TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

BÁO CÁO ĐỒ ÁN MÔN HỌC:

PHÂN TÍCH & TRỰC QUAN DỮ LIỆU

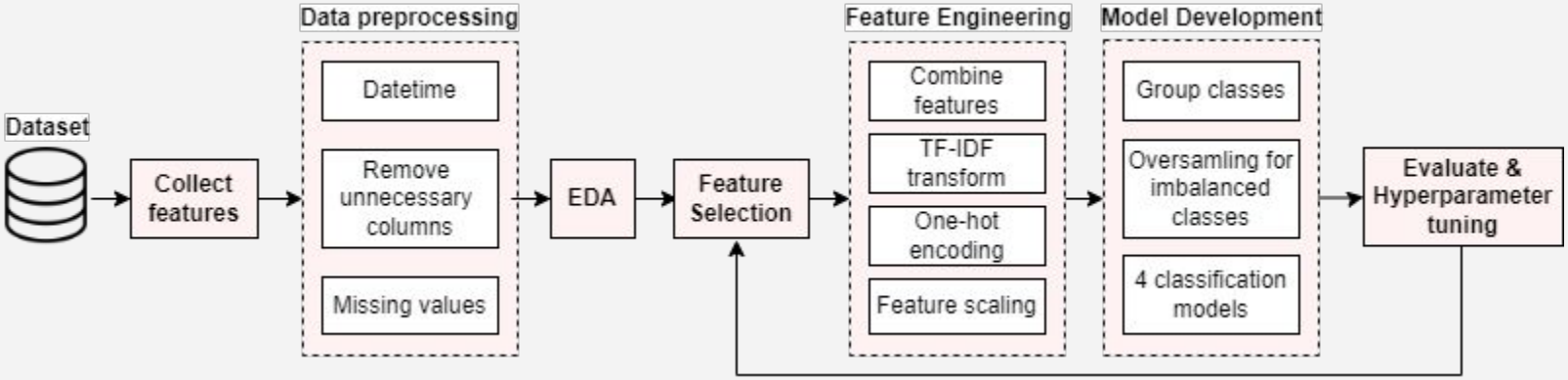
1. 19522065 – Nguyễn Thị Minh Phương
2. 19521882 – Chu Hà Thảo Ngân
3. 19522397 – Thái Minh Triết

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN QUY MÔ SẠT LỞ ĐẤT

Table of Contents

1. Introduction

WORKFLOW



Input: natural environment features.

Output: landslide size.

2. Dataset

Original Dataset: Global Landslide Catalog (GLC)

Size: 11033 x 31

Source: NASA Goddard Space Flight Center (GSFC)

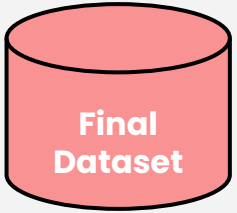
source_name	location_ description	landslide_ setting	event_import_ source	gazeteer_ closest_point
source_link	Location_ accuracy	fatality_count	event_import_ id	gazeteer_ distance
event_id	event_ description	injury_count	country_name	submitted_date
event_date	landslide_ category	storm_name	country_code	created_date
event_time	landslide_ trigger	photo_link	admin_ division_name	Last_edited_ date
event_title	**landslide_size**	notes	admin_division_ population	latitude, longitude

Collected Dataset

Size: 11033 x 36

Source:

- **Weather**: Visual Crossing API
- **Elevation**: Airmap Elevation API
- **Continent**: PyPi pycountry-convert
- **Season**: referenced from National Geographic
- **Population Density**: Gridded Population of the World Version 4.11 Dataset
- **Tree Cover**: Hansen Global Forest Change v1.8 (2000-2020) Dataset
- **Soil Texture**: OpenLandMap Soil Texture Class (USDA System) Dataset



32 categorical features
35 numerical features

11033 x 67

3. Preprocessing

54 features having missing values
Missing value ratio: **12.18%**

Preprocessing datetime: *'event_date', 'created_date', 'last_edited_date', 'submitted_date', 'event_time'*

Dropping unnecessary features: *'source_link' , 'photo_link' , 'event_id', 'event_import_id', 'submitted_date', 'created_date', 'last_edited_date', 'storm_name' 'notes' , 'event_time', 'event_import_source' , 'source_name'*

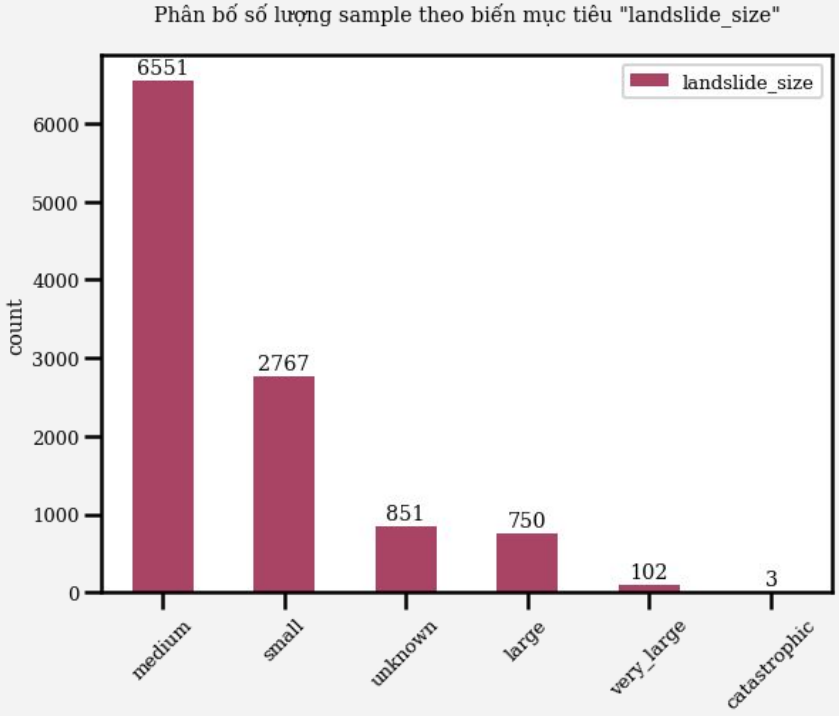
Dealing with missing values:

