PO_64 Project RichterX
Earthquake Pattern Analysis & Magnitude Prediction

Intern: Hetal Patel Dholu
Organization: Outsource360

Abstract

This project focuses on analyzing earthquake and volcanic datasets to identify patterns in magnitude, depth, time, and location, with the aim of deriving insights that can contribute to seismic hazard preparedness. Machine learning methods were applied, including supervised learning for magnitude prediction and unsupervised clustering for pattern discovery. The study emphasizes the importance of realistic temporal validation and highlights limitations, key insights, and recommendations for future research.

1. Introduction

Earthquakes are complex natural phenomena that cause devastating impacts on human society and infrastructure. Despite decades of research, earthquake prediction remains one of the most challenging problems in geosciences. This project applies machine learning techniques to explore whether data-driven models can provide useful signals about earthquake magnitudes and event clustering. By leveraging supervised learning and clustering techniques, we aim to uncover patterns that may contribute to risk understanding and preparedness.

2. Dataset Description

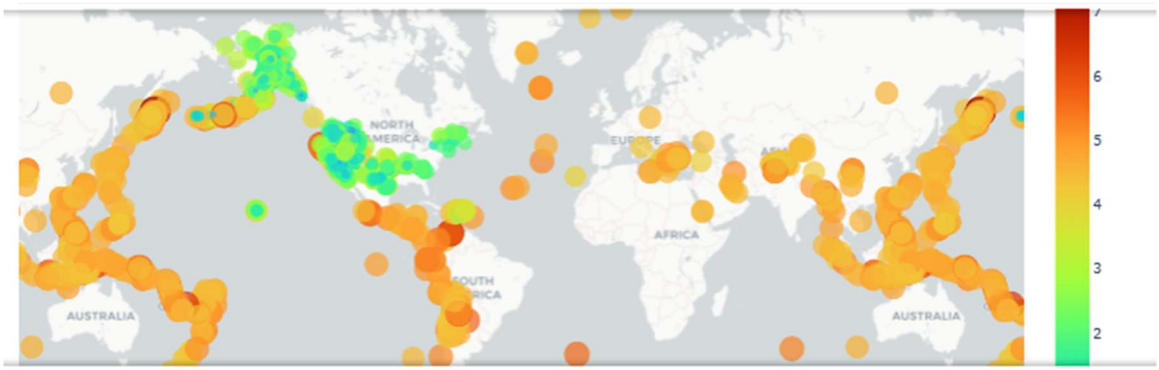
The dataset consists of approximately 9,000–9,500 earthquake and volcanic records, containing variables such as magnitude, depth, latitude, longitude, and event time. Some volcanic-related attributes were also present. Preprocessing steps included handling missing values, standardizing timestamps, normalization, and outlier detection.

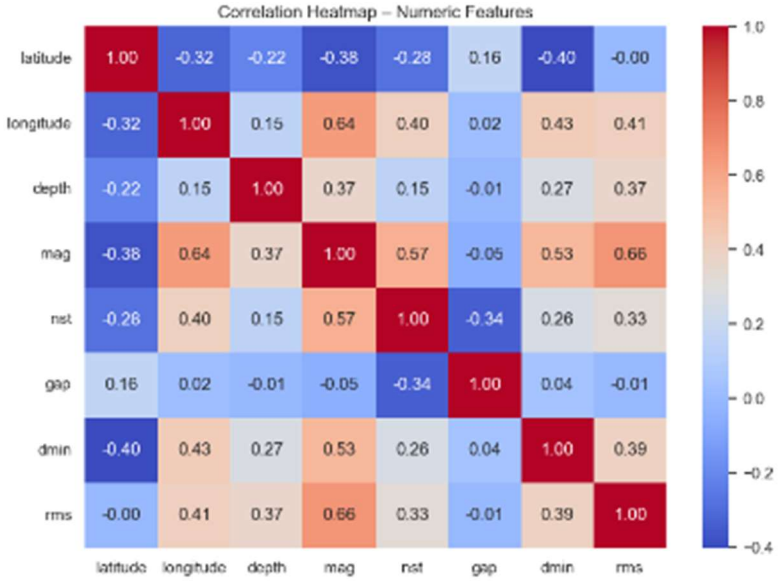
⚠ Due to system resource limitations, only ~1 month of data could be processed (~9.5k records). Extracting and processing larger yearly datasets required significant time and often caused system shutdowns. The program is, however, designed to handle full yearly datasets given sufficient computational resources.

