INFO 6105 FINAL PROJECT

Earthquake - Tsunami Prediction System

Xinru Zhang - 002560736

Zhan Tang - 002580461

Northeastern University

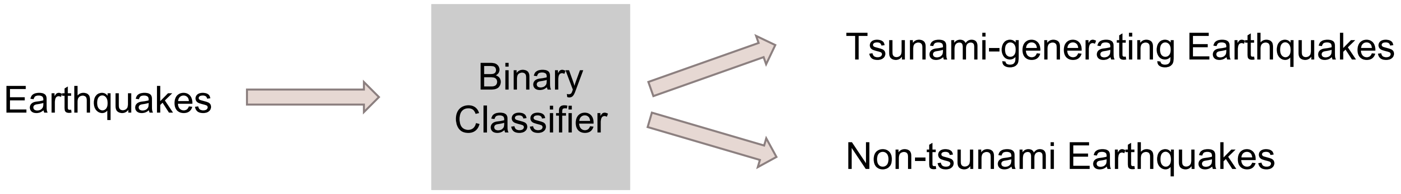
Problem

Why the Earthquake-Tsunami Prediction System?

- Earthquakes are the most common cause of tsunamis
- Tsunami brings huge damage to coastal communities and loss of life
- Predicting whether an earthquake would produce a tsunami is critical for minimizing loss

Our Goal:

Build a system to classify earthquakes to support faster and reliable early tsunami warnings



Earthquakes

Binary Classifier

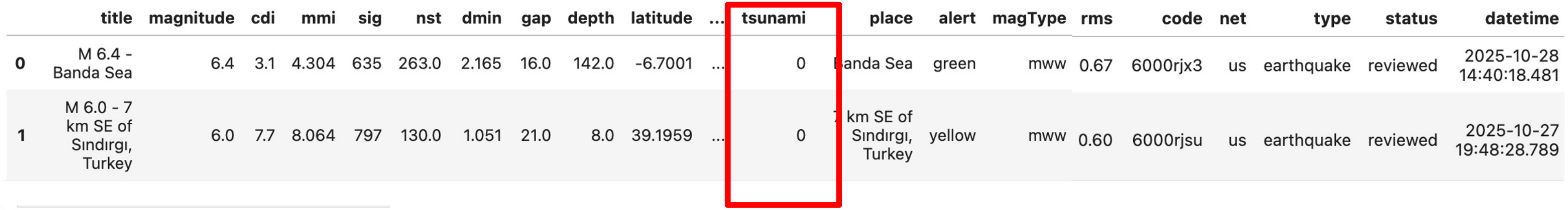
Tsunami-generating Earthquakes

Non-tsunami Earthquakes

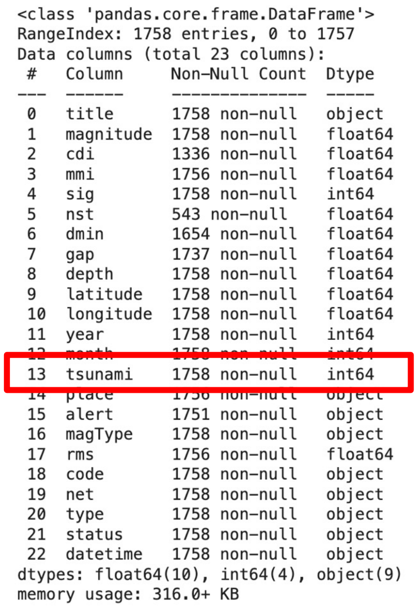
ML Pipeline Overview

	Highlights
1. Data Collection	USGS API (2013-2025, mag≥6)
2. Exploratory Data Analysis	Distributions, Outliers, Correlations
3. Data Processing	Data Cleaning, Imputation, Feature Engineering, Feature Scaling
4. Baseline Model Training	Decision Tree, Random Forest, XGBoost, KNN, Logic Regression
5. Hyperparameter Tuning	Optuna (RF, XGB)
6. Best Model Deployment and Prediction	F1 Score: 0.91, Accuracy: 0.91

Data Overview



	title	magnitude	cdi	mmi	sig	nst	dmin	gap	depth	latitude	...	tsunami	place	alert	magType	rms	code	net	type	status	datetime
0	M 6.4 - Banda Sea	6.4	3.1	4.304	635	263.0	2.165	16.0	142.0	-6.7001	...	0	Banda Sea	green	mww	0.67	6000rjx3	us	earthquake	reviewed	2025-10-28 14:40:18.481
1	M 6.0 - 7 km SE of Sındırgı, Turkey	6.0	7.7	8.064	797	130.0	1.051	21.0	8.0	39.1959	...	0	7 km SE of Sındırgı, Turkey	yellow	mww	0.60	6000rjsu	us	earthquake	reviewed	2025-10-27 19:48:28.789

