National College of Ireland

Accurate Prediction of Significant Earthquakes using Machine Learning Algorithms

MSc Research Project
Data Analytics

Samrudhi Hawalli Ramachandra
Student ID: x23242361

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Samrudhi Hawalli Ramachandra
Student ID:	x23242361
Programme:	Data Analytics
Year:	2024-2025
Module:	MSc Research Project
Supervisor:	Teerath Kumar Menghwar
Submission Due Date:	12/12/2024
Project Title:	Accurate Prediction of Significant Earthquakes using Machine Learning Algorithms
Word Count:	6432
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Samrudhi Hawalli Ramachandra
Date:	29th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Accurate Prediction of Significant Earthquakes using Machine Learning Algorithms

Samrudhi Hawalli Ramachandra
x23242361

Abstract

Earthquakes are, probably, the most damaging of all the natural disasters. Due to the nature of seismic activities, it's tough to predict them more or less accurately because of the various inherent limitations that are present within the traditional statistical methods. This paper aims to explore the probability of machine learning algorithms in helping the accuracy of earthquake prediction by using the attributes of geospatial and temporal data. To meet this requirement, three datasets sourced from Kaggle have been used: "Earthquakes 2023 Global", "Earthquake Dataset," and "Global Earthquake and Aftershock Data (January 2023)". All three have rich data characters such as magnitude and depth as well as spatial coordinates along with temporal data that aids for more complex models. The machine learning models tested include ensemble methods like Random Forest and Gradient Boosting; linear models, including Linear Regression and Lasso Regression; and even more complex models, like Support Vector Regression. Among all the models, Random Forest always has been the best performer where R^2 values have been greater than 0.90 and low Mean Squared Error value. Geospatial and time-related features such as latitude, longitude, and recurrence interval dominated the model's performance. Further insight into the interplay between the features used and the generated seismic patterns was captured through visualizations including heatmaps and scatter plots. This study reveals unprecedented machine learning-based potential in predicting earthquakes while underlining the need to combine spatial and temporal aspects for better performance. Therefore, the results call for real-time embedding of predictive models into seismic monitoring frames for better preparedness of disasters. More work in advanced techniques of deep learning, richer data sets, and multi-hazard frameworks will further develop earthquake prediction methodologies.

1 Introduction

Seismic events, which have been characterized by earthquakes among others, are the most destructive natural phenomena worldwide, causing severe damage to infrastructure and economies as well as massive loss of human life. Despite these broader impacts of seismic events, earthquake scientists, policymakers, and emergency managers still face the challenge of forecasting such unpredictable nature of the disaster. The accurate forecasting of seismic activity is very crucial because it opens pathways that may assist in the conservation of lives or decrease the damage finally enhancing the preparedness activities (Fox et al.; 2022). Though some conventional methods applied in seismology have been productive in contributing significant discoveries, their disorder has sparked interest in

alternative solutions in dealing with this universal problem. Latest developments in machine learning revealed new horizons toward enhancement of precision in the forecasting of earthquakes. Though not founded on traditional approaches in historical experience or statistical structure, machine learning algorithms prove very powerful in analyzing large sets of data and identifying intricate patterns or learning from new data (Rundle and Crutchfield; 2022) which is a characteristic that could make this a promising approach for model improvements in predictive modeling, enabling much more accurate prediction of time and location of such seismic events.

Including geospatial and temporal characteristics is fundamental in the process of seismic event forecasting. Geospatial characteristics provide information regarding the spatial distribution of these seismic occurrences, including fault line locations, the dynamics of interaction between tectonic plates, and historical patterns that have been linked to the incidence of seismic activities. Temporal attributes include time-series data that contains historical intervals of earthquake recurrence, current seismic activity, and environmental factors such as the build-up of stress along fault lines (Tanaka and Matsumoto; 2024). All these factors put together give an overall view of the complex interactions involved in the phenomenon of earthquakes. However, the current earthquake forecasting models do not use the full potential of this high-dimensional input data. Most traditional methods do not account for relationships and nonlinear links that constitute geospatial and time series data sets, hence making the forecasts less reliable (Şengöz; 2024). This therefore creates a gap that requires a study on systematically assessing the implications of adding these variables in machine learning models to achieve higher precision and reliability of forecasts. This study is motivated by the vast potential to accurately forecast earthquake which would help in better disaster preparation and response. Improving the geospatial and temporal variables in the machine learning model opens up the resource-allocation opportunity, allowing the design of advanced early warning systems and even some pre-event evacuation plans. The major gap that should be addressed is still a huge aspect concerning the incorporation of an overall feature set within the forecasting system of earthquakes. The present report discusses the research question:

- How do the addition of geospatial and temporal attributes to machine learning frameworks impact the accuracy of when seismic activities will happen and whether they will happen?

It would be through this question that the research would try to improve earthquake forecasting techniques towards a more secure and solid future.

2 Related Work

2.1 The Role of Machine Learning in Earthquake Damage Assessment

Earthquake damage assessment has changed much with machine learning, especially in city areas, due to thorough evaluation that is needed to plot recovery. In this regard, Moradi and Shah-Hosseini (2020) presented an example with CNNs, especially on the UNet model on very high-resolution satellite imagery analysis in assessing building damages after an earthquake. Thus, their approach tackled some shortcomings of traditional manual inspections that entail much man-hours and time. The precision was at 68.71%

by the model. Again, it hinted towards the possibility of fast self-automatic damage evaluations. The other important issue of the problem-situation is the reliance on predicative data and failure to detect some of the partly damaged buildings. Updates about the future might involve the integration of multispectral and LIDAR data along with the development of its usability. It can be generalized toward earthquakes with the aid of transfer learning.

2.2 Hybrid Approaches to Earthquake Prediction

Hybrid machine learning techniques combine strengths from different algorithms that can balance the inherent difficulties inherent with earthquake prediction. Salam and Abdelminaam (2021) proposed a hybrid architecture integrating Support Vector Machines, Artificial Neural Networks, and Decision Trees. This approach leveraged the synergy among the components, such as SVM, which is efficient in processing data with high dimensions and the ANN, which captures complex nonlinear relationships. Hybrid performed much better than the single algorithm, especially for early warning signals. However, each model has problems such as computational overhead and hyperparameter optimization. Thus, the present study emphasizes tuning models to the specificity of seismicity conditions, and even domain knowledge can be used to fine-tune predictive models.

2.3 Spatial and Temporal Dynamics in Earthquake Forecasting

Detailing spatial and temporal aspects has to be done for appropriate earthquake prediction. Yousefzadeh and Farnaghi (2021) highlighted the need to incorporate fault density and seismic recurrence intervals into DNNs. The experiment conducted by them explained how the use of spatial and temporal relationships, such as proximity to tectonic fault lines and aftershock sequences, enhances the predictions related to major earthquakes. Integrating these aspects, the paper moves away from the conventional techniques, where these dimensions have been taken care of separately. Limitations include the computational intensity to assimilate large datasets and enforcing data consistency over different regions. Real-time integration of global seismic networks and hardware acceleration progress can be taken as work for the future.

2.4 Use of Geospatial and Remote Sensing Data

Remote sensing techniques have brought new avenues in the prediction of earthquakes by tracking ionospheric and crustal changes that occur before seismic events. Asaly and Reuveni (2022) employed SVMs along with GPS TEC data, which highlighted the role of ionospheric anomalies as a possible precursor to an earthquake. The model presented achieved high accuracy levels, which indicates the possibility of remote sensing in the development of early warning systems. However, there is a need to address the challenge of noise in TEC data caused by solar activities and geomagnetic storms. The techniques used for data pre-processing need to be improved upon, and other remote sensing techniques such as SAR will enhance confidence and accuracy in the detection of seismic precursors.

