INFO 6105 Final Project Summary

Problem

Predicting whether an earthquake will generate a tsunami is critical for minimizing loss of life and enabling early emergency response. Our project aims to experiment with different machine learning models and find the best model that automatically classifies earthquake events as “tsunami-generating” or “non-tsunami” using historical data from the United States Geological Survey (USGS).

Data

The data was collected from the USGS Earthquake API. We filtered the data to include only earthquakes of magnitude 6.0 or higher, since tsunamis are typically associated with strong earthquake events. The dataset covers events occurring between 2013 and 2025 and contains 23 features. During data preprocessing, we selected 11 meaningful features and dropped irrelevant columns. Missing values were imputed with column medians, and several features were engineered to capture spatial and temporal patterns. Finally, all numerical features were scaled to ensure consistent model performance.

Model

We trained several models including Logistic Regression, KNN, Decision Tree, Random Forest, and XGBoost to predict whether an earthquake would generate a tsunami. Tree-based models performed best, with Random Forest and XGBoost showing the strongest F1 scores and ROC-AUC values. We then performed hyperparameter tuning with the Optuna framework on both models to optimize their performances. Tuning significantly improved performance, especially for XGBoost. The tuned XGBoost model achieved the best overall results, with an F1 score of 0.87, a Recall of 0.872 and an Accuracy of 0.91, and was selected as the best model for prediction.

Results

The tuned XGBoost model achieved strong performance on the test set, reaching an overall accuracy of 0.91. It correctly identified 93% of non-tsunami earthquakes and 87% of tsunami-producing earthquakes. Further analysis of the results shows that the prediction errors are low and evenly distributed, which means that the model is not biased. Overall, the model provides stable and accurate predictions, making it a reliable system for supporting early tsunami warnings.

Limitations

Our project has several limitations. First, the dataset is relatively small, with only 1757 events from 2013 to 2025 due to the absence of tsunami records before 2013. Second, the analysis relies solely on features provided by USGS, lacking potentially important features like ocean floor topography and coastal proximity. Finally, the model supports historical analysis only and cannot be used for real-time early warning because live data feeds and operational validation are not available.

Next Steps

For future works, we can expand the dataset with additional global records and incorporate external features to improve accuracy. Integrating real-time data streams may also enable the system to be a live tsunami-warning application. We can also explore more advanced models or ensemble approaches for enhanced model performance and robustness.