Missing values	Preprocessing method	Missing values	Preprocessing method
'country_name'	sử dụng module *Nominatim* của thư viện *geopy* để định dạng toàn bộ tên quốc gia theo tiếng Anh	*'fatality_count', 'injury_count'*	median
'admin_division_name', 'location_description', 'gazeteer_closest_point'	tên quốc gia, tên đơn vị hành chính tương ứng	Other numerical features	***mean*** nếu tuân theo phân phối chuẩn ***median*** nếu xuất hiện outliers
'admin_division_population'	KNN Imputation theo *'population_density'*	Rows with almost values are missing	Drop that rows

Data size after preprocessing: *9345 x 54*

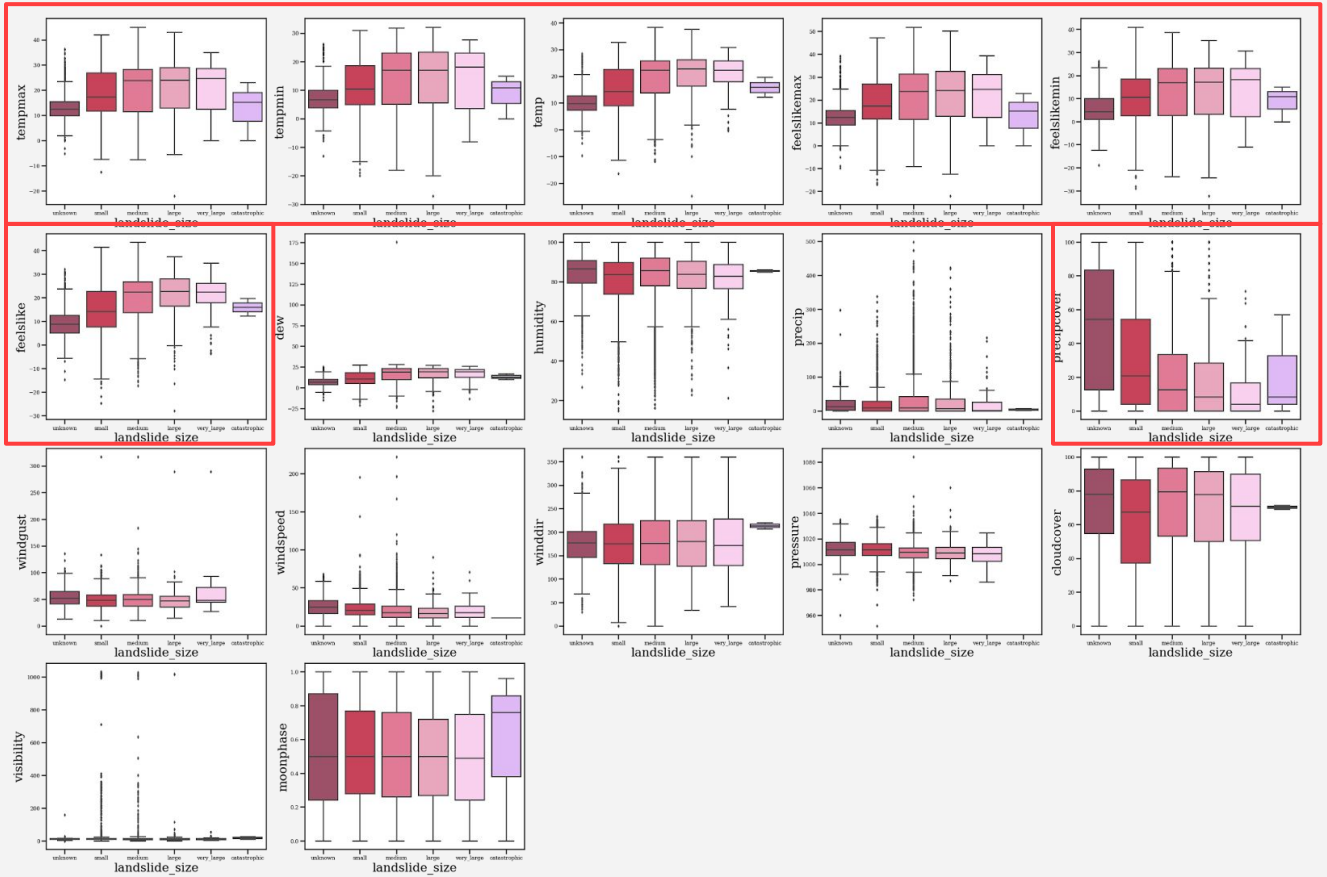
*Data size before preprocessing: **11033 x 67***

