ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN QUY MÔ SẠT LỞ ĐẤT

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Nguyễn Thị Minh Phương	19522065
2	Chu Hà Thảo Ngân	19521882
3	Thái Minh Triết	19522397

TP. HỒ CHÍ MINH – 12/2021

1. GIỚI THIỆU

Trong đồ án lần này, chúng tôi đã áp dụng các kỹ thuật phân tích và trực quan dữ liệu để phân tích sự ảnh hưởng của các yếu tố tự nhiên lên các sự kiện sạt lở đất đã diễn ra, từ đó xây dựng mô hình máy học dự đoán quy mô (kích thước) sạt lở đất.

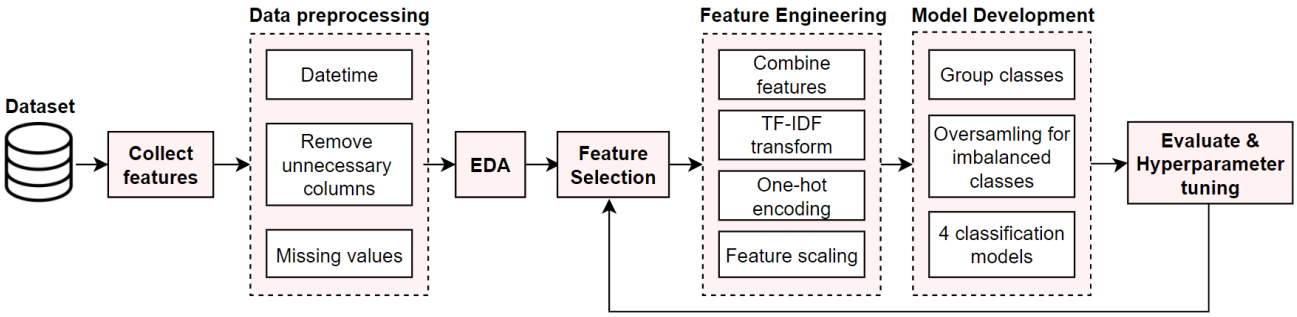
Chúng tôi đã thu thập thêm các thuộc tính về thời tiết, độ cao, mùa, mật độ dân số, kết cấu đất,... để bổ sung vào bộ dữ liệu gốc nhằm hỗ trợ cho quá trình phân tích. Sau đó sử dụng các thư viện *Scikit-Learn, Matplotlib và Seaborn* để tiền xử lý dữ liệu và phân tích thăm dò. Những insights sau khi EDA được kết hợp cùng với hai phương pháp thống kê *Chi-Square* và *ANOVA* để lựa chọn ra được các thuộc tính quan trọng, thực hiện Feature Engineering (*one-hot encoding, TF-IDF transform, MaxAbsScaler,...*) để xử lý trước khi đưa vào mô hình.

Đây là bài toán phân lớp trên dữ liệu mất cân bằng, nên chúng tôi đã áp dụng kỹ thuật *ADASYN* để oversampling cho những lớp có ít điểm dữ liệu hơn nhằm giải quyết vấn đề này. Thử nghiệm trên 4 mô hình phân lớp: *Logistic Regression, Support Vector Machine, Random Forest và Passive Aggressive Classifier* và sử dụng độ đo *Macro F1-score* để đánh giá cho bài toán mất cân bằng. Cuối cùng chúng tôi sử dụng *GridSearch* tinh chỉnh siêu tham số để có được mô hình tốt nhất.

Kết quả thu được với mô hình tốt nhất là *Logistic Regression* với accuracy bằng 0.717189 và macro F1-score bằng 0.522588.

2. NỘI DUNG

Chúng tôi thực hiện đồ án theo quy trình phân tích dữ liệu cụ thể cho bài toán như sau:



Hình 1. Quy trình phân tích dữ liệu

2.1. Giới thiệu các bộ dữ liệu

2.1.1. Bộ dữ liệu gốc

- Tên bộ dữ liệu: Global Landslide Catalog (GLC).
- Nguồn dữ liệu: https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog-Export/dd9e-wu2v
- Kích thước: **11033 x 31**, bao gồm 22 categorical features và 9 numerical features.
- Mô tả bộ dữ liệu: bộ dữ liệu GLC được biên soạn năm 2007 tại NASA Goddard Space Flight Center. Bộ dữ liệu chứa thông tin liên quan đến các sự kiện sạt lở đất trên toàn thế giới khoảng từ năm 1988 đến năm 2017.

