Model Masters

CE 784 Project Report: Semester 2024-25 (I)

Earthquake Prediction

Aashika Gupta(210015), Mudit Sengar(210638), Shambhavi Agarwal(210956), Shrey Patel(210999)

Abstract

Predicting earthquakes is a longstanding scientific challenge due to the complex and dynamic nature of seismic activity. Recent advancements in machine learning have opened new pathways for understanding and forecasting these events by leveraging vast amounts of seismic data. This project explores the application of machine learning models to predict earthquake occurrences by analyzing patterns within historical seismic records. By employing supervised learning techniques, including classification and regression algorithms, the study aims to identify correlations between various geological factors, such as tectonic movements, fault line pressures, and past seismic activities, that can serve as predictors of future earthquakes. Data preprocessing, feature extraction, and model training are critical components in refining predictive accuracy. The project also discusses the challenges of false positives, data sparsity, and model interpretability in the context of earthquake forecasting. Ultimately, this approach holds promise for contributing to early-warning systems that could mitigate the impact of earthquakes on at-risk communities, representing a significant stride in natural disaster preparedness and risk reduction.

Keywords: Earthquake prediction, Machine Learning, Seismic Activity, Supervised Learning, Regression Models

1. Introduction

Earthquakes rank among the most destructive natural disasters, leading to widespread loss of life, severe infrastructure damage, and significant economic impacts globally. The sudden, often unanticipated nature of earthquakes amplifies their potential for devastation, particularly in densely populated or vulnerable regions. Despite decades of scientific research, accurately predicting earthquakes remains a formidable challenge due to the highly complex and dynamic interactions between tectonic plates and the vast array of factors influencing seismic events. Traditional seismology methods, which often rely on historical data and geological indicators, have encountered limitations, as seismic activity does not always follow predictable patterns. This unpredictability, combined with the complexity of geological systems, has made earthquake forecasting an elusive goal.

In recent years, however, advancements in data science and machine learning have introduced new possibilities for seismic analysis and earthquake prediction. Machine learning algorithms, capable of processing and identifying patterns in large-scale datasets, have shown promise in capturing relationships within historical seismic records that may not be evident through traditional methods. By analyzing various geological parameters, such as tectonic movements, fault line pressures, depth, and recurrence intervals, machine learning models can potentially recognize subtle indicators of impending seismic activity, contributing valuable insights for early-warning systems.

This report explores the application of machine learning techniques for earthquake prediction, focusing on the models and methodologies employed, the challenges encountered, and the potential for improving disaster preparedness. Our project applies several machine learning approaches, including **Linear and Polynomial Regression**, **SVR**, **Random Forest Regressor**, **RFTS**, and **ANN**, to evaluate their accuracy in earthquake prediction. Comparing their results helps identify the most reliable approach for forecasting seismic events.