4. EDA

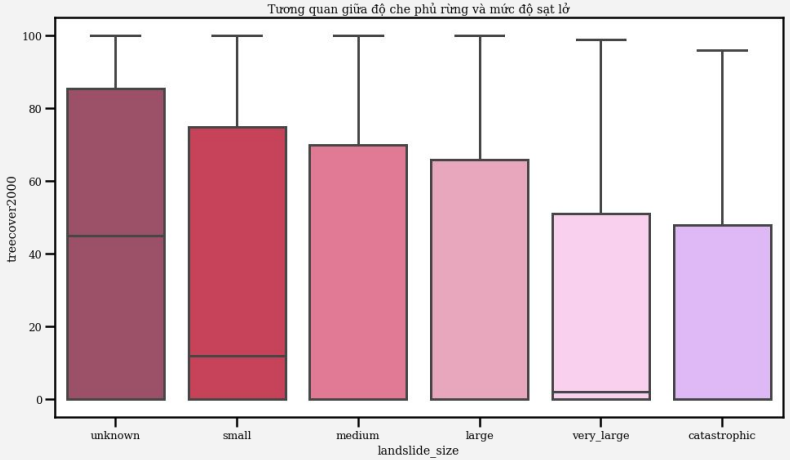


➜ **Imbalanced**

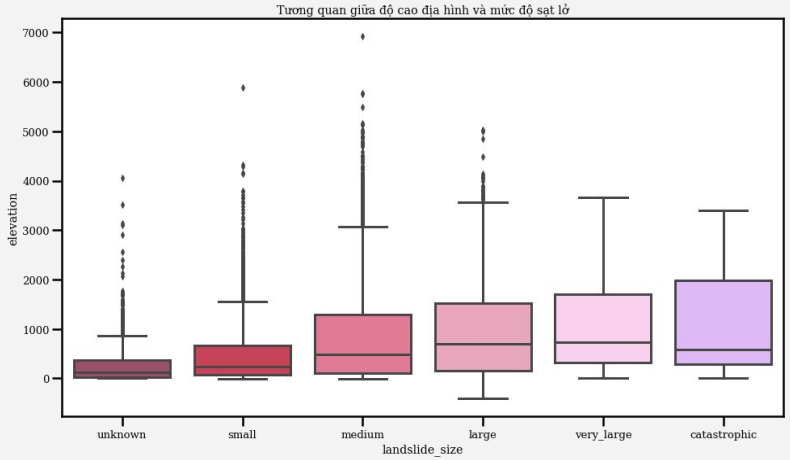
4. EDA


tempmax
tempmin
temp
feelslikemax
feelslikemin
feelslike
dew
humidity
precip
precipcover
windgust
windspeed
winddir
pressure
cloudcover
visibility
moonphase
landslide_size
unknown
small
medium
large
very_large
catastrophic

4. EDA

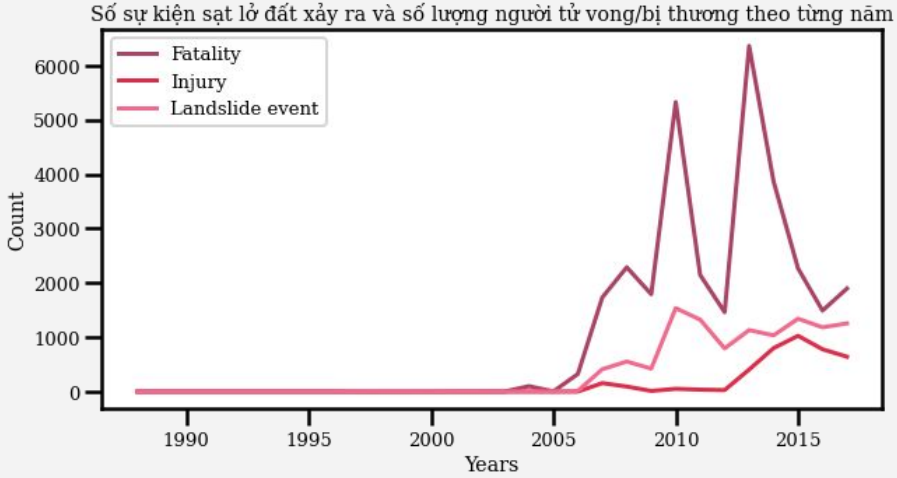


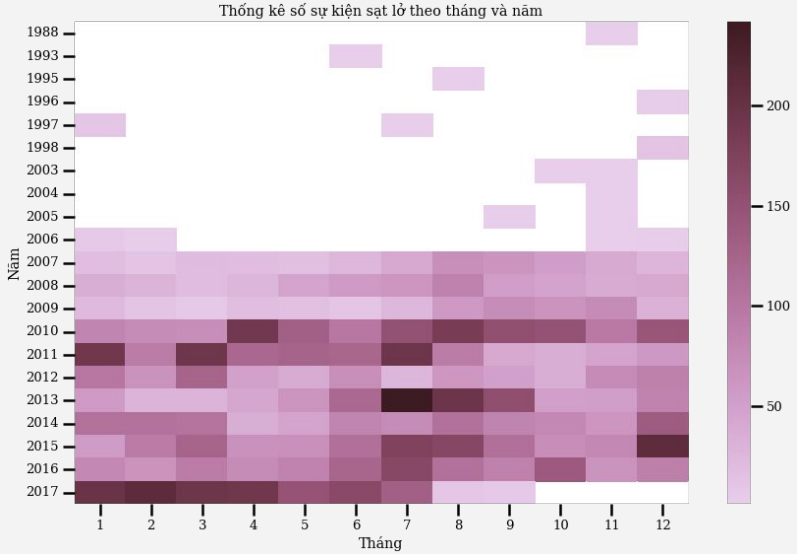
Độ che phủ rừng càng giảm thì sạt lở càng nghiêm trọng.