STT	Tên thuộc tính	Kiểu dữ liệu	Mô tả	Miền giá trị
1	source_name	object	Tên báo đưa tin	9News, SkyNews,…
2	source_link	object	Liên kết dẫn đến tin	
3	event_id	int64	Mã sự kiện sạt lở đất	
4	event_date	object	Giờ/ngày/tháng/năm diễn ra sạt lở đất	1988-11-07 to 2017-09-28
5	event_time	float64	Giờ diễn ra sạt lở đất	
6	event_title	object	Tiêu đề tin tức sạt lở đất	Landslide in Shazi,…
7	location_description	object	Mô tả thông tin vị trí sạt lở	Tay Tra district,…
8	location_accuracy	object	Khoảng cách chênh lệch giữa vị trí ghi nhận so với vị trí thực tế	unknown, exact, 5km, 10km, 25km, 100km…
9	event_description	object	Mô tả sự kiện sạt lở đất	LANDSLIDES hit Surat …
10	landslide_category	object	Loại sạt lở đất	landslide, mudslide, debris flow,…
11	landslide_trigger	object	Nguyên nhân gây ra sạt lở đất	downpour, rain, continuousrain
12	landslide_size	object	Mức độ sạt lở đất	large, small, medium,…
13	landslide_setting	object	Môi trường xung quanh vị trí sạt lở đất	mine, above_road, natural_slope,...
14	fatality_count	float64	Số lượng người tử vong	0 to 5000
15	injury_count	float64	Số lượng người thương vong	0 to 374

16	storm_name	object	Tên cơn bão xảy ra trước khi sạt lở	Agaton
17	photo_link	object	Đường dẫn tới hình ảnh khu vực bị sạt lở	
18	notes	object	Ghi chú	
19	event_import_source	object	Nguồn cung cấp sự kiện sạt lở	glc, test, …
20	event_import_id	float64	Mã cung cấp sự kiện sạt lở	
21	country_name	object	Tên quốc gia nơi xảy ra sự kiện	Vietnam,…
22	country_code	object	Mã quốc gia	US, PH…
23	admin_division_name	object	Tên đơn vị hành chính	New York,…
24	admin_division_population	float64	Dân số của đơn vị hành chính	0 to 13M
25	gazeteer_closest_point	object	Vị trí trên bản đồ gần nơi xảy ra sạt lở nhất	Morongo Valley
26	gazeteer_distance	float64	Khoảng cách từ *"gazeteer_closest_point"* tới nơi xảy ra sạt lở	3e-5 to 215.45 (km)
27	submitted_date	object	Ngày nộp/hoàn thành sample trên dataset	2014-04-01 to 2017-11-21
28	created_date	object	Ngày tạo sample trên dataset	2017-11-20 to 2017-12-20
29	last_edited_date	object	Ngày cuối cùng chỉnh sửa sample trên dataset	2018-02-15
30	latitude	float64	Vĩ độ nơi xảy ra sạt lở	-46.77 to 72.62
31	longitude	float64	Kinh độ nơi xảy ra sạt lở	-179.98 to 179.99

2.1.2. Dữ liệu thu thập thêm

Kích thước các thuộc tính thu thập thêm: ***11033 x 36***, bao gồm 10 categorical features và 26 numerical features.

2.1.2.1. Nguồn thu thập và tham khảo:

- Thời tiết: visualcrossing.com/weather-api/
- Độ cao: developers.airmap.com/docs/elevation-api/
- Châu lục: pypi.org/project/pycountry-convert/

- Mùa: nationalgeographic.org/encyclopedia/season/
- Mật độ dân số: bộ dữ liệu *Gridded Population of the World Version 4.11* sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11/
- Độ che phủ rừng: bộ dữ liệu *Hansen Global Forest Change v1.8 (2000-2020)* data.globalforestwatch.org/documents/134f92e59f344549947a3eade9d80783/explore/
- Kết cấu của đất: bộ dữ liệu *OpenLandMap Soil Texture Class* developers.google.com/earth-engine/datasets/catalog/OpenLandMap_SOL_SOL_TEXTURE-CLASS_USDA-TT_M_v02

2.1.2.2. Phương pháp thu thập:

Đối với dữ liệu thời tiết, chúng tôi sử dụng api từ website *VisualCrossing* để lấy chỉ số thời tiết tại vị trí và thời điểm xảy ra sạt lở. Một vài chỉ số có tỉ lệ giá trị bị khuyết rất lớn và ít quan trọng nên chúng tôi quyết định không thu thập chúng. Về độ cao so với mực nước biển, chúng tôi sử dụng api từ *websiteAirmap* để lấy dữ liệu độ cao tại vị trí xảy ra sạt lở.