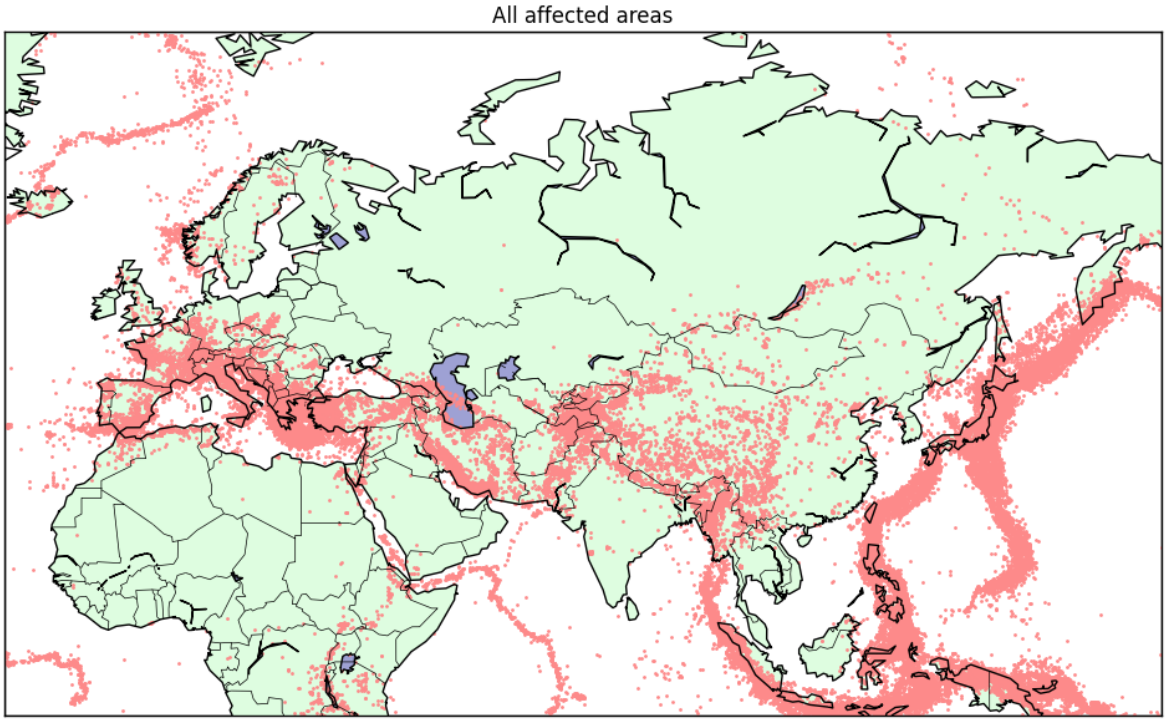

Figure 1: Affected Areas

2. Methodology: *GitHub Repository Link*

2.1. Feature Extraction

From the available features—time, latitude, longitude, depth, magnitude, type, nst, gap, dmin, rms, net, updated, place, horizontalErrors, mag, magNst, status, location, and MagSource—we selected only those without missing values. Replacing missing values with zero could lead to significant misinterpretations by the model. Therefore, the final features chosen for model training were Unix time, latitude, longitude, and depth.

2.1.1. Unix time

This feature represents the timestamp of each earthquake event in seconds since January 1, 1970, and is chosen to capture the exact moment of occurrence. Including Unix time enables the model to analyze temporal patterns, helping identify trends or clusters in seismic activity over time. This feature is essential for predicting recurrence intervals, as it allows the model to recognize potential periodicity in earthquake events. Missing values for time would disrupt the sequence of events, leading to incomplete temporal analysis and less accurate predictions.

2.1.2. Latitude and Longitude

These features denote the geographic coordinates where each earthquake occurs, critical for identifying seismic hotspots and regional trends. By analyzing the location, the model can detect areas with frequent activity and potential correlation with tectonic boundaries or geological features. Latitude and longitude are indispensable for spatial analysis, allowing predictions to reflect the spatial nature of earthquakes accurately. Filling missing values with zeros would inaccurately place events, leading to spatial distortions and skewed outputs in location-based analyses.

2.1.3. Depth

This feature measures how deep below the Earth's surface the earthquake originated, significantly influencing the event's surface impact. Depth is chosen because shallow earthquakes typically result in more surface damage,

making it a key factor in assessing potential severity. Accurate depth data helps the model understand the nature and impact range of different seismic events. Missing depth values can lead to misinterpretation of earthquake effects, and substituting with zero would create a false sense of shallow, highly impactful events, skewing predictions.

2.2. Feature Transformation

In the code, feature transformation is achieved through normalization using 'MinMaxScaler', which scales each feature to a range between 0 and 1. This process ensures that different features, such as time, latitude, longitude, and depth, all contribute equally during model training by eliminating disparities in scale. Normalization improves the model's performance and convergence speed, particularly for algorithms sensitive to feature magnitudes, like gradient descent. In earthquake prediction, this step allows the model to recognize patterns more accurately by treating each feature fairly and avoiding biases caused by varying data scales.

2.3. Model Training

- Linear Regression: The results from these models establish a foundational understanding of how certain seismic variables (such as depth and location) may relate to the target variables (e.g., magnitude or likelihood of occurrence). While Linear Regression may not capture all the intricacies due to the inherently non-linear nature of seismic data, Polynomial Regression results indicate the potential for better approximations. However, given the limited predictability observed, these models underscore the need for more complex, non-linear algorithms.

