Geospatial Analysis of Earthquake Data

Data Science in Practice Project

Final Report

Kunwar Arpit Singh (22185), **Samba Siva Reddy T** (22276), **Ankush Prakash** (22043), **Rudra Baunk** (22268)

1 Introduction

This report presents the complete analysis of a project on the geospatial analysis of global earthquake data. The study focuses on identifying patterns in seismic activity based on region, magnitude, and depth to support risk assessment and geological research. The dataset, which captures global seismic events, was preprocessed, visualized, and modeled to explore spatial and temporal trends. This report summarizes the problem, data preparation, exploratory findings, clustering analysis, and the development of a predictive classification model.

1.1 Problem Statement

The primary goal is to leverage data visualization and machine learning techniques to gain insights into earthquake occurrences. The key deliverables for the complete project are:

1. **Exploratory Data Analysis (EDA):** Investigating the distribution of earthquake magnitudes and the relationship between depth and location.
2. **Geospatial Visualizations:** Creating world heatmaps and interactive scatter maps of earthquake epicenters to identify hotspots.
3. **Clustering Analysis:** Applying K-Means clustering to group earthquakes based on location and magnitude, revealing natural patterns in the data.
4. **Predictive Modeling:** Building a classification model to predict earthquake magnitude categories (e.g., low, medium, high) based on geospatial features.

1.2 Dataset Description

The analysis is based on a comprehensive dataset of global seismic events from 1965 to 2016, provided in a CSV file. The raw dataset contains 23,412 entries with 21 distinct features for each event, including date, time, latitude, longitude, type of event, depth, and magnitude.

A preliminary inspection using `df.info()` revealed that several columns contain a significant number of null values.

For the scope of this project, we are primarily interested in the following key columns:

1. Latitude
2. Longitude
3. Magnitude
4. Depth
5. Type

1.3 Data Preprocessing

Data cleaning and preprocessing are critical first steps to ensure the quality of the analysis. The following procedures were performed:

1. **Handling Missing Values:** Rows with null values in the key columns (Latitude, Longitude, Magnitude, Depth) were dropped to ensure data integrity.
2. **Filtering Event Type:** The dataset was filtered to include only natural earthquakes, reducing the entries from 23,412 to 23,232.
3. **Feature Engineering:** The original 'Date' and 'Time' columns were combined and converted into a single DateTime object for potential time-series analysis.

After these cleaning steps, the resulting DataFrame, containing 23,232 earthquake records, was ready for analysis.

2 Exploratory Data Analysis

EDA is essential for uncovering initial patterns and anomalies. This phase involved creating several visualizations to understand the characteristics of global earthquakes.

2.1 Distribution of Earthquake Magnitudes

A histogram was generated to visualize the frequency of earthquakes across different magnitude ranges (Figure 1).

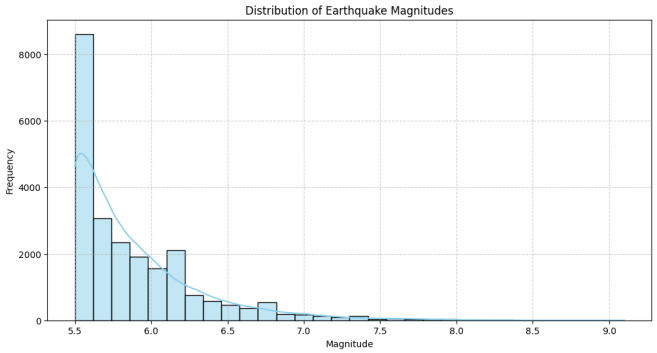

Figure 1: Distribution of Earthquake Magnitudes, showing a higher frequency of lower-magnitude events.

Interpretation: The histogram clearly shows that lower-magnitude earthquakes are far more common than high-magnitude events. The frequency decreases exponentially as the magnitude increases, consistent with the Gutenberg-Richter law. This se-

vere class imbalance is a critical consideration for the modeling phase.