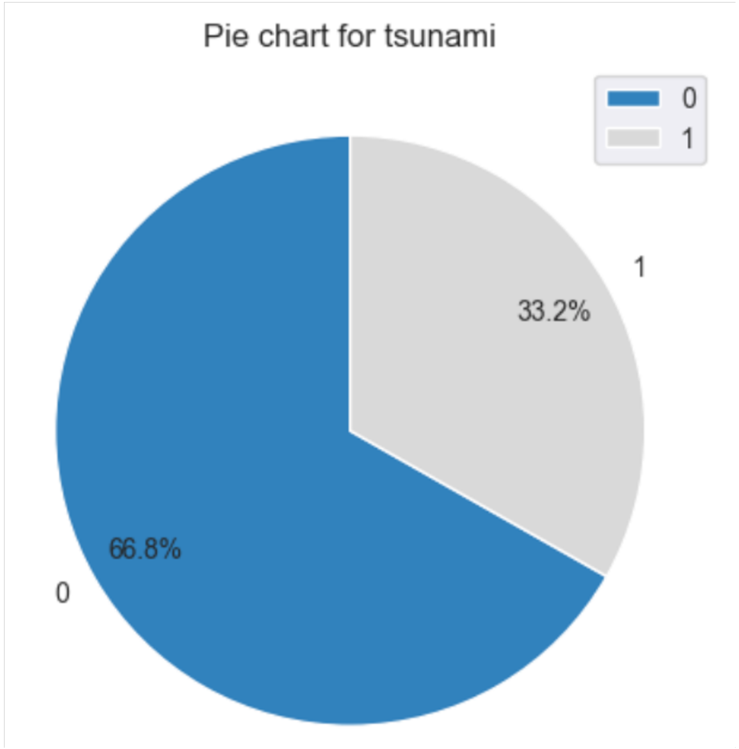


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1758 entries, 0 to 1757
Data columns (total 23 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   title      1758 non-null   object 
 1   magnitude  1758 non-null   float64
 2   cdi        1336 non-null   float64
 3   mmi        1756 non-null   float64
 4   sig        1758 non-null   int64  
 5   nst        543 non-null    float64
 6   dmin       1654 non-null   float64
 7   gap        1737 non-null   float64
 8   depth      1758 non-null   float64
 9   latitude   1758 non-null   float64
 10  longitude  1758 non-null   float64
 11  year       1758 non-null   int64  
 12  month      1758 non-null   int64  
 13  tsunami    1758 non-null   int64  
 14  place      1758 non-null   object 
 15  alert      1751 non-null   object 
 16  magType    1758 non-null   object 
 17  rms        1756 non-null   float64
 18  code       1758 non-null   object 
 19  net        1758 non-null   object 
 20  type       1758 non-null   object 
 21  status     1758 non-null   object 
 22  datetime   1758 non-null   object 
dtypes: float64(10), int64(4), object(9)
memory usage: 316.0+ KB
```

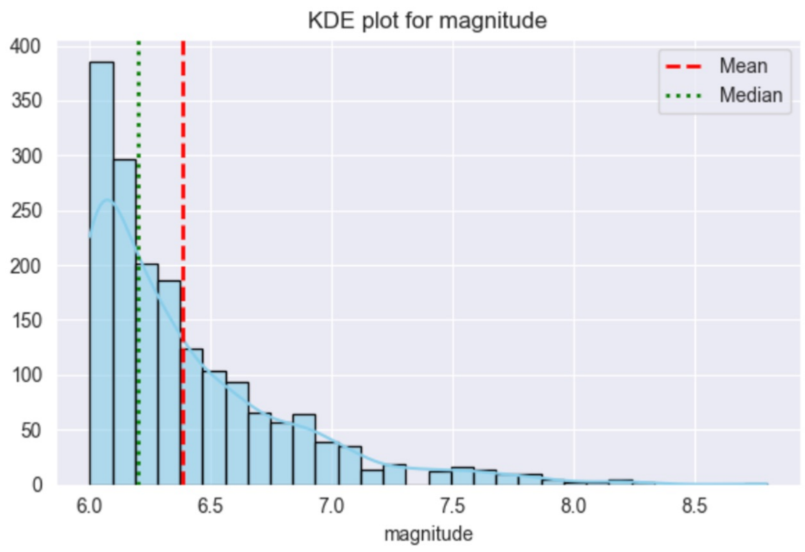
- Data collected from USGS Earthquake API (2013-2025, mag≥6)
- Total 1758 records
- Total 23 features (14 numerical, 9 categorical/textual)
