WHY DO BUILDINGS COLLAPSE?

A PREDICTIVE AND DESCRIPTIVE ANALYSIS OF EARTHQUAKE DAMAGE

Álvaro Bernal, Álvaro González, Antonio Rodríguez, Adrián Romero

FID – Mon. 16 Dec. 2024

Available at

`https://github.com/FID2425/ml-olympiad-earthquake-dmg-24`

WHAT WE'LL COVER

INTRODUCTION

INTRODUCTION - DATASET

- **Context:** Gorkha earthquake in Nepal on 25th April 2015.
  - Magnitude 7.8.
  - +8,000 deaths.
  - +20,000 injured.
  - ~500k buildings affected.
- Featured at **ML Olympiad 2024.**

Figure 1: Aftermath of the Gorkha Earthquake [1]

[1] **Source**: National Geographic - Nepal Earthquake Strikes One of Earth's Most Quake-Prone Areas

INTRODUCTION - DATASET

- **Context:** Gorkha earthquake in Nepal on 25th April 2015.
  - Magnitude 7.8.
  - +8,000 deaths.
  - +20,000 injured.
  - ~500k buildings affected.
- Featured at **ML Olympiad 2024.**

Figure 1: Aftermath of the Gorkha Earthquake [1]

[1] **Source**: National Geographic - Nepal Earthquake Strikes One of Earth's Most Quake-Prone Areas

INTRODUCTION - DATASET

- **Context:** Gorkha earthquake in Nepal on 25th April 2015.
  - Magnitude 7.8.
  - +8,000 deaths.
  - +20,000 injured.
  - ~500k buildings affected.
- Featured at **ML Olympiad 2024.**

Figure 1: Aftermath of the Gorkha Earthquake [1]

[1] **Source**: National Geographic - Nepal Earthquake Strikes One of Earth's Most Quake-Prone Areas

INTRODUCTION - DATASET

- **Context:** Gorkha earthquake in Nepal on 25th April 2015.
  - Magnitude 7.8.
  - +8,000 deaths.
  - +20,000 injured.
  - ~500k buildings affected.
- Featured at **ML Olympiad 2024.**

Figure 1: Aftermath of the Gorkha Earthquake [1]

[1] **Source**: National Geographic - Nepal Earthquake Strikes One of Earth's Most Quake-Prone Areas

INTRODUCTION - DATASET

- **Target:** `damage_grade`
  - Ordinal variable.
  - 1, 2, 3.
- **Features:** 38 features.
  - Row → building.
  - Regarding buildings' structure & legal ownership.
  - Categorical, ints and binary.

INTRODUCTION - TEAM

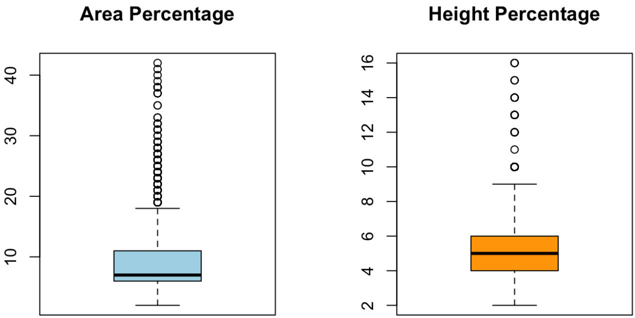
- Divided into **pairs**:
  - **EDA** and **Unsupervised learning**:
    - Adrián Romero.
    - Antonio Rodríguez.
  - **Preproc.** and **Supervised learning**:
    - Álvaro Bernal.
    - Álvaro González.