Đối với dữ liệu về mật độ dân số, độ che phủ rừng và kết cấu của đất, chúng tôi sử dụng thư viện *ee* trong package *geemap* để truy xuất dữ liệu từ các dataset tương ứng trên nền tảng Google Earth Engine.

Thuộc tính châu lục được tạo thêm dựa vào thuộc tính cơ sở là *'country'*. Thuộc tính mùa được tạo thêm dựa vào thuộc tính cơ sở là *'event_date'*, *'latitude'* và *'longitude'* theo các khái niệm về xuân phân, hạ chí, thu phân, đông chí ở hai nửa bán cầu.

*(Bảng mô tả thuộc tính của dữ liệu thu thập thêm tại **phụ lục A**)*

2.1.3. Kết hợp các bộ dữ liệu

- Bộ dữ liệu cuối cùng được kết hợp từ các bộ dữ liệu thành phần trên, dựa trên khóa là vị trí và thời điểm xảy ra sự kiện sạt lở đất *("latitude", "longitude", "event_date"*).
- Kích thước toàn bộ của bộ dữ liệu: 11033 x 67, bao gồm 32 categorical features và 35 numerical features.

2.2. Tiền xử lý dữ liệu

Chúng tôi sử dụng bộ dữ liệu được kết hợp cuối cùng cho bước phân tích thăm dò và phát triển mô hình. Trước khi tiền xử lý, bộ dữ liệu có tới 54 thuộc tính chứa giá trị bị khuyết và tỉ lệ dữ liệu bị khuyết giá trị là 12.18% với 90051 giá trị.

Chúng tôi tiến hành các bước tiền xử lý dữ liệu như sau:

2.2.1. Xử lý các thuộc tính dạng *datetime*

Định dạng lại các thuộc tính dạng datetime: '*event_date'*, '*created_date', 'last_edited_date', 'submitted_date', 'event_time'* thành kiểu dữ liệu datetime trong Python.

2.2.2. Loại bỏ các thuộc tính không cần thiết

Chúng tôi tiến hành loại bỏ các thuộc tính không cần thiết ứng với từng lí do sau:

Loại bỏ thuộc tính	Lý do loại bỏ
'source_link' (2), *'photo_link'* (17)	Vì là liên kết trang web
'event_id' (3), *'event_import_id'* (20)	Vì là mã định danh
'submitted_date' (27), *'created_date'* (28), *'last_edited_date'* (29)	Chỉ chứa các thông tin tracking sample trên dataset
'storm_name' (16), *'notes'* (18)	Số lượng sample là null trên 90%
'event_time' (5)	Trong 'event_date' đã chứa thông tin thời gian
'event_import_source' (19), *'source_name'* (1)	Chứa những thông tin không quan trọng.

2.2.3. Xử lý các giá trị bị khuyết

- *'country_name'* (21): sử dụng module *Nominatim* của thư viện *geopy* để định dạng toàn bộ tên quốc gia theo tiếng Anh cũng như xử lý tên quốc gia bị thiếu dựa trên toạ độ.
- *'admin_division_name'* (23) và '*location_description'* (7): điền giá trị bị khuyết bằng tên quốc gia tương ứng.
- *'gazeteer_closest_point'* (25): trong bộ dữ liệu, địa điểm này thường là tên thành phố hoặc tên thị trấn, tên ngôi làng… - nơi gần với sự kiện sạt lở xảy ra. Do đó, chúng tôi giải quyết địa điểm bị khuyết bằng tên đơn vị hành chính tương ứng tại nơi xảy ra sạt lở đó.
- Ở các thuộc tính chứa thông tin mô tả còn lại, thay thế giá trị bị khuyết là một chuỗi rỗng.
- *'admin_division_population'* (24): có mức độ tương quan vừa (~0.50) với mật độ dân số qua các năm, sử dụng phương pháp *KNN Imputation* để điền giá trị bị khuyết dựa trên mật độ dân số qua các năm.
- *'fatality_count'* (14), *'injury_count'* (15): giá trị trên 2 thuộc tính này là số nguyên và xuất hiện giá trị ngoại lai (outliers), do đó dùng giá trị *median* để thay thế cho những giá trị bị khuyết.