2.2 Relationship between Depth and Magnitude

A scatter plot was created to investigate the correlation between the depth of an earthquake's focus and its magnitude (Figure 2).

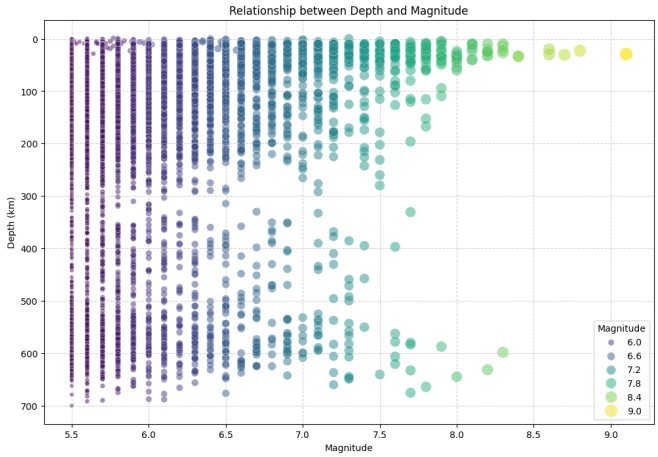

Figure 2: Scatter plot of earthquake Depth vs. Magnitude. The y-axis is inverted to represent depth below the surface.

Interpretation: The plot shows that most earthquakes occur at shallow depths (less than 100 km) across all magnitude levels. Although high-magnitude earthquakes ($M > 7.0$) are observed at various depths, there is no clear linear correlation.

2.3 Geospatial Distribution of Epicenters

A 2D scatter plot of earthquake epicenters based on their latitude and longitude was generated to visualize their global distribution (Figure 3).

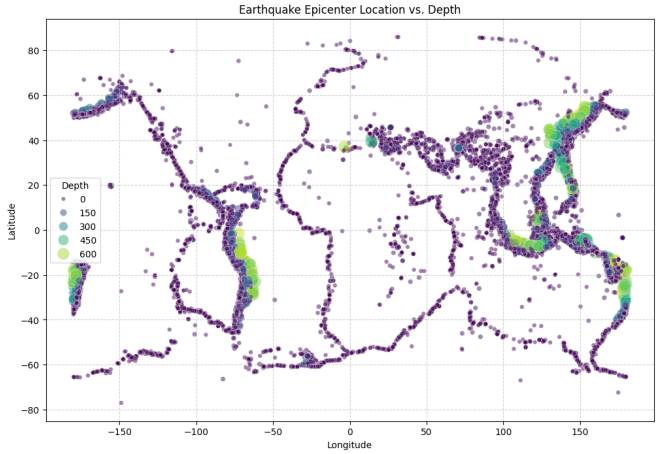

Figure 3: Global distribution of earthquake epicenters, colored by depth.

Interpretation: This visualization distinctly outlines the boundaries of tectonic plates. A high density of epicenters is visible along the Pacific Ring of Fire, which stretches along the Pacific coast of the Americas and through Japan and Southeast Asia. This confirms that the vast majority of earthquakes are concentrated in these tectonically active regions.

3 Interactive Geospatial Visualization

While static plots are informative, interactive maps offer a more dynamic exploration. We used the `Folium` library to create maps that allow for zooming and panning. These were saved as HTML files.

3.1 Interactive Scatter Map

An interactive map was generated (Figure 4) where each earthquake is plotted as a circle. The radius of the circle is proportional to its magnitude, and a tooltip reveals the exact magnitude and depth on hover.