EXPLORATORY DATA ANALYSIS

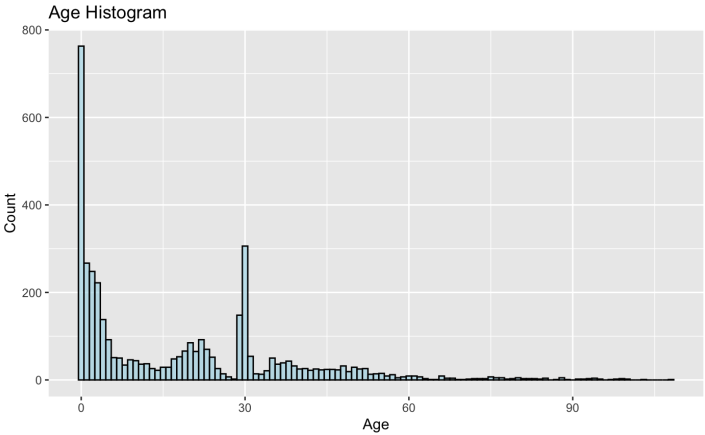
EXPLORATORY DATA ANALYSIS

1. **Data Overview**:
   - 36 columns × 4000 rows.
2. **Data Consistency**:
   - No inconsistencies found.
   - Clean, with no missing values.
3. **Data Visualization**:
   - Percentages.
   - Numerical.
   - Categorical.
   - Binaries.
4. **Data Correlation**.

