Khadiga Abdelkarim Hamid Idris - AIU22102199
Ridwan Afolabi Muhammad -AIU23102185
Thet Hnin Eain - AIU24032003
Bashar Ibrahem Ahmed Hagag - AIU22102319

ALBUKHARY INTERNATIONAL UNIVERSITY
SCHOOL OF COMPUTING AND INFORMATICS

SEMESTER 2 2024/2025
CCS2213
MACHINE LEARNING GROUP PROJECT
LECTURERS -
PROF ZURINAHNI & DR. UMMI KULTHOM

GLOBAL FRAMEWORK FOR MACHINE-LEARNING-DRIVEN EARTHQUAKE IMPACT ANALYSIS

PROBLEM STATEMENT

In March 2025, a 7.7 magnitude earthquake struck Myanmar, causing massive loss of life and infrastructure. Many countries remain vulnerable due to limited tools for accurate earthquake prediction. This project introduces a machine learning framework in a way to contribute to earthquake mitigations by analyzing risk zones and making classification prediction based on seismic events for better disaster planning

OBJECTIVES

- Build ML model to predict earthquake severity
- Forecast secondary impacts (e.g., aftershocks, landslides)
- Strengthen early warning systems for faster response
- Create a scalable, adaptable framework for global use

UNSUPERVISED LEARNING

EXPERIMENT

▶ Clustering Methodology (USV)
- Data Preprocessing:
  - Handled missing values using the median strategy.
  - Standardized numerical features for fair clustering.
  - Encoded categorical features (magType, place) into numeric form.

▶ Choosing K:
- Used the Elbow Method to determine the optimal number of clusters.
- Selected K = 3 as the most suitable choice.

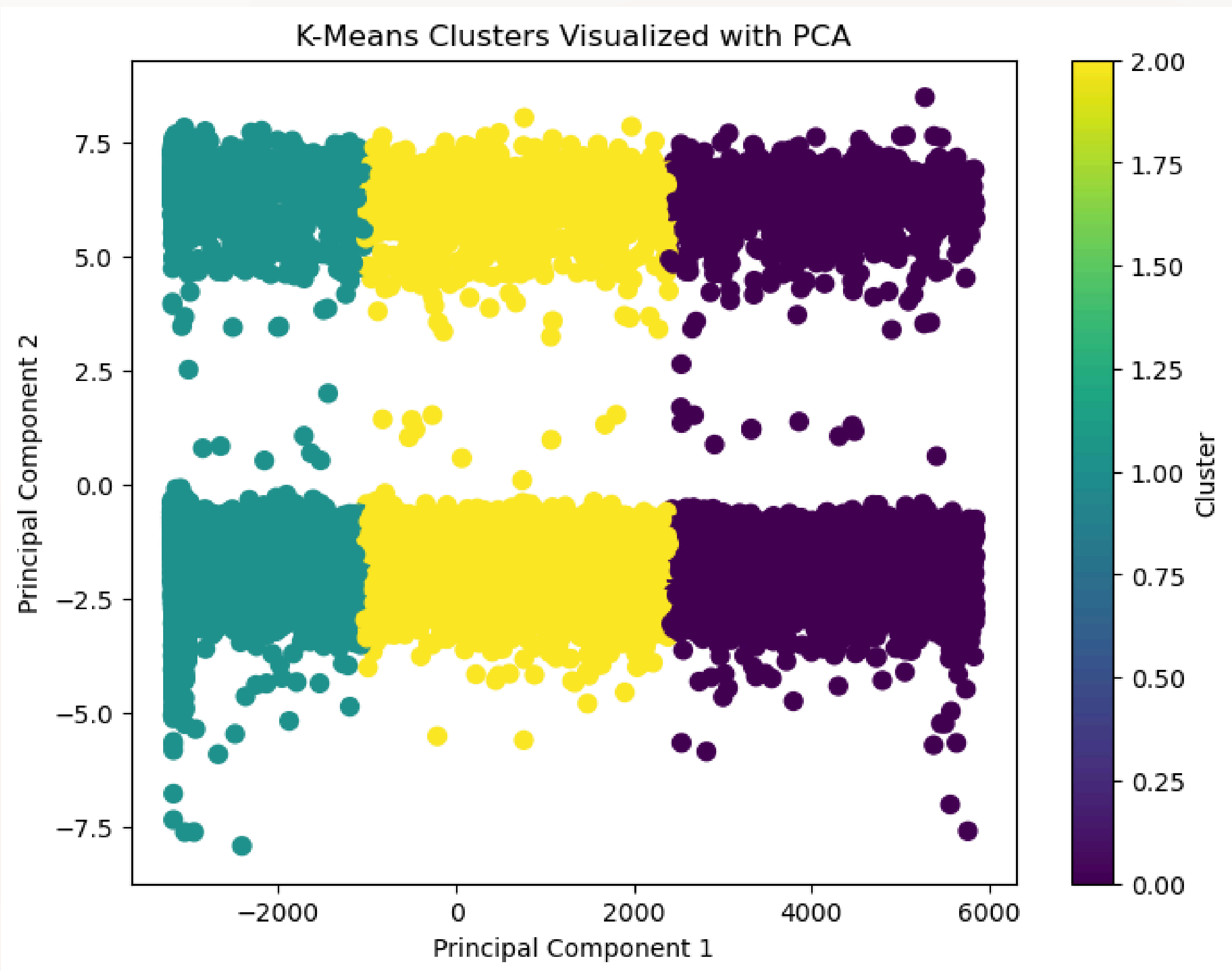
▶ K-Means Clustering:
- Classified earthquake data into three risk levels:
  - High-risk zone
  - Medium-risk zone
  - Low-risk zone

