine learning (ML), which automates procedures, increases accuracy, and overcomes the drawbacks of conventional techniques. Methods such as random forests and gradient boosting offer improved classification reliability, tolerance against noisy data, and transparency. There are still issues, though, like the requirement for high-quality labelled seismic datasets because of geographical variations, data protection regulations, and the processing requirements of CNNs and other deep learning models. Because ML models' decision-making processes are frequently opaque and difficult for seismologists to test, interpretability is still crucial. Effective early warning systems also depend on real-time data processing and the incorporation of machine learning models into current seismic monitoring systems. Consideration must be given as well to ethical and social issues, such as data security, equity, and the economical effects of false alarms.

# 2.7 Literature Gap Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Authors | Dataset used | Techniques and Algorithms | Accuracy and quantitative result achieved | Limitations |
| Abdalzaher, M.S., Soliman, M.S., Krichen, M., Alamro, | Time-series seismic data(INSTANCE) | KNN, Decision Tree | 94.85% | Less diverse dataset and lack of detailed alert levels |
| Mohammed AlHamaydeh | NGA-west 2 | CNN, RNN | 5.41 × 10−5 1.96 × 10−5 | Single prediction and covers only seismic signals and not disaster management |
| Bao, Z., Zhao, J., Huang, P., Yong, S. and Wang, X.A | Electromagnetic signals | Deep learning | High | Multi-level alert classification not done |
| Hasib, M., Bagas Anwar, A.N., Huffaz Muhammad | Volcano related dataset | K -means | Classification model | Not focused on seismic data |
| Aden‐Antoniów, F., Frank, W.B. and Seydoux, L | Earthquake catalogs (declustering). | CNN | Random forest | No comparison or evaluation done |
| Huang, W., Gao, K. and Yu, F | fault dynamics data using FDEM | Light GBD | - | Model evaluation is low |
| Muhammad Ary Murti | ESM database | ANN, random forest, Decision tree | 67%-96% | Low accuracy |
| Akyol, A.A., Arikan, O. and Arikan, F | ionospheric data. | Machine learning for precursor detection. | 88% | Low Data quality and not eligible for real-time classification |
| Peters, S., Liu, J., Keppel, G., Wendleder, A | Multispectral S2 data | SVM, Random forest | 89% | Less diverse with data |
| Zhou and Lok | Dispersion curves data | Decision Tree, Random Forest | 73.33% | Less impactful domain |
| Zhu, F., Zhu, C., Lu, W., Fang, Z., Li, Z. and Pan, J. | Soil data, remote sensing data | XGBoost, Rpart | 70%-80% | Low accuracy |
| Apriani, M., Wijaya, S.K. and Daryono | Seismological data | DNN, Random forest | 80% | Less computationally efficient |

# CHAPTER 3: METHODOLOGY

## 3.1 Introduction

This chapter describes the procedure for classifying datasets of earthquakes via machine learning techniques. The objective of this research is to predict the likelihood of occurrence of earthquakes with their magnitude, depth, and position on the ground. The target variable for classification is the "alert" level, which includes four categories green, yellow, orange and red, where green represents low warning, orange and yellow intermediate and red as the highest alert level(high warning). This chapter introduces methods like research strategy, research approach, research design, method of data collection, analysis of data and lastly consideration of ethical issues in the study.

This is to establish a step-wise plan for the attainment of the various research objectives. Descriptive research methodology is used in the study to provide earthquake information because it illustrates the characteristics of the given dataset, while quantitative research methodology is used to analyse and generalize characteristics of the given data with the help of statistical and machine learning tools. The information used for this study is obtained from secondary data shared on the Kaggle platform that includes multiple characteristics that point out all possible earthquake classification.

## 3.2 Research Strategy

In this study the research strategy used is descriptive and the goal is to systematically find a variety of patterns in the provided dataset on earthquakes. Exploratory research is useful in investigating an event and trends since this type of research only captures and explains the event without manipulating the factors. Concerning earthquake classification, such a strategy helps in identifying the relationship between the cause and effects of the various attributes of the earthquake including magnitude, depth, location and the alert levels.

This strategy makes it possible to scrutinize the dataset in a manner that allows one to identify the relationship that might not be so obvious in the dataset. To achieve this, the study aims to perform an ambulance cross-sectional analysis of the various alert levels to establish the general patterns of earthquakes at various alert levels (Chen *et al.* 2024). Descriptive analysis enables the identification of significant variables that determine the degree of earthquakes to be shown to the learners.

The strategy aims at finding useful relationships in the given dataset, giving overviews of the earthquake events without the given goal of predicting further events and establishing given correlations. In addition, since this is a descriptive study, patterns of seismic history data are explained in a very clear and coherent manner, which is of paramount importance for the construction of a reliable classification model. The descriptive strategy is thus aligned with the overall goal of the research and thus, enabling the determination of the intensity of earthquakes by observed characteristics only and achieving a high predictive power through the methods of machine learning.

## 3.2.1 Justification of Research Strategy

Descriptive research is utilized as the strategy for studying earthquakes because it affords the opportunity for analysing the characteristics of the events and recognition of the patterns without alteration of the data. A number of prior works (e.g., Chen et al., 2023) have found that exploratory analysis done on earthquake data reveals important associations between the corresponding attributes like magnitude and depth as well as location. This is especially important in a discipline where the patterns with data are rarely simple or are directly traceable. One of the purposes of descriptive analysis is to help get a first quantitative grasp of the situation, so it is suitable for the task at hand.

## 3.3 Research Approach

This research employs a quantitative method to model and predict earthquake intensity using machine learning. The quantitative approach helps in determining numerical data and the interactivity between different measured parameters within the dataset including magnitude, depth and geographical location (Choi *et al.* 2024). The major strength of this approach is to deal with a large volume of data and use it to build models through statistical analysis instead of being based on experiences.

Quantitative research also deals with numbers and involves the use of statistical techniques. As for the earthquake events, they are characterised by more tangible characteristic(magnitude, depth and intensity) and these characteristics are calculated to determine their link to the represented severity levels in the “alert” column.

Therefore, algorithms that come under the use are Naive Bayes, Gradient Boosting and Random Forest, these are applied in developing the models which belong to the quantitative category since they involve numerical computations in establishing patterns from the data. The feasibility of these models is then evaluated with quantitative measures such as accuracy, precision, recall and F1-score in ways that the models are measured by tangible results (Fan *et al.* 2023). Furthermore, the quantitative approach is highly objective due to its scalability and its ability to have the same procedures undertaken using other datasets to prove its credibility.

This strategy is suitable for the study, which is earthquake classification because it requires the analysis of a considerable amount of data and the application of quantitative procedures in the data analysis and would reveal general trends that can hardly be observed when the analysis is done qualitatively. It grants one with a concrete conception of how earthquake characteristics play a role in the extent of their impact and therefore the approach to estimating the earthquake’s exponential nature and its concrete application in the prevention of disaster occurrences.

## 3.3.1 Justification of Research Approach

In particular, the quantitative approach, especially in the machine learning framework, has been shown to be efficient for classification tasks in earthquake engineering. Research like those conducted by Gu et al. (2024) show that powerful features like Random Forest and Gradient Boosting can be used to classify seismic events. These models can work on big data and can model non-linear interaction of parameters describing earthquake characteristics which are useful for predicting probability of future earthquake impacts. The large and varied data sets inherent in earthquake prediction require the scale and objectivity of current machine-learning techniques.

## 3.4 Research design

The research method used for this study is cross-sectional method where the data is captured and analysed at one point in time. Since there is no need to test or observe the before and after change of event the cross-sectional design is suitable for this study. This design allows the categorization of earthquakes with different features such as magnitude, depth, and the area of occurrences so that it can predict the level of alert(target variable) for each of the earthquakes.

In this research, this dataset offers specific details of earthquake events that have taken place in the past with each event characterised by various attributes. The aim is to categorise these events into severity levels of green, yellow, orange and red and the cross-sectional design is appropriate to examine the current correlations between the independent variables and the dependent variable (Feng *et al.* 2024). This type of design is useful for analysing a big set of data and does not require constant data collection, hence it is both time-saving and economical.

By using cross-sectional study, the machine learning models are trained as well as tested in a single dataset hence providing a generalised approach to earthquake classification. This is not a system designed to forecast future changes in earthquake behaviour but to analyse patterns from the coarse data and give efficient output(early warnings). The models derived from this design can be used on the newest data of earthquakes which makes them beneficial for classifying the effect of earthquakes in real-time mode concerning the data obtained in the past and also the machine learning models give better accuracy. In summary, the cross-sectional research design makes the kind of research plan possible and systematic adequately to serve the aims of the study.