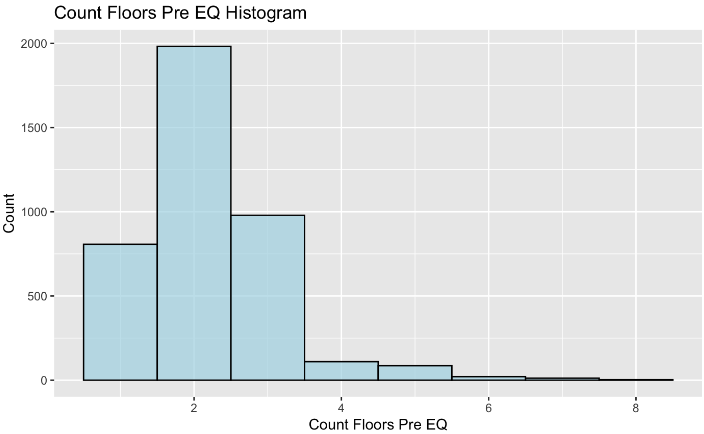
EDA - PERCENTAGES VISUALIZATION



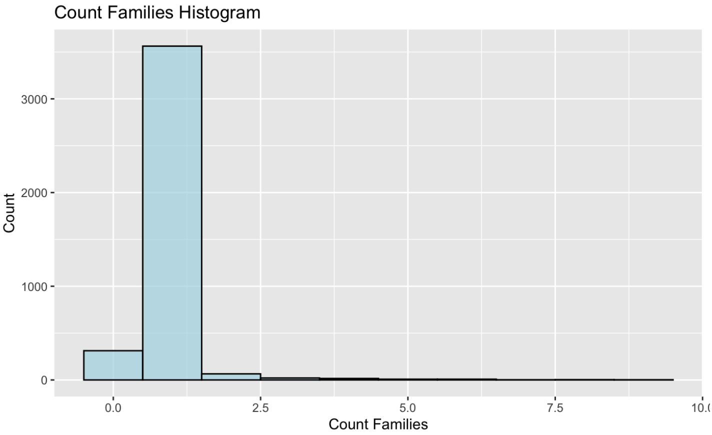
EDA - NUMERICAL VISUALIZATION



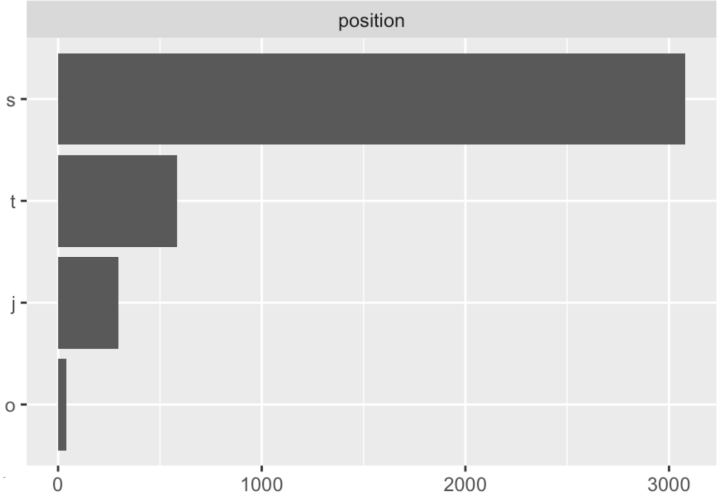
EDA - NUMERICAL VISUALIZATION



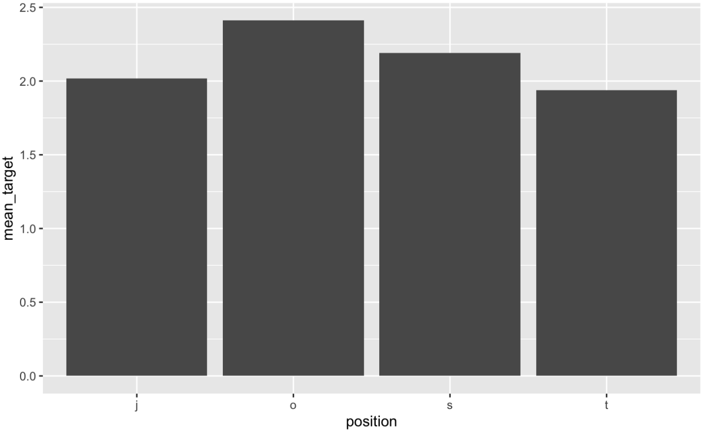
EDA - NUMERICAL VISUALIZATION



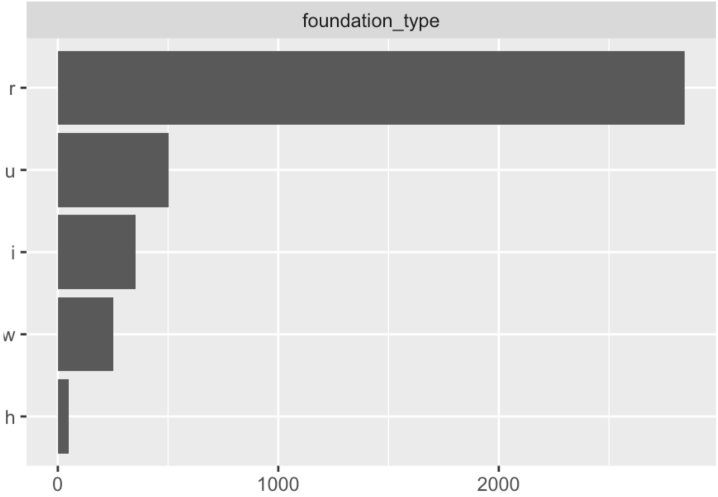
EDA - CATEGORICAL VISUALIZATION



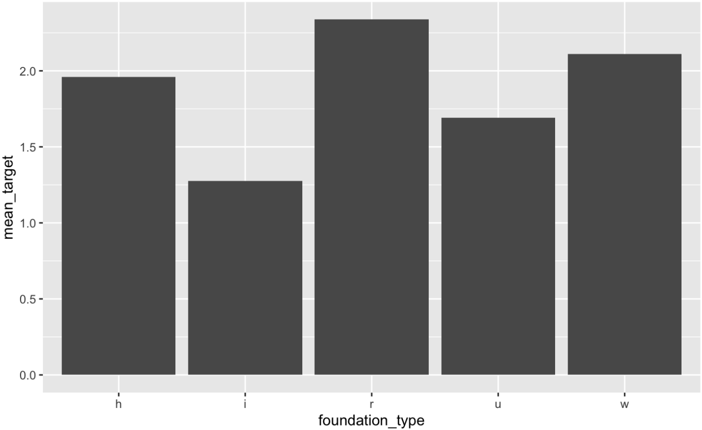
EDA - CATEGORICAL VISUALIZATION



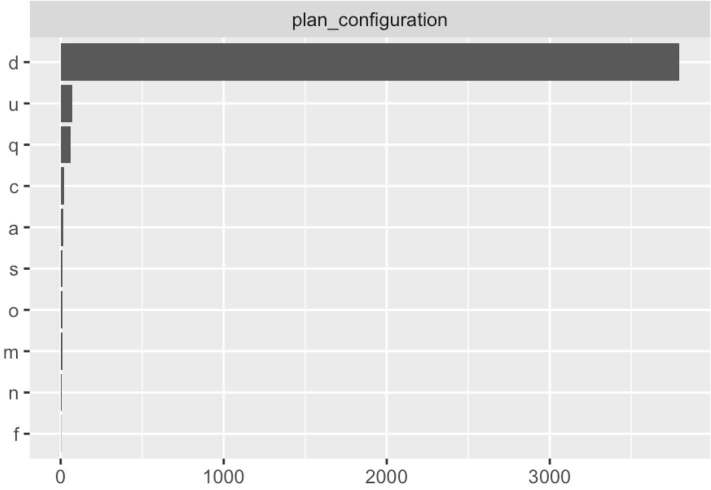
