Modeling Earthquake Damage

MAY 16

PRESENTED BY

BHAVYA JAIN, ADITYA MASUTEY & VANSH KOHLI

PROBLEM?

Approximately 20,000 earthquakes occur worldwide each year. While many are too small to be felt, around 100,000 are felt, and about 100 cause damage. (USGS)

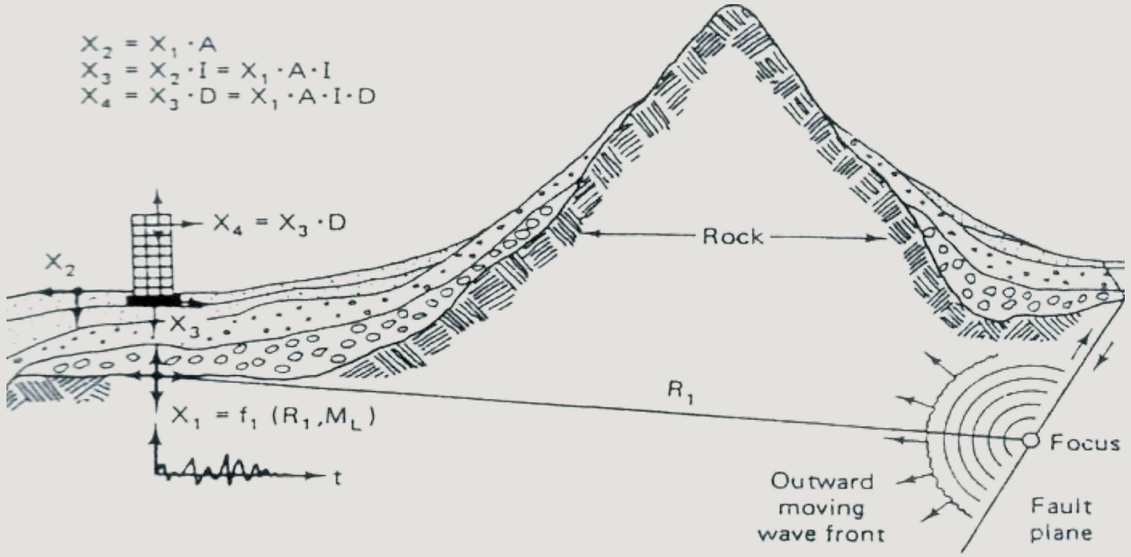
TODAY, WE TURN OUR FOCUS TO THE DEVASTATING EARTHQUAKE THAT STRUCK GORKHA, NEPAL IN 2015. A TRAGEDY THAT SHOOK THE NATION TO ITS CORE AND LEFT LASTING SCARS ON COUNTLESS LIVES.

8,896 dead,
198 missing,
&
22,302 injured - millions left without a home!

Chhetri, M. B. P. (2018). Aftermath of Gorkha–Nepal Earthquake 2015: Lessons Learnt. Oceanography & Fisheries Open Access Journal, 8(2), Article 555735. https://doi.org/10.19080/OFOAJ.2018.08.555735

Problem we are trying to solve

We aim to develop a predictive model to estimate the damage grade of buildings affected by an earthquake. Accurately predicting damage levels is crucial for efficient resource allocation, rapid disaster response, and post-earthquake recovery efforts.



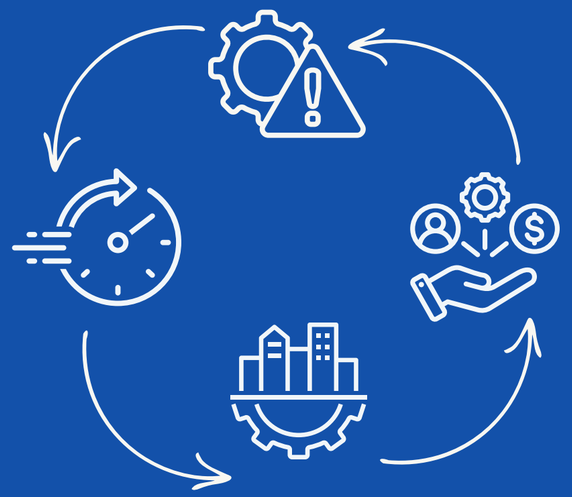
Why?

- Earthquakes cause massive destruction, leading to loss of life, injuries, and loss of homes.
- Traditional methods of seismic design, while effective to a degree often rely on static models and empirical data that may not fully capture the complexities of modern construction and the dynamic nature of earthquakes.
- A data-driven approach can make the process faster, more accurate, and more efficient—helping people get the support they need quickly.

Kolhe, A. S., & Rathi, V. R. (2024). AI-powered earthquake resilience: Predictive modeling and design optimization for seismic-resistant structures. Nanotechnology Perceptions, 20(6), 4099–4113. http://www.nano-ntp.com/

Potential Applications

- Disaster Response: Quickly identify high-risk areas for aid and rescue.
- Infrastructure Planning: Support earthquake-resistant building designs.
- Insurance & Risk Assessment: Help insurers assess risks and estimate damages.