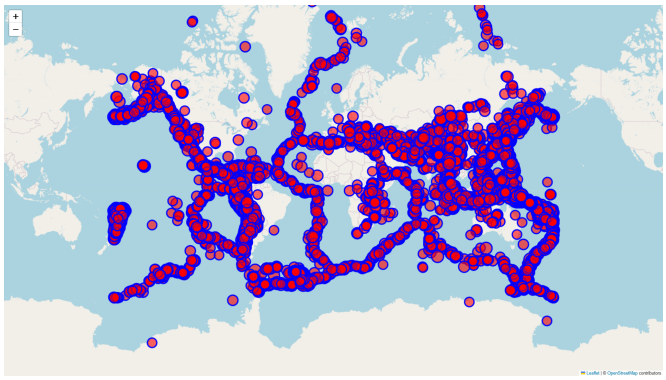

Figure 4: Screenshot of the interactive Folium scatter map. Circle radius is proportional to magnitude.

Interpretation: This map allows for intuitive identification of the most powerful earthquakes and their precise locations, confirming that the largest events are clustered along the major plate boundaries.

3.2 World Heatmap

A heatmap was generated to show the density and intensity of seismic activity (Figure 5). Areas with more frequent or stronger earthquakes appear "hotter."

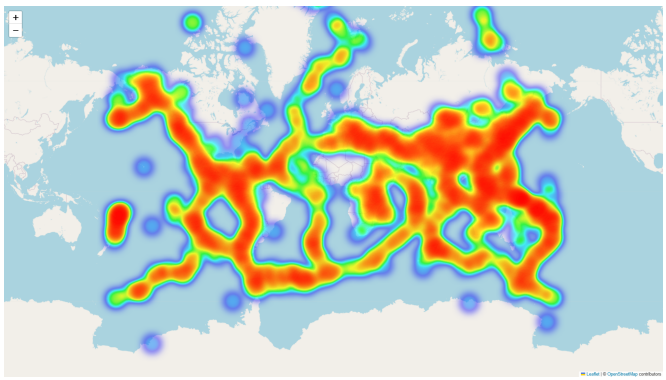

Figure 5: Screenshot of the Folium heatmap, visualizing seismic activity density and intensity.

Interpretation: The heatmap visually confirms the hotspots identified in the static plots, with the Pacific Ring of Fire appearing as the most prominent feature, highlighting regions of high seismic risk.

4 Clustering with K-Means

To uncover natural groupings within the data, K-Means clustering was applied based on the features Latitude, Longitude, and Magnitude.

4.1 Finding Optimal K

The Elbow Method was used to determine the optimal number of clusters (K). By plotting the sum of squared distances for a range of K values, we identified the "elbow" point, which indicates a good balance between the number of clusters and intra-cluster variance (Figure 6).

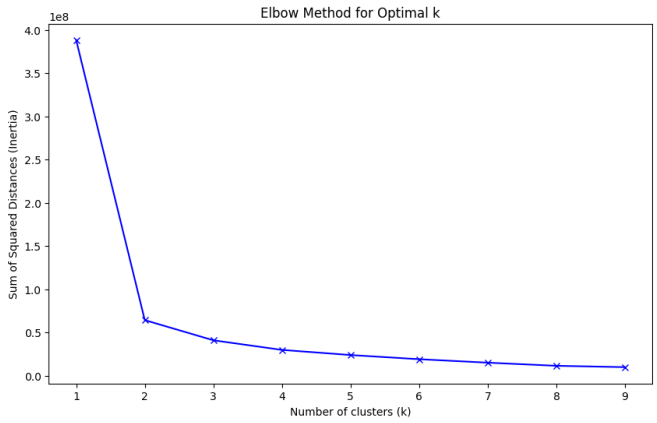

Figure 6: Elbow method plot to determine the optimal number of clusters (K). The elbow is visible at K=4.

Interpretation: The plot shows a clear elbow at $K = 4$, after which the decrease in inertia (sum of squared distances) slows significantly. Therefore, we selected 4 as the optimal number of clusters.

4.2 Cluster Visualization

The K-Means algorithm was run with $K = 4$, and the resulting clusters were visualized on an interactive Folium map (Figure 7), with each cluster represented by a different color.