A diagram of a model

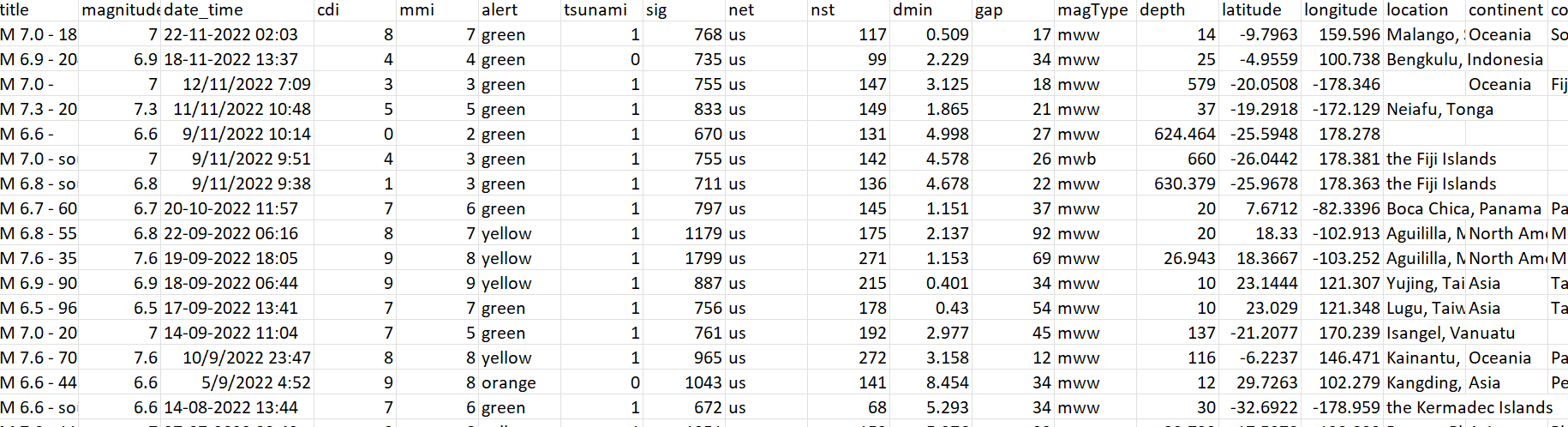
Description automatically generated

#### Figure 3.4.1: Flow Chart showing the experimental pipeline

## 3.5 Data Collection Method

This work employs secondary research data and downloads an earthquake dataset from the Kaggle. Secondary data can be defined as data collected and analysed by other people and it is appropriate for this research due to the following reasons: The first advantage of using secondary data is that it is relatively easier to obtain several methods which are time and cost-saving. Besides, secondary data is generally collected from credible sources and thus there is always a sense of quality and standard that helps develop more consistent findings.

In the context of this study, Kaggle’s dataset contains records of earthquake occurrence from different parts of the world with all important attributes such as magnitude, depth, and location. By its nature, this dataset is selected because it is very valuable and it contains all the variables required for classifying machine learning models making it relevant to the study’s goal. Also, such an approach to the analysis of datasets is beneficial because the data are public, and others can conduct similar analyses or expand upon the research findings.



#### Figure 3.5.1: Dataset based on earthquake alert

The dataset includes key features relevant to earthquake classification, such as:

Title: The name of the Seism which is the name of the event which took place, Magnitude: The earthquake’s strength, Date and Time: That means they reveal the exact time of occurrence, CDI (Community Determined Intensity) and MMI (Maximum Instrumental Intensity): Standard results of earthquake intensity, Alert: The variable that needs to be classified into “green,” “yellow,” “orange,” and “red”, Tsunami, its importance referred to as Significance (Sig), and other geological factors including depth and location.

## 3.6 Data analysis

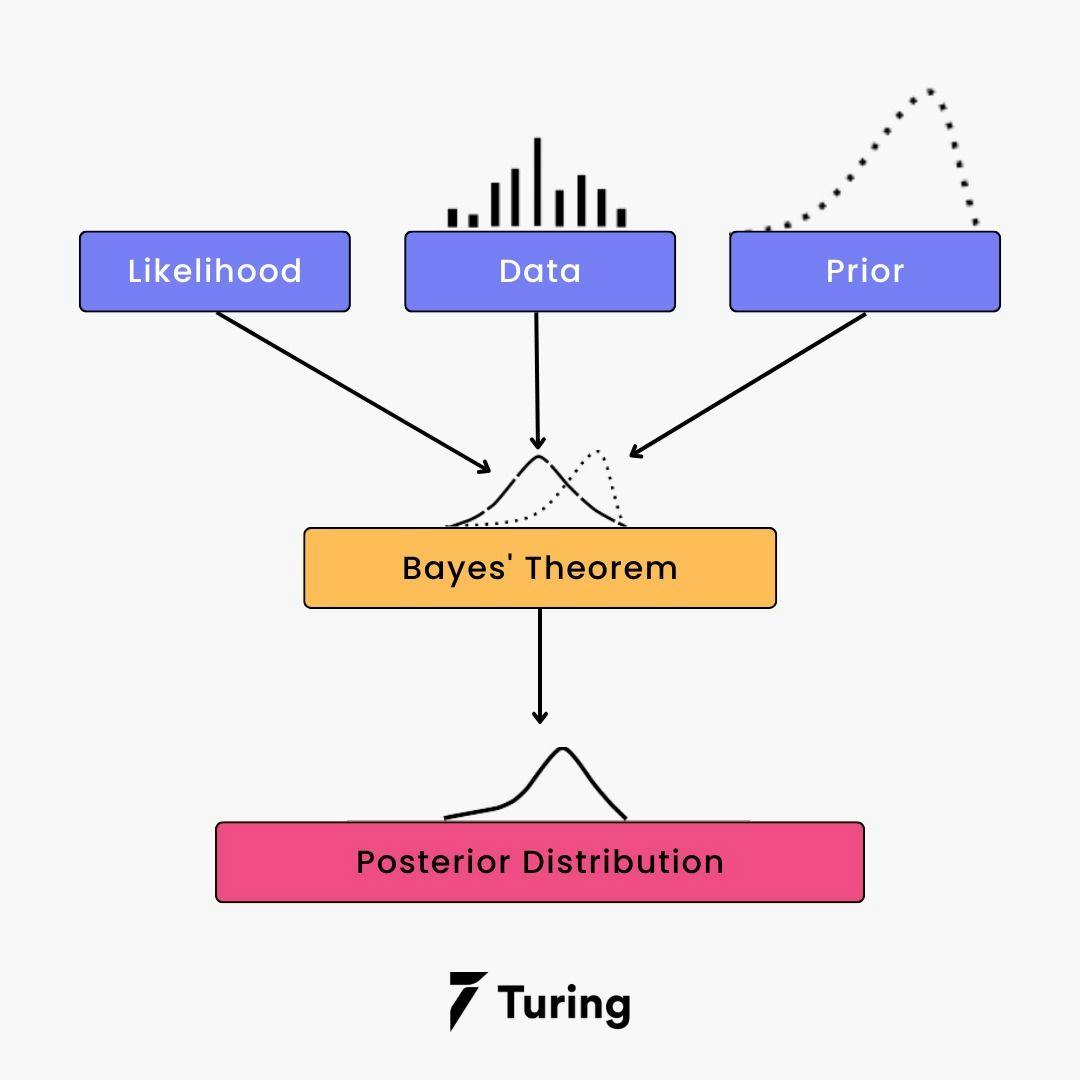
The steps for data analysis for this study are data preprocessing, data exploration, and finally, the classification using machine learning to determine the attendees of the earthquake(an accurate image of the working of this study in shown in fig3.5.1). First data cleaning is performed on the dataset so that the data is fit for analysis and modelling process before the current analysis is done. Some of the processing steps in this are missing values, normalization, and encoding of the features. We use Exploratory data analysis which involves examining the data and preparing an exploratory report of the uploaded data, Concerning the handling of missing values, this is done using some imputations like the mean or median in such important columns like CDI, MMI and NST. Normalization is used on features such as magnitude, depth and significance to make them comparable and help prevent prejudicing the performance of an algorithm due to a feature’s range (Gu *et al.* 2024). Some of the outcomes of the data preprocessing step are the conversion of categorical features “Alert” level and “MagType” indicator using label encoding so that the machine learning algorithms can respond to them adequately.

This is followed by Exploratory Data Analysis (EDA) which is used to discover the possible correlation between the features. Histograms, scatter plots, and heat maps are adopted for Visualisation and analysis of characteristics for the given variables in terms of magnitude, depth, and location, as well as, their relationship with the alert levels. EDA further assists in recognizing hidden trends in the data that may have an impact on the outcome of the model. For instance, it may unveil information that greater magnitude and deeper earthquakes are inclined to acute alert level rates. After data pre-processing and data exploration machine learning algorithms are used to categorize earthquakes into four different alert levels (Hosseinzadeh *et al.* 2024). The study employs three models: Naive Bayes, Gradient Boosting and Random forest. Each model is been trained and evaluated with the newly encoded(numerical) data and the results are obtained in terms of accuracy, precision, recall and the F1 Score.

Evaluation Metrics: Accuracy gives an overall percentage of correct predictions, however, it can be quite misleading especially where data is imbalanced. From the confusion matrix shown in chapter 4, one can observe that Naive Bayes made some classification errors, however, Random Forest and Gradient Boosting were slightly better.