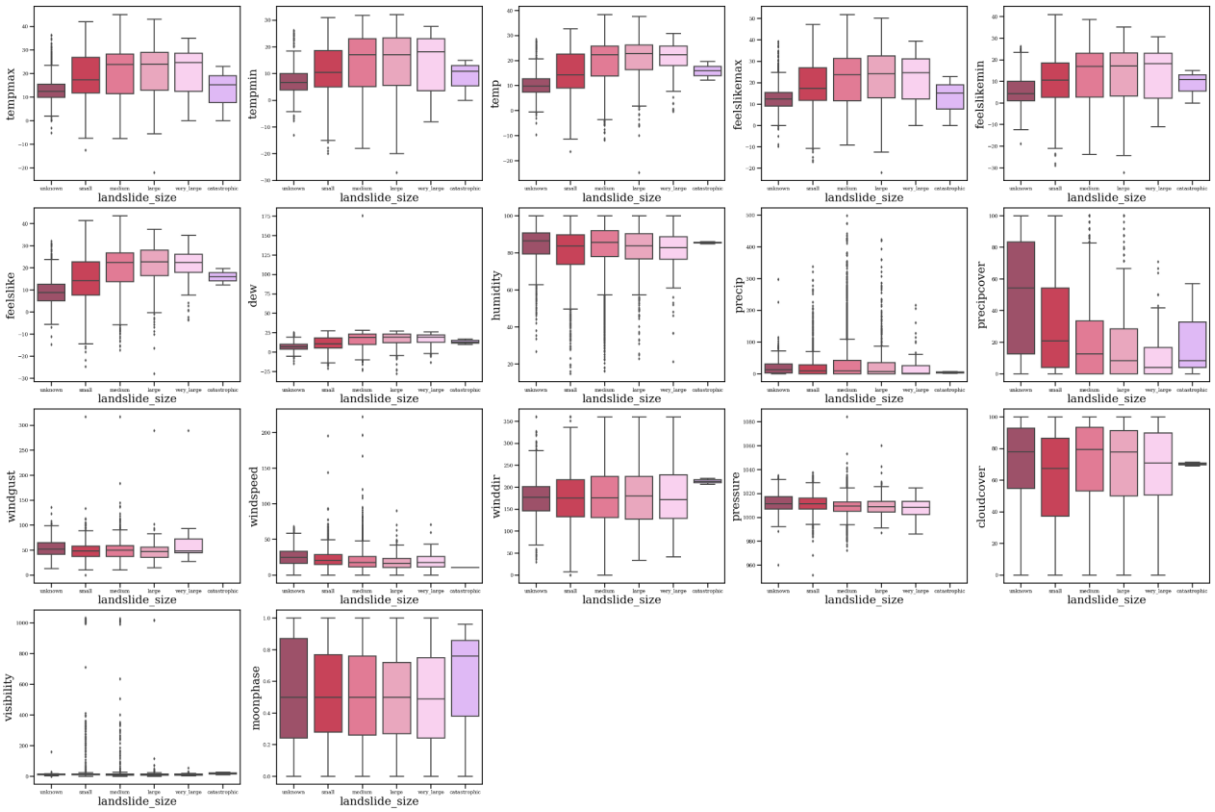
Với các thuộc tính là biến định lượng liên tục, chúng tôi plot phân bố giá trị của từng thuộc tính. Nếu thuộc tính tuân theo phân phối chuẩn, điền giá trị bị khuyết là giá trị *mean*. Nếu thuộc tính có nhiều outliers thì lựa chọn điền khuyết bằng giá trị *median* sẽ phù hợp hơn.

Cuối cùng, chúng tôi quyết định loại bỏ những sample có giá trị bị khuyết ở hầu hết các thuộc tính, cũng như vì chưa có cách điền khuyết phù hợp.