2.5 Advances in Artificial Intelligence for Earthquake Prediction

Application of Artificial Intelligence The most innovative and changeable approach which has shown high effectiveness in dealing with intricacy in earthquake prediction challenges more than the usual methods is in seismic predicting areas. According to Banna et al. (2020), artificial intelligence methodologies encompass rule-based systems, machine learning strategies, and deep learning architectures, which have been shown to have exceptional abilities in identifying patterns in extensive and complex datasets. Such an ability makes it possible for artificial intelligence to predict the magnitude, location, and time of earthquakes with an accuracy that exceeds that of traditional statistical methods. However, this field remains largely open for improvement. Some of the major limitations in this regard lately have been the limitation due to not enough abundant data for very large to extreme earthquakes; therefore, it's difficult for AI to learn how to predict values toward that rare moment, yet probably key instances of high activity. A further important issue is concerning the overfitting issues and limited interpretation, especially among many deep learning algorithms, thus undermining the whole integrity as decision-support tools under critical choice conditions. Infusion of domain knowledge about the seismic domain into the design of AI models and standardization of datasets across regions would obviate such barriers. Such approaches augment the capabilities, accuracy, and feasibility of the AI models developed within this earth science disaster readiness framework. By integrating AI advancements in domains with modern expertise, it would potentially enable this discipline to achieve an extremely high amount, especially with increased safety through overall region exposure in exposed regions.

2.6 Predicting Earthquake Timing and Size

Corbi and Lallemand (2019) reported that there was a big step forward in earthquake prediction by employing machine learning techniques for the laboratory-scale simulation of subduction zones. Their model, that is, to study spatiotemporal deformation dynamics has a higher resolution than the slip-deficit techniques, and it was considered one of the important steps toward predicting when and how big the seismic events would be. It just hints that such understanding on some complicated physical mechanism in earth-quakes has the capabilities in enhancement of their predictive approaches but the transfer of it is too challenging for practical approach in implementation. Most challenges lay around variabilities, disturbances that usually come around from various types of geologic variability to environmental as well as disturbances in characteristic tectonics. Furthermore, laboratory experiments usually reduce the complexity that naturally exists in the phenomenon under study, thus reducing the direct relevance of the conclusions derived. These challenges may necessitate, on one hand, the inclusion of high-resolution satellite geodesy data and the development of adaptive algorithms that can adapt to the natural variability that exists in the systems under study. The implementation of these upgrades should make it possible to construct practical, robust, and actionable earthquake forecasting tools.

2.7 Integration of Deep Learning for Tsunami and Earthquake Modelling

The deep learning techniques have been very effective in applications, for instance, the modelling of tsunamis, with critical insights gained from such analyses that can be used in earthquake predictions. Mulia and Satake (2020) converted low-resolution tsunami simulation data into high-resolution inundation maps using neural networks and reduced the computation cost by 90% over traditional methods. This methodology not only demonstrates deep learning efficiency and scalability but also its applicability in real-time disaster forecasting. Analogous techniques applied in earthquake prediction can revolutionize early warning systems, making fast and accurate predictions possible even in areas with scarce data. Such systems would be especially helpful in regions at risk from both seismic and tsunami threats because prompt predictions could greatly reduce casualties and damage to property. Future studies might work on multi-hazard models where earthquake and tsunami models can be integrated together to form a more holistic approach for prediction, which may ultimately lead towards better disaster management and mitigation programs.

2.8 Spatiotemporal Neural Networks for Earthquake Prediction

Kail and Zaytsev (2021) developed the earthquake prediction model by using a hybrid model of spatiotemporal dependencies, incorporating CNNs and RNNs. They targeted midterm predictions through events that took place in specific regions using historical seismic data from Japan within 10 to 60 days. Within their structure, convolutional layers have very efficiently been used for recognizing spatial patterns, and recurrent layers are highly efficient at picking up the temporal dynamics which provide comprehensive inspection of seismic trends. This architecture had performed better than standard approaches over medium term predictions for seismicity. However, the model still had a higher rate of false positives, which, more importantly, emphasized the need for improvement in input features and quality of data. Adding mechanisms like attention would further help the model focus on related patterns and filter out noise. Inclusion of probabilistic models would lead to the accuracy of predictions made by the model through the quantification of uncertainties that help make better choices. That notwithstanding, it still emphasizes the applicability of deep learning to earthquake forecasting and further improvement in its use of space-time dynamics.

2.9 Regional Risk Assessment Using Deep Learning

Jena et al. (2021) made groundbreaking research by applying convolutional neural networks (CNNs) with geospatial analysis for the assessment of earthquake hazards in Northeast India, which is one of the highly seismically active regions in the world. The innovative framework adopted a holistic approach, wherein hazard, vulnerability, and coping capacity metrics were used to generate complex risk maps. Such maps are regarded as quite powerful tools for enhancing the strategies regarding disaster preparedness and response. It depicted how CNNs can process robust amounts of multi-dimensional geospatial data that could possibly not be unearthed using the usual methods and, importantly proved practical insight in regions. Despite its success, the study faced several challenges, to which most were attributed to data quality and scalability. Accuracy and applicability

is an issue in rural areas; very extensive geospatial information is lacking. Improvement of the deficiency by integrating crowd-sourced data along with efficiency improvements regarding computation might boost the robustness of the model and expand the deployment to even more high-risk areas across the world for better disaster preparedness.

2.10 Real-time Seismic Source Mechanism Analysis

The Focal Mechanism Network is a new framework from deep learning developed by Kuang et al. (2021), which transforms real-time assessment of earthquake source mechanisms. Innovation mainly deals with accuracy in fault geometry and prediction within stress distribution in milliseconds after the actual acquisition of seismic data; hence, it significantly improves response times regarding an emergency. Unlike classical approaches requiring much human interaction and time, FMNet provides a fully automated, highly scalable solution, so it is suitable for seismic areas. Framework was validated for earthquakes having magnitudes larger than Mw 5.4 and shown to be feasible and potentially viable for widescale application. But such theory in the real world, in the case of usually changing seismicity, creates a big issue: FMNet relies on artificially generated training data. Improving robustness and flexibility means supplementing this with richer kinds of real-world data put into its training schedule. In future research studies, FMNet's integration into international frameworks for seismic monitoring may potentially become one of the promising routes that could increase predictive power for proactive disaster mitigation. However, in general, there's potential for AI techniques toward better observation and handling of seismic risks.

3 Methodology

This section describes the approach that is applied in the analysis of seismic data and the development of predictive models using machine learning algorithms. The process is divided into five basic steps: Collection of data, Data cleaning and preprocessing, Data visualization and transformation, Modeling and testing, and Evaluation. The phases are well designed and implemented in this study to ensure precise and actionable predictions. The overall methodology is given in the below figure.

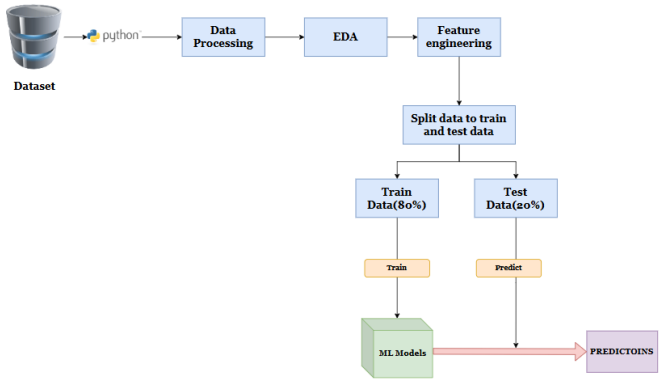

Figure 1: Overall Methodology

The datasets give below have been downloaded from Kaggle, based on these datasets the entire task is done:

- Earthquakes 2023 Global (Dataset1)
- Earthquake Dataset (Dataset2)
- Global Earthquake and Aftershock Data (January 2023) (Dataset3)

Above data sets were collaborated in forming a holistic and rugged pipeline of processing. These datasets acquired from Kaggle that are put together in this analysis, as a matter of generalization, provided reports about earthquake occurrence around the world and including the location, time, depth along with magnitudes. Of course, the earthquakes 2023 global one was such a source used to depict the patterns that occurred behind the latest ones. This compilation of historical seismic event occurrence together with their immediate aftershocks has improved the temporal dimension and provided even more profound insight into trends of earthquake activity. The Global Earthquake and Aftershock Data for January 2023 have concentrated on a realistic presentation of seismic records for January 2023, detailing recent events in relation to sequence earthquakes and aftershock sequences. Initially, they were in CSV file formats. Further on, they were loaded to a dataframe with the help of using a pandas library within the python environment. The data provided wide and inclusive sets of features ranging from geospatial coordinates, event time, and intensity of earthquakes. Exploratory Data Analysis was done on the data so that there is uniformity in the integrity of data and presentation of all parameters, magnitude and depth. This therefore developed an excellent methodology in data acquisition that built a strong base for modelling and subsequent analysis.