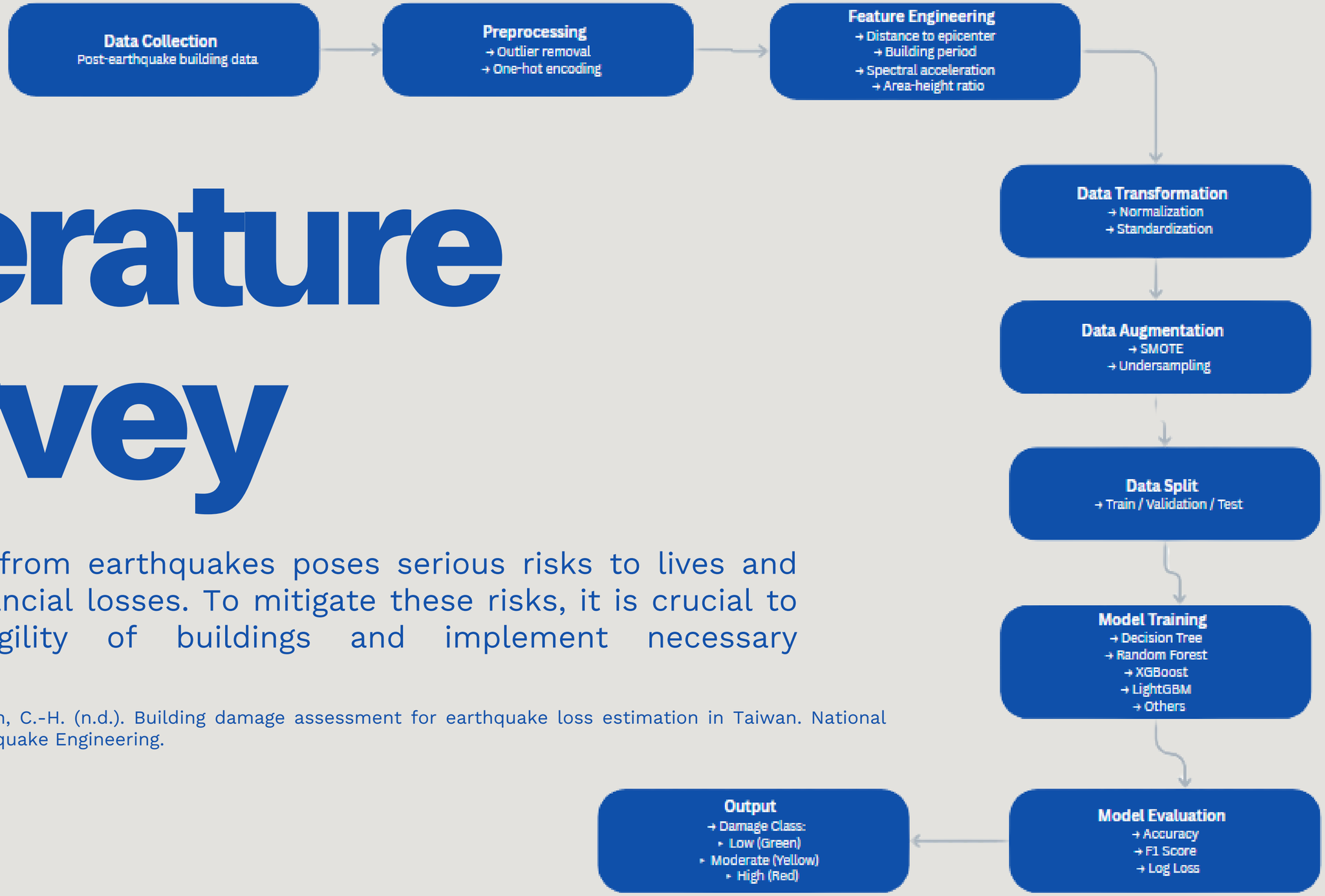


Impact of the Solution

- Faster rebuilding with data-driven decisions.
- Lower financial losses through better planning.
- Safer buildings by identifying weaknesses.
- Improved disaster preparedness for future earthquakes.

Yavas, C. E., Chen, L., Kadlec, C., & Ji, Y. (2024). Improving earthquake prediction accuracy in Los Angeles with machine learning. Scientific Reports, 14, 24440. https://doi.org/10.1038/s41598-024-76483-x

Literature Survey



Building damage from earthquakes poses serious risks to lives and causes major financial losses. To mitigate these risks, it is crucial to assess the fragility of buildings and implement necessary precautions.

Yeh, C.-H., Jean, W.-Y., & Loh, C.-H. (n.d.). Building damage assessment for earthquake loss estimation in Taiwan. National Center for Research on Earthquake Engineering.