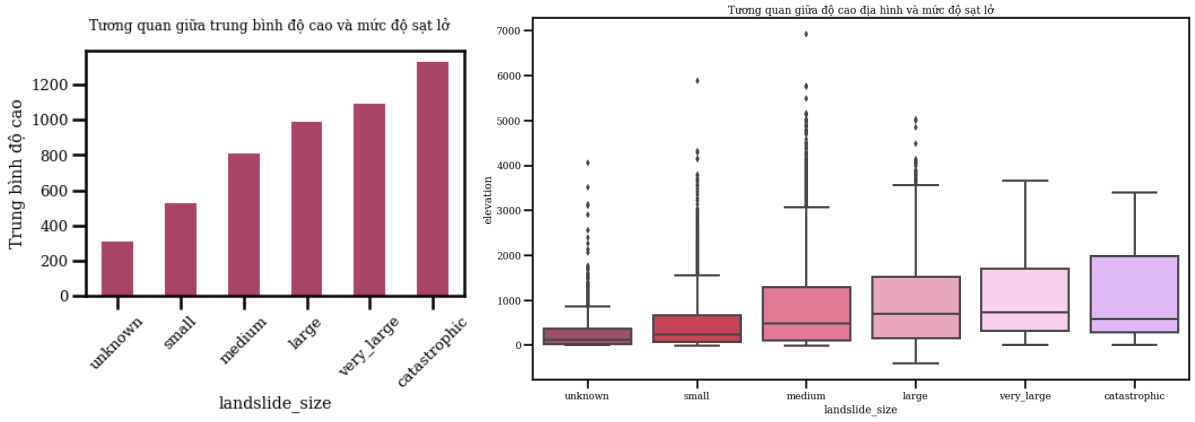
Kích thước của bộ dữ liệu sau khi xử lý các giá trị bị khuyết có được là ***9345 x 54*** so với bộ dữ liệu ban đầu là 11033 dòng và 67 cột.