- Target feature: tsunami (binary, 0 stands for non-tsunami events; 1 stands for tsunami-generating events)

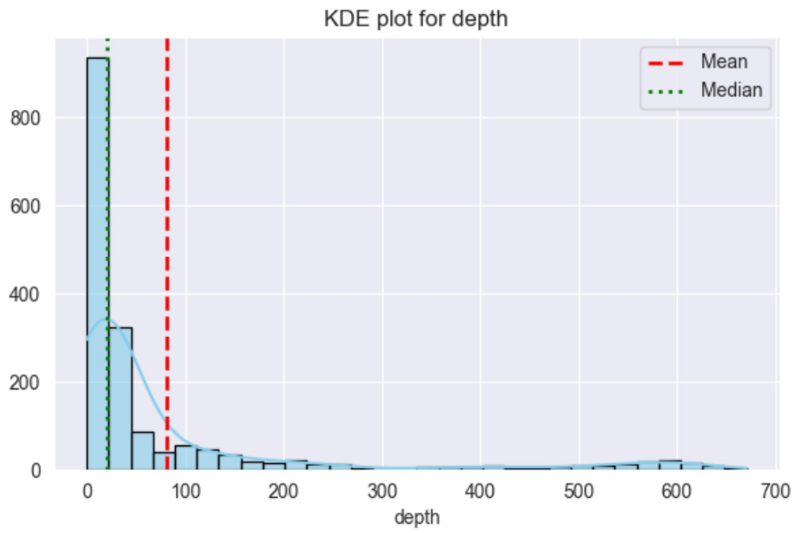
Exploratory Data Analysis Example

- 67% non-tsunami events; 33% tsunami events
- Not very balanced but consistent with natural laws



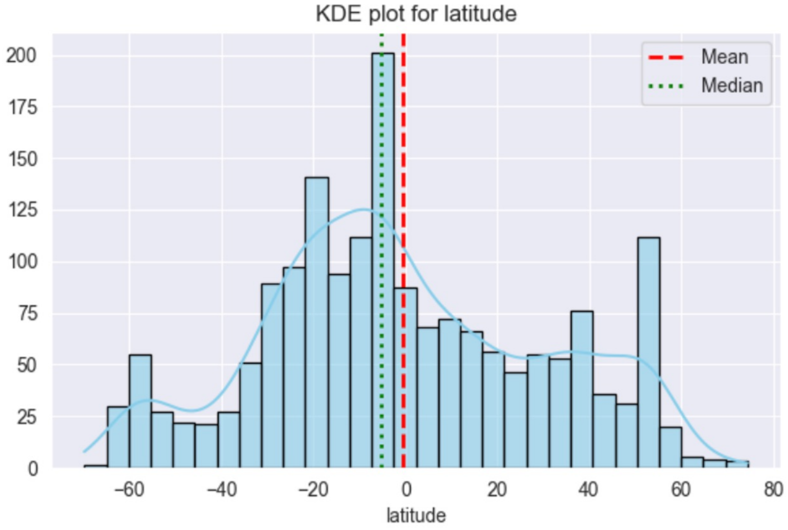
Exploratory Data Analysis Example

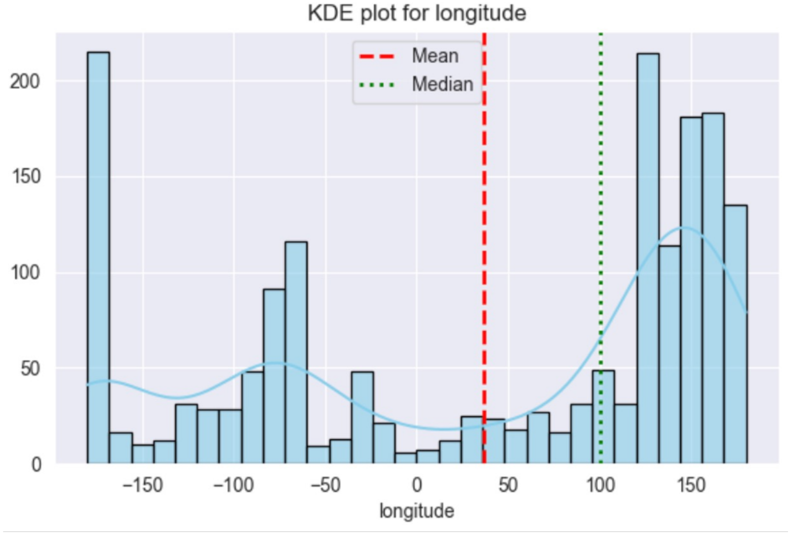




- Magnitude & depth: right-skewed
- Most earthquakes happen at shallow depth with magnitude 6.5 or lower

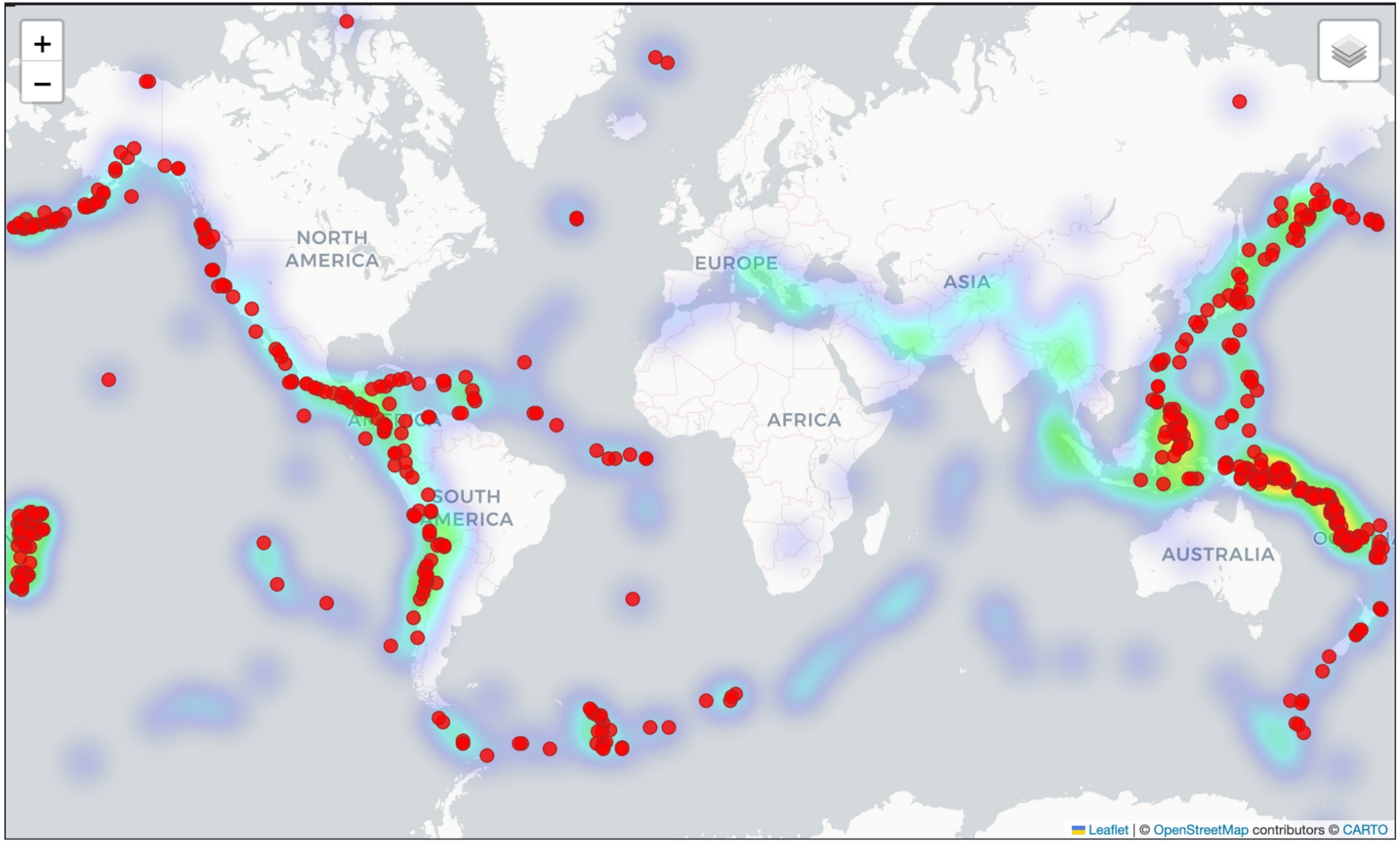
Exploratory Data Analysis Example





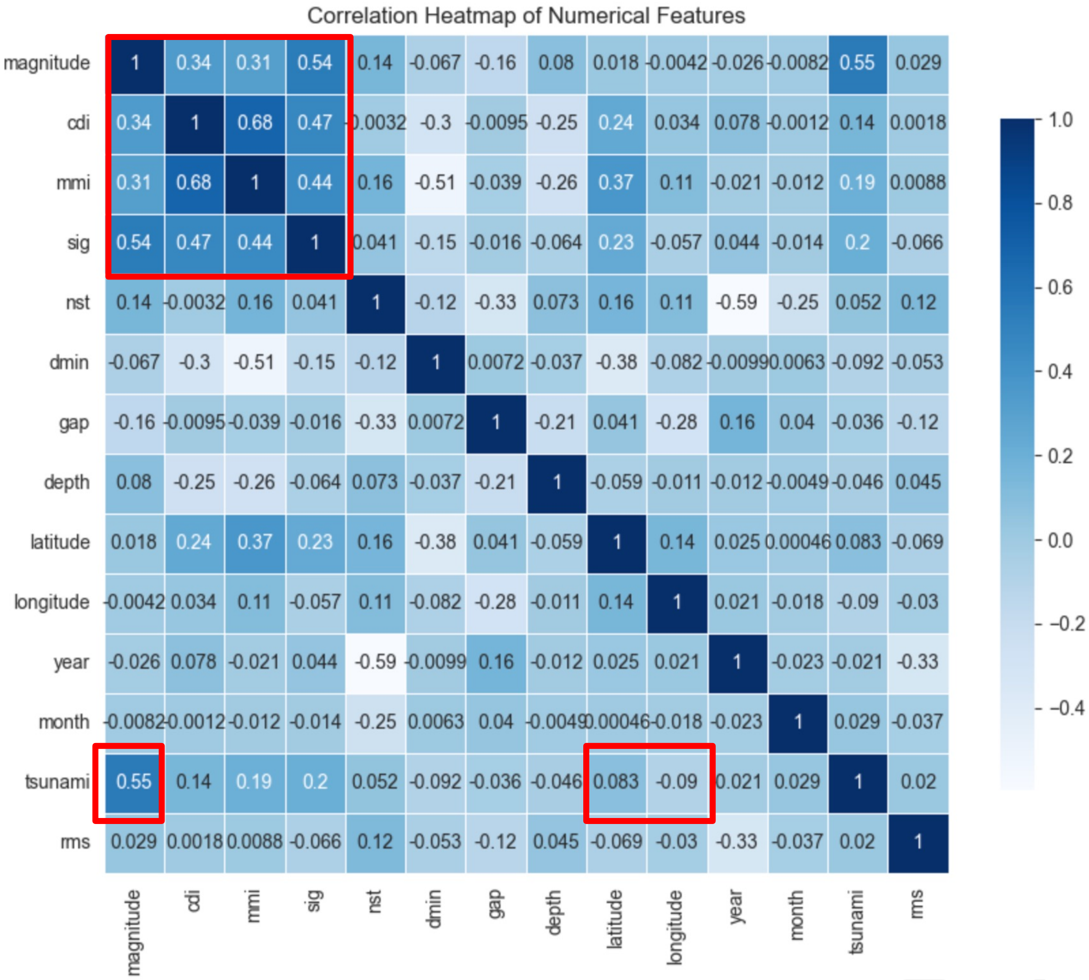
- Latitude: peak around 0 (equator)
- Longitude: peak around -150 & 150
- The Pacific Ring of Fire

Exploratory Data Analysis Example



Exploratory Data Analysis Example

- Magnitude, cdi (reported intensity), mmi (measured intensity), sig (significance score): strongly correlated because all measure earthquake intensity
- Magnitude: strong correlation with tsunami
- Latitude & longitude: weak linear correlation with tsunami; suggest further non-linear relationships



	magnitude	cdi	mmi	sig	nst	dmin	gap	depth	latitude	longitude	year	month	tsunami	rms
magnitude	1	0.34	0.31	0.54	0.14	-0.067	-0.16	0.08	0.018	-0.0042	-0.026	-0.0082	0.55	0.029