Lit Review 1

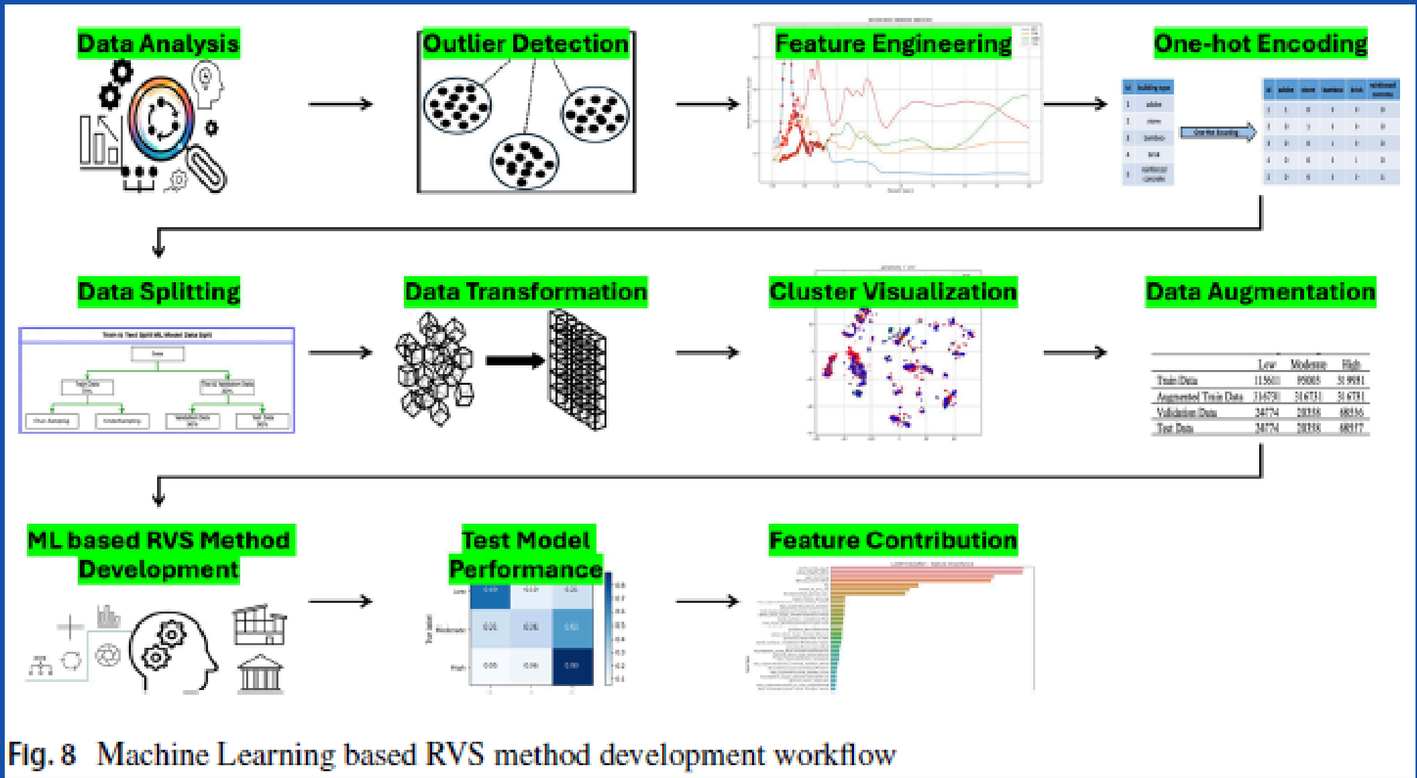


Fig. 8 Machine Learning based RVS method development workflow

[1] Bektaş, N., & Kegyes-Brassai, O. (2024). Developing a machine learning-based rapid visual screening method for seismic assessment of existing buildings on a case study data from the 2015 Gorkha, Nepal earthquake. Bulletin of Earthquake Engineering.

Methodology

- Used Gorkha 2015 data (762k+ buildings);
- Preprocessed with scaling, encoding, SMOTE;
- Engineered features; trained 9 ML models.

Observations

- RC buildings had least damage;
- Stone/mud had most severe damage;
- Engineered features boosted accuracy.

Performance Metrics

- Used F1, accuracy, recall, precision;
- 3-class damage prediction.

Analysis

- 73% accuracy (40% ↑ vs. RVS);
- XGBoost, RF best performers;
- Supports scalable risk assessment.

Lit Review 2

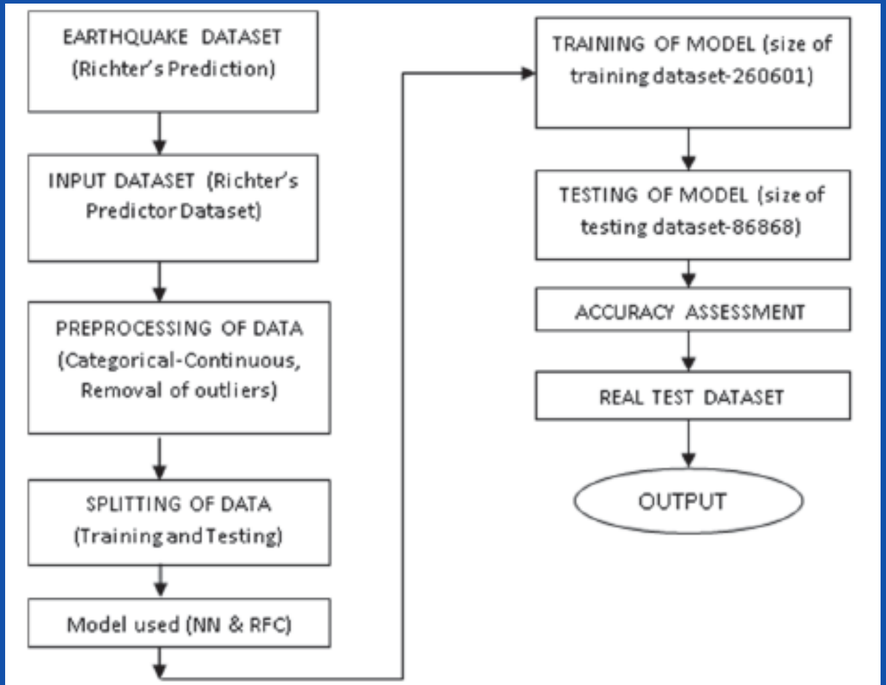


Fig 1. Methodology

[2] Chaurasia, Kuldeep & Kanse, Samiksha & Yewale, Aishwarya & Singh, Vivek & Sharma, Bhavnish & Burle, Dattu. (2019). Predicting Damage to Buildings Caused by Earthquakes Using Machine Learning Techniques. 81-86. 10.1109/IACC48062.2019.8971453.

Methodology

- Used DrivenData Nepal dataset (38 features);
- Preprocessing: cleaning, encoding, splitting;
- Models: Neural Network (6 layers), Random Forest (750 trees).