3. Methodology

The methodology consisted of several stages:

- Data preprocessing and cleaning.
- Exploratory Data Analysis (EDA) including scatterplots, correlation heatmaps, and geospatial mapping.
- Feature engineering such as lag features and outlier flags.
- Supervised modeling for magnitude prediction, testing Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and LightGBM.
- Temporal validation using sliding-window cross-validation (train=2000, test=500).
- Clustering using KMeans and DBSCAN for unsupervised analysis.





4. Results & Observations

Supervised Learning Results

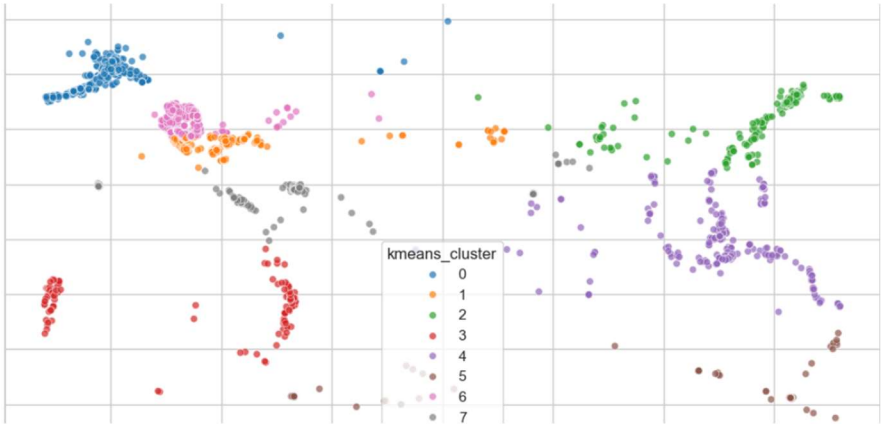
- Random Forest achieved the best balance between predictive performance and computational efficiency.
- Metrics: ROC-AUC ≈ 0.99, PR-AUC ≈ 0.85, F1 ≈ 0.76, with execution time ≈ 3.4s.
- Boosted models (LightGBM, XGBoost) achieved comparable metrics but were significantly slower.
- Logistic Regression was efficient but weaker in precision-recall performance.

```
Logistic         ROC_AUC=0.9805, PR_AUC=0.7248, F1@0.5=0.5815, best_thresh=0.897, F1@best=0.6900, Time_Taken=0.752
RandomForest     ROC_AUC=0.9898, PR_AUC=0.8528, F1@0.5=0.7487, best_thresh=0.393, F1@best=0.7606, Time_Taken=3.422
GradientBoosting ROC_AUC=0.9834, PR_AUC=0.7960, F1@0.5=0.6632, best_thresh=0.139, F1@best=0.7241, Time_Taken=36.471
XGBoost          ROC_AUC=0.9871, PR_AUC=0.8338, F1@0.5=0.7418, best_thresh=0.399, F1@best=0.7636, Time_Taken=50.286
LightGBM         ROC_AUC=0.9879, PR_AUC=0.8387, F1@0.5=0.7511, best_thresh=0.339, F1@best=0.7692, Time_Taken=33.766
```