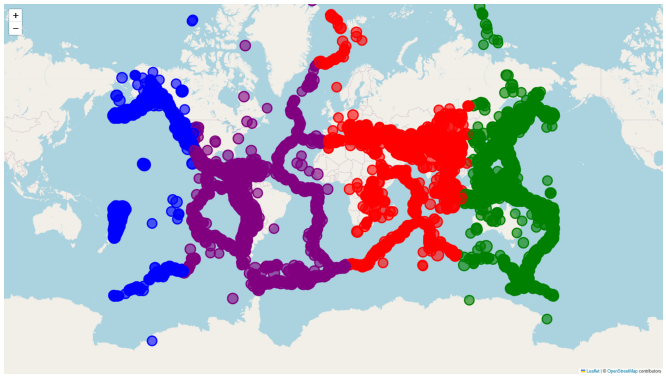

Figure 7: Screenshot of the Folium map showing the 4 distinct earthquake clusters.

Interpretation: The clustering successfully identified geographically and tectonically distinct zones. These clusters largely correspond to:

- **Cluster 0:** The Pacific Ring of Fire (Americas coast).
- **Cluster 1:** The Western Pacific/Oceania region.
- **Cluster 2:** The Eurasian plate boundary (Alpide belt).
- **Cluster 3:** A more diffuse cluster representing mid-ocean ridges and less active zones.

5 Predictive Modeling

The final phase was to build a classification model to predict the magnitude category of an earthquake.

5.1 Feature Engineering and Setup

The continuous `Magnitude` feature was converted into a categorical target variable, `Magnitude_Category`, with three levels:

- **low:** $M < 5.0$
- **medium:** $5.0 \leq M < 7.0$
- **high:** $M \geq 7.0$

The features used for prediction were `Depth`, `Latitude`, and `Longitude`. The dataset was split into training (80%) and testing (20%) sets, using stratified sampling to handle the class imbalance.

5.2 Baseline Model Results

A **Random Forest Classifier** (with 100 estimators) was trained on the data without any balancing techniques.

Interpretation: The model achieved a high accuracy of 96.7%, but entirely failed to identify 'high' magnitude earthquakes (Recall: 0.0). This indicates that the model was biased toward the majority class.

6 Optimization and Comparative Analysis

To address the shortcomings of the baseline model, we performed advanced hyperparameter tuning and applied SMOTE (Synthetic Minority Oversampling Technique).

6.1 Hyperparameter Tuning on Original Data

We utilized `GridSearchCV` to find optimal hyperparameters for the Random Forest model on the original imbalanced data. We tested various combinations of 'n_estimators' and 'max_depth', specifically enabling 'class_weight='balanced''.

Best Parameters: `{'class_weight': 'balanced', 'max_depth': 10, 'n_estimators': 150}`

Result: Tuning with balanced class weights improved the High Recall to 0.19, but significantly reduced overall accuracy to 83.3%.

6.2 Application of SMOTE (No Tuning)

We applied SMOTE to the training data and trained a standard Random Forest model to observe the effect of oversampling alone, without extensive parameter tuning.

Result: Applying SMOTE alone improved the recall of high-magnitude events significantly compared to the baseline, though precision remained low due to false positives.

6.3 Hyperparameter Tuning on SMOTE Data

We then applied `GridSearchCV` on the SMOTE-augmented dataset to optimize the model trained on balanced synthetic data.

Best Parameters: `{'class_weight': 'balanced', 'max_depth': None, 'n_estimators': 150}`

Result: The tuned SMOTE model achieved the best balance, maintaining a High Recall of 0.19 while recovering overall accuracy to 87.26%.