**3.6.1 Justification of Algorithms**

a) Naive Bayes:



#### Figure 3.6.1.1: Naïve Bayes

Naive Bayes is mainly used because of the effectiveness of this algorithm in the probabilistic classification. As is the case with features that are highly correlated, such as magnitude and depth, it was not very good at determining between those features nonetheless, it serves as a starting reference point (Fan *et al.* 2023).

b) Random Forest:

A diagram of a tree

Description automatically generated

#### Figure 3.6.1.2: Random Forest Algorithms

As a decision tree based BOF, Random Forest is also capable of handling both categorical and numerical features, and also help decrease overfitting. It has been applied in earthquake prediction (Hosseinzadeh *et al.* 2024) due to its ability to construct intricate correlational structures within features as well as guarantee high accuracy.

c) Gradient Boosting:

A diagram of a tree

Description automatically generated

#### Figure 3.6.1.3: Gradient Boosting Working

In each stage of Gradient Boosting the models are constructed in a sequential way to overcome mistakes made in the previous stage. This method is especially useful for a large number of samples with an imbalance problem as demonstrated by Feng *et al.* (2024) making it ideal for earthquake classification where imbalances occur with respect to alert levels.

## 3.7 Ethical Consideration

By following the best practices in data science concerning model development, the research process is done in a black-box style. Data pre-processing, model training, as well as model performance measurements are done in strict accordance with scientific standards with no over-emphasis or repression of results. This earthquake dataset under analysis is accessed from Kaggle and is open for use under a non-commercial license. One has to abide by the Data License and it should also be noted that the given data set is for Research purposes only. Also, the proper citation of the data source used in the study is well observed throughout the study.

Even though there are no personal or sensitive data, it is crucial to adhere to ethical issues while working with the datasets. The models presented in this research work are as designed for research use, but should not be applied directly to real-world decision-making processes without the validation and supervision of seismological indications. Society needs to forecast the degree of the earthquake as it may have an impact on people’s lives and their property. At the same time, it should be stated that machine learning models are not ideal and should be used in combination with the specialists’ knowledge and the data from monitoring systems is in real-time mode (Jain *et al.* 2023). The models described in this study are not supposed to provide a basis for a new completely different system from the current existing Seismic Early Warning systems, but as other tools that may be used in conjunction with the current systems.

Common ethical challenges associated with the application of machine learning algorithms for earthquake classification include data privacy if the data involved is human related, and the problem of biases given that the algorithms will inform the public’s safety. It is also important to explain why an algorithm makes the decision, and what the consequences are for the environment.

Commercial and economic factors are related to costs which include data gathering, model building, and maintenance. Seismic risk reduction could potentially lower the likelihood of damage to structures, possibly in the millions, and misclassification could result in costly false alarms or underestimation which would affect insurance and construction businesses. A good compromise between cost and precise results might be one of the keys to expanding usage.

## 3.8 Summary

This chapter describes the process for predicting earthquake warnings (green, yellow, orange, and red) based on variables including magnitude, depth, and location and then classifying earthquake datasets using machine learning algorithms. The dataset, which was gathered from Kaggle, is analysed using a descriptive and quantitative research approach to find patterns and relationships. An effective and methodical technique is provided by the cross-sectional design, which permits research at a single moment in time. Normalisation, addressing missing numbers, and storing categorical information like alert levels come data preprocessing. By finding patterns and correlations, exploratory data analysis (EDA) reveals connections between characteristics like magnitude and alert levels. Basic ethical consideration and model requirements with respect to commercial and economical factors are also explained.

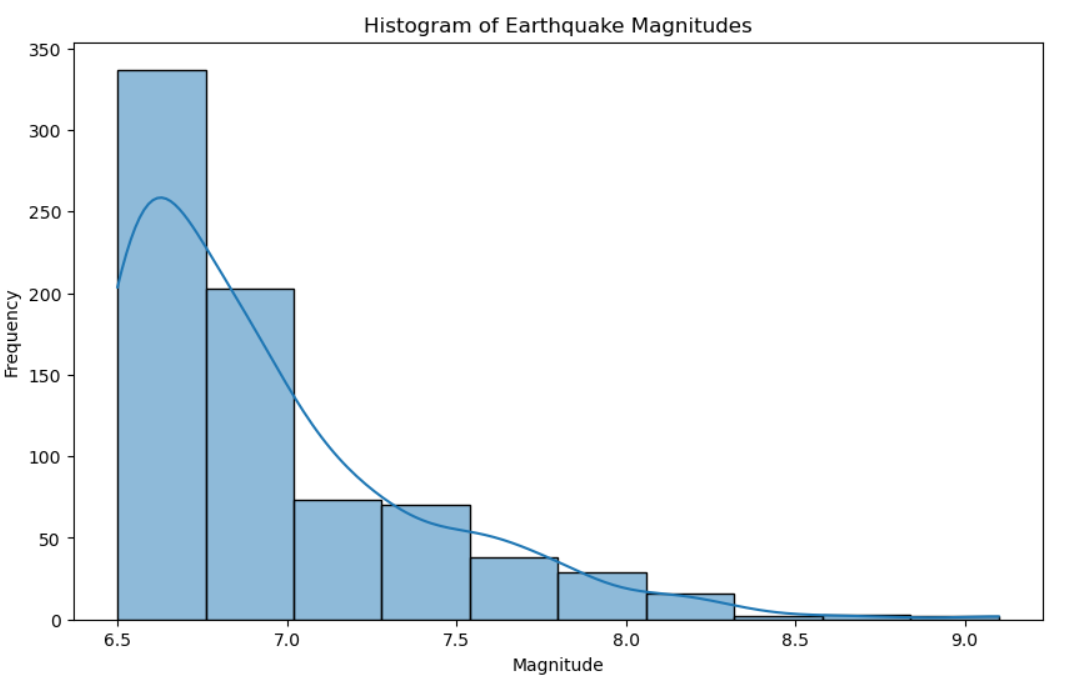
# CHAPTER 4: DATA ANALYSIS AND MODEL PERFORMANCE

## 4.1 Introduction

This chapter provides a detailed description of the earthquake dataset and the results generated from the application of different machine-learning algorithms used to forecast alert levels based on specified aspects. The data set belongs to the disaster domain and has seven eighty entries that show various aspects of the character of an earthquake. The primary objective of this analysis is to examine the dataset, perform exploratory data analysis (EDA), and evaluate the effectiveness of three different classifiers which are Naive Bayes, Random Forest and Gradient Boosting.

## 4.2 Data Exploration

The first move in the process was to read the data from an earthquake.csv file and place them in a data frame using pandas. The structure of the data, and the information contained here are closely examined to get acquainted with the data. Other attributes contained in the dataset included the magnitude and depth of the earthquake, CDI (Community Determined Intensity), MMI (Modified Mercalli Intensity), alert levels, tsunami data, importance(sig), coordinates, and many others (Iaccarino *et. al* 2021).



#### Figure 4.2.1: Histogram of Earthquake Magnitudes

When working with the dataset, it was established that the majority of the columns contained the full set of values for the coordinates while others had some attributes with missing values due to which the variances of the ‘Alert’ and ‘Location’ columns stated that some of the earthquakes were not given an alert level or that their geographical location was unspecified. To avoid this the missing values, the alert column were replaced with a default value of “red” as the formula above was used. However, missing values had to be filled to make further analysis possible and upon the elimination of duplicate entries, the current data set stands at 773 records.

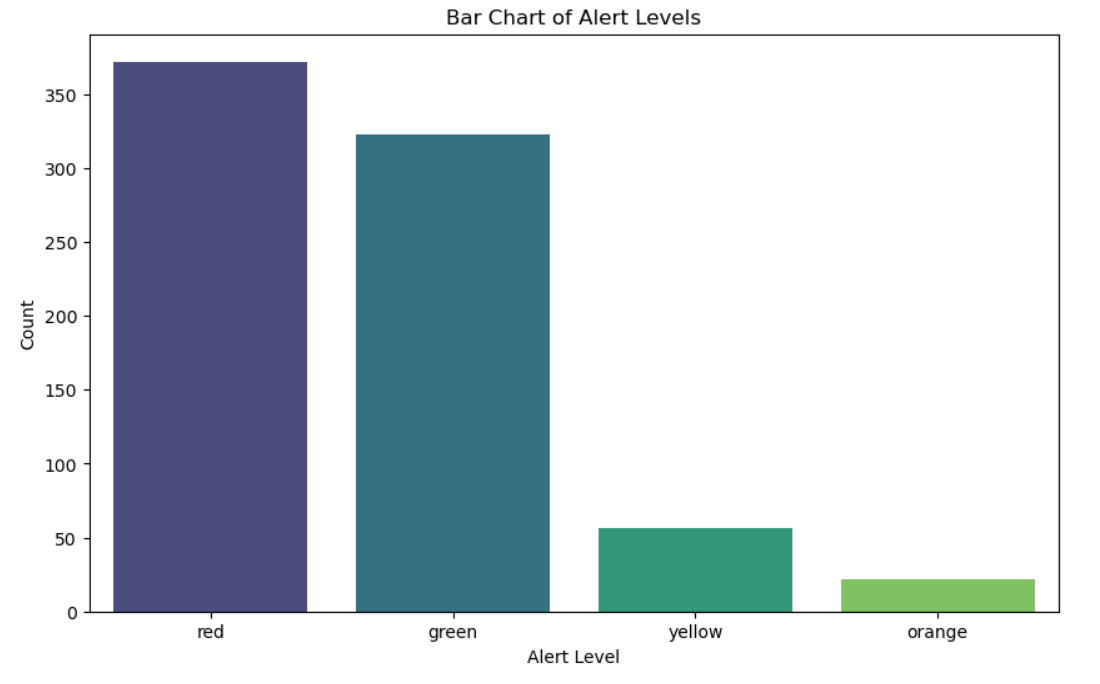
A screenshot of a computer

Description automatically generated

#### Figure 4.2.2 Selected features and missing values

From the analysis of the features derived from the earthquake dataset, it was possible to unravel a world of information regarding earthquakes. There are many attributes about every earthquake like magnitude and depth – these two are measures of the threat. The magnitude, which is continuous and ranges from 6.5 to 9.1 helps distinguish the intensity of each earthquake’s depth expressed in kilometers and reveals how far beneath the Earth’s crust the earthquake took place. Another clear discovery made during the analysis is the correlation between size and depth which is most often used in earthquakes of small depth and are more catastrophic than those of greater depth, although there are cases here as well (Ma *et. al* 2020).