3.1 Data Cleaning and Pre-Processing

The datasets were unrefined, hence full cleaning and preprocessing was needed to ensure data quality and usability. Different methods were applied in imputation of missing values, particularly in critical attributes of magnitude and location. In numeric columns, mean or median was used, and in categorical columns, imputation was done using mode. There were duplicate records that may skew the result, so these were detected and removed. The outliers were detected using any of the statistical methods , like interquartile range or by visual observations from box plots for extreme values in either magnitude or depth. Those were either excluded or limited their numbers to reduce the effect on the models. Depending on the characteristic of categorical features, one-hot encoding and label encoding transform were used to encode the features into the forms of numbers. Normalization and standardization were implemented to all features using Python StandardScaler from the sklearn library developed by (Budiman and Ifriza; 2021) This process normalized all the numerical values to a constant range. This is important because some of the machine learning algorithms like KNN and SVM are sensitive to feature scaling. Further the processed data was split into training and testing subsets with an 80:20 ratio, hence providing a fair assessment framework. These data cleaning and preprocessing activities ensured that the data were accurate, consistent, and ready for further analysis.

3.2 Data Visualization and Transformation

The role of visualization is very important to understand the distribution, trend, and relationship within the data. Different libraries were used including matplotlib, seaborn,

and plotly with which a number of graphical representations were developed including histograms, scatter plots, box plots, and heatmaps. The scatter plots allowed an investigation of the kind of relationship that exists between depth and magnitude that was helpful for discovering patterns for feature engineering. Heatmaps showed the correlation of various features, in particular, intensity with the aftershock frequency (Jia et al.; 2019) Time patterns were shown with line plots which revealed very useful information on the frequency and intensity of earthquakes over different periods of time. Geospatial visualization highlightes the geographical patterns and nature of data about earthquakes. The folium and geopandas libraries were applied to create cartographic maps showing the distribution of seismic activity across regions. These maps clearly highlighted the areas of increased seismicity, including tectonic plate boundaries and fault lines. Feature transformation was another essential stage of this process. The developed attributes were innovative, between successive seismic occurrences, with logarithmic modifications for skewed features like depth. Dimensionality was also reduced through Principal Component Analysis to eliminate superfluous features, thereby concentrating models around relevant information. The intercombination of visualization and transformation greatly improved the interpretability as well as the predictive capability of the dataset.

3.3 Modelling and Experimentation

A number of machine learning algorithms were used, and the best-performing method for predicting earthquakes would be determined. The models which were used were Linear Regression, Random Forest, Decision Tree Regression, K-Nearest Neighbors, Support Vector Machines, Gradient Boosting, Ada Boost, Lasso Regression, ElasticNet Regression, and Bayesian approaches, each of which was trained on the preprocessed data set and then tested against the test subset. Linear Regression was used as a baseline for comparison, while ensemble methods like Random Forest and Gradient Boosting provided the predictions by combining multiple decision trees. Decision Tree Regression performed very well in detecting non-linear relationships and thus was particularly suitable for this particular dataset. Hyperparameter tuning was carried out using grid search and random search techniques to enhance the performance of each model. For example, the accuracy can be enhanced by the number of estimators and max depth of Random Forest, while SVM's kernel and regularisation parameters were fine-tuned (Rasel and Meesad; 2019). Cross-validation was used such that the models were not overfitting to training data. Because this is an iterative process of experimentation, it could establish which models performed the best-the predominant models were Random Forest and Gradient Boosting because they were efficient at processing complex data patterns.

3.4 Evaluation

A broad set of evaluation metrics was used to perform model performance checks in order to measure both accuracy as well as reliability. Metrics including Mean Absolute Error and Root Mean Squared for regression models were checked during evaluation. R-squared values help measure how much of the variation in the dependent variable (y) can be explained by the independent variables (X) in a model. Put simply, it shows how well the model fits the data and how effectively it can predict future outcomes. It is calculated using below equation:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \tag{1}$$

- SSE (Sum of Squared Errors): This represents the total of the squared differences between the actual values of the dependent variable and the values predicted by the regression model. It shows how much of the variation in the data the model fails to explain.
- SST (Total Sum of Squares): This measures the overall variation in the dependent variable by summing up the squared differences between each actual value and the mean of the dependent variable. It represents the total variability in the data.

Mean Squared Error (MSE) measures the average of the squared differences between actual values and the predictions made by a model. In machine learning, it's a way to quantify how far off the model's predictions are, on average, from the true values. Mathematically, it's defined as:

$$\text{MSE} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \tag{2}$$

Here's what the terms in the MSE formula mean:

- n : The total number of data points.
- y_i: The actual value for the i-th data point.
- $\hat{y}_i$: The predicted value for the i-th data point.

Feature importance analysis was also part of the assessment. The importance of features was done by permutation and the importance score from Random Forest techniques have shown the most important used features and have been the magnitude, depth, and spatial coordinates, respectively. The results also confirm that models' ability was validated at the same time and it provided necessary information on main drivers of these earth predictions. Cross-validation means using algorithms that would predict and model seismic events over various sets. It used this general framework of assessment such that the models were feasible and accurate for actual real-time application in earthquake forecasting.

4 Design Specification

The architecture of the system for earthquake prediction is multi-layered and modular to enhance adaptability, scalability, and resilience. Data Aggregation at the starting point aggregates datasets collected from Kaggle into one dataset. The Data Pre-Processing phase includes filling in missing values, removing duplicates, normalizing data, and creating such features as time intervals between events. Exploratory data analysis also includes spatial distribution and interrelation visualization. The Modeling Framework employs multialgorithms: Linear Regression, Random Forest, Gradient Boosting, and Support Vector Machines, with hyperparameter tuning towards best performance. The Validation Layer includes cross-validation and importantly preventing overfitting. The Evaluation Module performs multimetric analysis using RMSE, MAE, precision, recall,

and F1-score. Finally, a Comparison Framework is used to assess the relative performances of models and even plot results. Python is used for coding purpose and uses various libraries such as pandas, scikit-learn, seaborn, and matplotlib and effectively computed on Jupyter Notebook.

5 Implementation

5.1 Data Preparation

In data preparation, this study incorporated three different datasets obtained from Kaggle: "Earthquakes 2023 Global," "Earthquake Dataset," and "Global Earthquake and Aftershock Data (January 2023)." These datasets contained various attributes that would describe earthquake events, such as magnitude, depth, geographical locations, and time-related factors. All the datasets were imported into Python using the pandas library to process and manipulate further. First, the data from the files were checked for consistency and the features such as time, mag, depth, and place are prioritized when standardizing columns across datasets. Standardization also resulted in uniformity of naming and data types of the columns. For instance, type of earthquake (type) was made consistent across the datasets such that labelling does not appear different. Extraction and standardization of time data through the datetime column is done in relation to temporal analysis. Duplicate rows are detected, and removed to eliminate the redundant entries that might cause overlaps that in the datasets. This stage ensured the integrated database maintained separated records and thus enhanced both storage and processing performance (Debnath et al.; 2021). At the same time, time intervals between consecutive earthquakes were included as features and magnitudes were classified into different classes-for example, low, moderate, and high. The aim of such feature engineering was to extract information regarding time dependence and intensity distributions of the seismic activities. With the merged and enriched dataset, the stage was now set for further cleaning and analysis.