Observations

- Damage graded into 3 levels (low to complete);
- NN showed limited improvement;
- RF handled nonlinear patterns better.

Performance Metric

- Used Micro-averaged F1 Score (balances precision and recall across 3 classes).
- Neural Network: F1 Score = 62.8%
- Random Forest: F1 Score = 74.32%

Analysis

- RF outperformed NN in accuracy and F1;
- Robust and scalable for fast earthquake damage prediction.

Lit Review 3

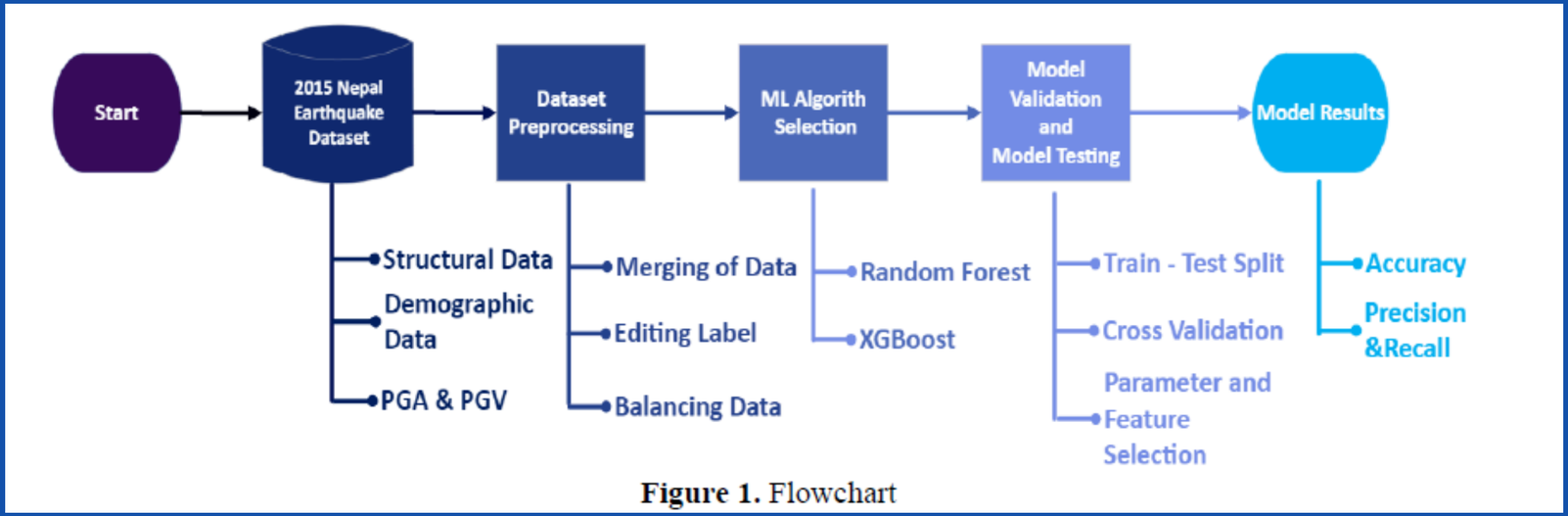


Figure 1. Flowchart

[3] Hasiloglu, Muhammed Ali & Tatar, Tuba. (2022). Prediction of Building Damage Caused by Earthquake with Machine Learning.

Methodology

- Used Gorkha 2015 data (~800k buildings);
- Preprocessed,encoded, balanced;
- Trained RF & XGBoost with/without demographics.

Observations

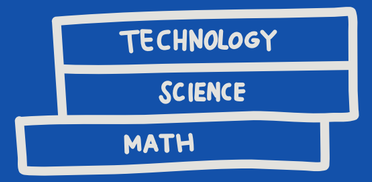
- Demographics improved results;
- RF favored numeric, XGB favored categorical;
- Balancing was crucial.

Metrics

- RF: 70.83% accuracy, 76.36% recall
- XGB: 70.72% accuracy, 75.92% recall

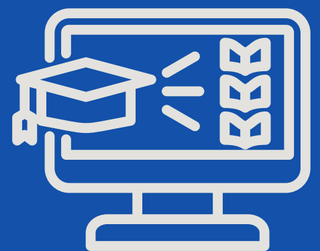
Analysis

- RF slightly better;
- Feature selection worked;
- Effective for quick, low-cost damage prediction.



Applying Key Learnings

- **Enhancing Accuracy**: Real-world applications require higher precision, especially in identifying severely damaged buildings to prevent loss of life. We aim to improve model accuracy.
- **Addressing Research Gaps**: We will work on identified gap and solve it by making a machine learning model.
- **Expanding Model Exploration**: Since previous research has tested a limited number of models, we will explore additional classification techniques for better performance.



About the Dataset

General fact: In case of earthquake, run for cover before facebooking about it!

Dataset Overview

- **Source:** Collected by Kathmandu Living Labs and the Central Bureau of Statistics under Nepal's National Planning Commission Secretariat.
- **Scope:** One of the largest post-disaster datasets, encompassing information on earthquake impacts, household conditions, and socio-economic-demographic statistics.
- **Objective:** Predict the level of damage to buildings caused by the earthquake based on aspects of building location and construction.

https://www.drivendata.org/competitions/57/nepal-earthquake/

Dataset

train_values (2).csv

train_labels (2).csv

test_values (2).csv

2015 Nepal Earthquake Open Data Portal

The totality of the data is available through the 2015 Nepal Earthquake Open Data Portal.