To complement these findings, the analysis considers other impacts of earthquakes, such as the intensity of shaking based on CDI and MMI indexes, both of which scale the intensity of shaking felt at different locations. In the communities, the CDI scale, which ranges: between 0 and 9,, describes the level of shaking experienced while MMI give the seismic intensity at a particular place. One major finding from the data exploration assessment was that increased earthquake magnitudes seemed to be directly proportional to CDI and MMI which showed that the strength of the earthquakes increased the degree and extent of damage.



#### Figure 4.2.3: Bar Chart of Alert Level

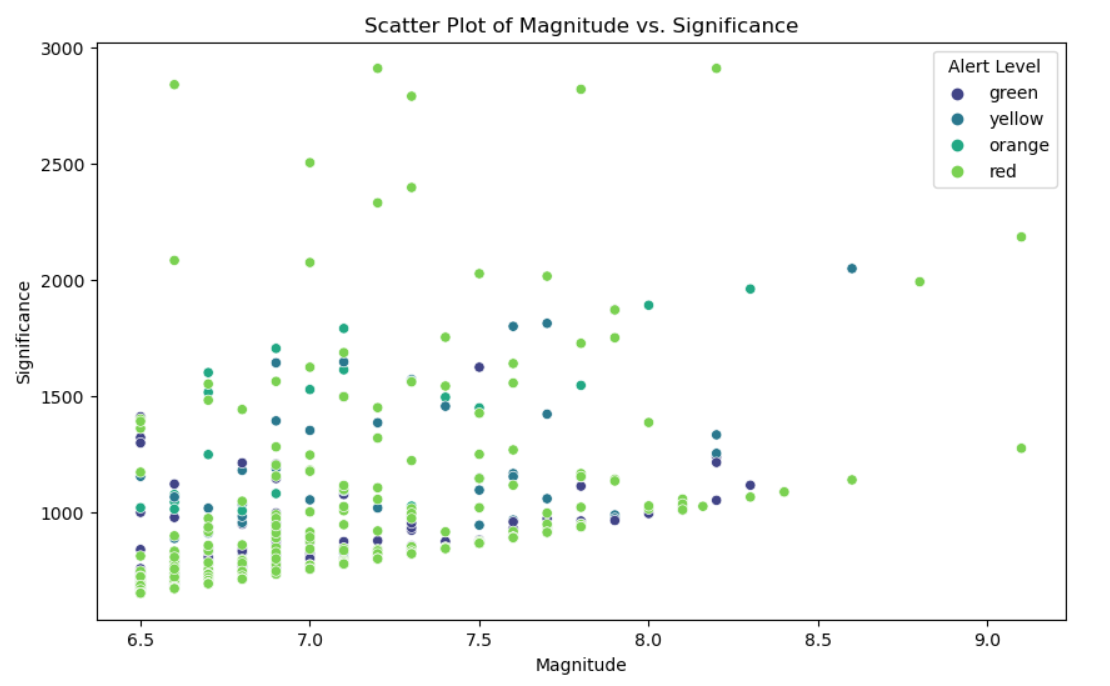
Furthermore, the data set contains nominal features like magType, alert and so on. The magType gives information about how the methods are used to come up with respect to the magnitude of the earthquake, and the alert level gives information on the level of danger posed by each earthquake. A similar distribution of the types of alerts demonstrated that the green, representing relatively low threat level which basically dominated the overall picture, while the red alerts signifying the presence of immediate threat which are exceptionally rare (Aden‐Antoniów *et. al* 2022). These differences in the alert level led to the need to explore the conditions leading to their activation, the location of the earthquakes and the effects they have on the societies in the regions. In aggregate, the data exploration phase was important to figure out pertinent details and connections that would be used to guide the subsequent modelling phase.

## 4.3 Exploratory Data Analysis (EDA)

The exploratory analysis was performed to understand the nature of the chosen dataset. The first visual we received was a histogram that described the distribution of the earthquakes by their magnitude. Precisely from this histogram, it was shown that the majority of the earthquakes recorded were of low values with few higher values (Hussain *et. al* 2022).

The analysis of the alert log, the bar chart was created to display the count of the various alert levels based on the histogram constructed above. This was in agreement with the results highlighted in Figure 4.2.1 that more instances of ‘green’ alert levels, of low risk were observed, as compared to ‘red’ and ‘yellow’ alert levels of higher risks. By making this observation, it was possible to assert the safety of the regions under study in terms of several earthquakes. Subsequently, correlation analysis was performed to see whether there is a relation between the different magnitudes and the significance of earthquakes. The located scatter plot showed that most lower magnitude earthquakes have different levels of significance whereas high magnitude has higher significance. Alert level took a different color and hence, this provided the context of the relationship between an alert and the severity of an earthquake.

The last visual used in the EDA phase was the correlation heatmap. This heatmap depicted the inter-numerical features associations present in the dataset. It was noted that there lies a high, positive correlation rate between the variables of magnitude, CDI, MMI, and sig implying that these variables have a significant relationship as far as the intensity and effects of the earthquake are concerned. Such understanding was helpful especially when selecting features for use in the next modelling phases (Bijelić *et. al* 2020).



#### Figure 4.3.1: Scatter Plot of Magnitude and Significance

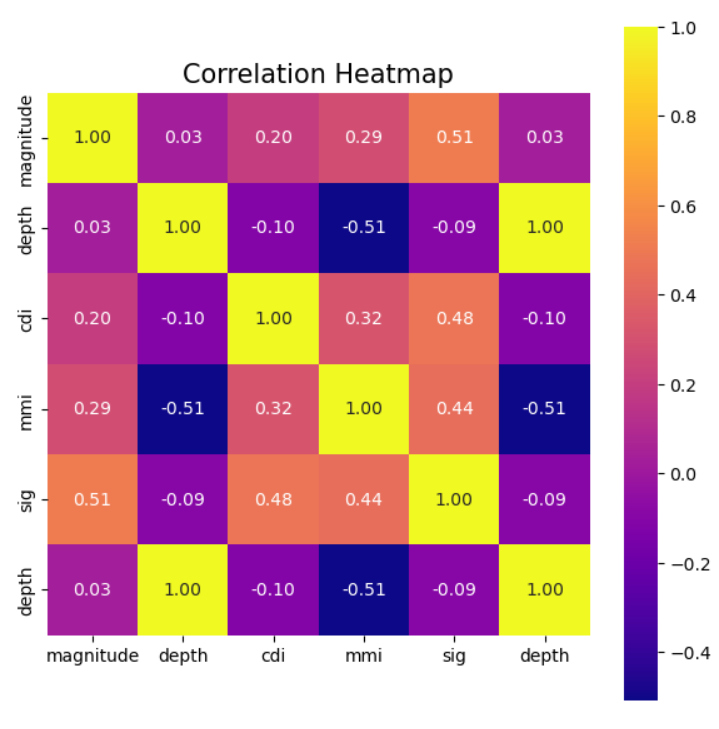
The EDA probed deeper into other relations involving the features of the datasets but with emphasis on the geographical properties of earthquake events. The occurrence of earthquakes was then represented on a geographical scatter plot against the latitude and longitude of the events. From this visualization, I was able to identify differences in the geographical distribution of earthquakes within these clusters- to some extent proving the hypothesis of tectonic plate boundaries affecting the regions in consideration. Such geographic information can be greatly useful for researchers and decision-makers in order to define zones which have a higher probability of experiencing seismic events.

Further, an alert level-wise box plot was prepared to understand the spread and distribution of earthquake magnitude more effectively. This box plot suggested that the relationships between higher alert levels and greater magnitude added to the learning that greater earthquakes are more dangerous. Even in the box plot, I noted that some earthquakes were extremely powerful in which few of them drastically affected the structures and response plans developed at national and global levels.

## 4.4 Data Preprocessing

The third stage of the study is data preprocessing where the data is used model training. The features being used in the decision-making process of the machine learning models were the magnitude, depth, CDI, MMI, magType, sig, and alert. The alert variable needs to be classified is converted into numerical values from its categorical form through label encoding.

The features matrix, represented by X, was created by excluding the alert column while the vector y contains only the alert column. The given dataset was further divided into training as well as testing data by using a method of split ratio of about 70:30. Since using separate datasets, the training set was used for the training of the models while the testing set was used to evaluate the models on some unseen dataset.