Độ cao tăng dần thì quy mô sạt lở càng nghiêm trọng

4. EDA





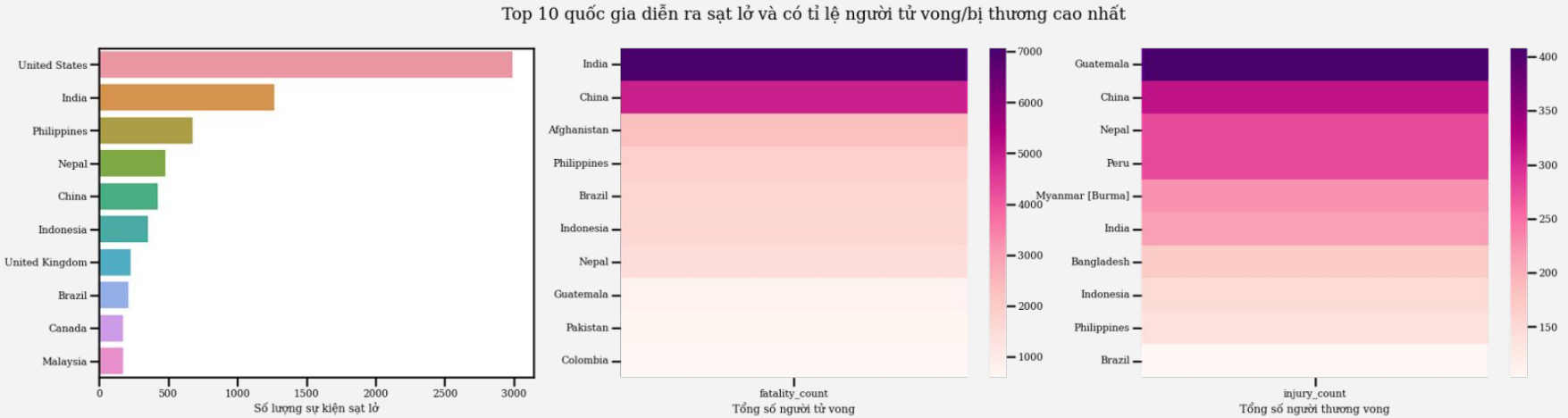
Số lượng sự kiện sạt lở tăng dần từ năm 2007.

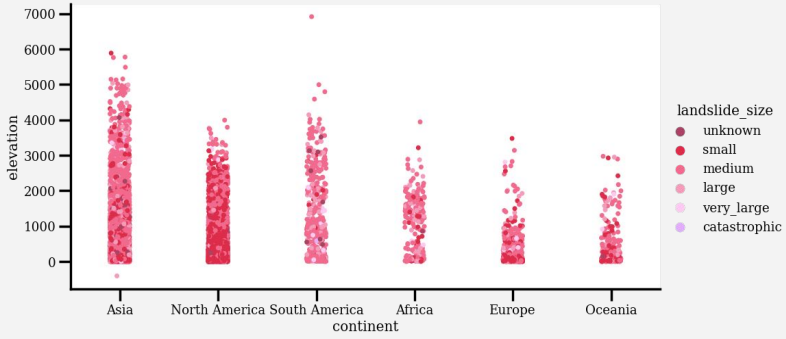
Tháng 7/2013 xảy ra hơn 200 vụ sạt lở.

Năm 2010 xảy ra nhiều sự kiện sạt lở nhất.

Năm 2010 và năm 2013 cho thấy số thương vong và tử vong cao.