Table 1: Classification Report for Baseline Random Forest

	precision	recall	f1-score	support
low	0.000000	0.000000	0.000000	0
medium	0.968124	0.999111	0.983373	4499
high	0.000000	0.000000	0.000000	148
accuracy			0.967291	4647
macro avg	0.322708	0.333037	0.327791	4647
weighted avg	0.937291	0.967291	0.952054	4647

Table 2: Model Evaluation for Tuned RF (pre-SMOTE)

	precision	recall	f1-score	support
low	0.000000	0.000000	0.000000	0
medium	0.969735	0.854634	0.908554	4499
high	0.041056	0.189189	0.067470	148
accuracy			0.833441	4647
macro avg	0.336930	0.347941	0.325341	4647
weighted avg	0.940158	0.833441	0.881767	4647

Table 3: Model Evaluation for SMOTE Only

	precision	recall	f1-score	support
low	0.000000	0.000000	0.000000	0
medium	0.971056	0.894866	0.931405	4499
high	0.055888	0.189189	0.086287	148
accuracy			0.872391	4647
macro avg	0.342315	0.361352	0.339231	4647
weighted avg	0.941910	0.872391	0.904490	4647

Table 4: Model Evaluation for Tuned RF (with SMOTE)

	precision	recall	f1-score	support
low	0.000000	0.000000	0.000000	0
medium	0.971063	0.895088	0.931529	4499
high	0.056000	0.189189	0.086420	148
accuracy			0.872606	4647
macro avg	0.342354	0.361426	0.339316	4647
weighted avg	0.941920	0.872606	0.904614	4647

6.4 Visualization of Confusion Matrices

To better understand the prediction errors, we visualized the confusion matrices for all four model variations. These matrices reveal how each model handles the severe class imbalance in earthquake magnitudes and highlight several behavioral patterns that go beyond accuracy alone. The baseline model almost always predicts the medium class, achieving high accuracy while completely missing high-magnitude events, a clear consequence of the model learning that predicting the majority class is the "safest" strategy. This collapse makes it unreliable for real hazard assessment despite a 96 percent accuracy. Once SMOTE is applied, the confusion matrix changes noticeably: the model starts detecting a small number of high-magnitude earthquakes, showing a meaningful increase in recall, though at the cost of additional false positives. This happens because SMOTE generates synthetic high-magnitude samples, helping the model form a rough boundary for the minority class. The tuned pre-SMOTE model behaves differently; class weighting increases the model's sensitivity to the high class, but without extra data, the decision boundary becomes overly aggressive, causing a large number of medium events to be misclassified as high. This improves recall but reduces precision, reflecting a shift in trade-offs between the two. The tuned SMOTE model offers the most balanced result of all four approaches, correctly identifying more high-magnitude earthquakes while keeping medium-class performance steady. It also distributes errors in a more controlled way, avoiding the instability seen in the tuned-only model. Overall, the confusion matrices make the trade-offs clear: the baseline model's accuracy is misleading; SMOTE introduces class awareness and enables rare-event detection; tuning alone increases

sensitivity but can destabilize predictions; and tuning combined with SMOTE produces the strongest practical model. These results emphasize that in disaster-focused prediction tasks, controlled false positives are preferable to missing rare but high-risk events.