#### Figure 4.4.1: Correlation Heatmap

The dataset was checked for missing values and found lacking 27 values which were imputed accordingly After removing the duplicates; the study included 773 records (Cano *et. al* 2021). This step made it possible to have only one instance of each earthquake in the training of the model so that there is no favoritism towards some instances because they have been repeated.

After the data cleaning process, encoding was done so that categorical variables could be changed to numerical format. MagType and alert were two columns for which label encoding was used to convert string data type values into integer type. This change is crucial in many machine learning algorithms because most of these algorithms work optimally when they receive numerical inputs to compute the desired predictions.

Lastly, the data was split into feature matrix X excluding the alert and targeted vector or list y including the alert levels (Ogunjinmi *et. al* 2022). This rigorous approach to data preprocessing provided a well-formatted dataset towards the machine learning models so that more sound predictions of the alert levels of earthquakes can be accomplished.

## 4.5 Model Training and Evaluation

Three classification schemes were used to predict the alert levels given the selected features. The first used model was the Naïve Bayes Classifier which is more appropriate under probabilistic classification. With the data obtained in the work, the Naive Bayes model was trained on the training data set and then made predictions on the testing data set (Saad *et. al* 2021).

The Naive Bayes classifier demonstrated a very low accuracy rate of 53.45% which indicates this classifier does not perform well for this type of classification. This is evidenced by the confusion matrix as shown by high accuracy rate on “alert level 0” (green) matrix value of 0.91. However, the classifier has more problems classifying the other classes of alert levels. For example, low precision and recall have been obtained for the “alert level 3” (red): precision = 0.07; recall = 0.18. This imbalance indicates that Naive Bayes has high bias towards the majority class alert level 0 while minority classes are not well classified. The classification report supports this observation indicating that the model is skewed by a high accuracy on class 0 with relatively low precision and recall on other classes. The macro average precision is 0.45, recall is 0.53, and the F1-score is 0.40, and it also does not correctly handle multi-class classification when it is highly imbalanced.

A screenshot of a computer

Description automatically generated

#### Figure 4.5.1: Naïve Byes Classification Report

In comparing the two classifiers, the Random Forest classifier did far better, with an accuracy of 84.48%. The confusion matrix shows that it generally does a good job in identifying alert level 0 (green), with high value of precision = 0.86 and recall = 0.91. But it is somewhat problematic for alert levels 1 (yellow) and 3 (red). In particular, the recall of the model for the example of alert level 1 is rather low, 0.17, while the precision is 0.33. Likewise, where alert level 3 is concerned, both precision equal 0.50 and recall equal to 0.35 are lower, which suggests that it is hard to accurately predict such classes. Nonetheless, when taking an average of the scores achieved for each level by the Random Forest classifier, the overall score was 0.84 F1-score and this signifies that the classifier is very efficient in alert level classification, especially for the majority classes..

A screenshot of a graph

Description automatically generated

#### Figure 4.5.2: Random Forest Classification Report

The Gradient Boosting classifier too had a classification accuracy of 84.05% which is slightly lower than the Random Forest model. Speaking of the alert level of 0 (green), Gradient Boosting performs excellently with the highest value both precisely equal to 0.84 and recall that equals 0.93. The model can also classify alert level 2 (orange) considerably successfully with an accuracy of 0.89 and relevancy of 0.87. As with Random Forest, this handler has issues with alert levels 1 and 3. For example, alert level 1 (yellow) yielded 0.75 precision and 0.50 recall, and alert level 3 (red) yielded much lesser recall of only 0.24 but 0.40 precision. However, the proposed model offers reasonable performance for major classes; the weighted average of the F1-score is 0.83. The Gradient Boosting also seems to correctly identify the patterns in the majority classes but remains a challenge in predicting the alert levels 1 and 3 earthquakes.

A screenshot of a graph

Description automatically generated

#### Figure 4.5.3 Gradient Boosting Classification Report

The model with the highest accuracy is been tuned by using hyperparameters which in this study is Random Forest and it shows an accuracy of 83.19% up from 84.48%. Other hyperparameters include the number of estimators which is 50, max depth which is None, min samples split which is 5, min samples leaf which is 2 and max features as “sqrt” which generalizes better. The methods of confusion matrix analysis show that the model has a higher success rate in most of the classes but a very low success rate in alert level 1, especially the rare classes. The classification report shows a fairly high precision and recall for levels 0 and 2 for alerts but the performance on alert level 1 is relatively low, and it seems that fine-tuning of details or balance of data could improve the results for minor classes.

**A screenshot of a computer screen

Description automatically generated**

#### Figure 4.6.2: Hyperparameters tuning for best model

## 4.6 Model Comparison

As to the need for a clear comparison of the model performances, a bar chart that depicts the accuracy scores of the 3 classifiers was developed. By looking at the chart below, the performance of the Random Forest and Gradient Boosting models is much higher compared to the Naive Bayes classifier. This comparison has shown that it is crucial to choose the correct model which will predict the levels of an alert in dependence on a characteristic of a dataset. Also feature selection and comparison is done efficiently.

In general, the Random Forest classifier provided the highest results for accuracy and model reliability in the categorization of alert levels of earthquakes. The cross-system analysis successfully confirmed that by using machine learning methodologies the comprehension of disaster areas and future effects of an earthquake could be improved, thus improving the risk evaluation and emergency management planning (Akyol *et. al* 2020).

A graph of a graph showing different colors

Description automatically generated with medium confidence

#### Figure 4.6.1: Model Accuracy Comparison

For this particular analysis, the Naive Bayes classifier proved to be effective only in certain areas achieving 53.45% accuracy. These observations were also evident from the confusion matrices of each model. Several observations can be made, specifically Random Forest model was better placed at classifying the lower alert levels while the Gradient boost model was slightly better in reducing false positives across the various categories. This means that the ensemble methods not only increased the efficiency of the model’s predictions but also increased the stability of predicting such critical levels of alerts (Hernández *et. al* 2021). Altogether, the comparison highlighted the role of model selection in the enhancement of predictive performance and confirmed the usefulness of the ensembles for complicated classification problems.

## 4.7 Conclusion

This chapter provided a comprehensive analysis of the earthquake dataset to enable the reader to understand the importance of data preprocessing techniques and exploratory data analysis. By using Naive Bayes, Random Forest and Gradient Boosting machine learning models, the work showed that it is possible to make accurate predictions of higher alert levels appropriately. As the feature selection and analysis is very accurate due to robust data preprocessing including Handling of high quality data for better accuracy. The study showed that complex models based on the ensemble of decision trees such as Random Forest and Gradient Boosting perform better than other models, thus proving the usefulness of these models in situations where classification and decision making is critical. The use of hyperparameter has also been shown above of how tuning plays an important role for training a model.

The knowledge derived from this study not only advanced the literature on the ability to forecast an earthquake but also has application in practice for the management of such an event. Future work can be centered on using more databases as well as improving the examined algorithms with the aim of better feature selection and more accurate prediction in case of a natural disaster.

# CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

## 5.1 Conclusion

In this work, an attempt was made to forecast alert levels of an earthquake, using recorded seismic events as a data input and then, evaluate the various ML algorithms for their efficiency. Among the dataset parameters, it was possible to identify earthquake ***“magnitude, depth, CDI (Community Internet Intensity), MMI (Modified Mercalli Intensity), and alert levels”***. To decide which of the 3 classifiers; **“*Naive Bayes, Random Forest, and Gradient Boosting classifiers”***, is optimal in distinguishing high, medium, and low alert levels based on these features, alert levels were predicted using each model.

From feature analysis using EDA, it was found that earthquake magnitude, rated significance, and depth had close or high correlations to the consideration of the alert level. The analysis indicated that most of the time was characterized by low levels of alert, a few being either a yellow or a red alert (Pilz *et. al* 2020). One of the most clearly seen relationships was between magnitude and the alert level, where larger earthquakes correspond to higher alert levels. Scatter plots, bar charts and histograms were other forms of information presented to show distribution and relations within the data.

In the random sample of 40 cases, the Random Forest and Gradient Boosting classifiers performed at an accuracy of approximately 84% while the Naive Bayes classifier was much lower at a 53.45% accuracy. Evaluation matrices mentioned that Random Forest and GBM models showed low alert level classification meaning improvement over the Naive Bayes model, which had problems with classifications of certain alert levels because of imbalanced data and dependent variables. Therefore, Finalized Random Forest model is considered as the most competent, reliable and highly accurate model as GBM had a slight overfitting problem.

In conclusion, based on this analysis, it will be possible to prove the effectiveness of machine learning in the subsequent improvement of earthquake prediction skills (Han *et. al* 2021). By identifying critical variables and the application of effective models, the study provided cues for enhanced earthquake risk assessment conducive to the reach of improved disaster management to minimize the effects of future earthquakes.

## 5.2 Recommendations

The following are proposed recommendations to improve the future application of earthquake prediction models from the discovery made in this study. First, the dataset that is used for training the models can be added to contain more data, preferably more general data. This expansion might consist of features like the presence of geological fault lines, population density near the epicenters, and constructions in the zones. Integration of these contextual factors may enhance the accuracy of when the models are predicting the higher alerts as well as the risks (Jeddi *et. al* 2022).