4. EDA



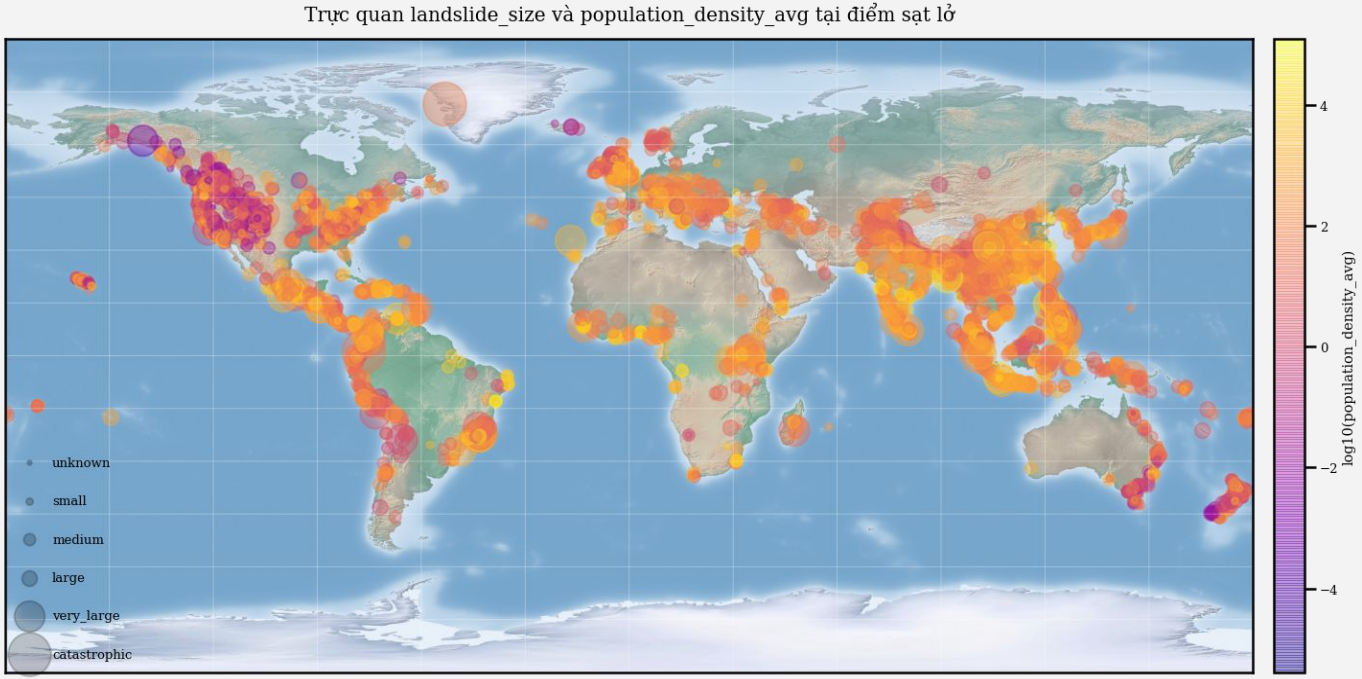


Hoa kỳ là quốc gia xảy ra nhiều sự kiện sạt lở nhất nhưng không nằm trong top quốc gia thương vong nhiều nhất.

Ở Bắc Mĩ sạt lở đa số xảy ra ở địa hình thấp, quy vừa, nhỏ hoặc không xác định.

4. EDA

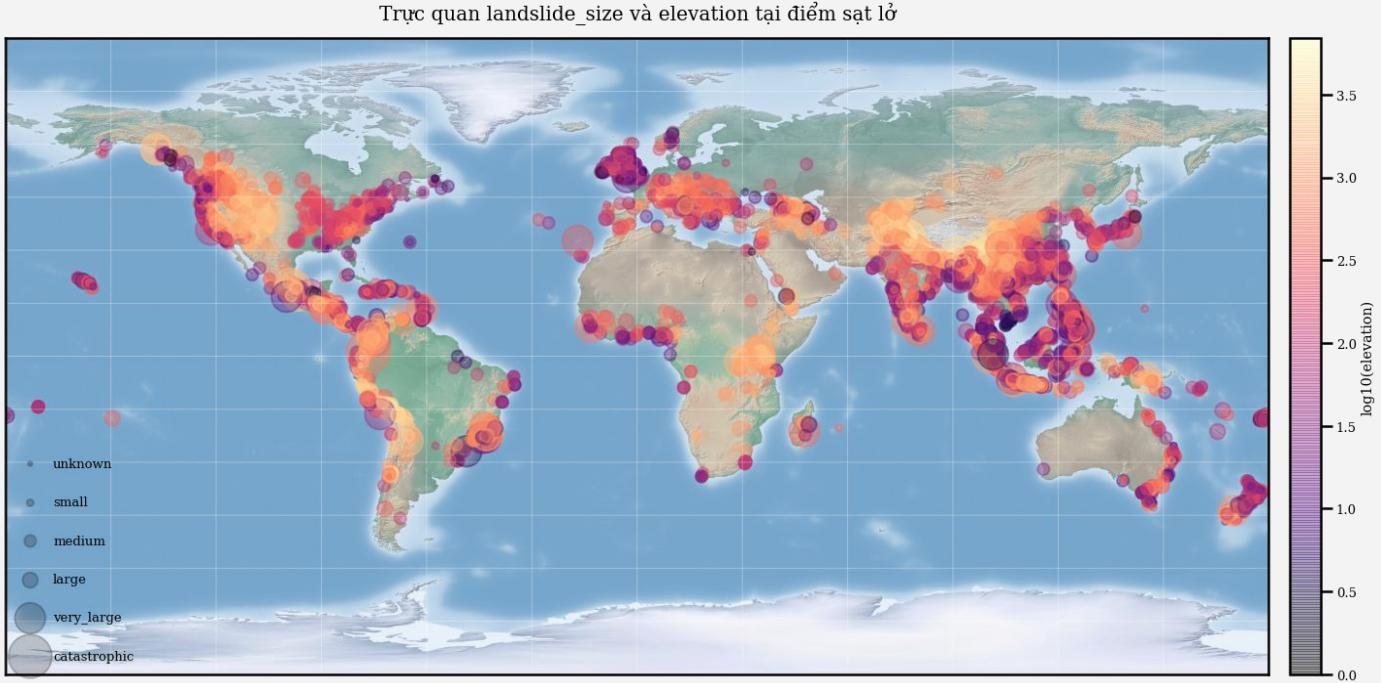
Mật độ dân số



Khu vực phía tây Hoa Kỳ, Đông Úc và New Zealand: Mật độ dân số thấp, quy mô sạt lở ít nghiêm trọng.