cdi	0.34	1	0.68	0.47	-0.0032	-0.3	-0.0095	-0.25	0.24	0.034	0.078	-0.0012	0.14	0.0018
mmi	0.31	0.68	1	0.44	0.16	-0.51	-0.039	-0.26	0.37	0.11	-0.021	-0.012	0.19	0.0088
sig	0.54	0.47	0.44	1	0.041	-0.15	-0.016	-0.064	0.23	-0.057	0.044	-0.014	0.2	-0.066
nst	0.14	-0.0032	0.16	0.041	1	-0.12	-0.33	0.073	0.16	0.11	-0.59	-0.25	0.052	0.12
dmin	-0.067	-0.3	-0.51	-0.15	-0.12	1	0.0072	-0.037	-0.38	-0.082	-0.0099	0.0063	-0.092	-0.053
gap	-0.16	-0.0095	-0.039	-0.016	-0.33	0.0072	1	-0.21	0.041	-0.28	0.16	0.04	-0.036	-0.12
depth	0.08	-0.25	-0.26	-0.064	0.073	-0.037	-0.21	1	-0.059	-0.011	-0.012	-0.0049	-0.046	0.045
latitude	0.018	0.24	0.37	0.23	0.16	-0.38	0.041	-0.059	1	0.14	0.025	0.00046	0.083	-0.069
longitude	-0.0042	0.034	0.11	-0.057	0.11	-0.082	-0.28	-0.011	0.14	1	0.021	-0.018	-0.09	-0.03
year	-0.026	0.078	-0.021	0.044	-0.59	-0.0099	0.16	-0.012	0.025	0.021	1	-0.023	-0.021	-0.33
month	-0.0082	-0.0012	-0.012	-0.014	-0.25	0.0063	0.04	-0.0049	0.00046	-0.018	-0.023	1	0.029	-0.037
tsunami	0.55	0.14	0.19	0.2	0.052	-0.092	-0.036	-0.046	0.083	-0.09	0.021	0.029	1	0.02
rms	0.029	0.0018	0.0088	-0.066	0.12	-0.053	-0.12	0.045	-0.069	-0.03	-0.33	-0.037	0.02	1

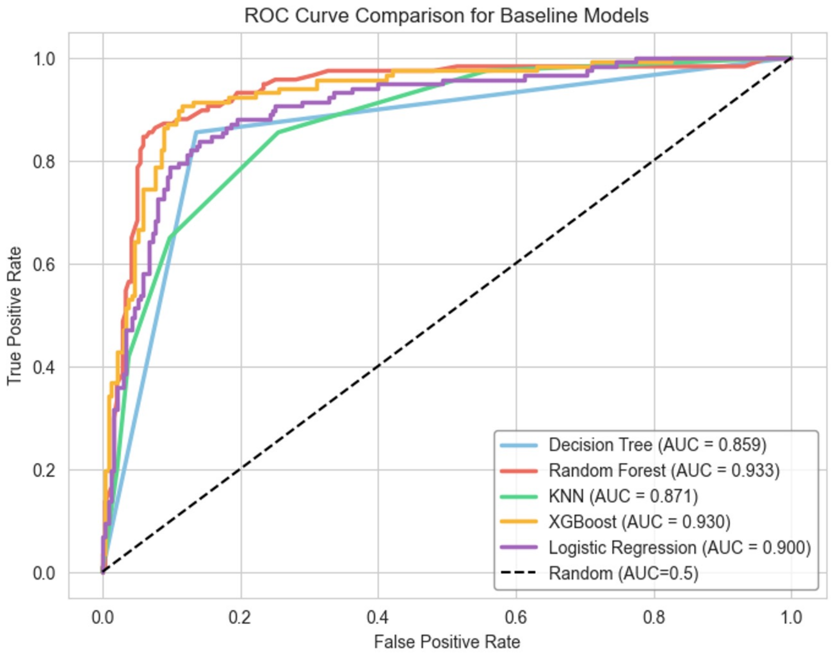
Data Processing

Missing Value Ratio	
nst	69.1%
cdi	24.0%
dmin	5.9%
gap	1.2%
alert	2.4%
rms	0.1%
mmi	0.1%
place	0.1%

	Details