▶ Dimensionality Reduction:
- Applied PCA for two-dimensional visualization.
- Helped in clearly observing cluster separation.

RESULT

▶ Model Evaluation:
- Achieved a Silhouette Score of 0.67, indicating good cluster quality.

▶ Cluster Interpretation:
- Cluster 0: Higher average magnitudes → High-risk
- Cluster 1: Moderate magnitudes → Medium-risk
- Cluster 2: Lower magnitudes → Low-risk



SUPERVISED LEARNING

EXPERIMENT

▶ Data Preprocessing

- **Target Creation** - Low (0–3.9), Medium (4.0–5.9), High (≥ 6.0)
- **Column Reduction** - Dropped IDs, URLs, timestamps, text fields & any columns leaking magnitude (e.g. alert, tsunami_*, sig)
- **Missing-Value Handling** - Imputed categorical (e.g. country, city) with **"Unknown"**, and Imputed numeric medians where needed
- **Encoding** - One-Hot: tsunami → tsunami_0/1, Ordinal: alert (green < yellow < orange < red), Label: type, magType, continent, country, subnational, timezone
- **Scaling** - Standardized numeric features (felt, cdi, mmi, nst, dmin, rms, gap, depth, latitude, longitude, sig) **with Z-score**

▶ Feature/Target Split & Sampling

- X & y Definition
- Train/Test Split - **70 / 30 stratified** split (random_state=42) to preserve class ratios
- Class Balancing - **Applied SMOTE on training set only** to oversample minority ("Low") class