In their own words:

> Following the 7.8 Mw Gorkha Earthquake on April 25, 2015, Nepal carried out a massive household survey using mobile technology to assess building damage in the earthquake-affected districts. Although the primary goal of this survey was to identify beneficiaries eligible for government assistance for housing reconstruction, it also collected other useful socio-economic information. In addition to housing reconstruction, this data serves a wide range of uses and users e.g. researchers, newly formed local governments, and citizens at large. The purpose of this portal is to open this data to the public.

Features: building_id, geo_level_1_id, geo_level_2_id, geo_level_3_id, count_floors_pre_eq, age, area_percentage, height_percentage , land_surface_condition , foundation_type, roof_type, ground_floor_type, other_floor_type , position, plan_configuration, has_superstructure_adobe_mud, has_superstructure_mud_mortar_stone, has_superstructure_stone_flag , has_superstructure_cement_mortar_stone, has_superstructure_mud_mortar_brick, has_superstructure_cement_mortar_brick, has_superstructure_timber has_superstructure_bamboo, has_superstructure_rc_non_engineered, has_superstructure_rc_engineered, has_superstructure_other legal_ownership_status, count_families has_secondary_use, has_secondary_use_agriculture, has_secondary_use_hotel , has_secondary_use_rental, has_secondary_use_institution , has_secondary_use_school, has_secondary_use_industry, has_secondary_use_health_post, has_secondary_use_gov_office, has_secondary_use_use_police, has_secondary_use_other

Data Files & More

Training Values (train_values.csv):

- Contains 38 features per building, including structural details (e.g., number of floors before the earthquake, age, foundation type) and legal information (e.g., ownership status, building use).
- Each building is identified by a unique building_id.DrivenData Labs

Training Labels (train_labels.csv):

- Provides the target variable damage_grade for each building_id:
  - 1: Low damage
  - 2: Medium damage
  - 3: High damage

Test Values (test_values.csv):

- Similar structure to the training values but without the damage_grade labels.DrivenData CommunityDrivenData Labs

Submission Format (submission_format.csv):

- Template for submitting predictions, listing building_id and a placeholder damage_grade (all set to 1). This file is for formatting purposes only and does not contain actual labels.

Any Considerations?

- **Data Quality:** As with many real-world datasets, there may be missing or inconsistent values. Proper data cleaning and preprocessing are essential.
- **Imbalanced Classes:** The distribution of damage grades may be skewed, necessitating techniques like resampling or class weighting.
- **Feature Engineering:** Creating new features or transforming existing ones can enhance model performance.
- **Evaluation Metric:** The competition uses the Micro F1 score to evaluate model prediction.

Data Preprocessing

Step 1 : Checking null values

building_id	0
geo_level_1_id	0
geo_level_2_id	0
geo_level_3_id	0
count_floors_pre_eq	0
age	0
area_percentage	0
height_percentage	0

fig: checking null values

to prevent errors and ensure the model receives complete and meaningful data.

Step 2 : Encoding

- One hot encoding is done when dealing with the categorical data.
- Creates a new binary column for each category.
- Puts a 1 in the column for the corresponding category, and 0 for others.

fig: dataset picture

land_surfa	foundation	roof_type	ground_flo	other_floor	position	plan_conf
t	r	n	f	q	t	d
o	r	n	x	q	s	d
t	r	n	f	x	t	d
t	r	n	f	x	s	d
t	r	n	f	x	s	d
t	r	n	f	q	s	d
n	r	n	x	q	s	d
t	w	q	v	x	s	u
t	r	q	f	q	s	d
t	i	n	v	j	s	d
t	r	q	f	q	t	d
t	u	n	v	j	t	d
t	r	n	x	x	s	d
t	r	q	f	q	s	d
t	r	n	f	x	s	d
n	r	n	f	q	s	d

Step3 : Standardization

	building_id	geo_level_1_id	geo_level_2_id	geo_level_3_id	count_floors_pre_eq	age	area_percentage	height_percentage
0	0.910312	-0.983414	-0.518705	1.629055	-0.178274	0.047100	-0.459460	-0.226419
1	-1.631438	-0.734459	0.481998	-0.945017	-0.178274	-0.224765	-0.004110	0.816109
2	-1.414337	0.883744	-0.819158	0.744612	-0.178274	-0.224765	-0.687135	-0.226419
3	0.214112	1.008221	-0.685893	1.216589	-0.178274	-0.224765	-0.459460	-0.226419
4	-1.063003	-0.361028	-1.381296	-1.308119	1.195989	0.047100	-0.004110	1.858636
...	...	...	...	...	...	...	...	...
260596	0.535096	1.381653	1.536007	-1.271644	-1.552536	0.386932	-0.459460	-1.268946
260597	0.472212	0.385835	0.033741	-1.151250	-0.178274	-0.360698	-0.459460	-0.226419
260598	0.252300	0.385835	-1.575137	0.522472	1.195989	0.386932	-0.459460	0.816109
260599	-1.228939	1.506130	-1.604213	-1.208568	-0.178274	-0.224765	1.361941	0.294845
260600	0.728690	0.883744	-1.676903	0.779715	1.195989	-0.224765	-0.231785	0.294845

260601 rows × 61 columns

fig: displaying encoded & standardized dataset

ensure all features have a common scale, improving model performance and convergence.

WE LEARN GEOLOGY THE MORNING AFTER THE EARTHQUAKE
-Ralph Emerson

Step4 : SMOTE for Oversampling

Why?

SMOTE is used for oversampling to address class imbalance by generating synthetic examples of the minority class.

