Indian Institute of Science Education and Research Bhopal

DATA SCIENCE IN PRACTICE PROJECT PRESENTATION

Geospatial Analysis of Earthquake Data

Presented By:
Kunwar Arpit Singh, 22185
Ankush Prakash, 22043
Samba Siva Reddy T, 22276
Rudra Baunk, 22268

INTRODUCTION

Goal: Analyze global earthquake data to identify spatial trends, clusters, and build a model to classify magnitude categories.

Key Components:

- **Data Cleaning and Preprocessing**
- **Exploratory Data Analysis**
- **Geospatial and interactive visualizations**
- **K-Means clustering**
- **Machine Learning Classification Model**

PROJECT OBJECTIVES

We aim to use data visualization and ML methods to:

01 Study magnitude and depth distributions to understand how earthquakes vary in size and origin.

02 Analyze global epicenter locations to reveal spatial patterns in seismic activity.

03 Use clustering to uncover natural groupings of earthquakes based on their characteristics.

04 Build a model that predicts an earthquake’s magnitude category from its geospatial features.

DATASET DESCRIPTION

Source: Global earthquake data (1965–2016)

Total Entries: 23,412

Features: 21 columns including date, time, latitude, longitude, depth, magnitude, and event type

Key Columns Used:

- Latitude
- Longitude
- Magnitude
- Depth
- Type

DATA CLEANING & PREPROCESSING

Steps performed:

- Removed rows with missing latitude, longitude, magnitude, or depth
- Filtered only natural earthquakes

Data Description

- Reduced entries: 23,412 → 23,232
- Combined Date + Time into a DateTime feature
- Prepared data for EDA, clustering, and modeling

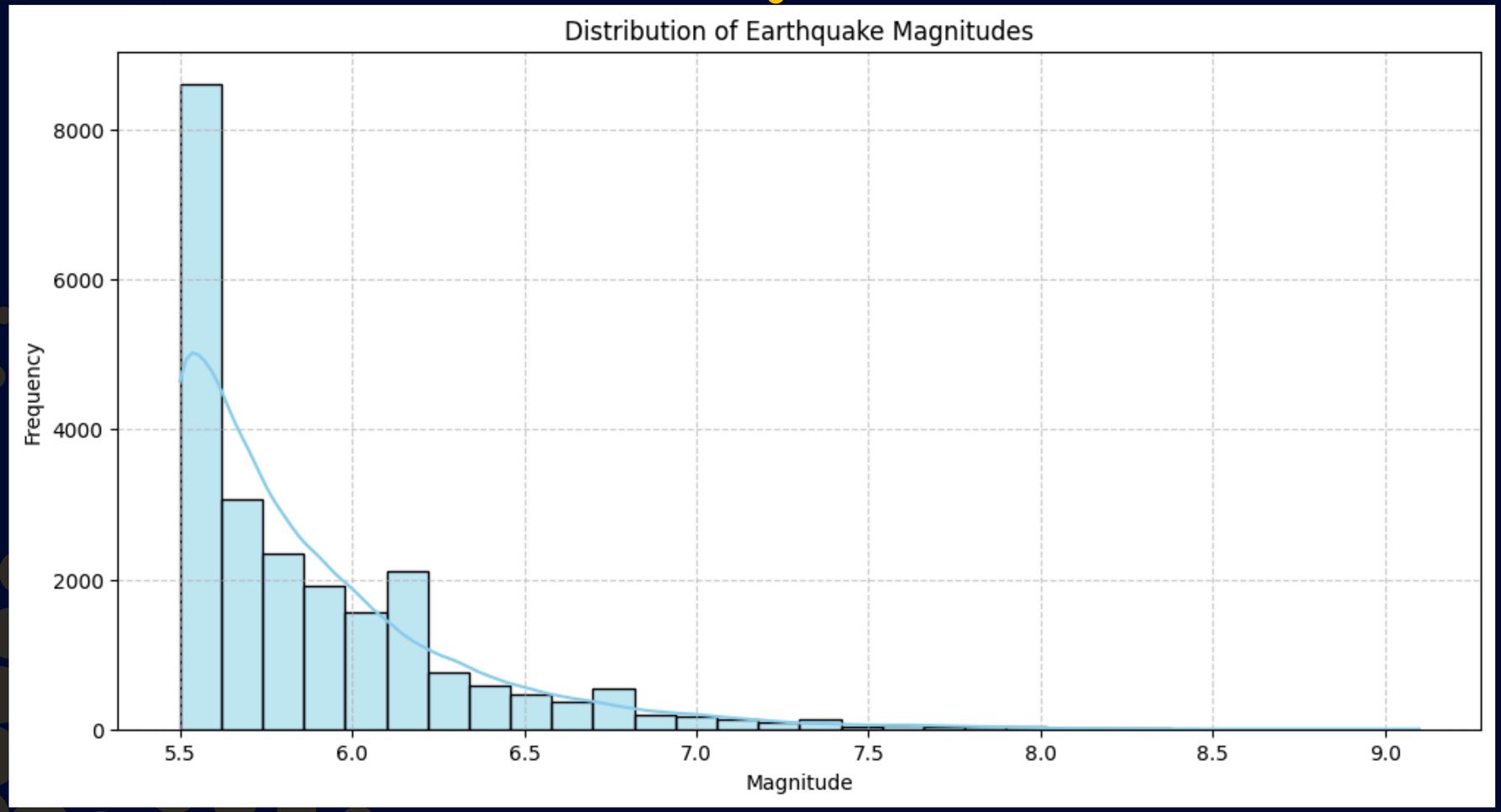
EXPLORATORY DATA ANALYSIS

EDA focused on:

- Magnitude distribution
- Depth–Magnitude Relationship
- Global Epicenter Distribution