2.3. Phân tích thăm dò

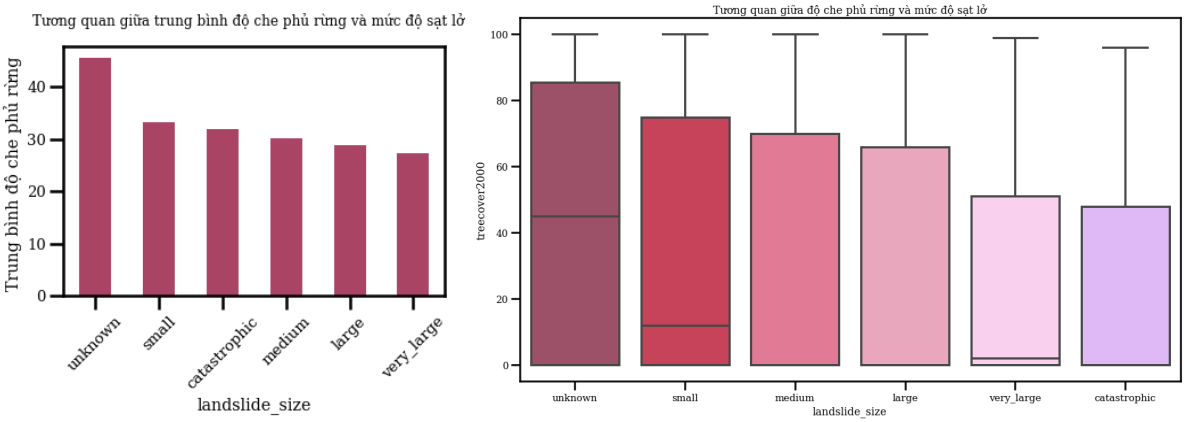
2.3.1. Thống kê mô tả



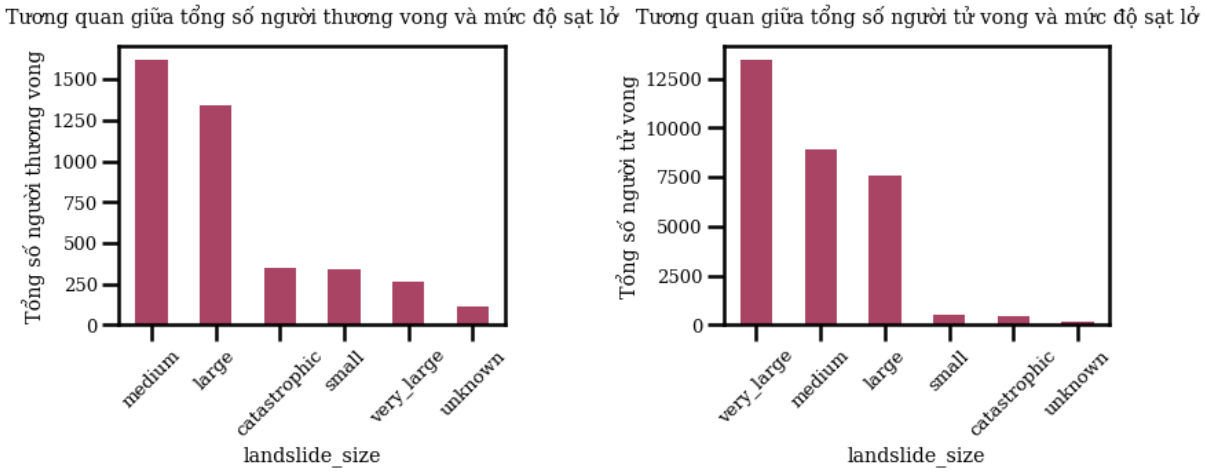
Boxplot các thuộc tính thời tiết đa số cho thấy có sự overlap lẫn nhau giữa các quy mô sạt lở, một số thuộc tính cho thấy một ít khác biệt như *'temp'*, *'feelslike'*, *'precipcover'*. Các khu vực có chỉ số *'temp'* và *'feelslike'* càng cao thì quy mô sạt lở xu hướng càng lớn, tuy nhiên vẫn xảy ra overlap giữa các quy mô sạt lở, ngoài ra chỉ số *'precipcover'* càng thấp thì có khả năng quy mô sạt lở càng nghiêm trọng.



Dựa vào boxplot và barplot tương quan giữa độ cao và quy mô sạt lở, chúng tôi nhận thấy có sự khác nhau về các độ cao địa hình giữa các quy mô sạt lở. Ở các khu vực có địa hình thấp dưới 800m đến dưới 1000m, quy mô sạt lở thường là vừa, nhỏ hoặc không xác định. Độ cao địa hình càng tăng thì quy mô sạt lở càng có xu hướng nghiêm trọng hơn, tuy nhiên chưa có sự chênh lệch đáng kể hay overlap về độ cao ở quy mô sạt lở vừa, lớn, rất lớn và thảm khốc.

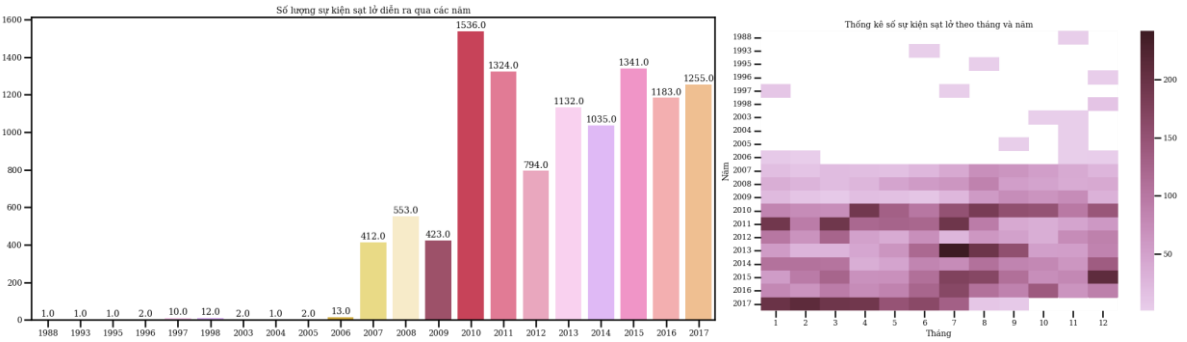


Nhìn chung quy mô sạt lở giảm dần ở các khu vực có độ che phủ rừng cao. Cho thấy được vai trò giữ đất của thảm thực vật, hạn chế nguy cơ xảy ra sạt lở nghiêm trọng.