4. EDA

Độ cao

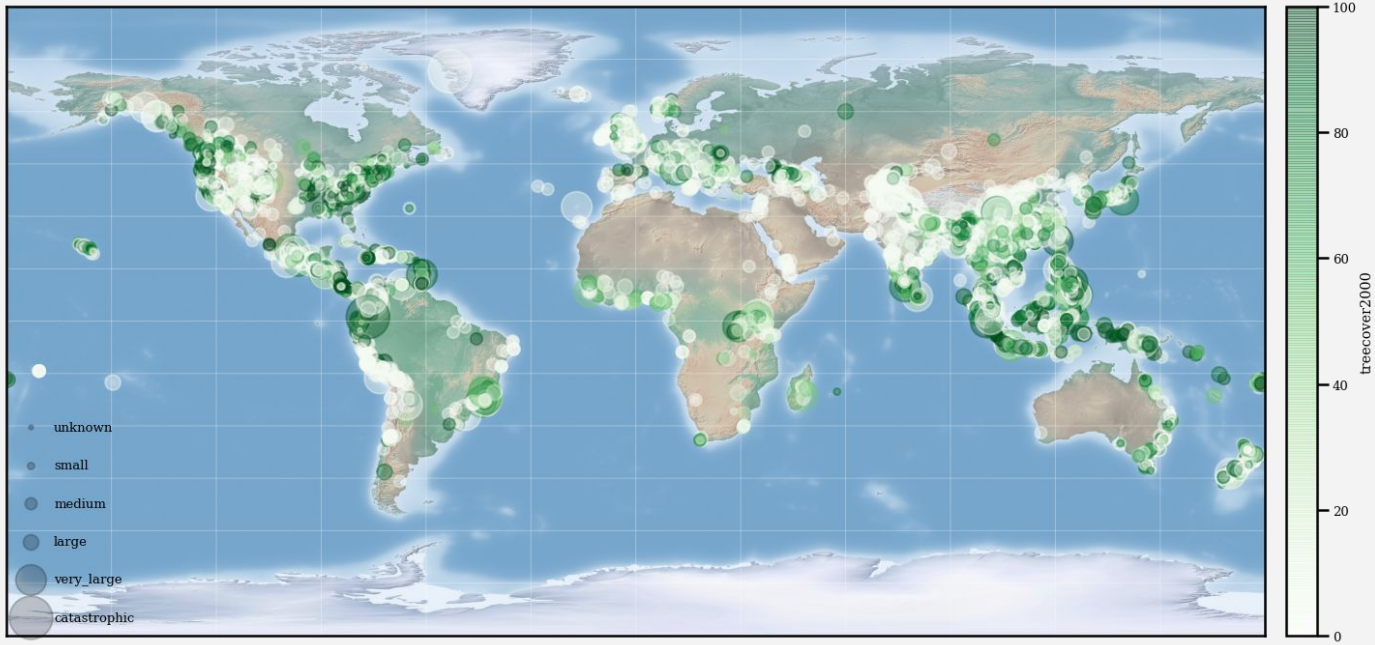


Địa hình đồi núi phía Tây Hoa Kỳ, phía Tây Nam Mỹ và xung quanh dãy Himalaya cho thấy quy mô sạt lở tương đối lớn.

4. EDA

Độ che phủ rừng

Trực quan landslide_size và treecover2000 tại điểm sạt lở



Khu vực có độ che phủ rừng cao cho thấy quy mô sạt lở ít nghiêm trọng

4. EDA

Kết luận ban đầu về các thuộc tính ảnh hưởng đến quy mô sạt lở:

- Địa hình tại khu vực sạt lở càng cao, quy mô sạt lở càng nghiêm trọng. Do sự khác biệt về độ cao địa hình, nên mức độ sạt lở ở các quốc gia cũng có sự khác nhau.
- Có sự khác nhau về quy mô sạt lở giữa các châu lục.
- Những khu vực có chỉ số nhiệt độ như ‘temp’, ‘feelikes’ cao thì mức độ sạt lở càng lớn.

- Các khu vực độ che phủ thảm thực vật thưa thớt có nguy cơ xảy ra sạt lở với quy mô lớn hơn so với những khu vực có thảm thực vật dày.

5. Feature selection

Target feature

'landslide_size': categorical

Numerical features

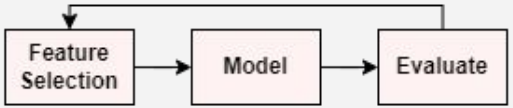
One Way ANOVA

Categorical features

Chi-Square

Threshold?