These helped understand seismic patterns and identify hotspots

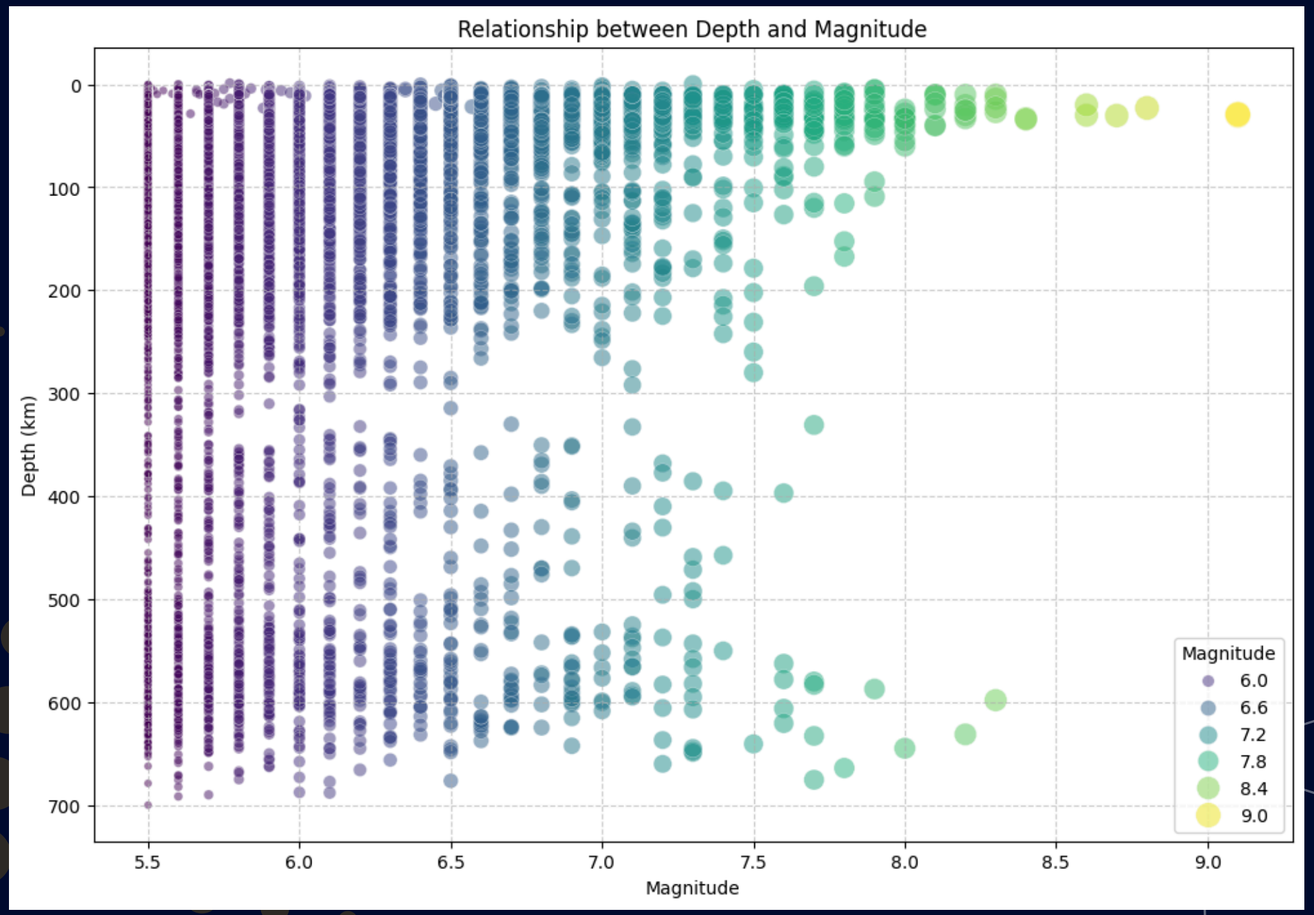
EDA: MAGNITUDE DISTRIBUTION



Observations:

- Most earthquakes have low magnitudes
- Frequency drops exponentially as magnitude increases
- Matches the Gutenberg–Richter law
- Strong indicator of class imbalance for modeling