EDA - CATEGORICAL VISUALIZATION



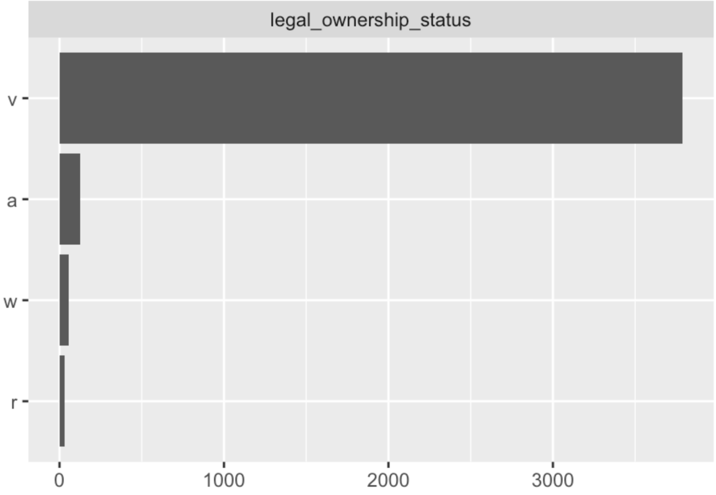
EDA - CATEGORICAL VISUALIZATION



EDA - CATEGORICAL VISUALIZATION



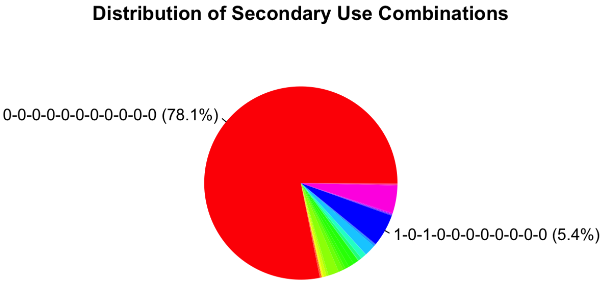
EDA - CATEGORICAL VISUALIZATION



EDA - BINARY VISUALIZATION

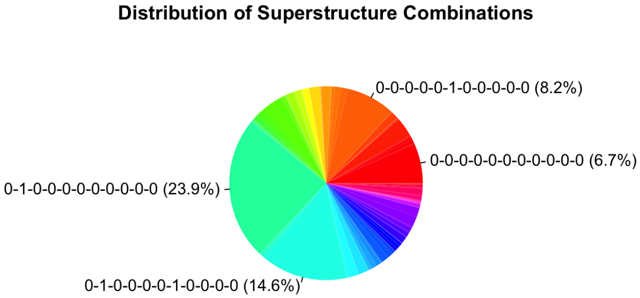
- Two main groups:
  - **Superstructure:** 11 features.
  - **Secondary use:** 11 features.
- **“Camouflaged” categorical feature**?

EDA - BINARY VISUALIZATION



No Secondary Use (78.1%) and Agriculture with Rental (5.1%).