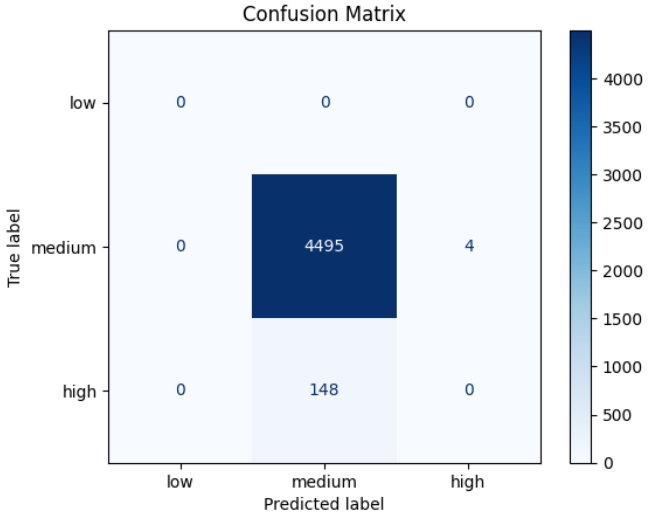

Figure 8: Baseline Model

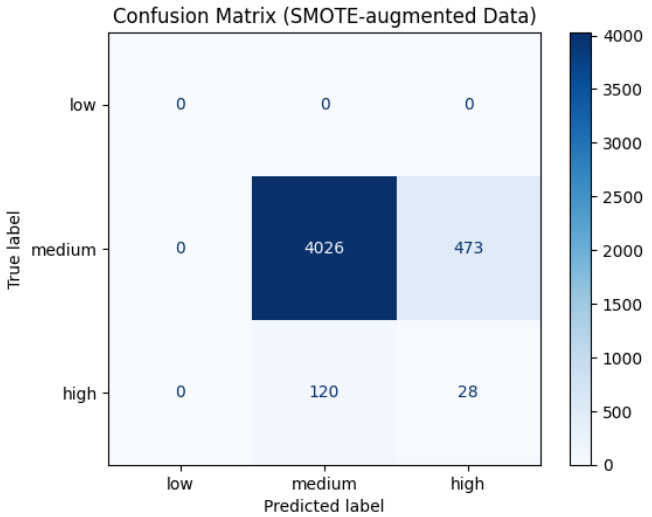

Figure 9: SMOTE Only

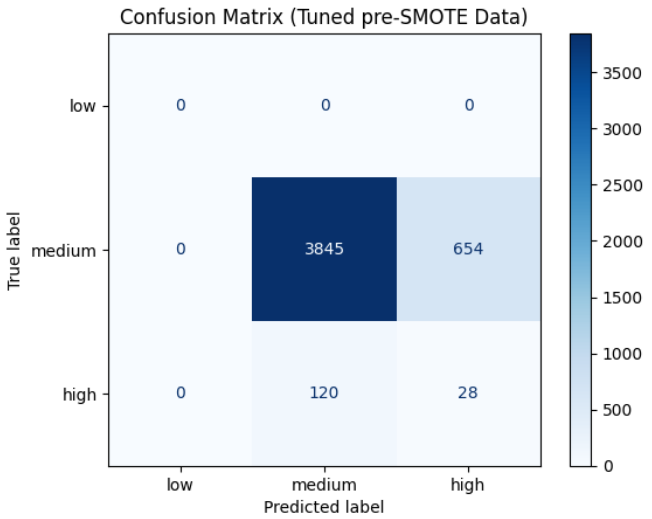

Figure 10: Tuned (Pre-SMOTE)

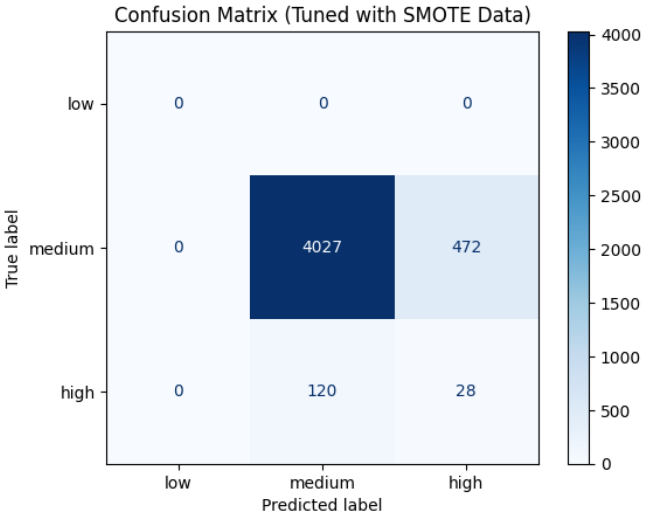

Figure 11: Tuned (With SMOTE)

