**PRCP-1015 – EquakeDamagePred: Earthquake Damage Prediction Project**

**Dataset Source:**

The dataset is sourced from a building damage assessment survey conducted after the 2015 Gorkha Earthquake in Nepal. It includes structural and geotechnical information about buildings and their corresponding damage grades. The dataset is widely used for multi-class classification tasks to predict the level of damage sustained by buildings during an earthquake.

**1. Introduction**

This project aims to build a predictive model to assess building damage levels in the aftermath of an earthquake using structural, material, and geographical data. The task is to classify buildings into three categories based on the damage they sustained:

* **Grade 1:** Low Damage
* **Grade 2:** Medium Damage
* **Grade 3:** High Damage

Accurate predictions can help governments, NGOs, and emergency services prioritize relief and retrofitting efforts, thereby saving lives and resources.

**2. Dataset Overview**

* **File Loaded:** train.csv (Features) + train\_labels.csv (Target)
* **Each row:** Represents a building affected by the 2015 earthquake
* **Features Categories:**
  + **Structural Attributes:** Building age, foundation type, roof and ground floor materials
  + **Geotechnical Info:** Soil type, ground floor type, area, height
  + **Ownership & Usage:** Legal ownership, building use type
  + **Location & Geo Data:** Region, district, seismic zone, geo-level codes
  + **Target Variable:**
    - damage\_grade: (1 = Low, 2 = Medium, 3 = High)

train\_values = pd.read\_csv('train\_values.csv')

train\_labels = pd.read\_csv('train\_labels.csv')

**3. Exploratory Data Analysis (EDA)**

**Initial Checks:**

* Inspected dataset dimensions, types, and joined labels:
* data = pd.merge(train\_values, train\_labels, on='building\_id')
* data.info()
* data.isnull().sum()
* No missing values found, but many categorical variables needed encoding.

**Distribution & Insights:**

* **Damage Grade Distribution:**
  + Imbalanced: Grade 2 was most common (~54%), Grade 3 (30%), Grade 1 (16%)
* **Key Insights from Visualizations:**
  + **Building Age:** Older buildings are more prone to severe damage.
  + **Material Quality:** Buildings with mud mortar and thatch roofs show higher damage grades.
  + **Building Height:** Taller buildings (with more floors) were more vulnerable.
  + **Foundation Type & Soil:** Certain foundation-soil combinations led to higher damage.

**Visual Tools Used:**

* Count plots and histograms for categorical distributions
* Boxplots for numerical feature vs damage grade
* Correlation analysis of ordinal features

**4. Feature Engineering**

**Key Steps:**

* **Label Encoding:** For categorical variables (e.g., material types, region codes)
* **One-Hot Encoding:** For low-cardinality categorical features (e.g., foundation type)
* **Feature Construction:**
  + Combined geo-level codes for region-based clustering
  + Created feature for building area per floor
* **Scaling:** Standardized numerical features like building age and area using StandardScaler
* **Class Imbalance Handling:** Used **SMOTE** in model training

**Important Features Identified:**

* age, area\_percentage, height\_percentage
* foundation\_type, roof\_type, ground\_floor\_type
* geo\_level\_1\_id, land\_surface\_condition

**5. Model Comparison**

Several classification algorithms were evaluated to predict damage\_grade.

**Evaluation Metrics:**

* **Accuracy:** General correctness
* **Precision, Recall, F1-Score (per class):** To handle class imbalance
* **Macro F1-Score:** Weighted performance across all damage classes
* **Confusion Matrix & Classification Reports**
* **Cross-validation:** 5-fold for model robustness

| **Model** | **Accuracy** | **Macro F1** | **Precision** | **Recall** |
| --- | --- | --- | --- | --- |
| Random Forest | 0.78 | 0.78 | 0.79 | 0.78 |
| XGBoost | 0.80 | 0.80 | 0.80 | 0.80 |
| Logistic Regression | 0.58 | 0.56 | 0.58 | 0.57 |

**Best Model Summary**

* **XGBoost** performed best with balanced class-wise performance and macro F1-score.
* **Random Forest** provided robust predictions but was slightly prone to overfitting.
* **Logistic Regression**, though interpretable, underperformed on non-linear patterns.

**Recommendation:**

* **XGBoost** are ideal for production use after hyperparameter tuning.
* **For model explainability**, use XGBoost with **SHAP** for local and global interpretation of damage drivers.

**6. Challenges Faced**

**Class Imbalance:**

* Grade 2 was overrepresented, requiring oversampling and class weighting strategies.

**Categorical Complexity:**

* Many categorical features (e.g., region, material type) with high cardinality demanded careful encoding.

**Feature Correlation:**

* Multicollinearity observed in some geo-level features; addressed via PCA and domain filtering.

**Model Generalization:**

* Risk of overfitting in tree-based models managed using pruning, cross-validation, and regularization.

**Domain Understanding:**

* Interpreting features like geo\_level\_1\_id, land\_surface\_condition, and foundation\_type needed contextual knowledge of Nepal’s geography and construction standards.

**Conclusion & Next Steps**

* **XGBoost** with engineered features delivers high accuracy and reliability in predicting earthquake damage levels.
* **Future Enhancements:**
  + Incorporate satellite or image-based features for richer context
  + Geo-spatial clustering to improve predictions at the community level
  + Use of ensemble stacking for performance boost
* **Deployment Strategy:**
  + Develop a web dashboard for real-time input and prediction
  + Integrate into disaster response planning systems for early prioritization