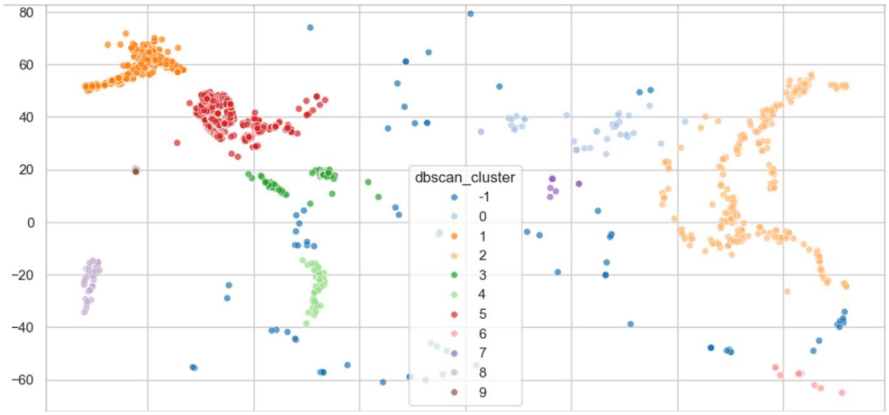
Clustering Results

- KMeans produced spherical clusters but forced some noisy points into groups.
- DBSCAN captured natural, irregular clusters and identified noise points effectively.
- DBSCAN was selected as the final clustering approach.

KMeans Clustering Output



DBSCAN Clustering Output



Observations
- Visual inspection confirms DBSCAN forms more meaningful clusters than KMeans.
- Later folds of sliding-window CV showed slightly higher errors, suggesting concept drift.

5. Limitations
- Earthquake and volcanic events are inherently stochastic and unpredictable.
- Dataset size (~9.5k records) limited model complexity and generalization.
- Feature space is limited; advanced seismic/geophysical variables (tectonic stress, GPS/InSAR deformation, gas emissions) were not included.
- Seismic waveform features such as **P-wave** and **S-wave** amplitudes could not be added but may improve predictive capability in future.
- Practical challenge: Extracting yearly datasets was computationally intensive and sometimes caused system shutdowns. The program supports yearly extraction, but execution requires patience and stronger compute resources.

6. Recommendations & Future Work
- Explore advanced sequence models such as LSTMs, GRUs, and Transformers.
- Engineer lagged, rolling, and spectral features, including seismic waveforms.
- Implement probabilistic forecasting (Bayesian models, quantile regression) to quantify uncertainty.
- Improve clustering with hierarchical DBSCAN (HDBSCAN) and semi-supervised approaches.
- Enrich the dataset with tectonic, geological, and volcanic gas-emission data.
- Operationalize the pipeline for real-time streaming seismic data on cloud platforms.