Second, it is suggested that deeper research of other advanced approaches in service of machine learning, for instance, deep learning or hybrid ensemble learning, should be analyzed. Nonetheless, the Random Forest and Gradient Boosting had promising results; deep learning architectures, using Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN), might better extract temporal and spatial features in the seismic data (Kwag et. al 2020). However, these models, when complemented with richer data sets, might provide more accurate and frequent forecasts than the existing methods.

Third, there should be a never-ending process of the model’s analysis and retraining because of changes in the data distribution (Majstorović *et. al* 2021). Attributes of an earthquake may alter because of some environmental factors such as the movement of tectonic plates or man made-activities which can involve the drilling of the ground. Whenever new seismic data is gathered, the models used should be altered faithfully to incorporate the changes hence retaining high alert level predictability.

Also, governments and disaster management agencies should adopt the use of predictive analytics to improve their preventative disaster management systems. Specialists said that incorporating ML models into early warning systems can help authorities produce automatic alarms and properly distribute resources. The author suggests that such systems should be accompanied by informational campaigns to familiarize the population with the meaning of different alert levels and their actions in case of a certain level activation (Bao *et. al* 2021).

## 5.3 Linking with Objectives

The objective of this project is to identify the best model to classify the earthquakes using Machine learning models such as decision trees, gradient boosting and Random forest. Therefore, Random Forest was ranked the best with an accuracy of 0.8319 after tuning the hyperparameters increasing the model’s predictive ability. To the objectives of the project, characteristics such as magnitude, depth, and alert level were used as the basis for evaluating the models, using such parameters as accuracy, precision, and recall. Hyperparameter optimization enhanced the model performance and answered the research questions related to the optimization of seismic data features and the performance assessment of the proposed model. The results indicate that integrating seismic-quality seeds with non-seismic sources, as well as further refinement of the models and methods, enhances earthquake identification. This improves disaster preparedness and the possibility of incorporating these models into early warning systems and also fulfils the general goal of increasing earthquake prediction and warning.

## 5.4 Future Work

The results and drawbacks of the present study suggest possibilities for further studies. All further studies should seek to increase the size of the dataset as well as use more various and detailed sources of information. Adding seismic information obtained in real-time with the geology characteristics and infrastructure data should improve the potential of the model to identify high-risk events. Subsequent datasets could additionally get information from satellites which also might give important data about ground deformations and other signs of seismic activities (Joshi *et. al* 2022).

Further research should be conducted to discover the ability of deep learning models including RNN and CNN for forthcoming earthquake prediction. These models can model spatial and temporal dependencies within the data and may hence provide better prediction. Further, blended technique models which include the useful aspects of different machine learning strategies could be tried out (Hamdy *et. al* 2022). For example, Random Forest when integrated with neural networks or Random Forest with other ensemble learning models that involve classifiers as the constitutive component could be more efficient.

One further domain that requires more attention at the theoretical and empirical levels is the creation of real-time forecast models. The alert produced by models of the earthquake is due to a short time given that earthquake prediction is sensitive to time. Other suggestions for future work may involve stream data technologies with which models can be updated when new data are received. This would add value to the predictions and set a bar for the responses making emergencies to be adequate and timely.

# References

Abdalzaher, M.S., Soliman, M.S., Krichen, M., Alamro, M.A. and Fouda, M.M. 2024, "Employing Machine Learning for Seismic Intensity Estimation Using a Single Station for Earthquake Early Warning", Remote Sensing, vol. 16, no. 12, pp. 2159.

Aden‐Antoniów, F., Frank, W.B. and Seydoux, L., 2022. An adaptable random forest model for the declustering of earthquake catalogs. Journal of Geophysical Research: Solid Earth, 127(2), p.e2021JB023254.

Akyol, A.A., Arikan, O. and Arikan, F., 2020. A machine learning-based detection of earthquake precursors using ionospheric data. Radio Science, 55(11), pp.1-21.

AlHamaydeh, M., Tellab, S. and Tariq, U. 2024, "Deep-Learning-Based Strong Ground Motion Signal Prediction in Real Time", Buildings, vol. 14, no. 5, pp. 1267.

Apriani, M., Wijaya, S.K. and Daryono 2021/06//, "Earthquake Magnitude Estimation Based on Machine Learning: Application to Earthquake Early Warning System", Journal of Physics: Conference Series, vol. 1951, no. 1.

Ayele, T.C. and Ray, R. 2023, "Assessing the Performance of Machine Learning Algorithms for Soil Classification Using Cone Penetration Test Data", Applied Sciences, vol. 13, no. 9, pp. 5758.

Bao, Z., Zhao, J., Huang, P., Yong, S. and Wang, X.A., 2021. A deep learning-based electromagnetic signal for earthquake magnitude prediction. Sensors, 21(13), p.4434.

Bijelić, N., Lin, T. and Deierlein, G.G., 2020. Efficient intensity measures and machine learning algorithms for collapse prediction of tall buildings informed by SCEC CyberShake ground motion simulations. Earthquake Spectra, 36(3), pp.1188-1207.

Blomeier, E., Schmidt, S. and Resch, B. 2024, "Drowning in the Information Flood: Machine-Learning-Based Relevance Classification of Flood-Related Tweets for Disaster Management", Information, vol. 15, no. 3, pp. 149.

Buchakchiev, V., Dimitrova, K., Georgiev, G., Georgieva, G. and Kozhuharov, V. 2023/12//, "Pattern recognition and signal parameters extraction using machine learning methods", Journal of Physics: Conference Series, vol. 2668, no. 1, pp. 012001.

Bukhowah, R., Aljughaiman, A. and Hafizur Rahman, ,M.M. 2024, "Detection of DoS Attacks for IoT in Information-Centric Networks Using Machine Learning: Opportunities, Challenges, and Future Research Directions", Electronics, vol. 13, no. 6, pp. 1031.

Cano, E.V., Akram, J. and Peter, D.B., 2021. Automatic seismic phase picking based on unsupervised machine-learning classification and content information analysis. Geophysics, 86(4), pp.V299-V315.

Castillo, E., Siervo, D. and Prieto, G.A., 2024. Colombian Seismic Monitoring Using Advanced Machine‐Learning Algorithms. Seismological Research Letters, 95(5), pp.2971-2985.

Chen, K.H., Chiu, H., Obara, K. and Liu, Y. 2023/12//, "Segmentation characteristics of deep, low-frequency tremors in Shikoku, Japan using machine learning approaches", Earth, Planets and Space (Online), vol. 75, no. 1, pp. 32.

Chen, Z., Cui, Y., Pu, Y., Rui, Y., Chen, J., Mengli, D. and Yu, B. 2024, "Classification of Microseismic Signals Using Machine Learning", Processes, vol. 12, no. 6, pp. 1135.

Choi, S., Lee, B., Kim, J. and Jung, H. 2024, "Deep-Learning-Based Seismic-Signal P-Wave First-Arrival Picking Detection Using Spectrogram Images", Electronics, vol. 13, no. 1, pp. 229.

Corbi, F., Bedford, J., Sandri, L., Funiciello, F., Gualandi, A. and Rosenau, M., 2020. Predicting imminence of analog megathrust earthquakes with machine learning: Implications for monitoring subduction zones. Geophysical Research Letters, 47(7), p.e2019GL086615.

Fan, X., Zhang, X., Wang, X. and Yu, X. 2023/12//, "A deep reinforcement learning model for resilient road network recovery under earthquake or flooding hazards", Journal of Infrastructure Preservation and Resilience, vol. 4, no. 1, pp. 8.

Feng, F., Ghorbani, H. and Radwan, A.E. 2024/02/16/, "Predicting groundwater level using traditional and deep machine learning algorithms", Frontiers in Environmental Science, .

Ghimire, S., Guéguen, P., Pothon, A. and Schorlemmer, D. 2023, "Testing machine learning models for heuristic building damage assessment applied to the Italian Database of Observed Damage (DaDO)", Natural Hazards and Earth System Sciences, vol. 23, no. 10, pp. 3199-3218.

Gu, J., Xie, Z., Zhang, J. and He, X. 2024, "Advances in Rapid Damage Identification Methods for Post-Disaster Regional Buildings Based on Remote Sensing Images: A Survey", Buildings, vol. 14, no. 4, pp. 898.

Hamdy, O., Gaber, H., Abdalzaher, M.S. and Elhadidy, M., 2022. Identifying exposure of urban area to certain seismic hazard using machine learning and GIS: A case study of greater Cairo. Sustainability, 14(17), p.10722.

Han, J., Nur, A.S., Syifa, M., Ha, M., Lee, C.W. and Lee, K.Y., 2021. Improvement of earthquake risk awareness and seismic literacy of Korean citizens through earthquake vulnerability map from the 2017 pohang earthquake, South Korea. Remote Sensing, 13(7), p.1365.

Hasib, M., Bagas Anwar, A.N., Huffaz Muhammad, A.B., Abdullah Mu’adz Muflih, Suhendi, C., Ghifari, B.F., Hafidz Bagus, P.A., Kriswati, E., Anggono, T., Febriani, F., Cinantya, N.D. and Aditya, D.P. 2024/03//, "Event classification of volcanic earthquakes based on K-Means clustering: Application on Anak Krakatau Volcano, Sunda Strait", IOP Conference Series.Earth and Environmental Science, vol. 1314, no. 1, pp. 012045.