1. Drop irrelevant and redundant rows and columns (23 → 11 features)	- Remove non-earthquake events. - Drop irrelevant columns for analysis: title, code, status, net, type - Drop redundant or similar, duplicated columns: place, datetime - Drop post-event indicators correlated with magnitudes: cdi, mmi, sig, alert - Drop high-missing column: nst (69%)
2. Imputing Missing Values	- Used **median** imputation for quality metrics: dmin, gap, rms
3. Feature Engineering	- Convert Geographic features(latitude and longitude) to Cartesian Coordinates - Cyclical encoding for feature - month. - One-Hot Encoding categorical feature - magType.
4. Feature Scaling	- Feature scaling with StandScaler

Baseline Model
Performance Comparison

Model	Accuracy	Class 1			ROC-AUC
		Precision	Recall	F1	
Random Forest	0.906	0.862	0.855	0.858	0.933
XGBoost	0.884	0.822	0.829	0.826	0.930
Decision Tree	0.861	0.758	0.855	0.803	0.859
Logistic Regression	0.835	0.817	0.650	0.724	0.900
KNN	0.818	0.768	0.650	0.704	0.871



❑ Random Forest and XGBoost Baseline Models are top performers, showing that they captures the most true tsunami events while maintaining excellent discriminative ability.

❑ Decision Tree shows moderate performance with an F1 Score even though it has a high Recall, but its ROC-AUC is notably lower than the other methods, suggesting limited discriminative power.

❑ Logistic Regression and KNN demonstrate weaker performance on Class 1 metrics.

