EARTHQUAKE DAMAGE PREDICTION

1.INTRODUCTION:

Earthquakes can cause severe destruction to buildings depending on their structural strength, materials used, and geographic characteristics.
The goal of this project is to analyze building-related features and predict the damage_grade (1 = low, 2 = medium, 3 = high).

This project focuses on:

- Performing complete Exploratory Data Analysis (EDA)
- Understanding structural and geographic patterns
- Developing a predictive machine learning model
- Identifying the most important features influencing damage
- Providing meaningful insights for structural safety improvement

2. PROJECT WORKFLOW OVERVIEW :

This section outlines the step-by-step flow of the project.

2.1 Importing Dependencies

The project uses a set of Python libraries for:

- Pandas: Data loading and manipulation

- NumPy: Numerical operations
- Matplotlib & Seaborn: Visualizations
- Scikit-Learn: Data preprocessing, encoding, scaling, and model building

These tools together create the full analysis workflow.

2.2 Loading the Dataset

Two CSV files are combined:

- train_values.csv → Building-level features
- train_labels.csv → Target variable (damage_grade)

Both files are merged using ***building_id***.
Total dataset size: 260,601 rows and 39 features.

2.3 Understanding the Dataset

Geographic Levels

- geo_level_1_id
- geo_level_2_id
- geo_level_3_id

These indicate hierarchical area segmentation.

Structural Features

- age
- count_floors_pre_eq
- area_percentage
- height_percentage
- count_families

Material Indicators

These flags show which material is used:

- mud-mortar
- adobe
- cement-mortar brick
- timber
- bamboo
- reinforced concrete (engineered / non-engineered)

Usage & Configuration

- land_surface_condition
- foundation_type
- ground/other floor types
- roof_type
- plan_configuration
- legal_ownership_status
- secondary uses (school, hotel, rental, etc.)

Target Variable

damage_grade (1, 2, 3)

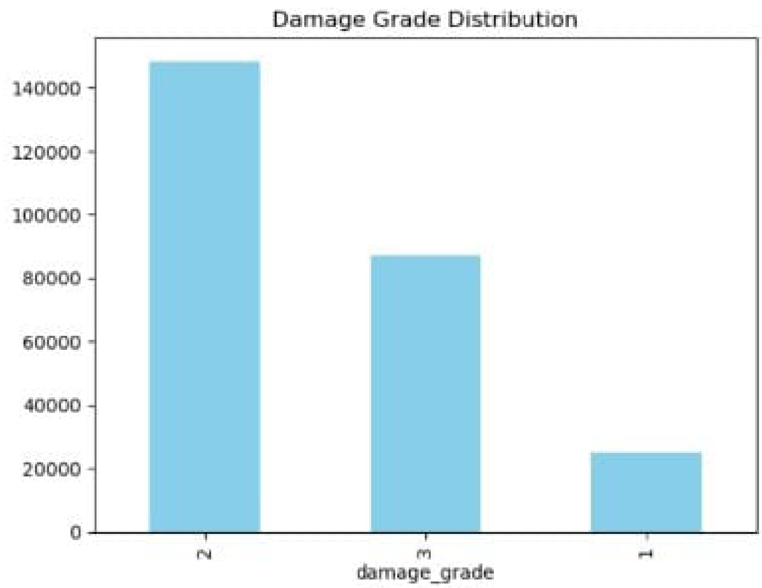
3. EXPLORATORY DATA ANALYSIS (EDA)

3.1 Missing Value Analysis

- No missing values found in the dataset
- All columns are complete and ready for modeling

3.2 Target Variable Distribution

- Damage Grade 2 occurs most frequently
- Grade 1 and 3 have moderate representation
- Indicates slight imbalance, handled later using Balanced Accuracy



3.3Numerical Feature Insights

Key observations:

- Age: Older buildings show more severe damage
- Height Percentage: Taller structures show higher vulnerability
- Area Percentage: Very small base area correlates with weaker stability

3.4 Material Usage Insights

Frequency analysis shows many buildings are constructed with:

- Mud mortar
- Adobe
- Stone
- Unengineered materials

These materials often correspond to higher damage levels.