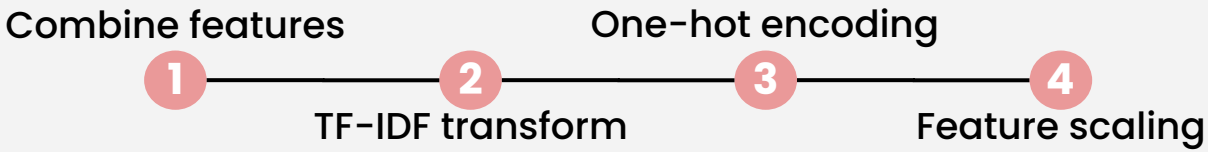
Important features after EDA



	Column name	F-Scores	P-values
0	longitude	517.812633	0.000000e+00
1	dew_temp_weather	331.813683	1.900757e-267
2	latitude	259.448519	9.540822e-212
3	fatality_count	247.823598	1.173878e-202
4	precipcover	153.893517	1.028695e-127
5	event_date	139.713610	3.638529e-116
6	elevation	98.518047	2.905317e-82
7	gazeteer_distance	96.384683	1.709599e-80
8	windspeed	71.052698	2.225879e-59
9	pressure	43.675437	2.252576e-36
10	cloudcover	41.649538	1.150167e-34
11	precip	35.607611	1.436220e-29
12	treecover2000	28.918354	6.316838e-24
13	humidity	24.424115	3.880514e-20
14	injury_count	23.895710	1.081302e-19
15	population_density_avg	20.415154	9.172612e-17
16	admin_division_population	16.827538	9.395941e-14
17	windgust	13.713358	3.757013e-11
18	visibility	13.593347	4.729512e-11
19	moonphase	2.063054	8.285606e-02
20	gain	1.115266	3.472597e-01
21	loss	0.982054	4.158636e-01
22	winddir	0.705781	5.878693e-01

	Column name	Chi-square	P-values
0	description	2.613296e+06	0.000000e+00
1	gazeteer_closest_point	3.660994e+04	0.000000e+00
2	country_name	2.420028e+04	0.000000e+00
3	admin_division_name	2.167065e+04	0.000000e+00
4	landslide_setting	5.060482e+03	0.000000e+00
5	landslide_trigger	4.531832e+03	0.000000e+00
6	location_accuracy	1.055746e+03	2.958066e-227
7	continent	4.692023e+02	3.063286e-100
8	soil_texture	1.651424e+02	1.153039e-34
9	landslide_category	9.754010e+01	3.284016e-20
10	conditions	6.629212e+01	1.374639e-13
11	season	5.013935e+01	3.376873e-10

6. Feature engineering

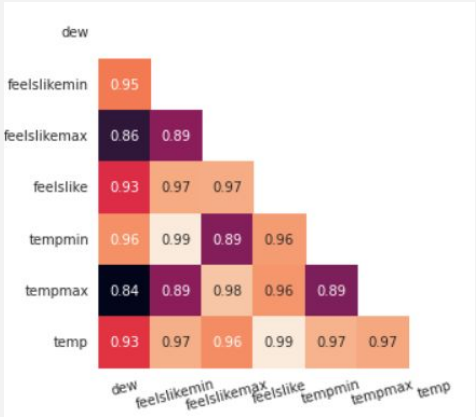


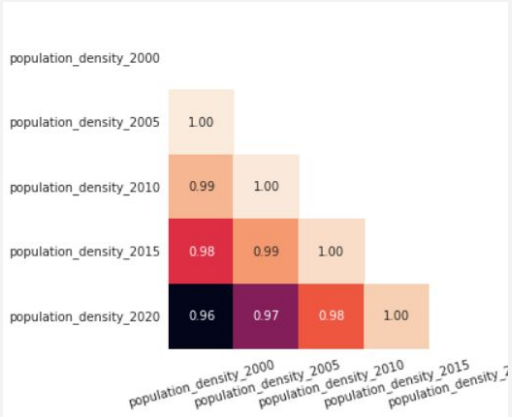
6. Feature engineering

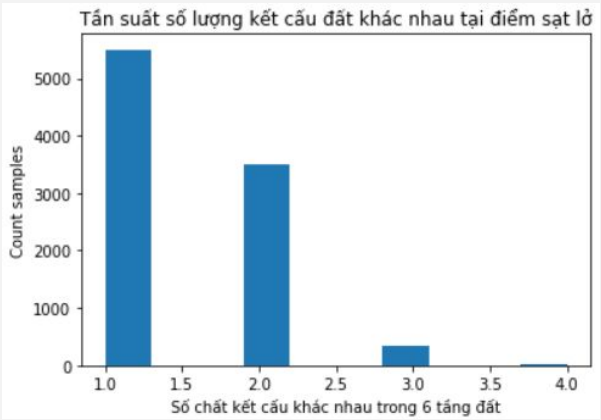
Combine features → One-hot encoding → TF-IDF transform → Feature scaling

1 — 2 — 3 — 4

'soil_texture_0', 'soil_texture_10', 'soil_texture_30', 'soil_texture_60', 'soil_texture_100', 'soil_texture_200'



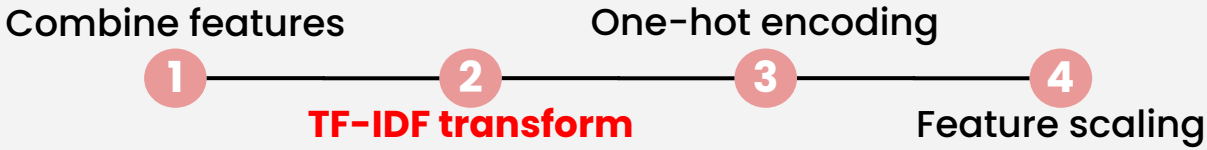




Numerical features with high correlation -> *mean()*

'soil_texture_' -> mode()