Hyperparameter Tuning - Optuna

	Objectives	Best Params	F1 GAP (Train - Test)
Random Forest	- n_estimators: [100, 400] - max_depth: [5, 15] - min_samples_split: [5, 25] - min_samples_leaf: [3, 12] - max_features: ['sqrt', 'log2']	- n_estimators: 359 - max_depth: 14 - min_samples_split: 7 - min_samples_leaf: 4 - max_features: 'sqrt'	Train F1: 0.9202 Test F1: 0.8644 Gap: 0.0558
XGBoost	- n_estimators: [100, 400] - max_depth: [3, 10] - learning_rate: [0.005, 0.2] - subsample: [0.6, 0.9] - colsample_bytree: [0.6, 0.9] - gamma: [0, 5] - reg_lambda: [0.5, 5.0] - Min_child_weight: [1, 10]	- n_estimators: 295 - max_depth: 10 - learning_rate: 0.017 - subsample: 0.883 - colsample_bytree: 0.843 - gamma: 0.355 - reg_lambda: 4.336 - Min_child_weight: 4	Train F1: 0.9059 Test F1: 0.8681 Gap: 0.0379

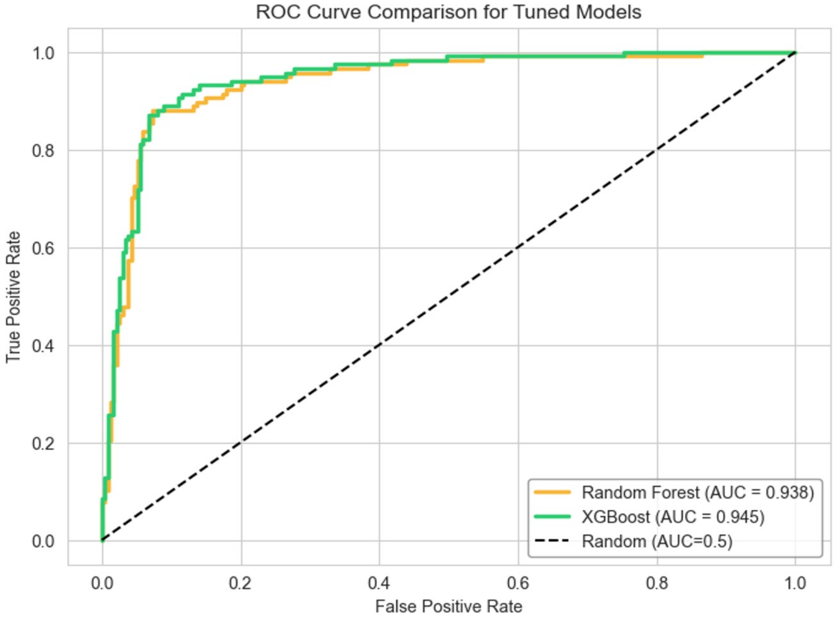
Framework: Optuna

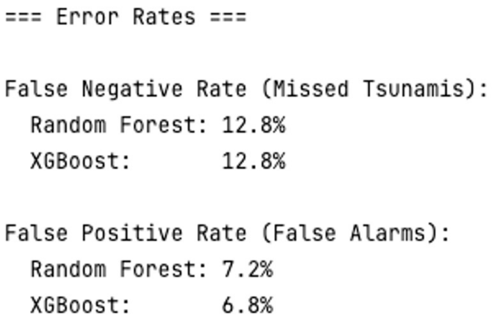
- ❑ Optimization Metric: **F1 Score** (Class 1 - Tsunami detection)
- ❑ Trials: 100 iterations per model
- ❑ Cross-Validation: 5-fold CV

Model	Baseline F1	Tuned F1	Improvement (Δ)	Improvement (%)
XGBoost	0.826	0.868	+0.043	+5.2%
Random Forest	0.858	0.864	+0.006	+0.7%

Model Evaluation

Model	Accuracy	Class 1		
		Precision	Recall	F1
XGBoost	0.912	0.864	0.872	0.868
Random Forest	0.909	0.857	0.872	0.864