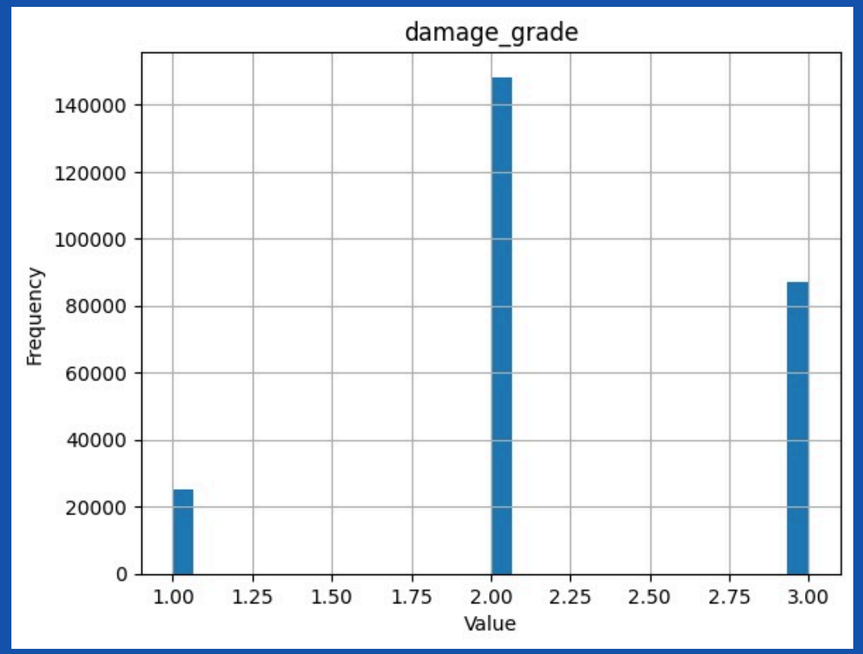


fig: before SMOTE - class imbalance

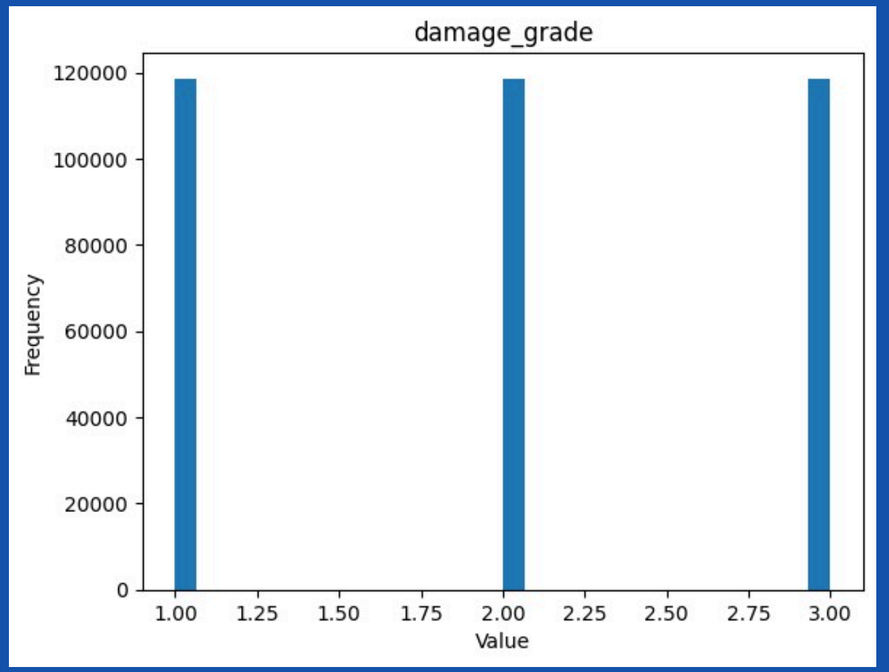


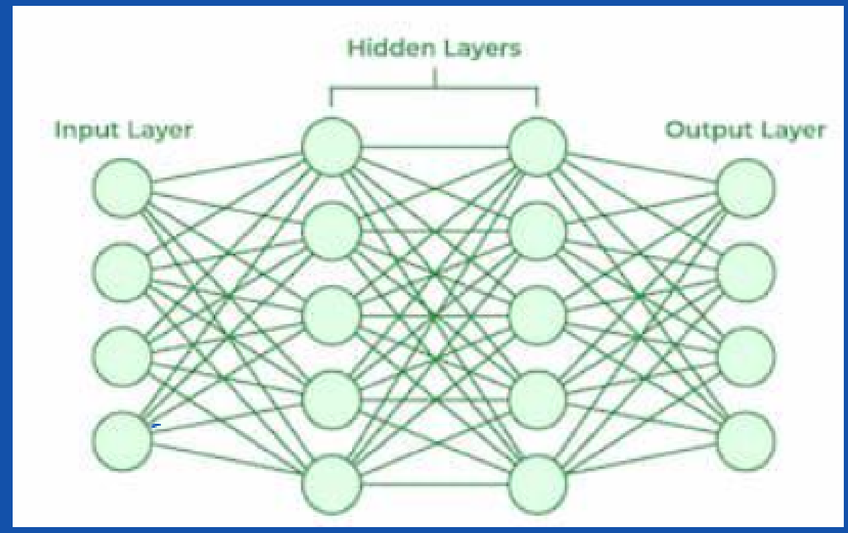
fig: after SMOTE

ML Methodology

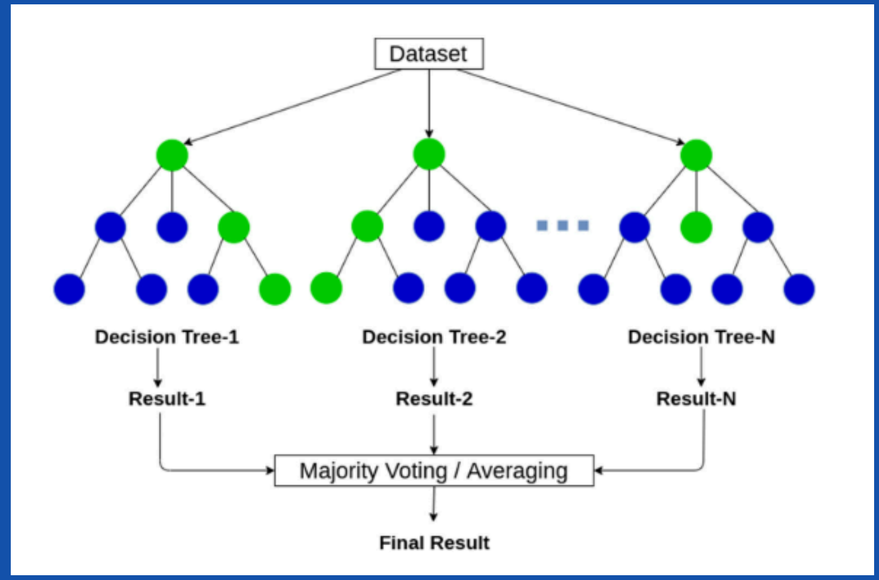
General Fact: When an earthquake gets emotional, its a really ‘faulty’ outburst!

ML Methods Used and Why?

1. RANDOM FOREST
2. NEURAL NETWORKS
3. XG BOOST
4. CATBOOST
5. LIGHTGBM



We chose to use boosting algorithms and other models based on **insights from prior research papers,** which highlighted these models as **top performers for damage classification tasks**. These studies provided a strong foundation for our approach, as our primary objective was to achieve a high micro averaged F1 score and improve our ranking on the leaderboard. Specifically, XGBoost and neural networks consistently emerged as the most effective models in similar contexts, guiding our model selection.

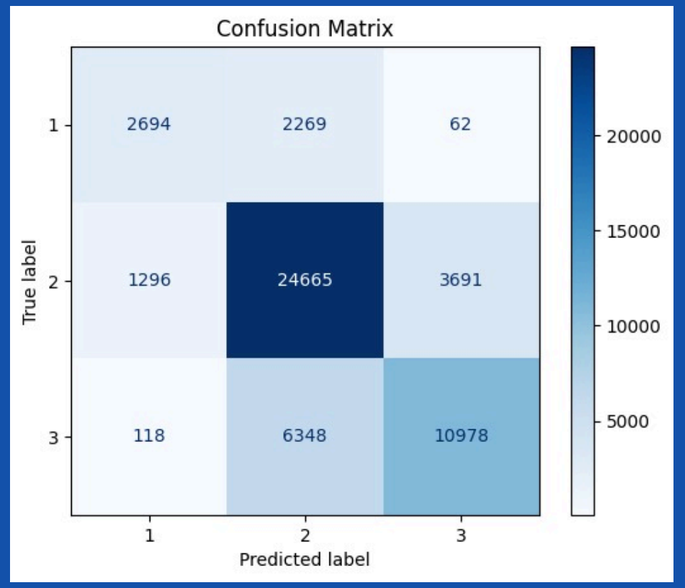