6. Feature engineering



'event_description'
'event_title'
'location_description'

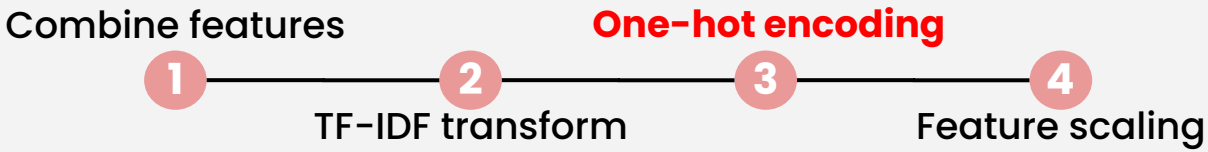
combine

'description'

TF-IDF
transform

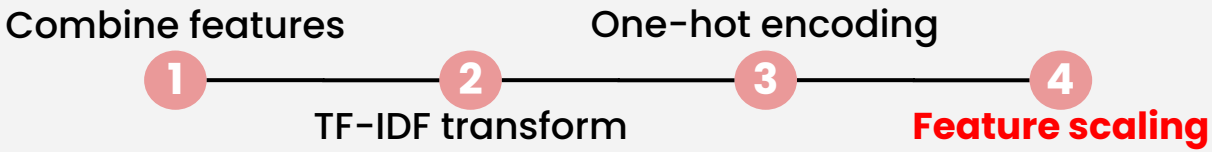
sparse matrix

6. Feature engineering



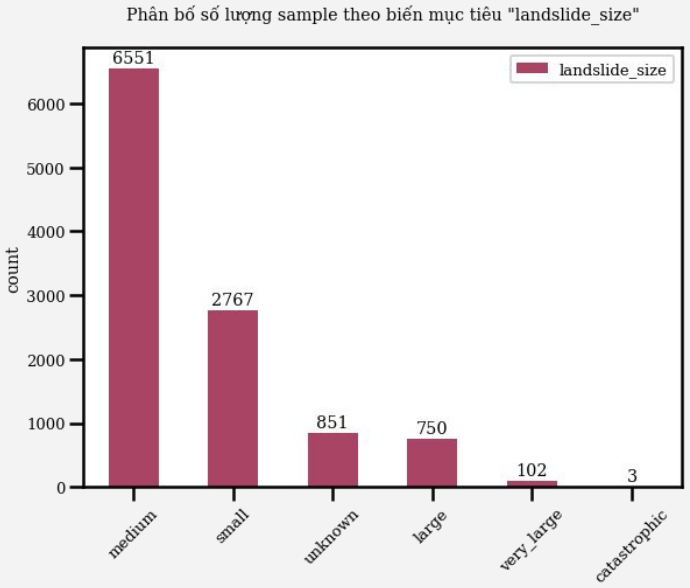
Categorical features: *pandas.get_dummies()*

6. Feature engineering



MaxAbsScaler()

7. Model development



Imbalanced classes

Combine *'catastrophic'* with *'very_large'*

Oversampling minority classes: *ADASYN()*

Models

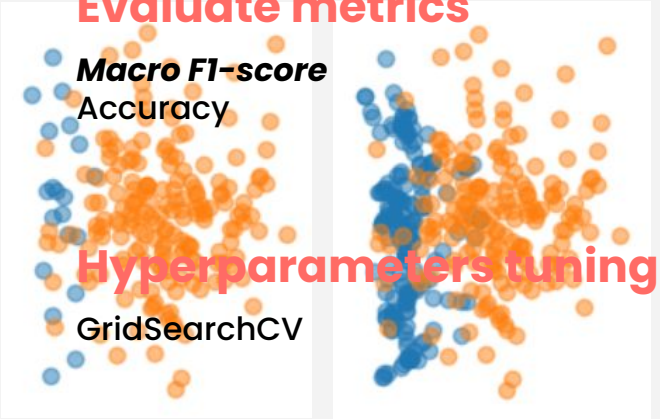
Logistic Regression
Support Vector Machine
Random Forest
Passive Aggressive Classifier

Evaluate metrics

Macro F1-score
Accuracy

Hyperparameters tuning

GridSearchCV



Method	Model	Metric	
		Macro F1	Accuracy
Before handling imbalanced classes	Random Forest	0.481892	0.713623
	Passive Aggressive Classifier	**0.519439**	**0.715407**
	Logistic Regression	0.511298	0.718973
	Support Vector Machine	0.472977	0.717190
After oversampling imbalanced classes	Random Forest	0.498321	0.716476
	Passive Aggressive Classifier	0.520779	0.711484
	Logistic Regression	**0.521931**	**0.715050**
	Support Vector Machine	0.477818	0.706847
Grid Search	**Logistic Regression**	**0.522588**	**0.717189**

8. Result & conclusion

Best model with performance:

Logistic Regression with macro F1 = 0.522588, Accuracy = 0.717189

Oversampling for handling imbalanced classes: better performance.

Hyperparameter tuning: slightly better performance.

Collected features: important and dependent on target feature.

Exploratory Data Analysis: useful insights about landslide events, affection of features on landslide size.

THANKS FOR LISTENING