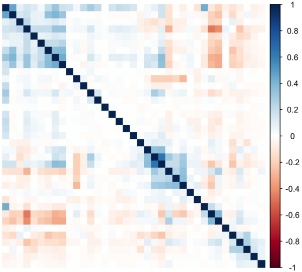
EDA - BINARY VISUALIZATION



Mud Mortar Stone (23.9%), Mud Mortar Stone with Timber (14.6%), Cement Mortar Brick (8.2%) and No Superstructure (6.7%).

EDA - DATA CORRELATION

- Correlation with **target** variable:
  - **Age** (0.269).
  - **Area Percentage** (-0.325).
  - **Roof Type** (-0.324).
  - **Sup.str. Mud Mortar Stone** (0.393).
  - **Sup.str. Cement Mortar Brick** (-0.415).
  - **Sup.str. RC No Engineered** (-0.221).
  - **Sup.str. RC Engineered** (-0.259).



EDA - PROPOSED PREPROCESSING

- **One-Hot encoding** for **categorical** features.
- Deal with **imbalanced data** on **target**.
- **Remove redundant columns** to reduce dimensions and noise.
- Combine **Secondary Use** columns.

PREPROCESSING

PREPROCESSING

- **4 steps**:
  1. **Data Cleaning**: Column Unification and Elimination.
  2. **Train-Test Split**: 80-20 split.
  3. **One-Hot Encoding.**
  4. **Correlation Analisis**:
     - `land_surface_condition_n` and `position_s`.
     - `land_surface_condition_t` and `position_t`.
- **Other techniques**:
  - **PCA.**
  - **Handle Imbalanced Data.**

PREPROCESSING - PCA

- Dimensionality reduction technique.
- Outstanding task: **Kaiser Criterion**:
  - Retained **13 principal components.**

PREPROCESSING - IMBALANCED DATA

- **Outstanding task**: **SMOTE**:
  - Synthetic samples based on nearest neighbours.
- **Weight Allocation**:
  - Assign weight based on class distribution.

PREPROCESSING - FINAL DATASETS

Train:

- **onehot_train.csv**
- **filtered_onehot_train.csv**
- **pca_train.csv**
- **smote_train.csv**
- **weighted_train.csv**

Test:

- **onehot_test.csv**
- **filtered_onehot_test.csv**
- **pca_test.csv**

SUPERVISED LEARNING

SUPERVISED LEARNING

- **Goal:** predict earthquake damage.
- **Data:** 5 previous datasets.
- **Algorithms:**
  - Random Forest.
  - Gradient Boosting Machine.
  - Stochastic Gradient Boosting.
- **Metrics:**
  - Confusion Matrix & associated metrics.
  - Outstanding task: **Multi-class AUC**[2].

---

[2] Jesús S Aguilar-Ruiz and Marcin Michalak. “Classification performance assessment for imbalanced multiclass data”. In: *Scientific Reports* 14.1 (2024), p. 10759

SUPERVISED LEARNING - RESULTS

- We obtained 15 different models.
- We chose the 6 best according to the metrics.

Metric	Best Models
Best Precision	`rf_filtered` & `rf_weighted`
Best Recall	`rf_weighted` & `gbm_pca`
Best F1 Score	`rf_onehot` & `gbm_smote`
Best AUC	`gbm_filtered` & `gbm_smote`

SUPERVISED LEARNING - PRECISION

Model	Class 1	Class 2	Class 3	Avg
rf_onehot	0.655	0.554	0.492	0.568
rf_filtered	**0.664**	0.548	**0.568**	**0.594**
rf_weighted	0.503	**0.567**	0.434	0.501
gbm_filtered	0.617	0.549	0.453	0.540
gbm_pca	0.000	0.518	0.070	0.197
gbm_smote	0.574	0.551	0.454	0.526