WORKING OF MODELS

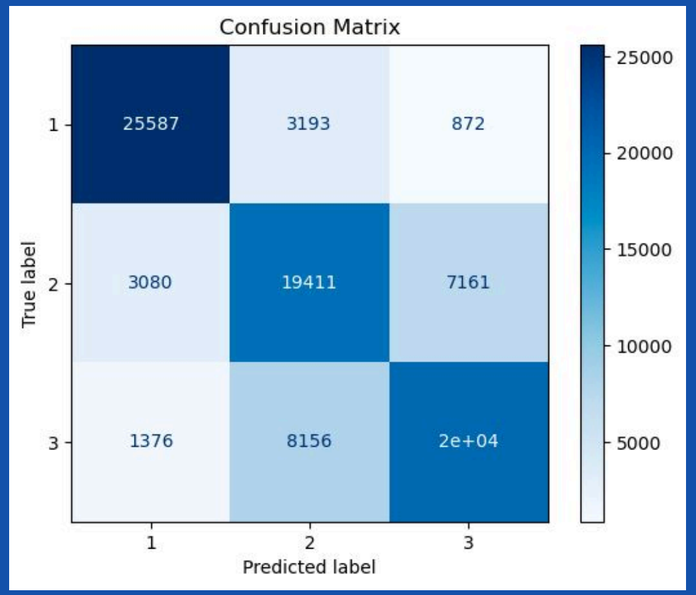
- **Random Forest:** An ensemble of decision trees that combines their predictions through majority voting (classification) or averaging (regression), improving accuracy and reducing overfitting.
- **Neural Networks:** Layered structures of interconnected nodes that learn patterns in data through weighted transformations and nonlinear activation functions, ideal for complex, high-dimensional tasks.
- **XGBoost:** A gradient boosting algorithm that builds trees sequentially, optimizing for speed and performance by minimizing errors of prior trees using second-order derivatives.
- **CatBoost:** A gradient boosting method designed to handle categorical features natively and reduce overfitting using ordered boosting and efficient encoding techniques.
- **LightGBM:** A fast, memory-efficient gradient boosting framework that grows trees leaf-wise instead of level-wise, improving accuracy and performance on large datasets.