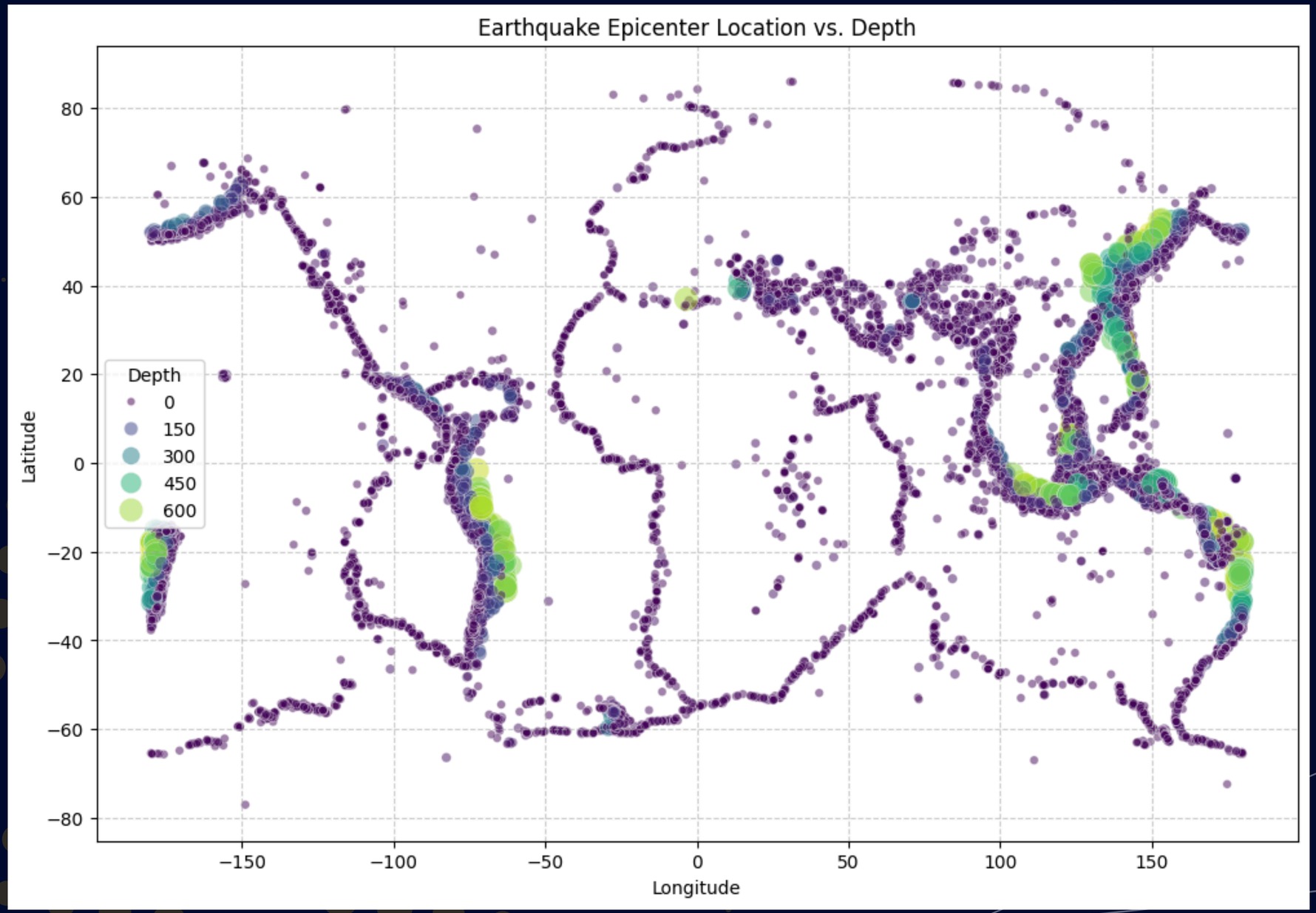
EDA: DEPTH VS. MAGNITUDE



Observations:

- Majority of earthquakes are shallow (<100 km)
- High-magnitude events occur across depths
- No clear linear correlation between magnitude and depth

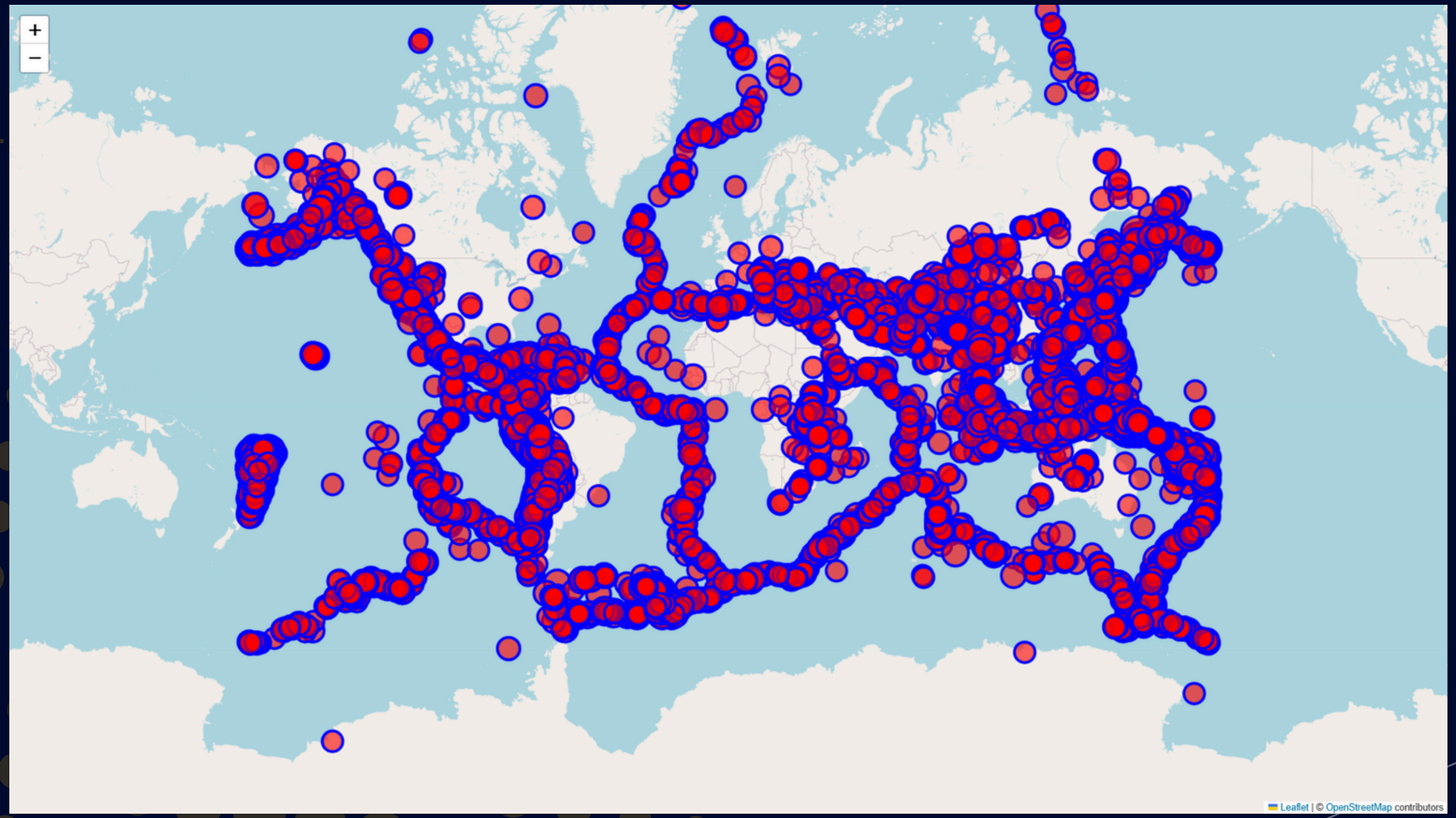
EDA: GLOBAL EPICENTER DISTRIBUTION



Insights:

- Earthquake epicenters outline tectonic plate boundaries
- Highest activity around the Pacific Ring of Fire
- Strong clustering in Japan, Southeast Asia, Western Americas

EDA: INTERACTIVE VISUALIZATION

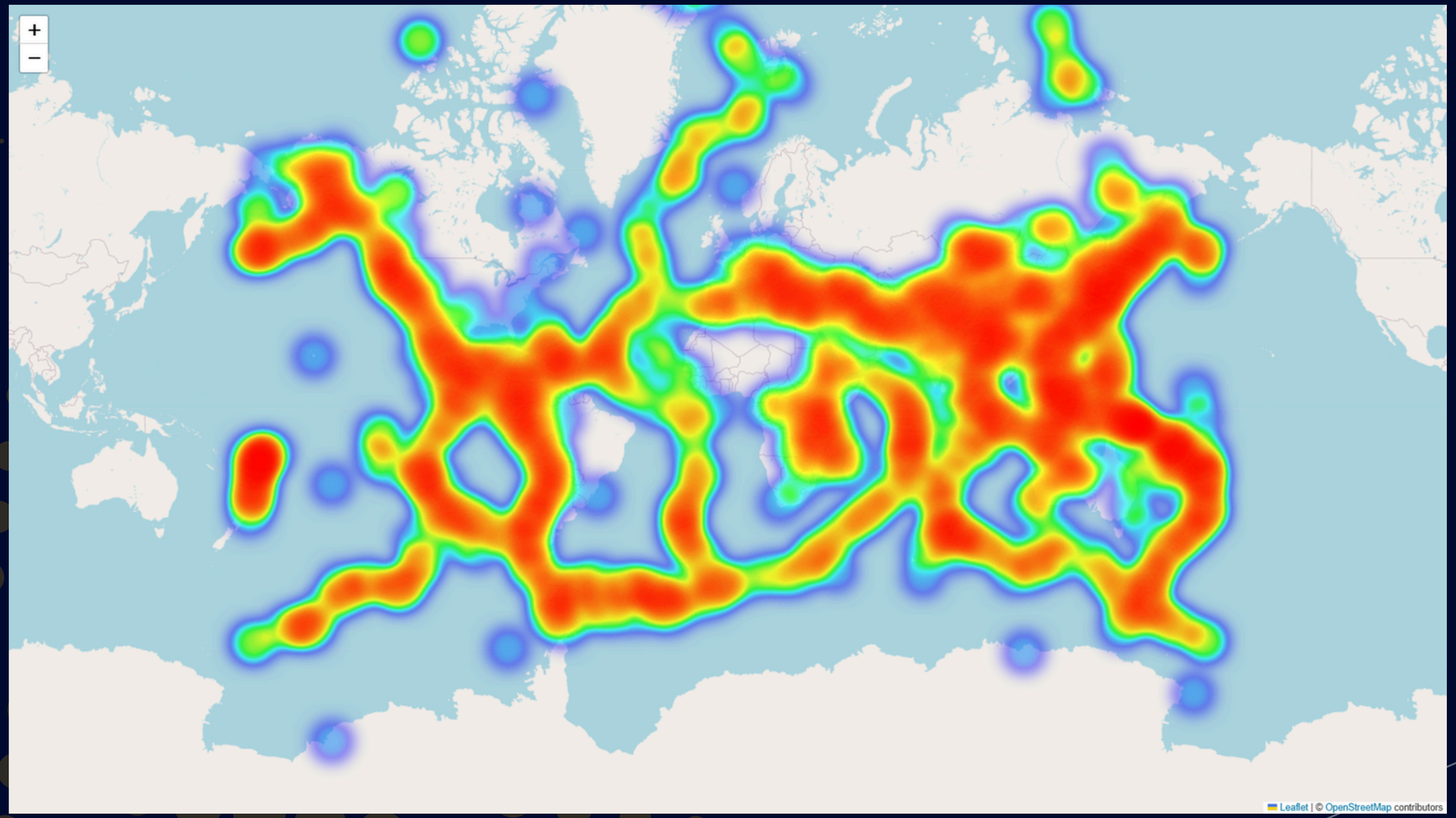


Explanation:

Built using Folium:

- Interactive scatter maps
- Circle radius proportional to magnitude
- Depth and magnitude shown on hover
- Helps explore exact locations of major events

EDA: HEATMAP VISUALIZATION



Purpose:

Highlight regions with dense seismic activity

Result:

- Pacific Ring of Fire is the hottest zone
- Eurasian plate belt also prominent
- Useful for risk visualization

CLUSTERING WITH K-MEANS

Objective:

Identify natural patterns using Latitude, Longitude, Magnitude

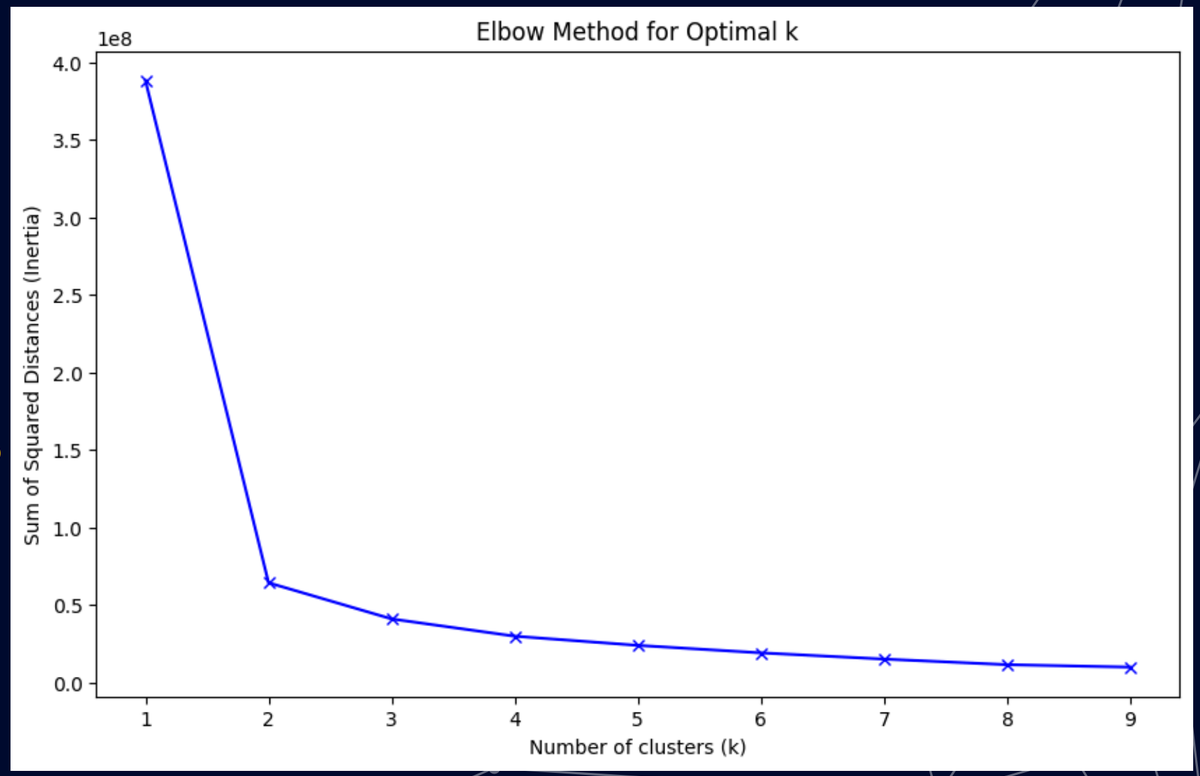
Method:

- Used Elbow Method to determine optimal K
- Compared inertia values for K = 1–10

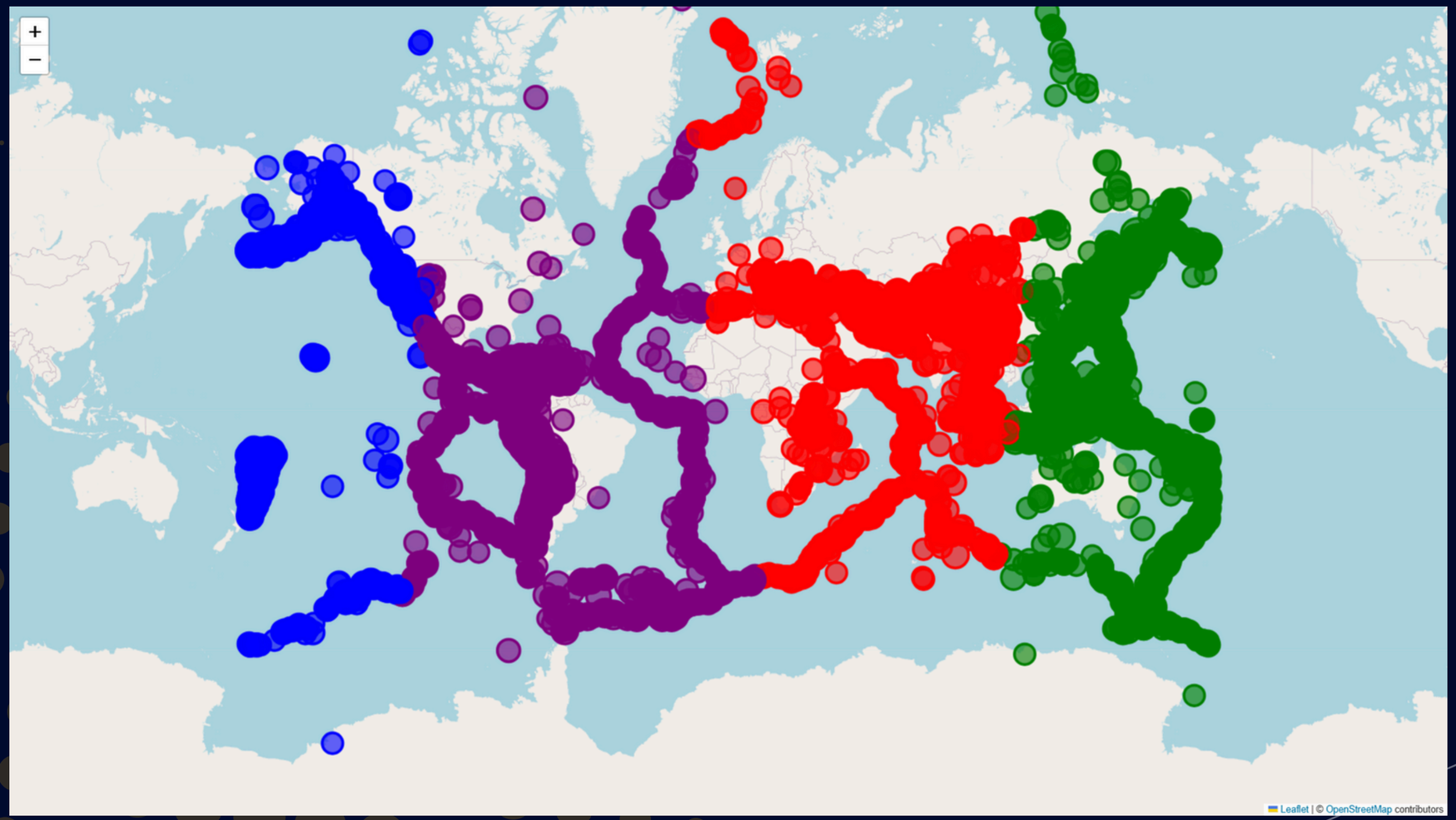
OPTIMAL K SELECTION

Result:

- Clear elbow at K = 4
- Balanced complexity and intra-cluster variance



CLUSTER INTERPRETATION



The four clusters align with major tectonic zones:

Cluster 0:
Pacific coast of the Americas

Cluster 1:
Western Pacific and Oceania

Cluster 2:
Eurasian plate boundary (Alpide belt)

Cluster 3:
Mid-ocean ridges and diffuse zones

PREDICTIVE MODELING

Goal:

Predict magnitude category:

- Low: <5.0
- Medium: 5.0–7.0
- High: ≥7.0

Features used:

- Depth
- Latitude
- Longitude

Train-test split: 80–20, stratified.