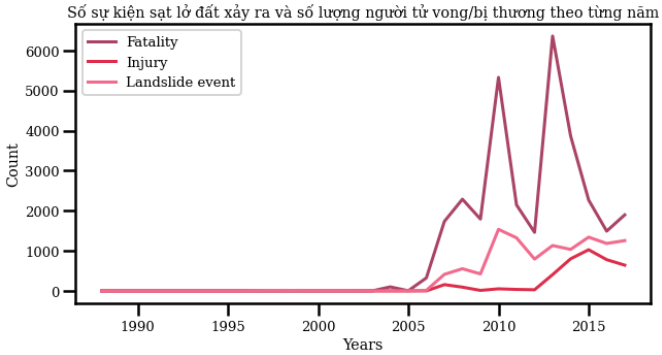


Các vụ sạt lở quy mô vừa và lớn khiến nhiều người thương-tử vong. Đặc biệt, sạt lở quy mô rất lớn gây thiệt hại về sinh mạng nhiều nhất trong khi số thương vong chỉ ở mức nhỏ.

2.3.2. Phân tích trực quan dữ liệu thời gian

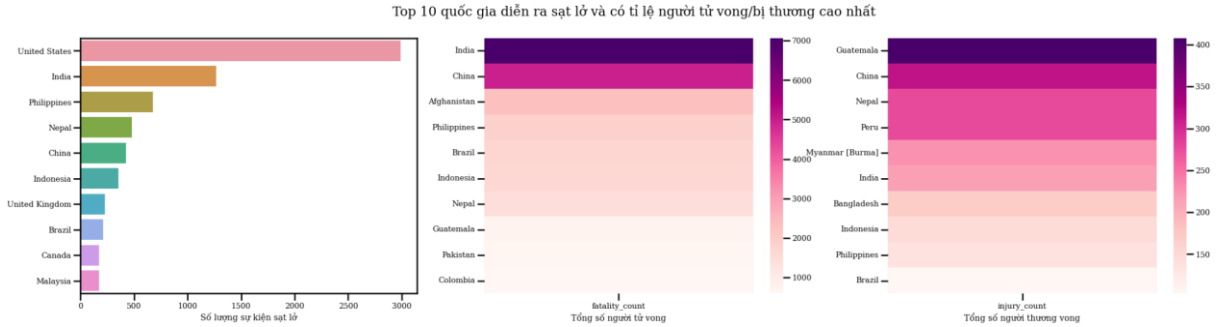


Như đã đề cập ở mục 2.1.1, trung tâm *Goddard Space Flight* của NASA bắt đầu biên soạn bộ dữ liệu vào năm 2007, do đó từ năm 2007 trở đi mới đầy đủ thông tin các sự kiện sạt lở. Đặc biệt, vào tháng 7 năm 2013 có hơn 200 sự kiện sạt lở trên toàn thế giới. Trước đó, đã diễn ra thảm họa sạt lở tại khu vực thung lũng Kedarnath, Uttarakhand, Ấn Độ vào ngày 16/06/2013 khiến hơn 5000 người thiệt mạng, nguyên nhân do mưa lớn kéo dài [4].

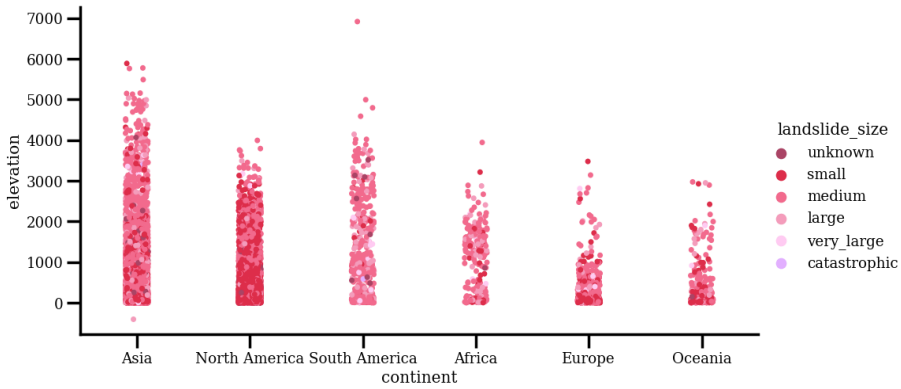


Năm 2010 và 2013 là năm có nhiều sự kiện sạt lở đất xảy ra, cũng như có nhiều số người tử vong.