WORKING OF MODELS

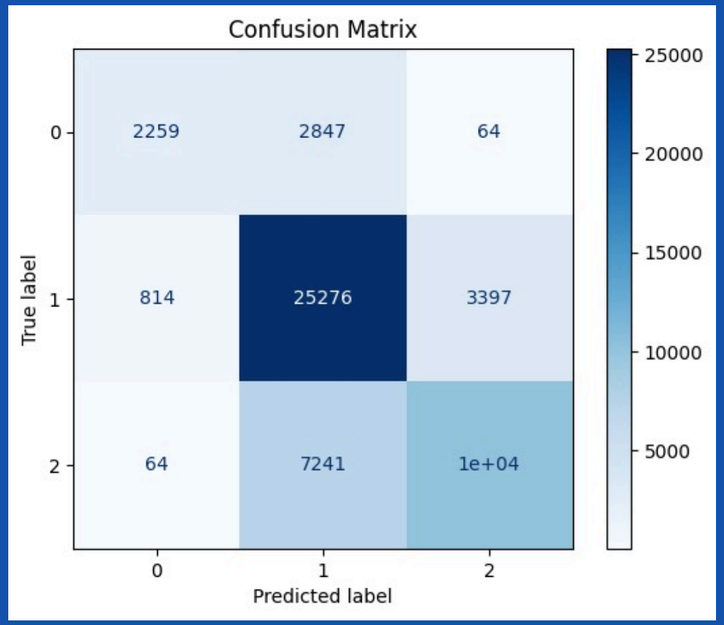
Model	Preprocessing Steps	F1 Score (Micro Average)
XGBoost + SMOTE	Null check, One-hot encoding, Standardization, SMOTE (multiclass balance)	0.7355
XGBoost	Null check, One-hot encoding, Standardization	0.7239
CatBoost	Null check, One-hot encoding, Standardization	0.7232
Random Forest	Null check, One-hot encoding, Standardization	0.7200
LightGBM	Null check, One-hot encoding, Standardization	0.6990
Neural Networks + SMOTE	Null check, One-hot encoding, Standardization, SMOTE (multiclass balance)	0.6476



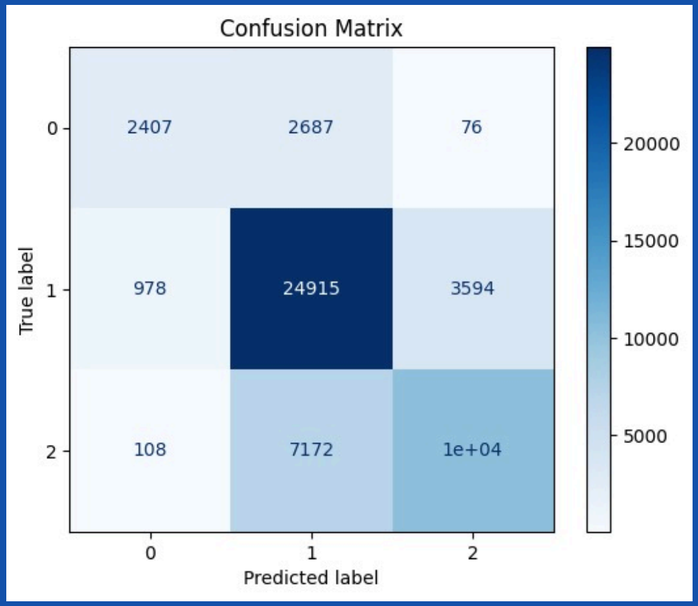
XG Boost + SMOTE



Neural Network



CatBoost



Random Forest

CHALLENGES FACED

We primarily encountered **two major** challenges:

1. **PCA Impact:**

When we applied Principal Component Analysis (PCA) for dimensionality reduction, the model's micro-averaged F1 score dropped significantly compared to training without PCA. This likely occurred because PCA discarded some important features that were crucial for accurate damage prediction.

Solution: We revised our preprocessing pipeline and removed PCA to retain all original features, which improved model performance.

2. **Class Imbalance:**

The dataset was imbalanced, with significantly fewer buildings labeled as damage grade 1 compared to grades 2 and 3.

Solution: We used SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset by generating synthetic examples for the minority class. This helped the model learn better representations for all classes and led to an increase in the F1 score.

Performance Metrics and Deployability

What were the performance metrics and how much were they?

- The competition used the **macro-averaged F1 score** as the primary evaluation metric.
- Macro-F1 score is the unweighted mean of F1 scores calculated per class (across the three damage severity classes: 1 – minor damage, 2 – moderate damage, and 3 – significant damage).

How do these performance metrics show that your solution works?

- The macro F1 score reflects both precision and recall across all damage classes:
  Precision: How many of the predicted damages were correct?
  Recall: How many actual damages were correctly identified?
- A high F1 score here means our model accurately predicted all damage levels, handled class imbalance well, and generalized across regions and building types without overfitting.

What may be some challenges for the deployed solution when it will scale up?

- **Data Quality & Availability:** Real-world data may be noisy, incomplete, or inconsistent compared to the training dataset.
- **Class Imbalance:** New regions might have different distributions of damage classes, making the model less effective without retraining.
- **Generalization:** The model may struggle to generalize to new building types, materials, or geographic contexts not seen during training.
- **Infrastructure Requirements:** Processing large volumes of data in real time (e.g., post-disaster) may require significant computational resources.
- **Model Drift:** Over time, changes in construction practices or damage patterns may reduce model accuracy if it's not regularly updated.
- **Interpretability & Trust**: In high-stakes situations, such as disaster response, decision-makers may require clear explanations of model predictions.

WHERE DO WE STAND IN THE LEADERBOARD?

Thank you!