- Polynomial Regression: Polynomial Regression improves upon linear models by fitting a non-linear curve to capture more complex relationships in seismic data, such as depth and magnitude. Its results show enhanced adaptability for non-linear interactions, revealing patterns missed by simple linear models. While not as complex as ANN or Random Forest, it underscores the value of non-linear approaches in earthquake prediction, aligning with the project's goal of enhancing forecasting accuracy for early-warning systems.

- Support Vector Regression: The SVR model's results demonstrate its effectiveness in handling noise and outliers, which are common in seismic datasets. The SVR model captures more complex relationships than linear models and, therefore, aligns more closely with the project's goal of improving predictive accuracy in earthquake forecasting. SVR's generalization capability suggests its utility for early warning, as it performs well on unseen data.

- Random Forest Regressor: The Random Forest model's high accuracy and stability in prediction make it particularly useful for large, noisy datasets like those in earthquake prediction. The results highlight the model's ability to interpret feature importance, aiding in identifying key factors associated with seismic activity. This model's predictive power suggests its potential in recognizing underlying patterns in seismic records, contributing to the project's aim of identifying patterns that traditional models may miss.

- Artificial Neural Network: The Artificial Neural Network (ANN) model effectively captures complex, non-linear patterns in seismic data, enhancing earthquake prediction accuracy. Its ability to handle high-dimensional features and adapt to new data supports real-time applications, aligning with the project's goal of developing reliable early-warning systems. ANN's strong performance suggests it could be transformative for improving disaster preparedness in earthquake-prone regions.