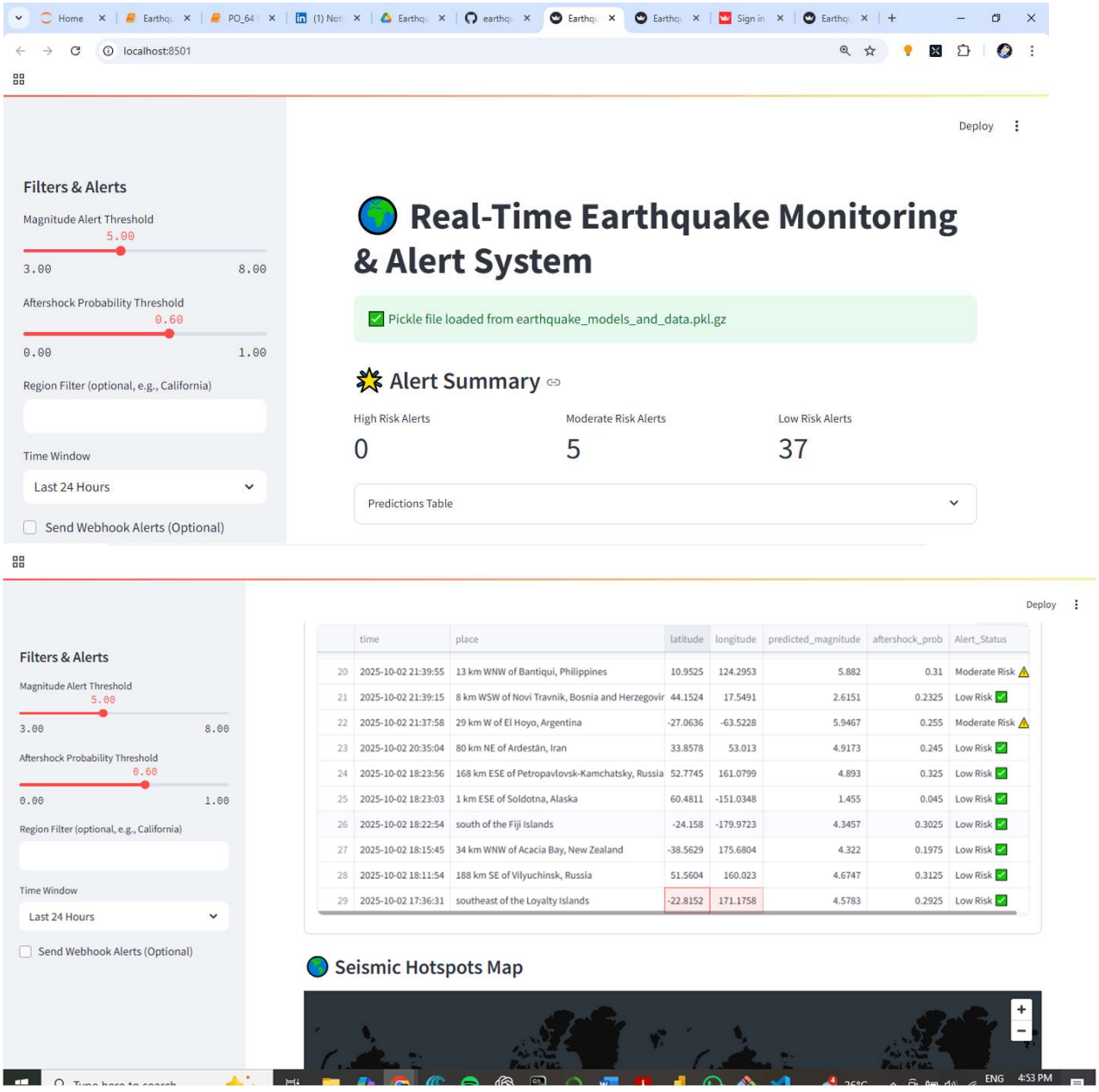
7. Conclusion
This project demonstrates that while machine learning cannot deterministically predict earthquakes, it can uncover patterns that support hazard preparedness. Sliding-window CV provided realistic temporal validation, Random Forest was chosen as the final predictive model, and DBSCAN as the final clustering method. With enhanced feature sets and advanced models, future research may yield more robust insights. The study highlights the importance of computational resources, proper validation, and integration of domain-specific knowledge.