5.2 Data Cleaning and Processing

Data cleaning and processing ensured that the data sets were clean, consistent, and ready for modeling. Missing values were also addressed, where missing values were very dominant in features like continent and country. Entries with missing values in those fields were imputed with the value "Unknown" to preserve as much data as possible without introducing bias. For the numeric columns like depth and magnitude missing values were imputed using median and mean as respective imputation algorithms where applicable and based on data distribution, ensuring the dataset does not introduce skewness during filling (Han et al.; 2020). The columns that add no value to the modeling prediction included identifiers, like alert and net, as well as metadata such as title are dropped. The statistical approaches like IQR and graphical representation in the form of box plots were used to acquire quantitative anomalies with high values or depths. These obtained anomalies were kept in acceptable ranges so that this does not influence the model traning. The categorical variable type and magType was also converted to numerical representation for compatibility purpose by using LabelEncoder. For example, two types of earthquakes, namely "earthquake" and "explosion," were assigned integers. The sanitized data set was then split for independent variables (X), and the dependent variable, y, represented how

strong the earthquake was (mag). After this, the split was done in the 80:20 ratio using the train_test_split function in the sklearn library, the actual dataset is split into subsets of training and testing. The scaling feature was performed to ensure that the numeric features, such as depth and latitude, have the same scale (Aden Antoniów and Seydoux; 2022). The standardization was performed using the StandardScaler function. The features are scaled to a mean of zero and a standard deviation of one. This is particularly important for algorithms like SVM and KNN, which have very specific requirements in terms of scaling. These steps in cleaning and processing made sure the dataset was clean from errors, properly structured, ensuring easy roll-out of the machine learning models.

5.3 Data Visualization and Transformation

5.3.1 Dataset 1: Earthquakes 2023 Global

This histogram represents the frequency distribution related to earthquake magnitudes. Most of the occurrences are between 3 and 5, meaning that low to moderate levels of seismic activity are dominant. The overlaid line represents a normal distribution that is skewed toward higher magnitudes.

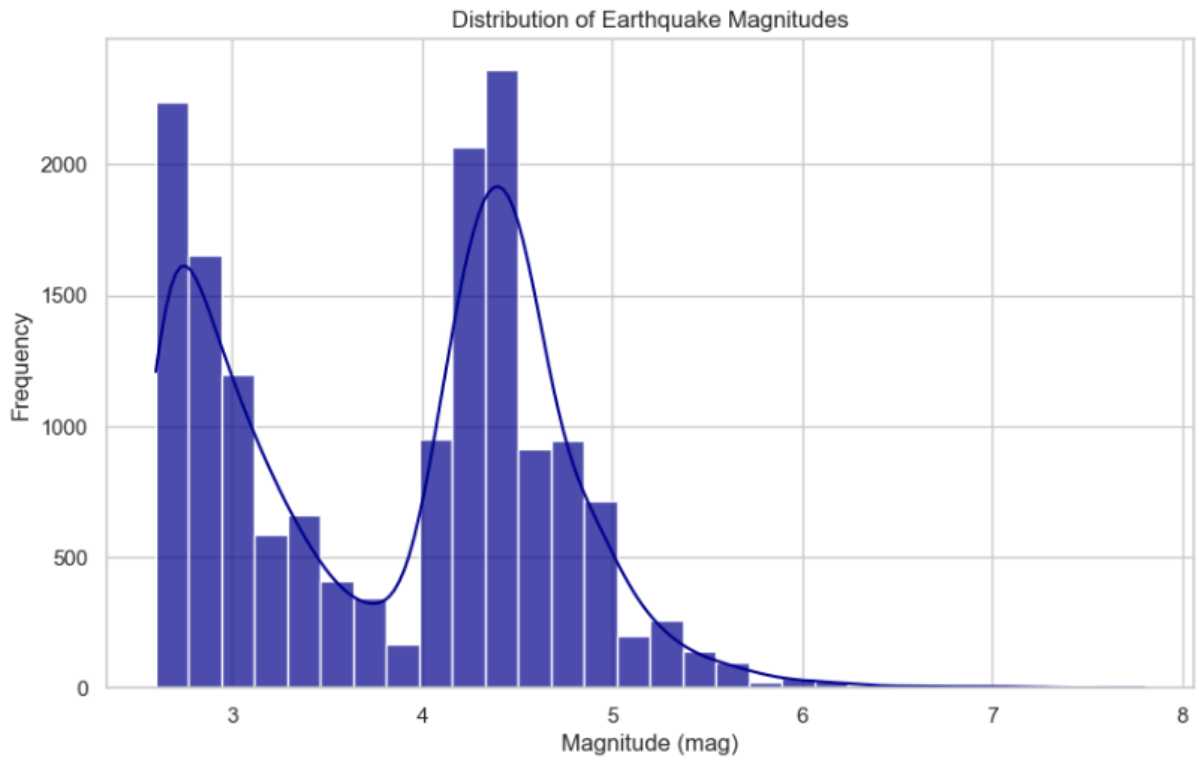

Figure 2: Magnitude vs depth scatter plot

The scatter plot below plots the magnitude of the earthquake against depth and categorizes by event type. Most earthquakes occur at shallower depths, with variability in magnitude. This gives the appearance of the existence of clusters, indicating possibly depth-dependent relationships for some classifications of events, such as "earthquakes" versus "explosions."

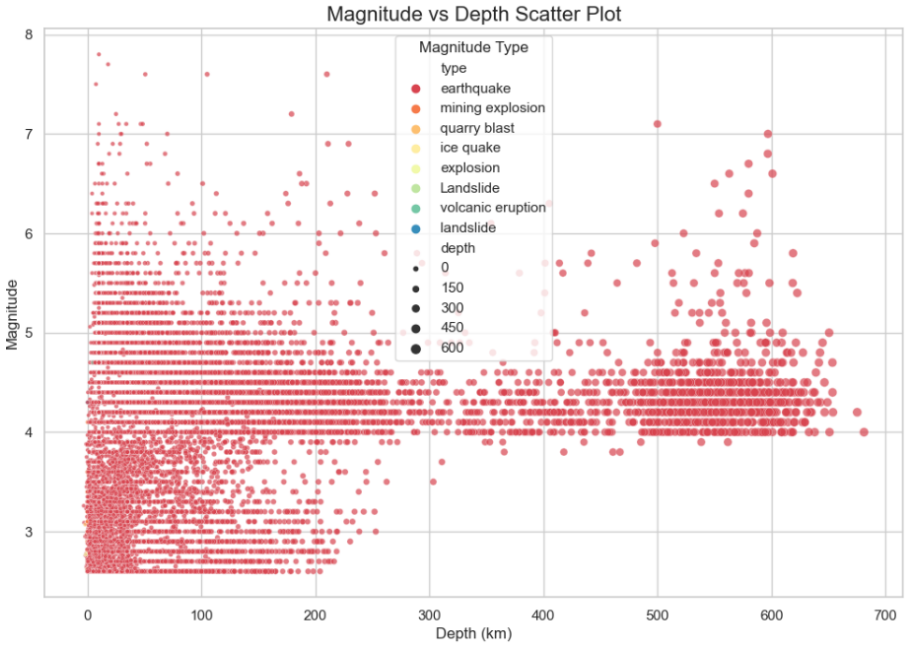

Figure 3: Magnitude vs depth scatter plot

The heatmap shows correlations between features such as magnitude, depth, and latitude. Strong correlations are represented by darker shades of blue or red. For instance, the magnitude feature is positively correlated with depth and negatively correlated with latitude, and is suggesting strong patterns for feature engineering.