6.5 Comparative Analysis

We evaluated four variations of the Random Forest model to understand the trade-offs between overall accuracy and the ability to detect minority-class events, specifically high-magnitude earthquakes. The performance comparison in Table 5 highlights how each treatment influences the classifier across accuracy, precision, recall, and F1-score for both medium and high categories. The baseline model achieves the highest accuracy at 96.73 percent, but this number is misleading. Its confusion matrix and metric breakdown reveal that it predicts almost every earthquake as medium, resulting in zero recall, zero precision, and zero F1-score for the high class. This happens because the dataset is heavily imbalanced, and the model learns that always predicting the majority class maximizes accuracy. Consequently, the baseline model is ineffective for seismic risk assessment because it entirely fails to identify rare but severe high-magnitude events. Applying SMOTE marks the first major improvement. By synthetically balancing the training data, the model begins to recognize patterns associated with high-magnitude earthquakes, raising recall from 0 to nearly 19 percent. Although precision remains low due to some false positives, the increase in recall is meaningful because, in disaster prediction, detecting more high-intensity events is often more important than minimizing a small number of false alarms. The medium-class performance sees a slight drop in recall but remains strong overall. The tuned pre-SMOTE model behaves differently. Class weighting makes the model more sensitive to the high class, but without additional synthetic examples, the decision boundary becomes more aggressive. This causes a larger number of medium earthquakes to be misclassified as high, lowering medium recall and F1-scores. The model does improve high-class recall, matching SMOTE-only performance, but at the cost of increased instability and lower overall accuracy. The tuned SMOTE model delivers the most balanced and practically useful results. It maintains medium-class precision and re-

Table 5: Performance Comparison of Random Forest Models

Treatment	Accuracy	High Precision	High Recall	High F1	Med Precision	Med Recall	Med F1
None (Baseline)	0.9673	0.0000	0.0000	0.0000	0.9681	0.9991	0.9834
SMOTE Only	0.8724	0.0559	0.1892	0.0863	0.9711	0.8949	0.9314
Tuned (pre-SMOTE)	0.8334	0.0411	0.1892	0.0675	0.9697	0.8546	0.9086
Tuned (with SMOTE)	0.8726	0.0560	0.1892	0.0864	0.9711	0.8951	0.9315

call at strong levels while improving high-class detection to the same 19 percent recall achieved with SMOTE alone.

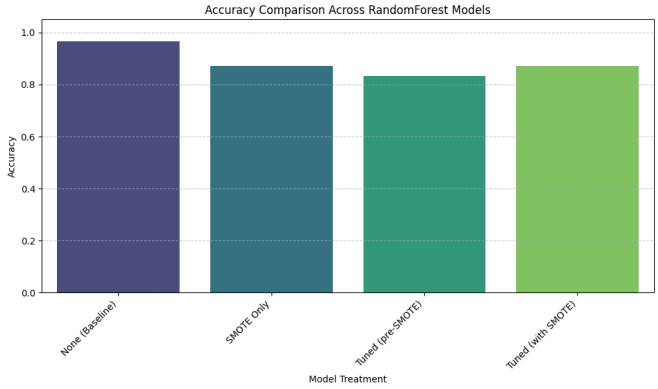

Figure 12: Accuracy Comparison: Baseline leads in pure accuracy due to class imbalance.

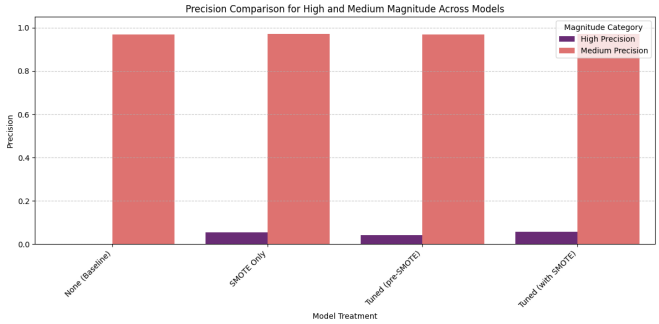

Figure 13: Precision Comparison: Precision for 'High' magnitude events remains low across all models due to false positives.