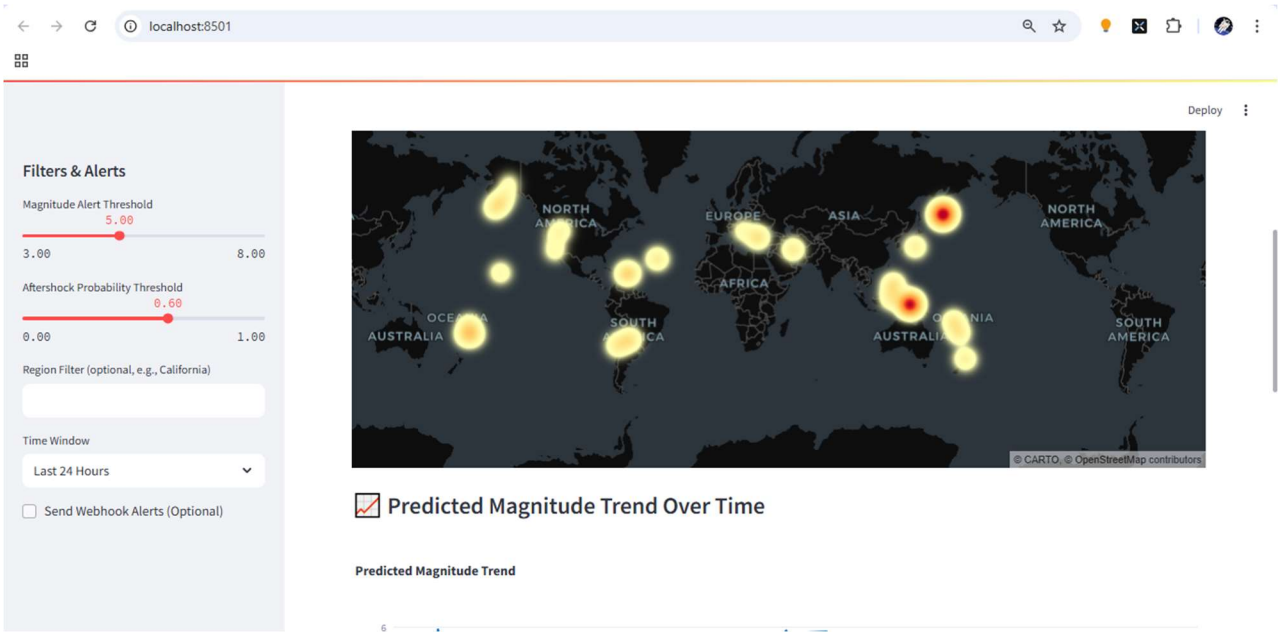
8. References

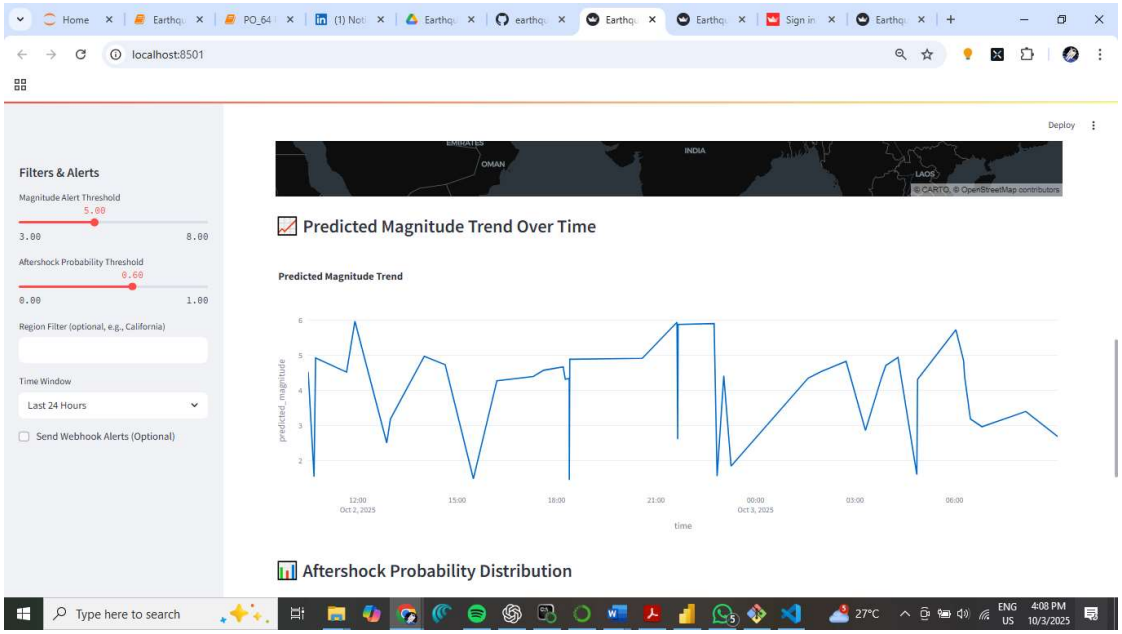
References have been compiled from the project background document and relevant literature. These include seismic hazard studies, machine learning research in time series forecasting, and documentation for scikit-learn and clustering methods.