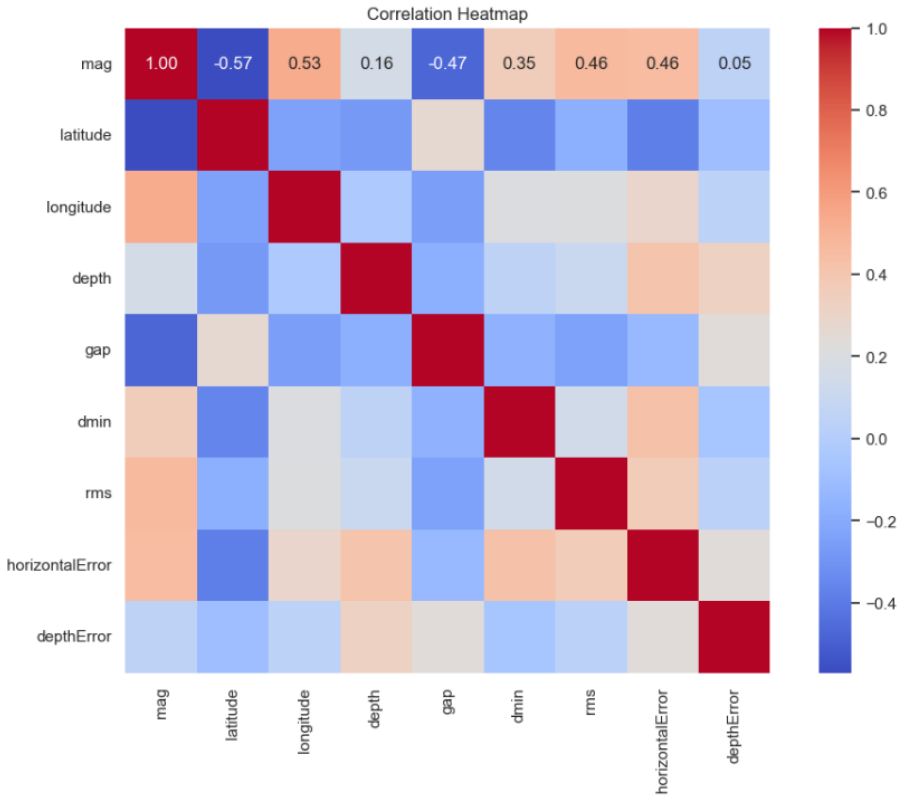

Figure 4: Correlation plot for Earthquakes 2023 Global dataset

5.3.2 Dataset 2: Earthquake Dataset

The following histogram illustrates the distribution of earthquake magnitudes. Majority of the earthquakes are in the range of 6.5 and 7.5 magnitudes, while the frequency tapers off with increasing magnitude. The curve demonstrates that most earthquakes were in this moderate magnitude range rather than at very high magnitudes.

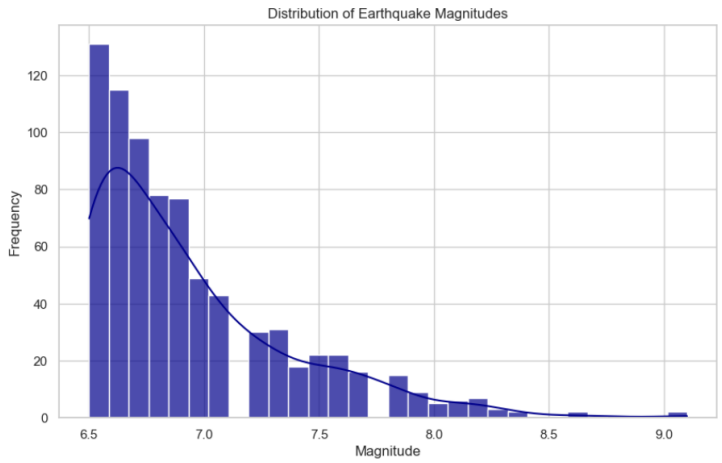

Figure 5: Histogram plot for magnitude

The graph shows an overall fluctuating trend in the number of earthquakes over the years, with noticeable peaks in 2011 and 2013, where earthquake counts were highest. After 2013, there is a gradual decline until 2018, followed by a recovery in earthquake numbers in the years leading up to 2022. Throughout the period, lower-magnitude earthquakes consistently dominate, while high-magnitude earthquakes remain relatively rare.

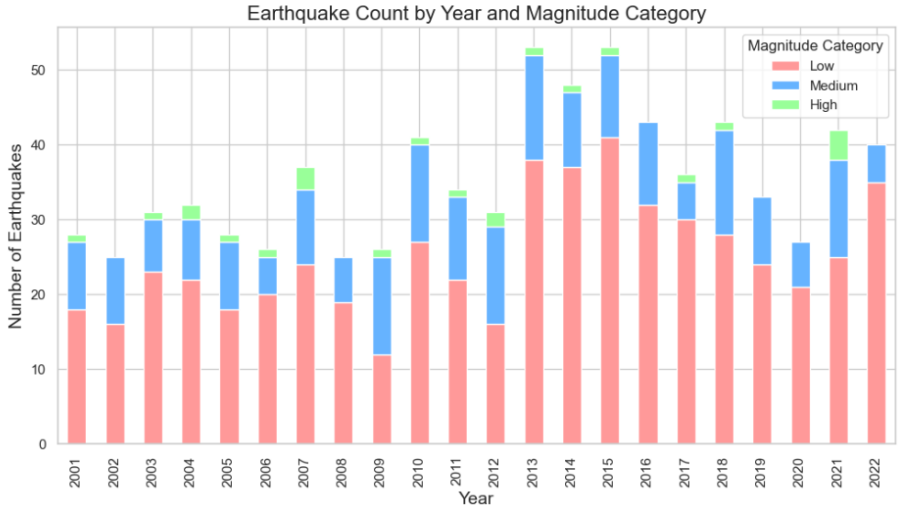

Figure 6: Earthquake frequency and magnitude category

The heatmap is for features in terms of magnitude, depth, latitude, and many other seismic attributes. Darker colors indicate high correlation values where variables like magnitude and sig (significance) are positively correlated. It is used to identify key features.

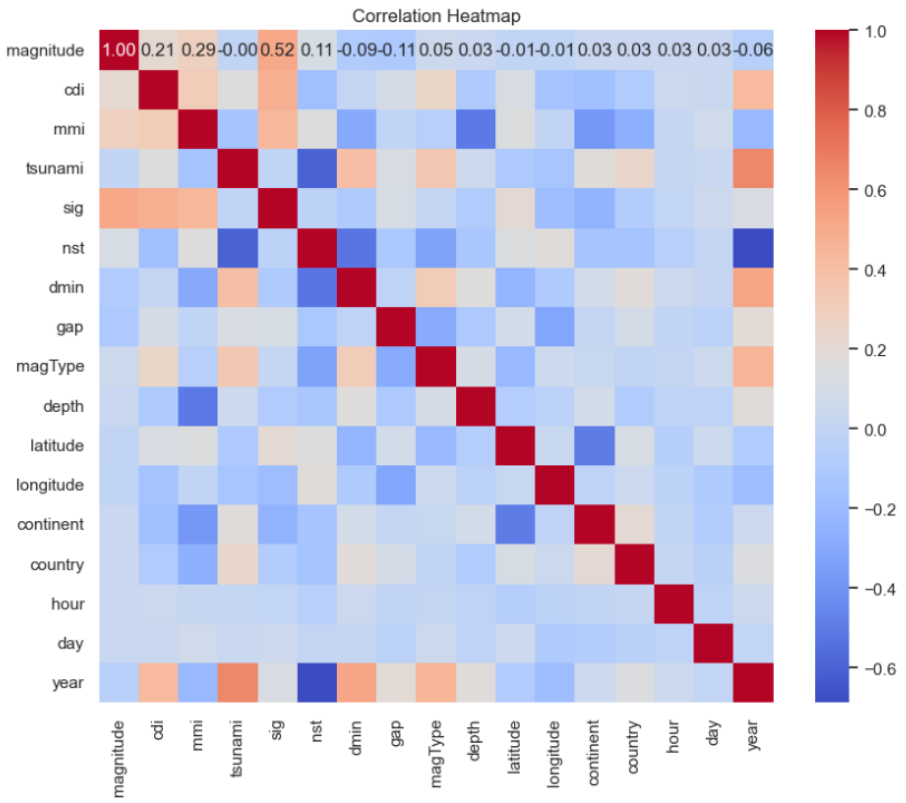

Figure 7: Correlation plot for Earthquake dataset

5.3.3 Dataset 3: Global Earthquake and Aftershock Data (January 2023)

This histogram shows the frequency of earthquake magnitudes. Most events cluster between 4.0 and 5.0. The frequency of higher magnitudes decreases because such significant seismic events are rare; understanding this is crucial in developing a prediction model based on magnitude-based distribution.