Hernández, P.D., Ramírez, J.A. and Soto, M.A., 2021. Deep-learning-based earthquake detection for fiber-optic distributed acoustic sensing. Journal of Lightwave Technology, 40(8), pp.2639-2650.

Hosseinzadeh, E., Anamaghi, S., Behboudian, M. and Kalantari, Z. 2024, "Evaluating Machine Learning-Based Approaches in Land Subsidence Susceptibility Mapping", Land, vol. 13, no. 3, pp. 322.

Hu, Y., Wang, W., Li, L. and Wang, F. 2024, "Applying Machine Learning to Earthquake Engineering: A Scientometric Analysis of World Research", Buildings, vol. 14, no. 5, pp. 1393.

Huang, W., Gao, K. and Yu, F. 2024, "Predicting Stick-Slips in Sheared Granular Fault Using Machine Learning Optimized Dense Fault Dynamics Data", Journal of Marine Science and Engineering, vol. 12, no. 2, pp. 246.

Hussain, M.A., Chen, Z., Wang, R., Shah, S.U., Shoaib, M., Ali, N., Xu, D. and Ma, C., 2022. Landslide susceptibility mapping using machine learning algorithm. Civ. Eng. J, 8, pp.209-224.

Iaccarino, A.G., Gueguen, P., Picozzi, M. and Ghimire, S., 2021. Earthquake early warning system for structural drift prediction using machine learning and linear regressors. Frontiers in Earth Science, 9, p.666444.

Jain, H., Dhupper, R., Shrivastava, A., Kumar, D. and Kumari, M. 2023/11/02/, "Leveraging machine learning algorithms for improved disaster preparedness and response through accurate weather pattern and natural disaster prediction", Frontiers in Environmental Science, .

Jeddi, A.B., Shafieezadeh, A., Hur, J., Ha, J.G., Hahm, D. and Kim, M.K., 2022. Multi‐hazard typhoon and earthquake collapse fragility models for transmission towers: An active learning reliability approach using gradient boosting classifiers. Earthquake Engineering & Structural Dynamics, 51(15), pp.3552-3573.

Joshi, A., Vishnu, C. and Mohan, C.K., 2022. Early detection of earthquake magnitude based on stacked ensemble model. Journal of Asian Earth Sciences: X, 8, p.100122.

Kim, S., Lee, K. and You, K., 2020. Seismic discrimination between earthquakes and explosions using support vector machine. Sensors, 20(7), p.1879.

Kwag, S., Hahm, D., Kim, M. and Eem, S., 2020. Development of a probabilistic seismic performance assessment model of slope using machine learning methods. Sustainability, 12(8), p.3269.

Ma, H., Liu, Y., Ren, Y., Wang, D., Yu, L. and Yu, J., 2020. Improved CNN classification method for groups of buildings damaged by earthquake, based on high resolution remote sensing images. Remote Sensing, 12(2), p.260.

Majstorović, J., Giffard‐Roisin, S. and Poli, P., 2021. Designing convolutional neural network pipeline for near‐fault earthquake catalog extension using single‐station waveforms. Journal of Geophysical Research: Solid Earth, 126(7), p.e2020JB021566.

Murti, M.A., Junior, R., Ahmed, A.N. and Elshafie, A. 2022, "Earthquake multi-classification detection based velocity and displacement data filtering using machine learning algorithms", Scientific Reports (Nature Publisher Group), vol. 12, no. 1.

Nguyen, D.T., Lin, W.C., Nguyen, M.T. and Nguyen, T.T. 2024, "Development of a Mango-Grading and -Sorting System Based on External Features, Using Machine Learning Algorithms", Agronomy, vol. 14, no. 4, pp. 831.

Ogunjinmi, P.D., Park, S.S., Kim, B. and Lee, D.E., 2022. Rapid post-earthquake structural damage assessment using convolutional neural networks and transfer learning. Sensors, 22(9), p.3471.

Peters, S., Liu, J., Keppel, G., Wendleder, A. and Xu, P. 2024, "Detecting Coseismic Landslides in GEE Using Machine Learning Algorithms on Combined Optical and Radar Imagery", Remote Sensing, vol. 16, no. 10, pp. 1722.

Pilz, M., Cotton, F. and Kotha, S.R., 2020. Data-driven and machine learning identification of seismic reference stations in Europe. Geophysical Journal International, 222(2), pp.861-873.

Rao, A., Jung, J., Silva, V., Molinario, G. and Sang-Ho, Y. 2023, "Earthquake building damage detection based on synthetic-aperture-radar imagery and machine learning", Natural Hazards and Earth System Sciences, vol. 23, no. 2, pp. 789-807.

Saad, O.M., Huang, G., Chen, Y., Savvaidis, A., Fomel, S., Pham, N. and Chen, Y., 2021. Scalodeep: A highly generalized deep learning framework for real‐time earthquake detection. Journal of Geophysical Research: Solid Earth, 126(4), p.e2020JB021473.

Sharma, A., Chopra, S.R., Sapate, S.G. and Bhagawati, P.B. 2024/01//, "Recent Trends and advances in deep learning techniques for the classification of landslides using satellite images: comprehensive survey", IOP Conference Series.Earth and Environmental Science, vol. 1285, no. 1, pp. 012024.

Song, J., Zhu, J., Wang, Y. and Li, S., 2022. On-site alert-level earthquake early warning using machine-learning-based prediction equations. Geophysical Journal International, 231(2), pp.786-800.

Tehseen, R., Muhammad, S.F. and Abid, A. 2020, "Earthquake Prediction Using Expert Systems: A Systematic Mapping Study", Sustainability, vol. 12, no. 6, pp. 2420.

Thomas, A.M., Inbal, A., Searcy, J., Shelly, D.R. and Bürgmann, R., 2021. Identification of low‐frequency earthquakes on the San Andreas Fault with deep learning. Geophysical Research Letters, 48(13), p.e2021GL093157.

Torres, E. and Dungca, J. 2024, "Prediction of Soil Liquefaction Triggering Using Rule-Based Interpretable Machine Learning", Geosciences, vol. 14, no. 6, pp. 156.

Wang, X., Zhong, Z., Yao, Y., Li, Z., Zhou, S., Jiang, C. and Jia, K. 2023, "Small Earthquakes Can Help Predict Large Earthquakes: A Machine Learning Perspective", Applied Sciences, vol. 13, no. 11, pp. 6424.

Wen, H., Zhou, X., Zhang, C., Liao, M. and Xiao, J. 2023, "Different-Classification-Scheme-Based Machine Learning Model of Building Seismic Resilience Assessment in a Mountainous Region", Remote Sensing, vol. 15, no. 9, pp. 2226.

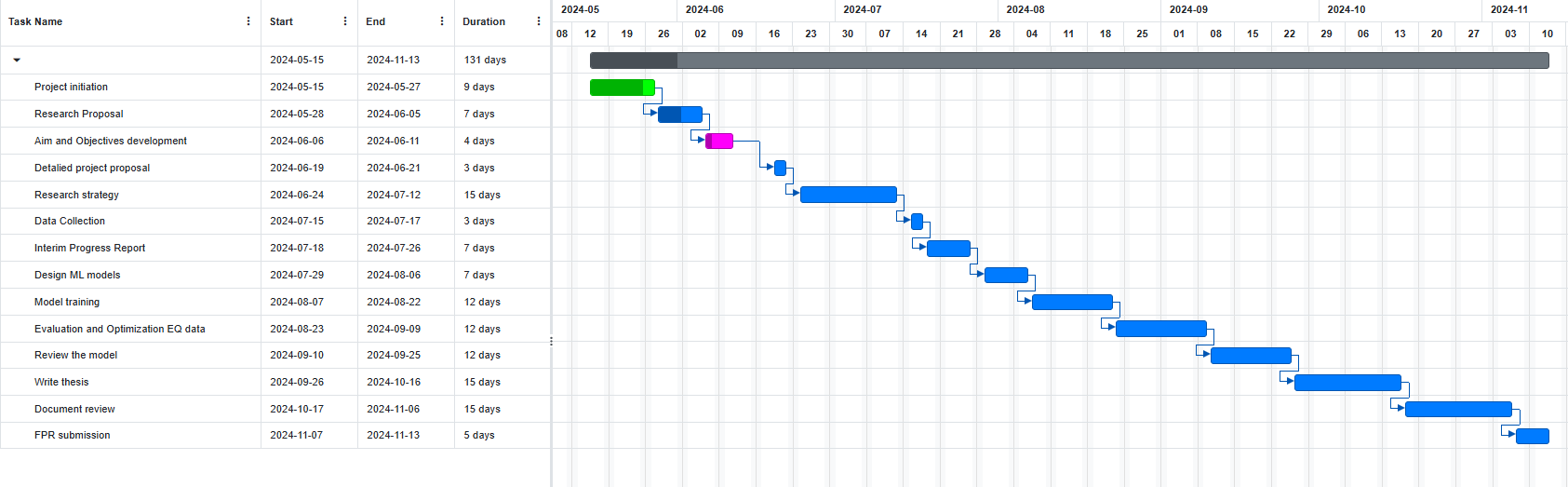
White, M.C.A., Sharma, K., Li, A., Kumar, T.K.S. and Nakata, N. 2023/12//, "Classifying seismograms using the FastMap algorithm and support-vector machines", Communications Engineering, vol. 2, no. 1, pp. 46.