```
=== Error Rates ===

False Negative Rate (Missed Tsunamis):
  Random Forest: 12.8%
  XGBoost:       12.8%

False Positive Rate (False Alarms):
  Random Forest: 7.2%
  XGBoost:       6.8%
```

- Both tuned models demonstrate strong performance with minimal differences. XGBoost shows marginally better performance than Random Forest.

- Two models' ROC–AUC scores are higher than 0.93. XGBoost shows slightly better.

- In real-world risk analysis, they have same False Negative Rate of 12.8%. However, for the False Positive Rate, the XGBoost model performed slightly better, with 6.8% compared to 7.2% for random forest.

Results

Test set: 352 records

Category	Precision	Recall	F1 Score	Accuracy
Non-tsunami events	0.94	0.93	0.93	0.91
Tsunami-generating events	0.86	0.87	0.87	

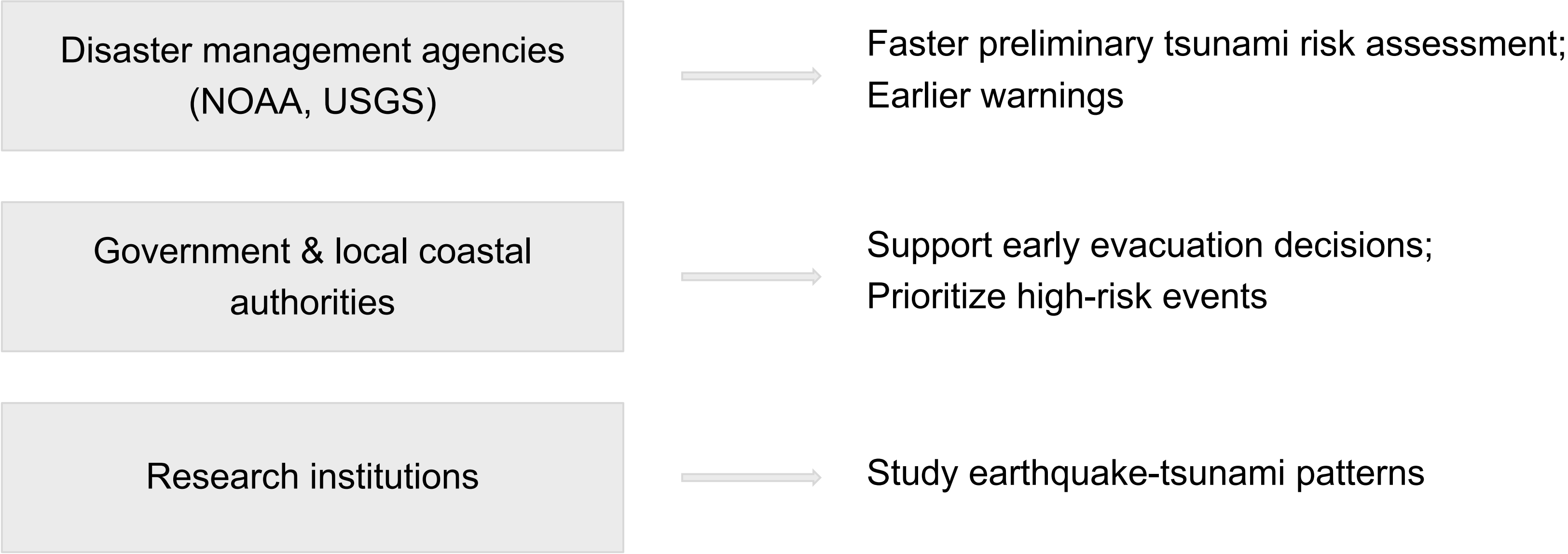
	Actual Tsunami (1)	Actual Non-tsunami (0)
Predicted Tsunami (1)	102 (True Positives)	16 (False Positives)
Predicted Non-tsunami (0)	15 (False Negatives)	219 (True Negatives)

- F1 score — Non-tsunami: 0.93; Tsunami-generating: 0.87
- Overall accuracy: 0.91
- Balanced prediction errors: 16 false positives & 15 false negatives