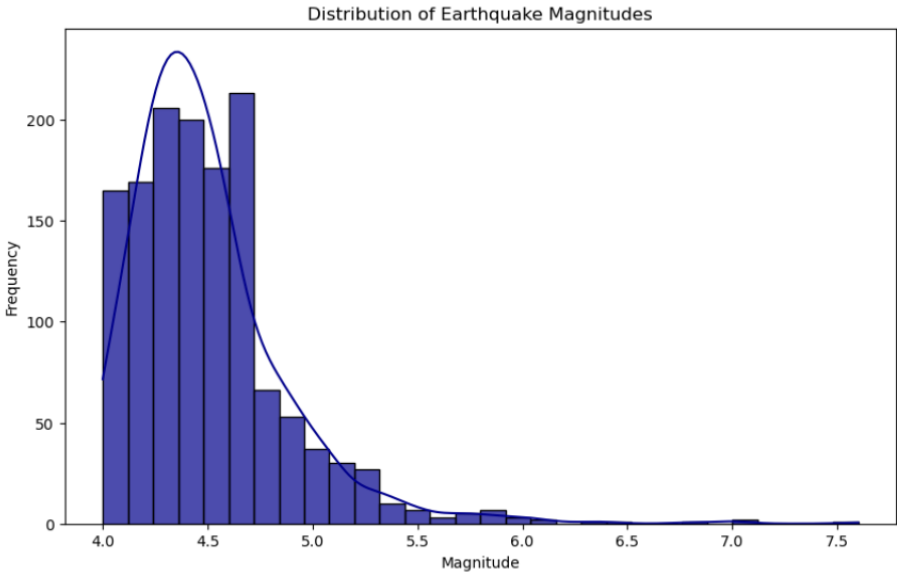

Figure 8: Histogram plot for magnitude

This scatter plot represents earthquake magnitude versus distance to the closest seismic station. A smaller distance to the stations has more significant magnitudes often characterizing areas with stronger seismicity and thus with stronger monitoring.

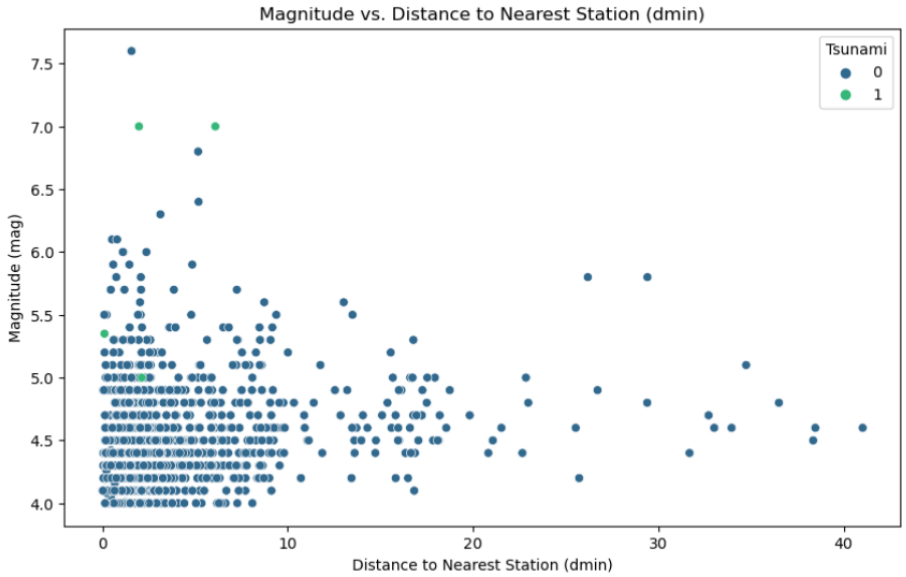

Figure 9: Magnitude vs nearest station plot

6 Evaluation

This chapter is based on detailed research using three different data sets: "Earthquakes 2023 Global," "Earthquake Dataset," and "Global Earthquake and Aftershock Data (January 2023)" in order to predict the intensity of earthquakes. Findings regarding the performance evaluation of the models and thorough inference based on other metrics concerning general implication will be presented for each experiment performed. Several algorithms are covered but not limited to the following: Linear Regression, Random Forest, Gradient Boosting, and Support Vector Machines. The metrics that can be used for performance are Mean Squared Error (MSE), R-squared (R^2), and time computation.

6.1 Experiment 1 on Earthquakes 2023 Global

The first experiment used the "Earthquakes 2023 Global" dataset. It had vast data covering all seismic events that had occurred around the world in 2023. The crucial parameters, including magnitude, depth, latitude, and longitude, have been used while predicting the magnitudes of the earthquake. This experiment was aimed at checking whether different types of machine learning models can identify linear as well as non-linear relationships in data. The evaluation metrics indicate that the performance of the models is different. Random Forest regressor is the best fit on this data set having a minimum MSE of 0.055 with maximum R^2 of 0.92, thus it is good enough for modeling complex relationships. The other is GradientBoostingRegressor that did fairly well at an MSE of 0.066 and an R^2 score of 0.90.

Linear models, Linear Regression and ElasticNet Regression did all right with MSEs of 0.156 and 0.173 and R^2 scores of 0.76 and 0.74 respectively. Those models can be used for comparison purposes but fail miserably to describe non-linear relationships. SVR

Table 1: Performance Metrics for Dataset 1 - "Earthquakes 2023 Global

Model	**MSE**	**R^2 Score**	**Computational Time (s)**
RandomForestRegressor	0.055	0.917	18.86
GradientBoostingRegressor	0.066	0.90	4.80
SVR	0.068	0.89	15.98
KNeighborsRegressor	0.084	0.87	0.23
DecisionTreeRegressor	0.107	0.839	0.206
AdaBoostRegressor	0.128	0.808	1.57
BayesianRidge	0.156	0.765	0.031
LinearRegression	0.156	0.765	0.067
ElasticNetModel	0.173	0.740	0.0165
LassoModel	0.191	0.713	0.001

obtained an MSE of 0.068 and R^2 score of 0.89 but cost around 15.98 seconds. This trade-off between accuracy and speed makes SVM lesser ideal for large-scale, real-time operations. Lasso Model naturally was the worst of the ones with an MSE value of 0.191 and an R^2 score value of 0.713. The simplistic assumptions taken by Lasso Model made its model not efficient for dealing with the complexity data of earthquakes. To sum it all, the Random Forest method is found to be the best ensemble predictor for the dataset "Earthquakes 2023 Global".