3.5 Correlation Analysis (Numeric Only)

Key findings:

- Strong relationship between the different geo_level IDs
- Moderate correlation involving height, area, and age
- These features influence damage_grade significantly

4. DATA PREPROCESSING

A structured preprocessing pipeline was used:

4.1 Numeric Data

- Median imputation
- Standard scaling

4.2 Categorical Data

- Frequent imputation
- OneHotEncoding

4.3 Model Input Preparation

- Combined using ColumnTransformer
- 80% Training & 20% Testing split with stratification

This ensures every feature is cleaned, encoded, and scaled correctly.

5. MODEL DEVELOPMENT

5.1 Selected Model: RandomForestClassifier

Reasons for selecting this model:

- **Works well with mixed numerical + categorical data**
- **Captures complex non-linear relationships**

5.2 Model Performance Summary

The model achieved strong results:

Metric	Score
Accuracy	71.8%
Balanced Accuracy	63.5%
Cohen’s Kappa	0. 468

Interpretation

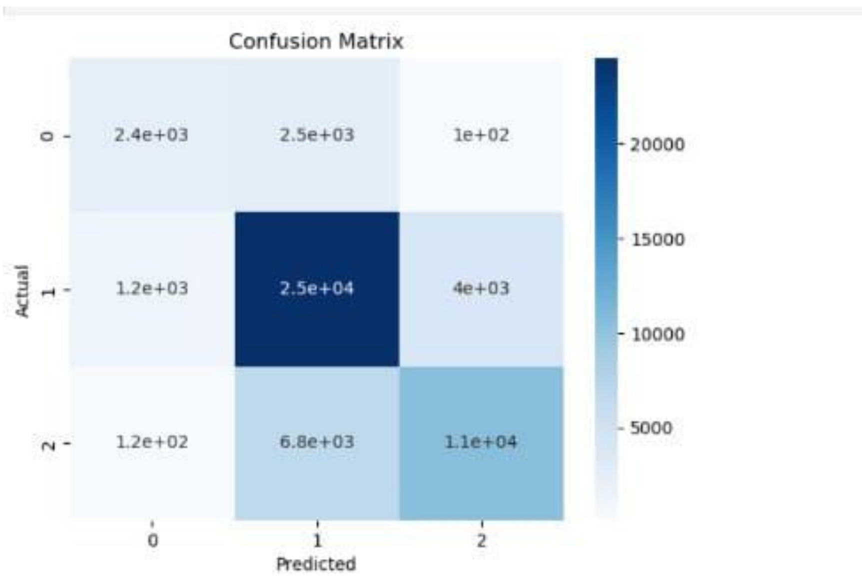
- **Model performs best on damage_grade 2**
- **Grades 1 and 3 also predicted moderately well**
- **Kappa shows the model handles ordinal nature reasonably**

5.3 Confusion Matrix Overview

The confusion matrix shows:

- **Grade 1 ↔ 2 misclassifications**
- **Grade 2 ↔ 3 misclassifications**

These are expected because the categories represent increasing levels of severity.



6. FEATURE IMPORTANCE ANALYSIS

Top contributors influencing damage:

1. geo_level_3_id
2. geo_level_2_id
3. geo_level_1_id
4. age
5. area_percentage
6. height_percentage
7. count_families
8. count_floors_pre_eq
9. foundation_type
10. material flags (mud-mortar, adobe, timber) 7. RECOMMENDATIONS BASED ON ANALYSIS

Interpretation

- Geographic region plays a major role
- Older buildings are highly vulnerable
- Weak materials increase failure probability
- Height and load-related features impact structural behaviour

7. RECOMMENDATIONS BASED ON ANALYSIS

✓ Retrofit older buildings
✓ Promote engineered reinforced materials
✓ Avoid weak construction materials
✓ Strengthen foundations
✓ Limit height in risk-prone zones
✓ Apply region-based safety standards
✓ Ensure public buildings follow strong engineering practices

8.CHALLENGES & SOLUTIONS

Challenge	Solution
Mixed numeric + categorical features	Used unified preprocessing pipeline
Slightly imbalanced classes	Used Balanced Accuracy metric
Ordinal target variables	Evaluate using Cohen's Kappa
Correlation errors	Performed numeric -only heatmap

9. CONCLUSION

This project successfully:

- **Completed detailed EDA**
- **Built a strong predictive model**
- **Identified the most influential building factors**
- **Provided actionable recommendations**

The project is clean, structured, and aligned with professional reporting standards.