SUPERVISED LEARNING - RECALL

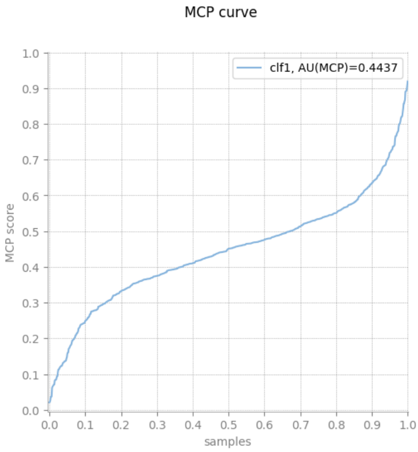
Model	Class 1	Class 2	Class 3	Avg
rf_onehot	0.548	0.760	0.250	0.519
rf_filtered	0.526	0.879	0.096	0.500
rf_weighted	**0.719**	0.295	**0.662**	**0.558**
gbm_filtered	0.526	0.671	0.331	0.509
gbm_pca	0.000	**0.884**	0.019	0.301
gbm_smote	0.519	0.584	0.435	0.513

SUPERVISED LEARNING - F1 SCORE

Model	Class 1	Class 2	Class 3	Avg
rf_onehot	**0.597**	0.641	0.332	**0.523**
rf_filtered	0.587	**0.675**	0.164	0.475
rf_weighted	0.591	0.387	**0.423**	0.501
gbm_filtered	0.568	0.603	0.382	0.518
gbm_pca	NA	0.653	0.030	0.341
gbm_smote	0.545	0.567	0.444	0.519

SUPERVISED LEARNING - AUC

Model	Value
rf_onehot	0.436
rf_filtered	0.432
rf_weighted	0.423
gbm_filtered	0.443
gbm_pca	0.342
gbm_smote	**0.444**



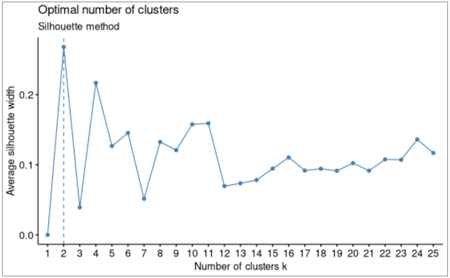
UNSUPERVISED LEARNING

UNSUPERVISED LEARNING

- **Clustering:**
  - **K-Means.**
  - **Hierarchical.**
  - **DBSCAN.**
- **High Dimensionality Visualization:**
  - **Outstanding task: Autoencoder.**
  - **Outstanding task: t-SNE.**

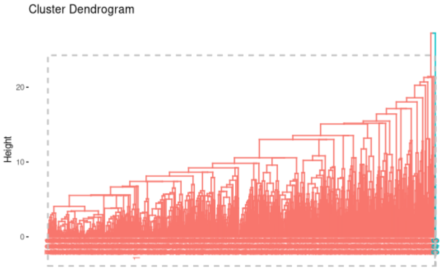
UNSUPERVISED LEARNING - CLUSTERING

- **K-means**
  - **Elbow** & **Silhouette** method.
    - 2 main clusters.
    - No patterns identified.
- **Hierarchical**
  - No good results.



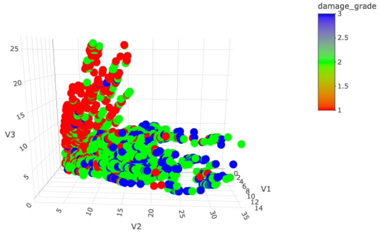
UNSUPERVISED LEARNING - CLUSTERING

- **K-means**
  - **Elbow** & **Silhouette** method.
    - 2 main clusters.
    - No patterns identified.
- **Hierarchical**
  - No good results.



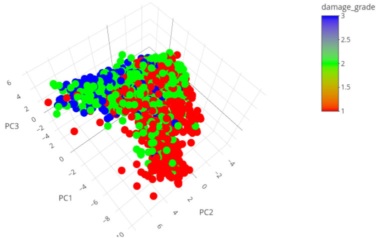
UNSUPERVISED LEARNING - AUTOENCODER

- Reduced **43 dimensions** into **8**:
  - **MAE 0,05**.
- Reduced **43 dimensions** into **3**:
  - **MAE 0,10**.