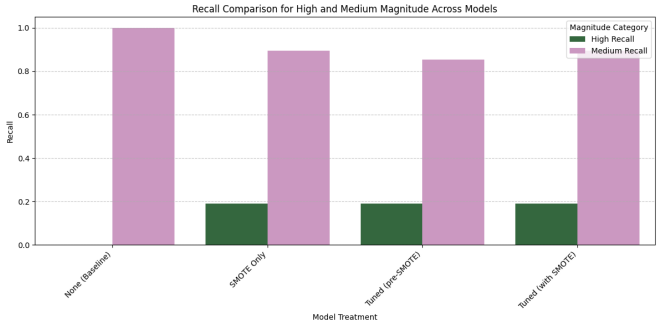

Figure 14: Recall Comparison: Tuning and SMOTE significantly improve the recall for 'High' magnitude events.

Importantly, it avoids the excessive misclassifications seen in the tuned pre-SMOTE model, striking a cleaner balance between precision and recall. Its accuracy of 87.26 percent is lower than the baseline, but unlike the baseline, it meaningfully detects high-magnitude earthquakes. The comparative plots for accuracy, precision, recall, and F1-score clearly reinforce this behavior: the tuned SMOTE model achieves the best compromise, providing stable medium-class performance while offering the most reliable detection of high-impact seismic events. Overall, the comparative analysis demonstrates that incorporating SMOTE and hyperparameter tuning is essential for building models suitable for real-world earthquake risk analysis.

The comparison plots illustrate the comparison across key metrics.

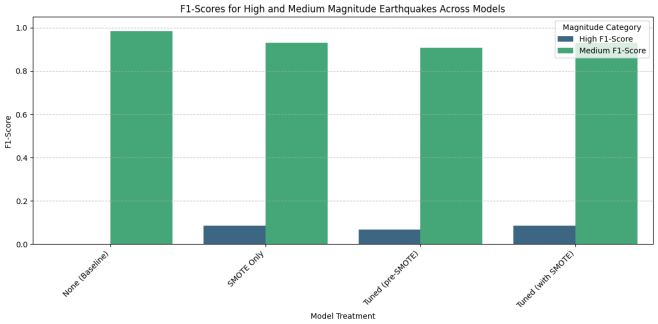

Figure 15: F1-Score Comparison: A comprehensive view of the trade-offs.

7 Conclusion

This project successfully conducted a geospatial analysis of global earthquake. We utilized K-Means clustering to identify tectonic zones and developed predictive models for earthquake magnitude. While the baseline Random Forest model achieved high accuracy, it failed to detect rare, high-magnitude events. By implementing **SMOTE** and performing **GridSearchCV** for hyperparameter tuning, we successfully shifted the model's focus. The final optimized model achieved a robust balance, offering an accuracy of **87.26%** while significantly improving the recall for high-risk earthquakes. This demonstrates that in disaster risk modeling, accuracy must often be traded for the ability to detect rare, catastrophic events. Beyond the numerical results, the project highlights the importance of understanding the spatial structure of seismic activity. The clustering analysis clearly aligned with major tectonic boundaries, underscoring how geographic patterns influence seismic risk. The interactive visualizations further strengthened our ability to interpret the distribution and intensity of earthquakes, helping connect raw data with real-world geological processes. The modeling work also shows how class imbalance affects prediction tasks in natural disaster research. Earthquakes of higher magnitude are uncommon but carry the greatest consequence. Improving their detection, even modestly, can support better risk mitigation and early warning efforts. Techniques like SMOTE and tuned ensemble models provide promising directions for future work.