CLASS DISTRIBUTION

After preprocessing:

- Medium: 22,494
- High: 738
- Low: 0 (none remain after filtering)

Severe imbalance affects model learning.

BASELINE RANDOM FOREST MODEL

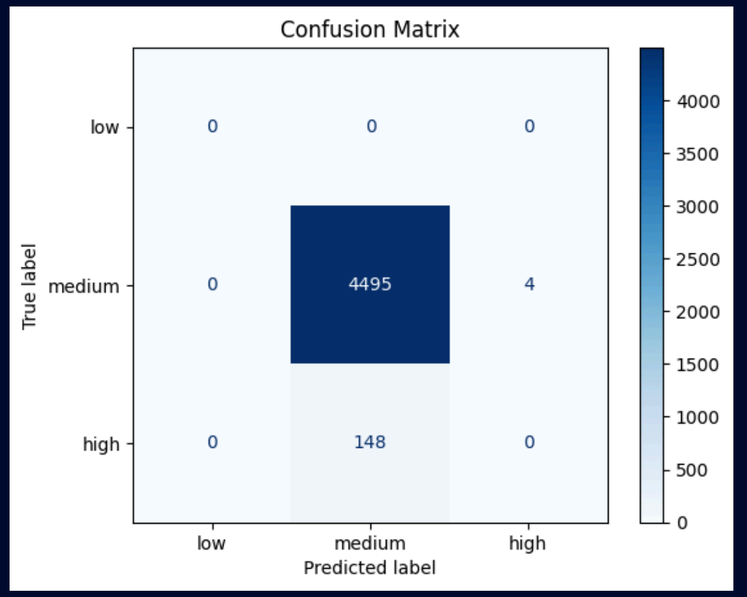
Baseline Performance:

- Accuracy: 96.73%
- Accuracy is misleading due to imbalance.

Interpretation:

- Model predicts medium class extremely well
- Completely fails to identify high-magnitude earthquakes
- No low-magnitude examples in test set
- Need better handling of rare events

	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



IMPROVING IMBALANCE WITH SMOTE

SMOTE is applied to training data:

Medium: 17,995 → 17,995

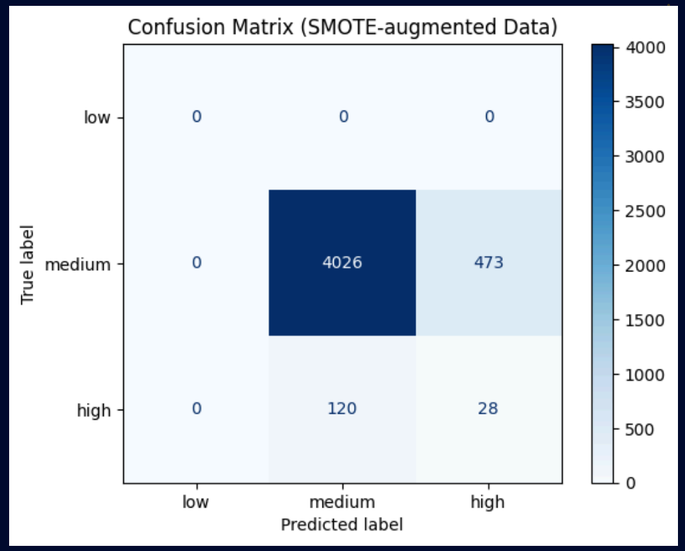
• High: 590 → 17,995 (synthetically balanced)

Accuracy: 87.24%

Interpretation:

- The model now detects some high-magnitude earthquakes
- Precision for high class remains low due to oversampling
- Medium class performance drops slightly
- Overall accuracy decreases, but high-class learning improves

	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



Trade-off: Better minority detection vs. Overall Accuracy.

HYPERPARAMETER TUNING ON BASELINE MODEL

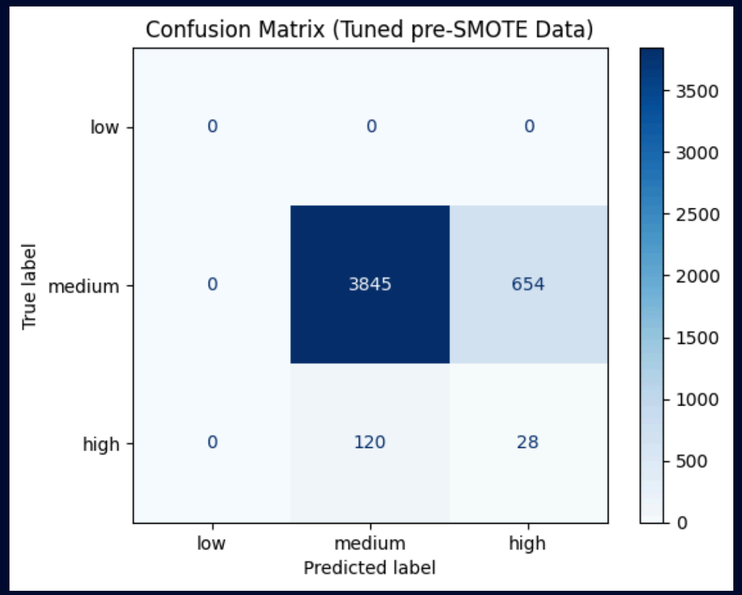
Best Parameters Found:

- n_estimators = 150
- max_depth = 10
- class_weight = balanced

Interpretation:

- The tuned model becomes more sensitive to the minority class.
- High-magnitude recall increases from 0 to ~19%, showing the effect of class weighting.
- However, overall accuracy drops to ~83% and medium-class errors increase, indicating an aggressive shift in the decision boundary.