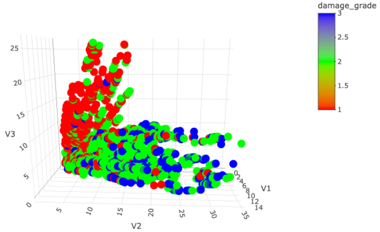


UNSUPERVISED LEARNING - AUTOENCODER VS PCA

PCA

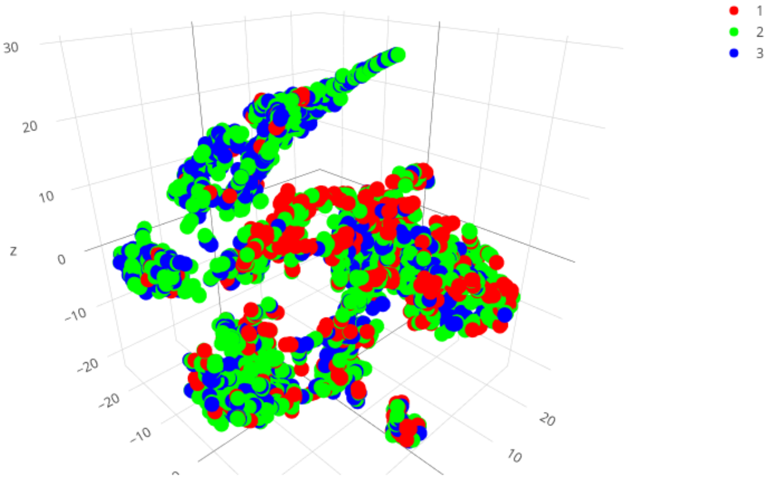


Autencoder

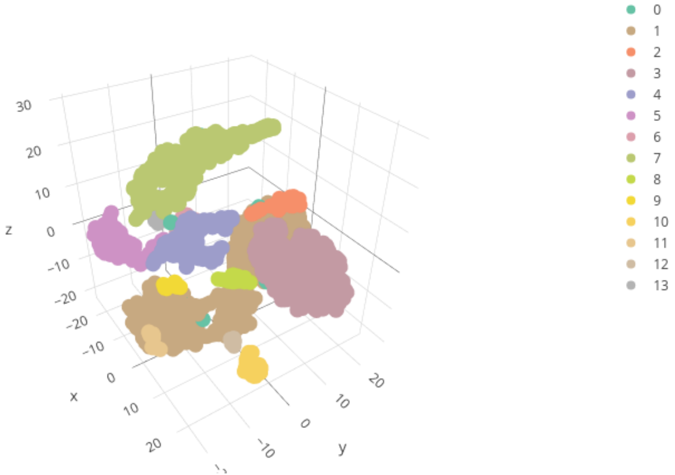


2 clusters, separating `damage_grade` **1** from **2-3**.

UNSUPERVISED LEARNING - T-SNE



UNSUPERVISED LEARNING - T-SNE & DBSCAN



13 clusters, with no relation to `damage_grade`.

DISCUSSION

DISCUSSION

- **Data Challenges:**
  - Low interpretability of categorical features.
  - Limited variability in features, reducing generalization.
  - Unbalanced class distribution, affecting accuracy for damage grades 2 and 3.
- **Dimensionality Reduction Insights:**
  - PCA, autoencoders, and t-SNE helped visualize clusters.
  - Clusters lacked correlation with target variable.
- **Missing Data:**
  - Distance from earthquake epicentre, adjacent structures, and environmental factors.

CONCLUSIONS

CONCLUSIONS

- This study highlights **significant limitations** in the dataset's influence on **earthquake damage prediction**.
- **Challenges:** Low interpretability and variability, unbalanced distribution.
- Clustering and dimensionality reduction techniques **failed** to provide robust segmentation.
- **Future Improvements:** Collect and integrate additional relevant features, avoid arbitrary selection biases in data preparation to ensure more representative distributions.

That's all!

Thanks for your attention.