→ High Performance

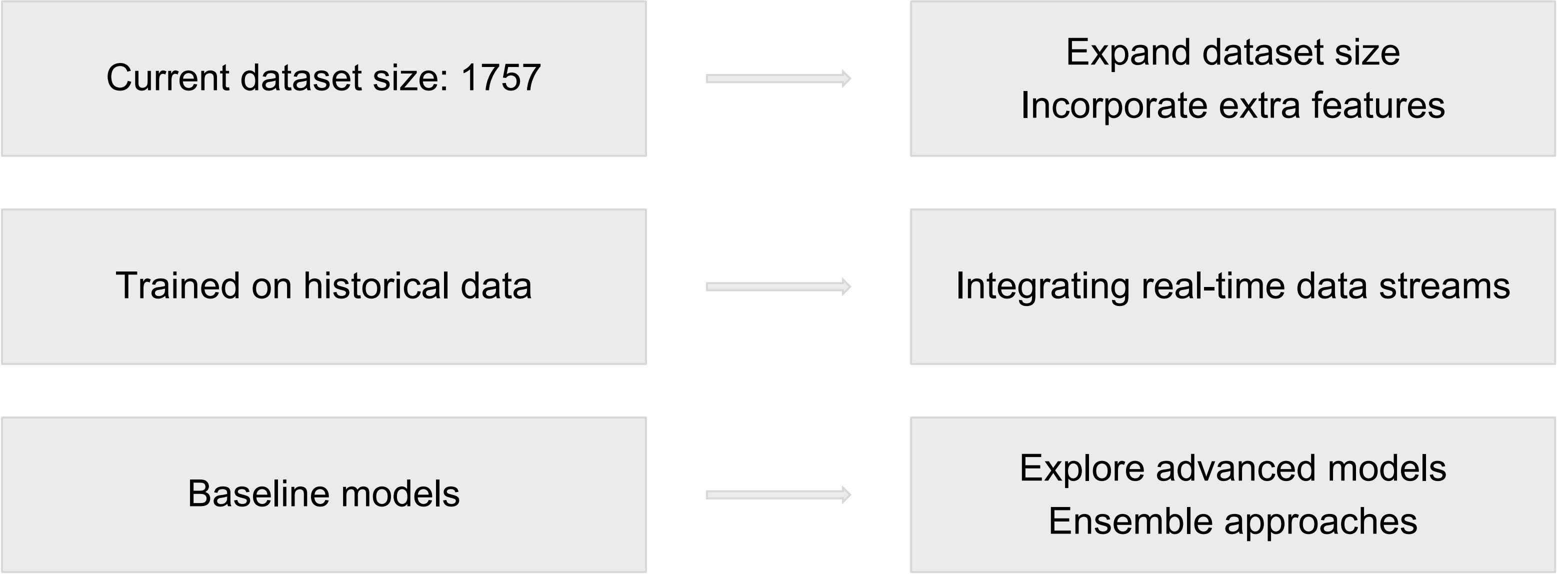
→ Not Biased

Usefulness



Disaster management agencies (NOAA, USGS)	Faster preliminary tsunami risk assessment; Earlier warnings
Government & local coastal authorities	Support early evacuation decisions; Prioritize high-risk events
Research institutions	Study earthquake-tsunami patterns

Future Work

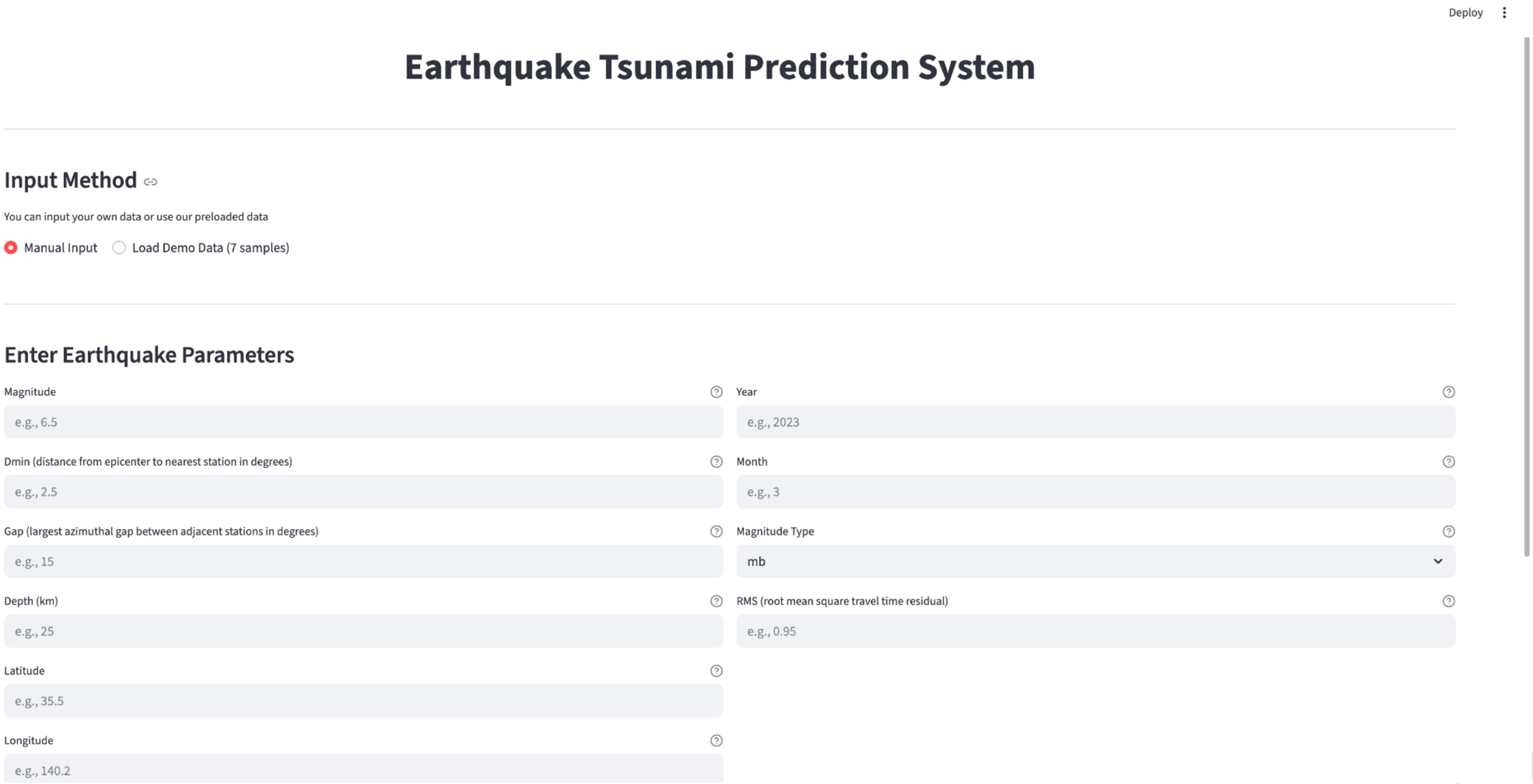


Current dataset size: 1757 → Expand dataset size
Incorporate extra features

Trained on historical data → Integrating real-time data streams

Baseline models → Explore advanced models
Ensemble approaches

Streamlit App Demo



Earthquake Tsunami Prediction System

Input Method

You can input your own data or use our preloaded data

● Manual Input ○ Load Demo Data (7 samples)

Enter Earthquake Parameters

Magnitude: e.g., 6.5

Year: e.g., 2023

Dmin (distance from epicenter to nearest station in degrees): e.g., 2.5

Month: e.g., 3

Gap (largest azimuthal gap between adjacent stations in degrees): e.g., 15

Magnitude Type: mb

Depth (km): e.g., 25

RMS (root mean square travel time residual): e.g., 0.95

Latitude: e.g., 35.5

Longitude: e.g., 140.2