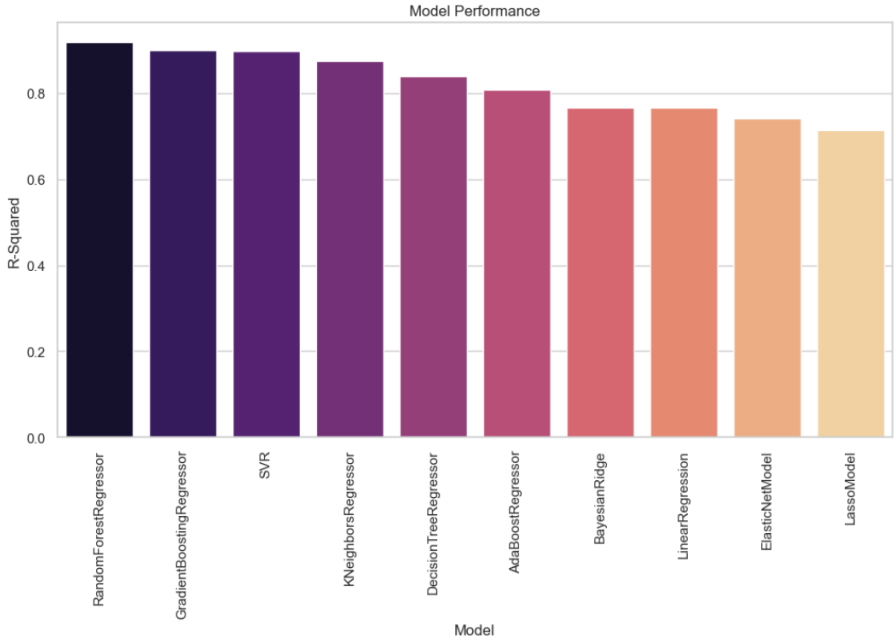

Figure 10: Performance of models on Dataset1

6.2 Experiment 2 on Earthquake Dataset

The "Earthquake Dataset" was utilized for the second experiment. This dataset consists of combined historical seismic events from all over the world, thereby widening the scope of time. Aftershock sequences and timing of occurrence and location are some of the attributes. This was to test the models for generalization over quite different earthquake

scenarios. Gradient Boosting outperformed all other models with an MSE of 0.046 and an R^2 score of 0.75. Random Forest also did well with an MSE of 0.048 and an R^2 score of 0.73, further establishing the utility of ensemble methods in this space. Both models showed their ability to learn temporal and spatial patterns in the data. The computation times of Gradient Boosting and Random Forest were 0.34 seconds and 0.90 seconds respectively, indicating that the two algorithms suit larger datasets of similar characteristics.

Table 2: Performance Metrics for Dataset 2 - "Earthquake Dataset

Model	**MSE**	**R^2 Score**	**Computational Time (s)**
GradientBoostingRegressor	0.046	0.751	0.244
RandomForestRegressor	0.0487	0.737	0.588
DecisionTreeRegressor	0.056	0.694	0.019
AdaBoostRegressor	0.077	0.585	0.141
SVR	0.092	0.501	0.050
LinearRegression	0.100	0.459	0.098
BayesianRidge	0.101	0.45	0.011
ElasticNetModel	0.124	0.327	0.003
LassoModel	0.138	0.254	0.015
KNeighborsRegressor	0.139	0.249	0.186

Linear Regression had an MSE of 0.1 and an R^2 score of 0.45, depicting that it is not a good fit for this data as these relationships are not linear in nature. ElasticNet Regression was also similar to the previous model with an MSE of 0.124 and an R^2 score of 0.32, which are useful for deriving interpretable results but lesser in accuracy than ensemble methods. SVM had an MSE of 0.092 and an R^2 score of 0.50, and thus was moderately accurate. In KNN, it has gone terribly wrong from SVM: 0.139 in terms of MSE and scored 0.249 as R^2 which is directly related to being sensitive to data distribution and scale. In general, Experiment 2 demonstrated that ensemble methods generally dominate data analysis on time-space mixtures by approaches such as Gradient Boosting and Random Forest.

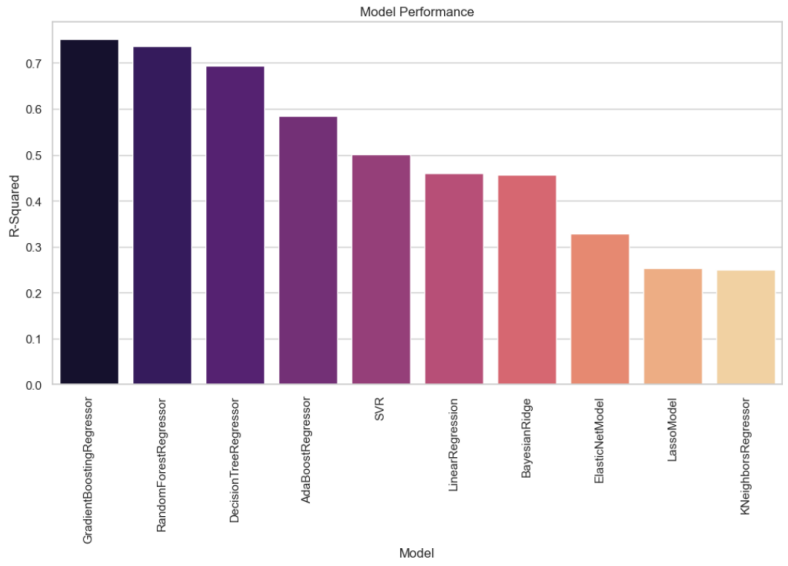

Figure 11: Performance of models on Dataset2

6.3 Experiment 3 on Global Earthquake and Aftershock Data (January 2023)

The third experiment utilizes the "Global Earthquake and Aftershock Data (January 2023)" dataset focused on recent seismicity and aftershock sequences, in order to be of particular utility in determining how well-functioning the models are for pretty localized and temporal data. It is the second experiment in succession where Random Forest Regressor proved to be the champion at MSE 0.012 and an R^2 of 0.94 points, hence it is again the best model for this dataset. The MSE was at 0.013, and the R^2 score was 0.93 with Gradient Boosting Regressor. The fitting of temporal dependencies and nonlinear relationships found in sequences of aftershocks was quite impressive for both models. Their computation times were at 0.7 seconds for Random Forest and 0.3 for Gradient Boosting, which was acceptable enough to be applied in real-time applications.

Table 3: Performance Metrics for Dataset 3 - "Global Earthquake and Aftershock Data

Model	**MSE**	**R^2 Score**	**Computational Time (s)**
RandomForestRegressor	0.012	0.939	0.353
GradientBoostingRegressor	0.013	0.93	0.174
DecisionTreeRegressor	0.0179	0.91	0.00
AdaBoostRegressor	0.019	0.904	0.122
KNeighborsRegressor	0.037	0.812	0.025
SVR	0.054	0.726	0.016
LassoModel	0.069	0.655	0.002
ElasticNetModel	0.078	0.607	0.002
BayesianRidge	0.107	0.465	0.005
LinearRegression	0.107	0.464	0.020

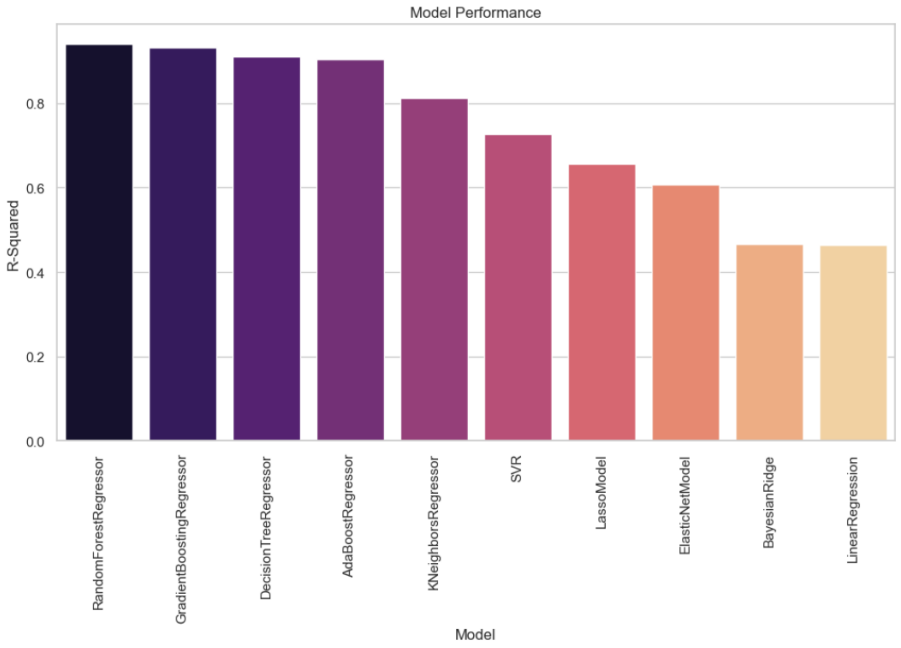

Figure 12: Performance of models on Dataset3

Linear Regression and Bayesian Ridge were poor models that had MSEs of 0.107,and achieved R^2 scores of 0.46. Neither of the models were really good at fitting the underlying non-linear relationships, but it was useful to see these results, which could help a lot in understanding the overall structure of the data. KNN was fragile with an MSE of 0.037 and an R^2 score of 0.81, indicating vulnerability to noise carried by the local pattern in the dataset.The experiment reiterated the strength of ensemble methods like Gradient Boosting and Random Forest in the case of earthquake prediction. These models can learn from nonlinear patterns and handle the dependencies in time for data based on aftershock sequences.