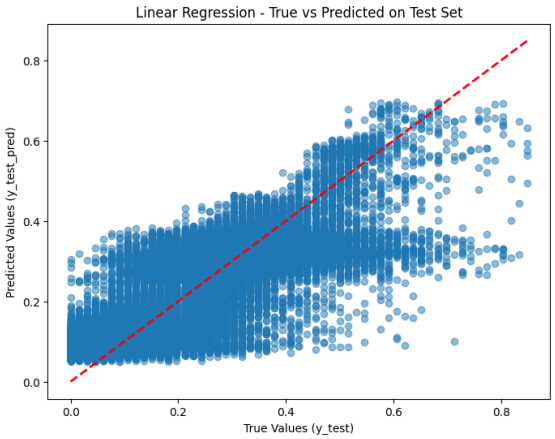

Figure 2: Linear Regression

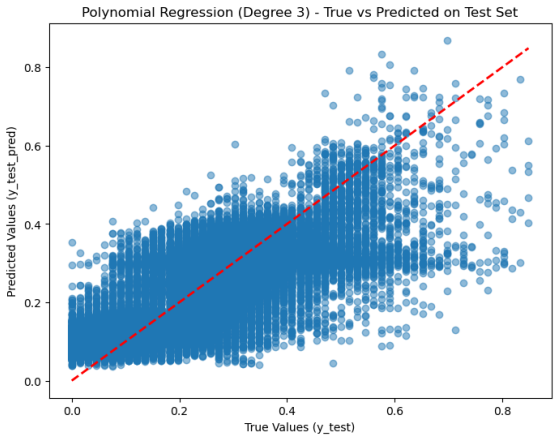

Figure 3: Polynomial Regression

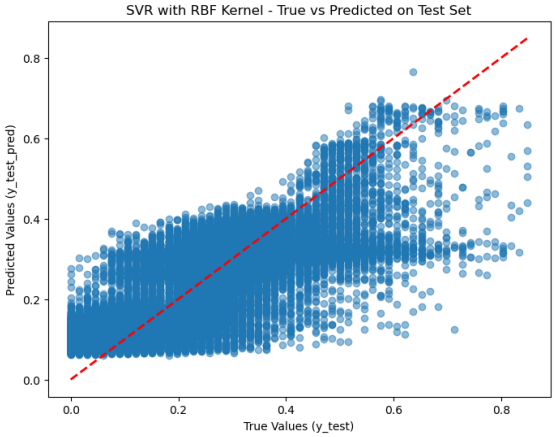

Figure 4: Support Vector Regression

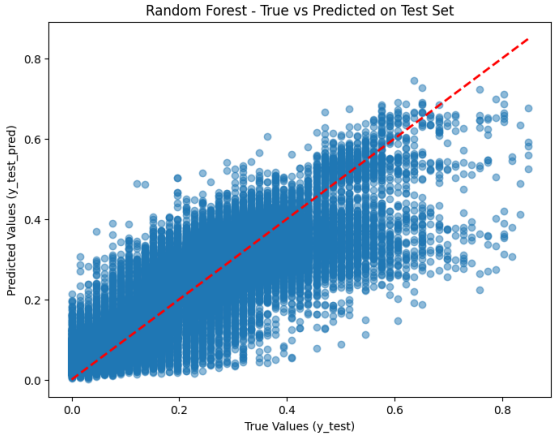

Figure 5: Random Forest Regressor

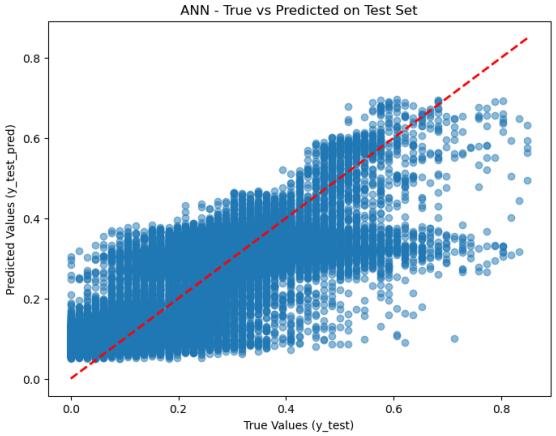

Figure 6: ANN

3. Result

The earthquake prediction project applied machine learning techniques to analyze historical seismic data and improve forecasting accuracy. Various models, including Linear Regression, Polynomial Regression, Support Vector Regression (SVR), Random Forest Regressor, and Artificial Neural Networks (ANN), were trained and evaluated. The Random Forest model achieved the highest accuracy with an R^2 score of 0.711, followed by ANN with 0.647. The project demonstrated that advanced machine learning models can capture complex patterns in seismic activity, showing potential for contributing to early-warning systems. Overall, this approach underscores the value of data-driven solutions in disaster preparedness and risk mitigation.

Table 1: Performance of the various models used on Test Set

Model	MSE	R^2
Linear Regression	0.008	0.445
Polynomial Regression	0.006	0.598
Support Vector Regression	0.006	0.618
Random Forest	0.004	0.711
Artificial Neural Network	0.005	0.647

4. References

[1] M. Norris, "Predicting Earthquakes Using Machine Learning," Medium, 11-Sep-2019. [Online]. Available: https://medium.com/marionete/predicting-earthquakes-using-machine-learning-21689435dc52.

[2] United States Geological Survey, "Earthquake Search," USGS Earthquake Hazards Program. [Online]. Available: https://earthquake.usgs.gov/earthquakes/search/.

[3] G. Marques, "Earthquakes for ML Prediction," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/gustavobmgm /earthquakes-for-ml-prediction.

[4] S. Mousavi, "STEAD: A Global Data Set of Seismic Signals for AI," GitHub, 2020. [Online]. Available: https://github.com/smousavi05/STEAD.

[5] T. Singh, "Earthquake Predictor," GitHub, 2021. [Online]. Available: https://github.com/noobtdbs/Earthquake-Predictor.

[6] J. Mohanty and B. Krishnan, "Machine Learning for Earthquake Prediction: A Review (2017–2021)," Earth Science Informatics, vol. 16, no. 1, pp. 101–120, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s12145-023-00991-z.

[7] M. Picozzi et al., "INSTANCE: The Italian Seismic Dataset for Machine Learning," Earth System Science Data, vol. 13, pp. 5509–5529, 2021. [Online]. Available: https://essd.copernicus.org/articles/13/5509/2021/ .

[8] S. Sah, A. U. Rahman, and J. Iqbal, "Deep Learning-Based Earthquake Prediction Technique Using Seismic Data," IEEE Access, vol. 10, pp. 38267–38275, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10185869.

5. Contribution

Member	**Contribution**
Aashika Gupta	Linear Regression Training, ANN Training and Optimization
Mudit Sengar	Feature Engineering & Scaling, Random Forest Regressor Training
Shambhavi Agarwal	Data Preprocessing, Polynomial Regression Training, Model Evaluation
Shrey Patel	Initial Feature Selection, Support Vector Regressor Training, Report Making

Table 2: Team Contributions