Zhang, A., Wang, X., Xu, C., Yang, Q., Guo, H. and Li, D. 2024, "Near-Real Prediction of Earthquake-Triggered Landslides on the Southeastern Margin of the Tibetan Plateau", Remote Sensing, vol. 16, no. 10, pp. 1683.

Zhou, Z. and Lok, T.M.H. 2024/05//, "Determining the number of soil layers using the surface wave method based on decision tree and random forest", IOP Conference Series.Earth and Environmental Science, vol. 1337, no. 1, pp. 012022.

Zhu, F., Zhu, C., Lu, W., Fang, Z., Li, Z. and Pan, J. 2024, "Soil Classification Mapping Using a Combination of Semi-Supervised Classification and Stacking Learning (SSC-SL)", Remote Sensing, vol. 16, no. 2, pp. 405.

# Gantt Chart



# Appendix

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

import warnings

warnings.filterwarnings('ignore')

from sklearn.model\_selection import train\_test\_split

from sklearn.model\_selection import GridSearchCV

from sklearn.naive\_bayes import GaussianNB

from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix

from sklearn.preprocessing import LabelEncoder

pd.pandas.set\_option('display.max\_columns', None)

EQ\_df = pd.read\_csv("earthquake\_data.csv")

EQ\_df.head()

EQ\_df.shape

EQ\_df.info()

EQ\_df.describe()

# Data preprocessing

features = ["magnitude", "depth", "cdi", "mmi", "magType", "sig", "depth", "alert"]

EQ\_df = EQ\_df[features]

EQ\_df.isnull().sum()

EQ\_df["alert"] = EQ\_df["alert"].fillna("red")

EQ\_df.duplicated().sum()

EQ\_df = EQ\_df.drop\_duplicates()

EQ\_df.shape

# EDA

# Histogram of earthquake magnitudes

plt.figure(figsize=(10, 6))

sns.histplot(EQ\_df['magnitude'], bins=10, kde=True)

plt.title('Histogram of Earthquake Magnitudes')

plt.xlabel('Magnitude')

plt.ylabel('Frequency')

plt.show()

# Bar chart of alert levels

plt.figure(figsize=(10, 6))

alert\_counts = EQ\_df['alert'].value\_counts()

sns.barplot(x=alert\_counts.index, y=alert\_counts.values, palette='viridis')

plt.title('Bar Chart of Alert Levels')

plt.xlabel('Alert Level')

plt.ylabel('Count')

plt.show()

# Scatter plot of magnitude versus significance

plt.figure(figsize=(10, 6))

sns.scatterplot(x='magnitude', y='sig', data=EQ\_df, hue='alert', palette='viridis')

plt.title('Scatter Plot of Magnitude vs. Significance')

plt.xlabel('Magnitude')

plt.ylabel('Significance')

plt.legend(title='Alert Level')

plt.show()

plt.figure(figsize=(7,7))

sns.heatmap(EQ\_df.select\_dtypes(include=[float, int]).corr(), vmax =1.0, fmt='0.2f', square = True, annot = True,cmap='plasma' )

plt.title('Correlation Heatmap',fontsize=15)

plt.show()

# Selecting features

features = ["magnitude", "depth", "cdi", "mmi", "magType", "sig", "depth", "alert"]

EQ\_df = EQ\_df[features]

# Encode categorical variables

le = LabelEncoder()

EQ\_df['magType'] = le.fit\_transform(EQ\_df['magType'])

EQ\_df['alert'] = le.fit\_transform(EQ\_df['alert'])

# Splitting the data into features and target

X = EQ\_df.drop("alert", axis=1)

y = EQ\_df['alert']

# Train-test split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Initialize models

nb\_model = GaussianNB()

# Train Naive Bayes

nb\_model.fit(X\_train, y\_train)

nb\_pred = nb\_model.predict(X\_test)

# Accuracy

nb\_accuracy = accuracy\_score(y\_test, nb\_pred)

print(f"Naive Bayes Accuracy: {nb\_accuracy:.4f}")

# Confusion Matrix

nb\_cm = confusion\_matrix(y\_test, nb\_pred)

plt.figure(figsize=(4, 4))

sns.heatmap(nb\_cm, annot=True, fmt="d", cmap="seismic", xticklabels=['0', '1', '2', '3'],

yticklabels=['0', '1', '2', '3'])

plt.title("Confusion Matrix - Naive Bayes")

plt.xlabel("Predicted Label")

plt.ylabel("True Label")

plt.show()

# Classification Report

nb\_cr = classification\_report(y\_test, nb\_pred)

print(f"Naive Bayes Classification Report:\n {nb\_cr}")

rf\_model = RandomForestClassifier(random\_state=42)

# Train Random Forest

rf\_model.fit(X\_train, y\_train)

rf\_pred = rf\_model.predict(X\_test)

# Accuracy

rf\_accuracy = accuracy\_score(y\_test, rf\_pred)

print(f"Random Forest Accuracy: {rf\_accuracy:.4f}")

# Confusion Matrix

rf\_cm = confusion\_matrix(y\_test, rf\_pred)

plt.figure(figsize=(4, 4))

sns.heatmap(rf\_cm, annot=True, fmt="d", cmap="seismic", xticklabels=['0', '1', '2', '3'],

yticklabels=['0', '1', '2', '3'])

plt.title("Confusion Matrix - Naive Bayes")

plt.xlabel("Predicted Label")

plt.ylabel("True Label")

plt.show()

# Classification Report

rf\_cr = classification\_report(y\_test, rf\_pred)

print(f"Random Forest Classification Report:\n {rf\_cr}")

gb\_model = GradientBoostingClassifier(random\_state=42)

# Train Gradient Boosting

gb\_model.fit(X\_train, y\_train)

gb\_pred = gb\_model.predict(X\_test)

# Accuracy

gb\_accuracy = accuracy\_score(y\_test, gb\_pred)

print(f"Gradient Boosting Accuracy: {gb\_accuracy:.4f}")

# Confusion Matrix

gb\_cm = confusion\_matrix(y\_test, gb\_pred)

plt.figure(figsize=(4, 4))

sns.heatmap(gb\_cm, annot=True, fmt="d", cmap="seismic", xticklabels=['0', '1', '2', '3'],

yticklabels=['0', '1', '2', '3'])

plt.title("Confusion Matrix - Gradient boosting")

plt.xlabel("Predicted Label")

plt.ylabel("True Label")

plt.show()

# Classification Report

gb\_cr = classification\_report(y\_test, gb\_pred)

print(f"Gradient Boosting Classification Report:\n {gb\_cr}")

# Comparison of models by accuracy

accuracy\_scores = {

"Naive Bayes": nb\_accuracy,

"Random Forest": rf\_accuracy,

"Gradient Boosting": gb\_accuracy

}

# Plotting accuracy comparison

plt.figure(figsize=(8, 6))

sns.barplot(x=list(accuracy\_scores.keys()), y=list(accuracy\_scores.values()), palette='viridis')

plt.title('Model Accuracy Comparison')

plt.ylabel('Accuracy Score')

plt.xlabel('Model')

plt.show()

# Hyperparameter for best model

# Hyperparameter grid for Random Forest

param\_grid = {

'n\_estimators': [50, 100, 200],

'max\_depth': [None, 10, 20, 30],

'min\_samples\_split': [2, 5, 10],

'min\_samples\_leaf': [1, 2, 4],

'max\_features': ['auto', 'sqrt', 'log2'],

'bootstrap': [True, False]

}

# Initialize Random Forest model

rf\_model\_ht = RandomForestClassifier(random\_state=42)

# Perform Grid Search with Cross Validation

grid\_search = GridSearchCV(estimator=rf\_model\_ht, param\_grid=param\_grid,

cv=5, n\_jobs=-1, verbose=2, scoring='accuracy')

# Fit the model

grid\_search.fit(X\_train, y\_train)

# Best parameters and best score

print("Best Parameters:", grid\_search.best\_params\_)

print("Best Cross-Validation Accuracy:", grid\_search.best\_score\_)

# Make predictions using the best estimator

best\_rf\_model\_ht = grid\_search.best\_estimator\_

rf\_pred\_ht = best\_rf\_model\_ht.predict(X\_test)

# Accuracy

rf\_accuracy\_ht = accuracy\_score(y\_test, rf\_pred\_ht)

print(f"Optimized Random Forest Accuracy: {rf\_accuracy\_ht:.4f}")

# Confusion Matrix

rf\_cm = confusion\_matrix(y\_test, rf\_pred)

plt.figure(figsize=(4, 4))

sns.heatmap(rf\_cm, annot=True, fmt="d", cmap="seismic", xticklabels=['0', '1', '2', '3'],

yticklabels=['0', '1', '2', '3'])

plt.title("Confusion Matrix - Naive Bayes")

plt.xlabel("Predicted Label")

plt.ylabel("True Label")

plt.show()

# Classification Report

rf\_cr\_ht = classification\_report(y\_test, rf\_pred\_ht)

print(f"Optimized Random Forest Classification Report:\n {rf\_cr\_ht}")