6.4 Discussion

In all three experimental studies, Random Forest and Gradient Boosting were the winners. These ensemble methods had higher precision as seen by the R^2 values that were often greater than 0.90. Their ability to capture non-linear relationships and with shorter computational times, make them the most reliable choices for earthquake prediction (Jia et al.; 2019). Though linear models were not very precise, they were highly useful for baseline comparison and interpretability. Their performance indicates that model complexity needs to be balanced with accuracy, particularly when explainable results are required. SVR had an average accuracy but was too computationally expensive and could not scale up to larger datasets (Rasel and Meesad; 2019). KNN was very simple to implement but failed to generalize well, especially if the patterns were very heterogeneous. Linear models like Linear Regression, ElasticNet Regression and Lasso Model always underperformed in all experiments, which indicates the inability of the algorithm to deal with complex, multi-dimensional data. It is unsuitable for earthquake prediction tasks due to the reliance on strong assumptions about feature independence.

7 Conclusion and Future Work

7.1 Conclusion

This work discusses the application of machine learning techniques in seismic events forecasting, based on exploiting both spatial and temporal features. Such features appear to greatly enhance performance as ensemble techniques, specifically Gradient Boosting and Random Forest, were seen to surpass all models in the compared setup. Very high accuracy metrics were achieved, characterized by R^2 values of above 0.90 with very low MSE on all datasets. Their capability of capturing non-linear relationships and temporal dependencies demonstrates their fair suitability for complex seismic data. The latitudes and longitudes, along with the temporal attributes such as timestamps and intervals between earthquakes, have been of prime importance for the enhanced performance of the model. Techniques such as heatmaps and scatter plots have produced robust correlations between these attributes and the magnitudes of earthquakes, hence giving evidence of their utility as predictors. More direct models like Linear Regression and Lasso Model provide useful baselines, although they are not more robust for accurate prediction. Although SVR achieved performance at an intermediate scale, higher computational costs are problems for applications that involve higher dimensions or near real-time applications. The machine learning models are applicable in disaster preparedness through

improving resource allocation and early warning systems. This will allow the stakeholders to strengthen their capability to predict and respond appropriately to the critical earthquakes therefore reducing the effects on human lives and infrastructure through the integration of these systems with real-time seismic monitoring.

7.2 Future Work

Even though this study has shown that geospatial and temporal characteristics are significant in the context of earthquake prediction, there are various potential leads for further exploration. It can be further enhanced with more data sources such as geological data, environmental data, or tectonic plate movements. The causes and sequences of seismic events will be better placed by the availability of these datasets. Further, more sophisticated feature engineering techniques include deep learning models such as CNNs and RNNs, which can be used for the detection of complex spatial and time patterns in seismic data. Although such models are computation-intensive, they have been quite useful for some applications where high-dimensional data requirements were involved. This paper has explored ensemble models whose performances would be further improved through automation of hyperparameter optimization and Bayesian optimization or genetic algorithms. More detailed understanding of the robustness and generalization abilities of the model would emerge from considering larger and much more diverse data sets along with extended periods of time and wider regions of interest. Ultimately, such prediction models may help in their application for real-time operation integration in seismic monitoring systems. The integration of predictive analytics into early warning systems would offer proactive management of disasters so that lives are saved, and economic losses are lessened. Future research work can, therefore, build on the findings of this study by making strides in the boundaries of earthquake prediction and mitigation technologies.

References

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

Asaly, S.and Gottlieb, L. I. N. and Reuveni, Y. (2022). Using support vector machine (svm) with gps ionospheric tec estimations to potentially predict earthquake events., *Remote Sensing* **14**(12): 2822.

Banna, A., M.H, T., K.A, Kaiser M.S, M. M. R., M.S.and Hosen, A. and Cho, G. (2020). Application of artificial intelligence in predicting earthquakes: state-of-the-art and future challenges., *IEEE Access* **8**.

Budiman, K. and Ifriza, Y. (2021). Analysis of earthquake forecasting using random forest., *Journal of soft computing exploration* **2**(2): 153–162.

Corbi, F.and Sandri, L. B. J. F. F. B. S. R. M. and Lallemand, S. (2019). Machine learning can predict the timing and size of analog earthquakes., *Geophysical Research Letters* **46**(3): 1303–1311.

Debnath, P., Chittora, P., Chakrabarti, T., Chakrabarti, P., Leonowicz, Z.and Jasinski, M. G. R. and E, J. (2021). Analysis of earthquake forecasting in india using supervised machine learning classifiers, *Sustainability* **13**(2): 921.

Fox, G. C., Rundle, J. B., Donnellan, A. and Feng, B. (2022). Earthquake nowcasting with deep learning., *GeoHazards* **3**(2): 199–226.

Han, J., Kim, J., Park, S., Son, S. and Ryu, M. (2020). Seismic vulnerability assessment and mapping of gyeongju, south korea using frequency ratio, decision tree, and random forest., *Sustainability* **12**(18): 7787.

Jena, R.and Pradhan, B. N., Alamri, S. and A.M. (2021). Earthquake risk assessment in ne india using deep learning and geospatial analysis., *Geoscience Frontiers* **12**(3): 101110.

Jia, H., Lin, J. and Liu, J. (2019). An earthquake fatalities assessment method based on feature importance with deep learning and random forest models., *Sustainability* **11**(10): 2727.

Kail, R.and Burnaev, E. and Zaytsev (2021). Recurrent convolutional neural networks help to predict location of earthquakes, *IEEE Geoscience and Remote Sensing Letters* **19**: 1–5.

Kuang, W., Yuan, C. and Zhang, J. (2021). Real-time determination of earthquake focal mechanism via deep learning., *Nature communications* **12**(1): 1432.

Moradi, M. and Shah-Hosseini, R. (2020). Earthquake damage assessment based on deep learning method using vhr images., *Environmental Sciences Proceedings* **5**(1): 16.

Mulia, I.E.and Gusman, A. and Satake, K. (2020). Applying a deep learning algorithm to tsunami inundation database of megathrust earthquakes., *Journal of Geophysical Research: Solid Earth* **125**(9).

Rasel, R.I.and Sultana, N. I. G. I. M. and Meesad, P. (2019). Spatio-temporal seismic data analysis for predicting earthquake: Bangladesh perspective., *Research, Invention, and Innovation Congress* pp. 1–5.

Rundle, J.B., D. A. F. G. and Crutchfield, J. (2022). Nowcasting earthquakes by visualizing the earthquake cycle with machine learning: A comparison of two methods., *Surveys in Geophysics* **43**(2): 483–501.

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

Tanaka, A. and Matsumoto, Y. (2024). Adaptive ai-driven earthquake simulation leveraging real-time geospatial data and advanced machine learning models., *IMET 2024* .

Yousefzadeh, M. H. S. and Farnaghi, M. (2021). Spatiotemporally explicit earthquake prediction using deep neural network., *Soil Dynamics and Earthquake Engineering* **144**.

Şengöz, M. (2024). Harnessing artificial intelligence and big data for proactive disaster management: Strategies, challenges, and future directions., *Haliç Üniversitesi Fen Bilimleri Dergisi* **7**(2): 57–91.