	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



HYPERPARAMETER TUNING ON SMOTE-AUGMENTED DATA

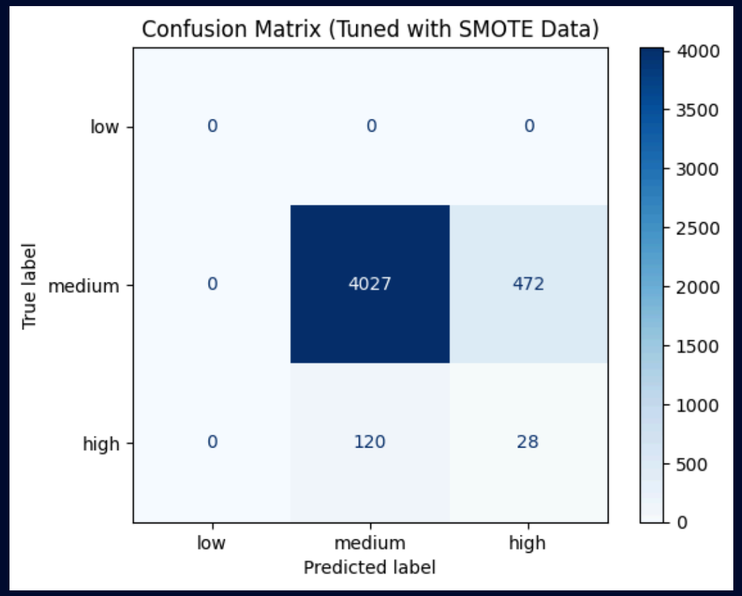
Best Parameters Found:

- n_estimators = 150
- max_depth = None
- class_weight = balanced

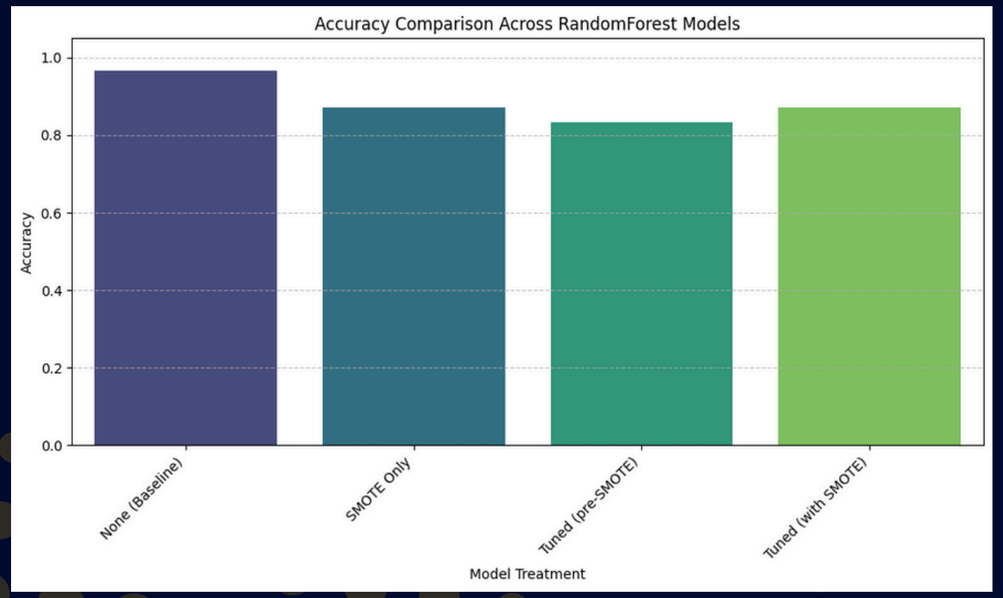
Interpretation:

- Tuning + SMOTE produces the most stable and reliable classifier.
- High recall remains ~19%, but medium-class performance is restored compared to the pre-SMOTE tuned model.
- This model strikes the best trade-off between accuracy and minority-class detection.

	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



F1-SCORE & ACCURACY COMPARISON

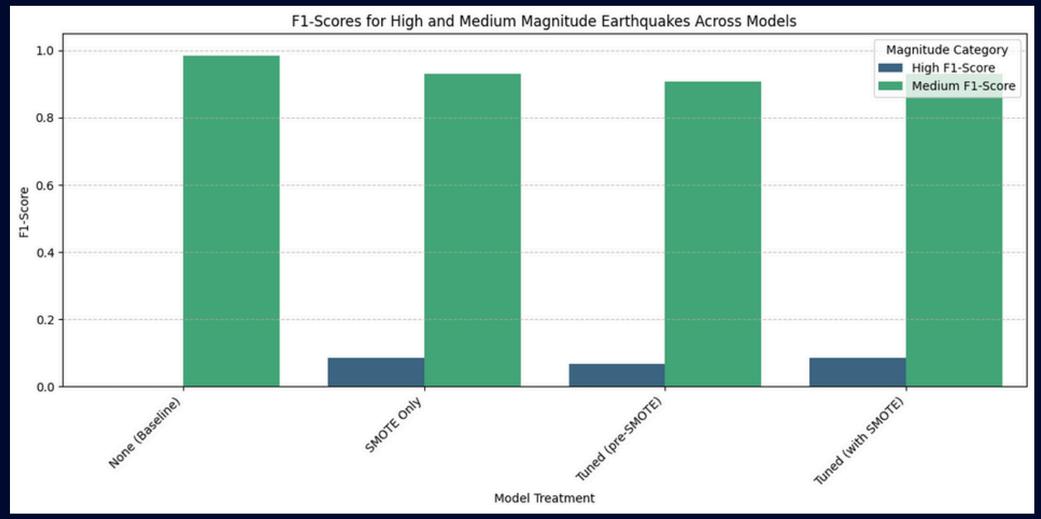


Accuracy:

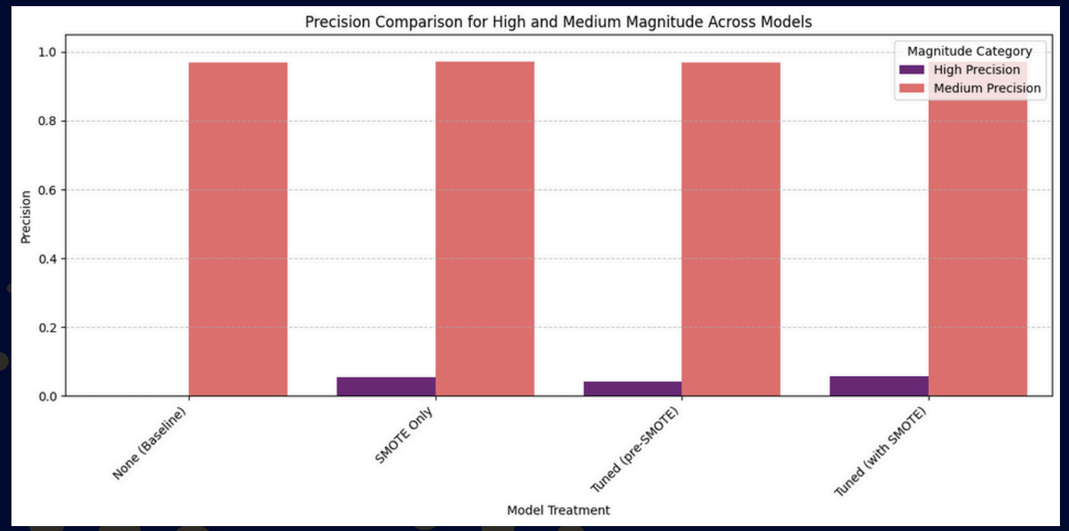
- The baseline model shows the highest accuracy, but this is misleading because it predicts only the majority class.
- All models that attempt to detect high-magnitude earthquakes show lower accuracy, reflecting the trade-off between overall accuracy and minority-class detection.

F1 Score:

- High F1-scores for the medium class across all models indicate stable performance for the dominant class.
- The high-magnitude F1-score remains low across all treatments due to rarity of events, but improves only in SMOTE-based models.
- The tuned SMOTE model offers the best balance, preserving strong medium-class F1 while slightly improving high-class F1.



PRECISION & RECALL COMPARISON

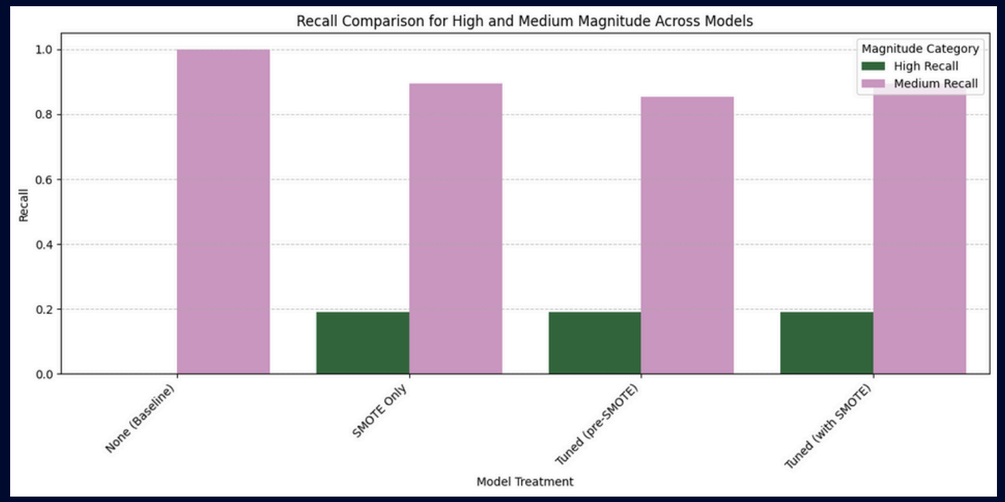


Precision:

- Precision for the high-magnitude class remains low across all models because rare-event detection leads to more false positives.
- SMOTE increases the model's tendency to predict high-magnitude events, which increases sensitivity but lowers precision.
- Medium-class precision stays consistently high across all treatments, showing that the models remain reliable for the dominant class.
- The tuned SMOTE model maintains strong medium precision while slightly improving high-class precision compared to the pre-SMOTE tuned model.

Recall:

- High-magnitude recall improves dramatically from 0% in the baseline model to around 19% when SMOTE is used, showing the model is finally able to detect rare events.
- The pre-SMOTE tuned model and SMOTE-only model achieve similar recall for the high class, but with different levels of stability.
- Medium-class recall decreases slightly in the SMOTE-only and tuned pre-SMOTE models due to the expansion of the high-class decision boundary.
- The tuned SMOTE model restores medium recall while preserving improved high recall, making it the most balanced option.



CONCLUSION

- Earthquake patterns clearly follow global tectonic plate boundaries, especially around the Pacific Ring of Fire.
- Visualizations and heatmaps effectively highlight major seismic hotspots and reveal strong regional clustering.
- K-Means successfully identified four distinct seismic zones that match known geological structures.
- The baseline Random Forest model showed high accuracy but completely failed to detect high-magnitude earthquakes due to severe class imbalance.
- Applying SMOTE and hyperparameter tuning significantly improved the model's ability to identify rare high-magnitude events, offering a more balanced and risk-focused prediction approach.

THANK YOU