RESULT

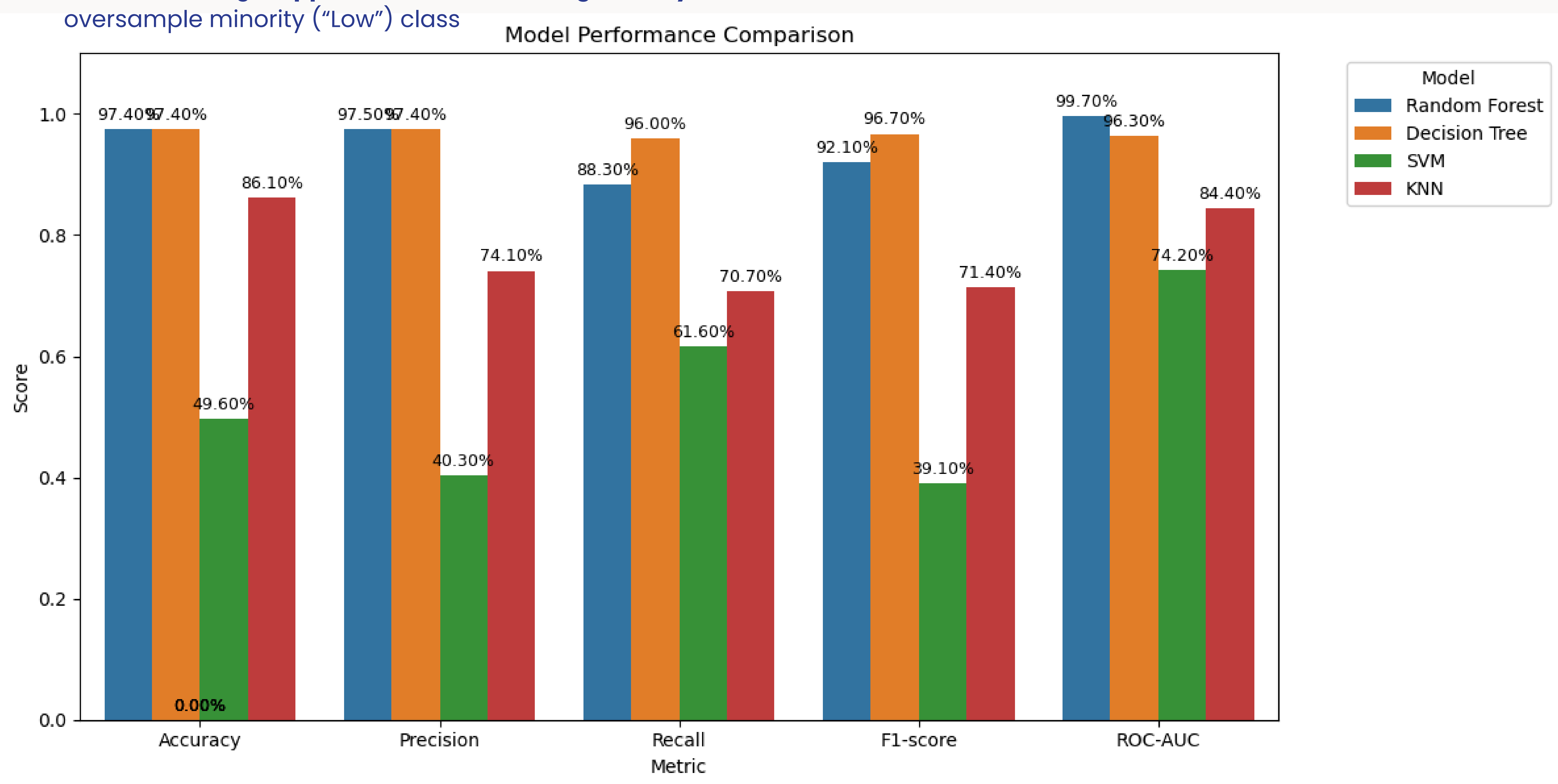
▶ Models Implemented

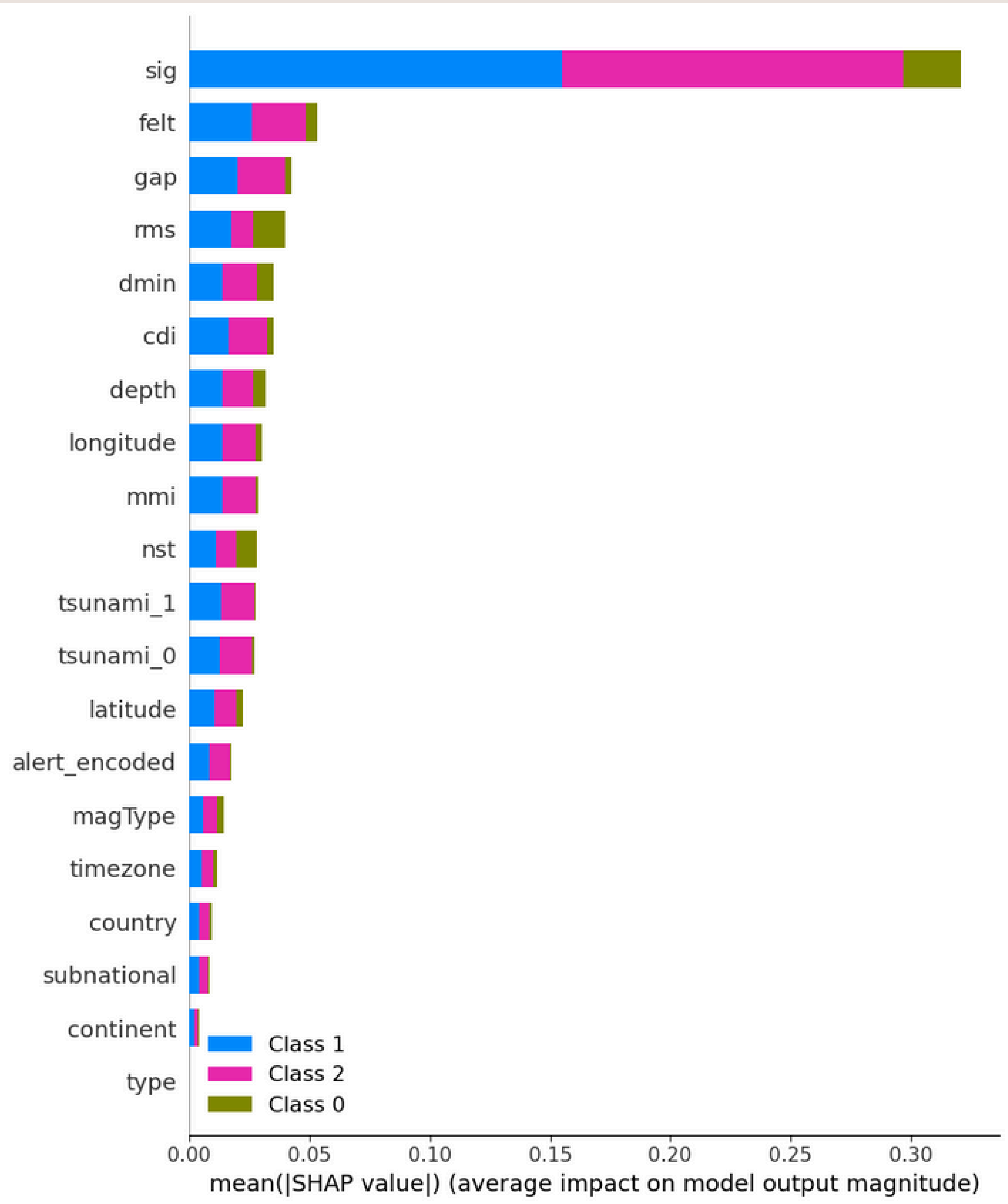
- **Decision Tree** (balanced weights)
- **Random Forest** (ensemble, feature importance)
- **KNN** (on PCA-reduced data)
- **SVM** (RBF kernel + PCA + hyperparameter tuning)
-

▶ Evaluation Metrics & Results Comparison

- ·Overall Correctness – **Accuracy**, High-Class Reliability – **Precision**, Low-Class Sensitivity – **Recall**, Balanced Detection – **F1-Score**, Magnitude Separability – **ROC-AUC**, Class-Aware Balance – **Weighted F1-Score**

This concise pipeline, from raw earthquake records to balanced, encoded features and multiclass classifiers demonstrates our approach to predicting Low/Medium/High seismic events.





FEATURE IMPORTANCE INSIGHTS

- Human Exposure (felt, cdi) - Action: Prioritize **early-warning** in densely populated areas
- Ground Shaking (mmi, depth) - Action: Enforce **stricter building codes** where MMI ≥ VII
- Catastrophic Events (tsunami, sig) - Action: **Enhance coastal preparedness in high-sig regions**
- Model Optimization - Exclude low-impact predictors (alert_encoded, magType) to reduce noise, **Focus on physical over political features,** risk driven by geophysics, not boundaries

CONCLUSION

This project successfully met its objectives through both learning approaches:

- **Unsupervised Learning** grouped global earthquakes into three clear risk zones **(High, Medium, Low)** using K-Means, validated by PCA and a strong Silhouette Score **(0.67)**.
- **Supervised Learning** identified **Random Forest** as the most accurate model **(97.4%)**, enabling precise severity classification based on seismic and impact features.

Real-World Impact:
The framework supports **early-warning systems**, **guides urban planning**, and **improves disaster response,** especially in vulnerable regions. It's scalable, interpretable, and ready for future deployment using real-time data.