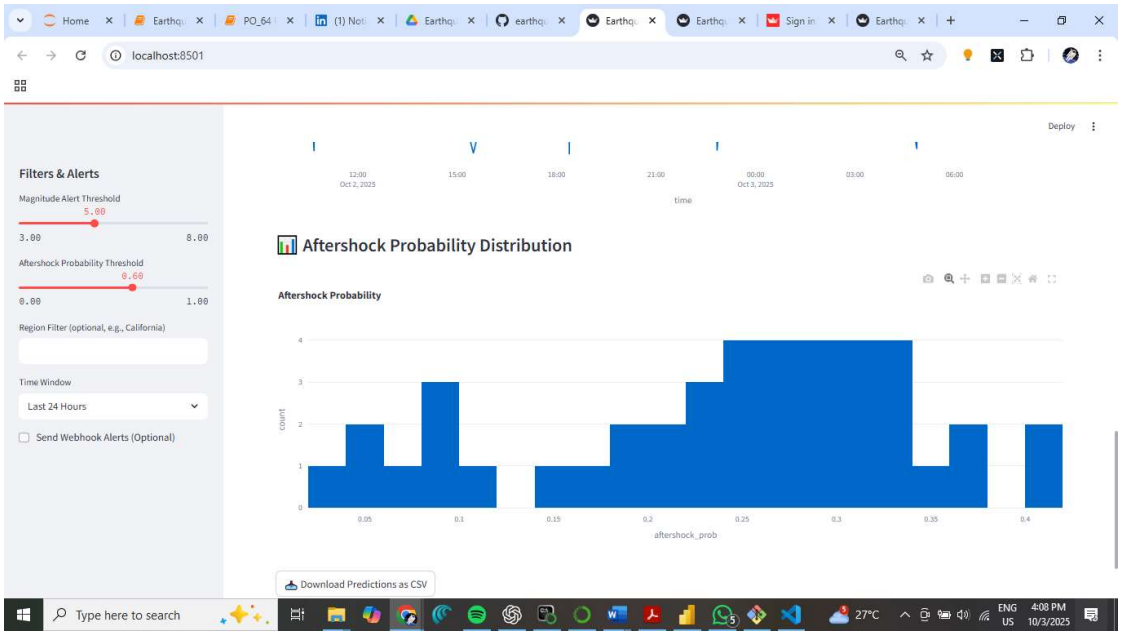
9. Streamlit Application

A Streamlit application was developed to provide an interactive interface for exploring seismic data and running predictions.




localhost:8501
Deploy
Filters & Alerts
Magnitude Alert Threshold
5.00
3.00
8.00
Aftershock Probability Threshold
0.60
0.00
1.00
Region Filter (optional, e.g., California)
Time Window
Last 24 Hours
Send Webhook Alerts (Optional)
NORTH AMERICA
EUROPE
ASIA
AFRICA
AUSTRALIA
SOUTH AMERICA
© CARTO, © OpenStreetMap contributors
Predicted Magnitude Trend Over Time
Predicted Magnitude Trend


localhost:8501
Deploy
Filters & Alerts
Magnitude Alert Threshold
5.00
3.00
8.00
Aftershock Probability Threshold
0.60
0.00
1.00
Region Filter (optional, e.g., California)
Time Window
Last 24 Hours
Send Webhook Alerts (Optional)
OMAN
INDIA
© CARTO, © OpenStreetMap contributors
Predicted Magnitude Trend Over Time
Predicted Magnitude Trend
predicted_magnitude
time
Aftershock Probability Distribution
Type here to search
27°C
ENG
US
4:08 PM
10/3/2025


localhost:8501
Deploy
Filters & Alerts
Magnitude Alert Threshold
5.00
3.00
8.00
Aftershock Probability Threshold
0.60
0.00
1.00
Region Filter (optional, e.g., California)
Time Window
Last 24 Hours
Send Webhook Alerts (Optional)
12:00
Oct 2, 2025
15:00
18:00
21:00
00:00
Oct 3, 2025
03:00
06:00
time
Aftershock Probability Distribution
Aftershock Probability
count
aftershock_prob
Download Predictions as CSV
Type here to search
27°C
ENG
US
4:08 PM
10/3/2025