2.3.3. Phân tích tổng hợp thuộc tính

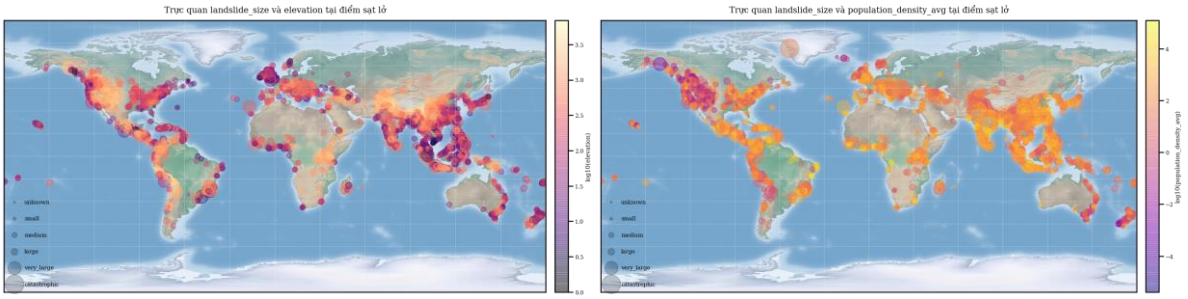


'United State' là quốc gia có nhiều sự kiện sạt lở đất diễn ra nhất, theo sau đó là *'India'* và *'Philippines'*. '*India'* và '*China'* là quốc gia có số lượng người tử vong do sạt lở đất nhiều nhất, '*Guatemala'* và '*Colombia'* có số lượng người bị thương do sạt lở đất nhiều nhất. Đặc biệt, '*United States'* có số lượng sự kiện sạt lở đất nhiều nhất nhưng lại không nằm trong top các quốc gia có số lượng người bị thương/tử vong nhiều vì đa số các sự kiện sạt lở diễn ra ở bờ Tây nước Mỹ - nơi dân cư thưa thớt (bản đồ mục *2.3.4*).



'North America' có số lượng sự kiện sạt lở nhiều thứ hai sau '*Asia'* nhưng mức độ ít nghiêm trọng hơn, đa số diễn ra ở quy mô vừa ('*medium'*), nhỏ ('*small'*) hoặc không xác định ('*unknown'*) ở địa hình thấp. Ở '*Asia'* cho thấy quy mô sạt lở trải đều ở các độ cao địa hình, chủ yếu quy mô vừa và lớn. *'South America'* và '*Africa'* cho thấy nhiều điểm sạt lở có quy mô lớn trở lên và trải đều các độ cao địa hình.

2.3.4. Phân tích trực quan trên bản đồ địa lý



Dựa vào bản đồ kết hợp với kết quả ở mục 2.3.1, một cách trực quan cho thấy được các điểm sạt lở ở khu vực Đông Nam Á, Đông Á, phía Đông Hoa Kỳ, Đông Úc, Vương quốc Anh, New Zealand xảy ra ở độ cao địa hình thấp, quy mô sạt lở phần lớn ở mức vừa và nhỏ. Xung quanh dãy Himalaya cho thấy quy mô sạt lở tương đối lớn.

Mật độ dân số cũng cho thấy ảnh hưởng nhất định đến mức độ sạt lở ở từng khu vực. Ở bờ Tây Hoa Kỳ, dân cư thưa thớt, sạt lở ít nghiêm trọng ở mức độ vừa và nhỏ. Khu vực Châu Á có mật độ dân số cao, quy mô sạt lở đa số từ mức độ vừa trở lên.

2.3.5. Tổng kết phân tích thăm dò

Sau đây là một số insights thu được từ bộ dữ liệu sau khi phân tích thăm dò:

- Kích thước của một sự kiện sạt lở đất hay gặp là *'medium'*, kích thước hiếm gặp nhất là *'very_large'* và *'catastrophic'*.
- Loại sạt lở đất hay gặp nhất là *'landslide'* và *'mudslide'*.
- Nguyên nhân gây ra sạt lở đất phần lớn là do mưa, nhiều nhất là *'downpour'*.
- Mùa hè là mùa thường xuyên xảy ra sạt lở đất nhất.
- North America và Asia là hai châu lục thường xuyên xảy ra sạt lở đất nhất.
- Tại United State, sạt lở đất diễn ra nhiều trong giai đoạn từ tháng 12 đến tháng 3, tại India là từ tháng 6 đến tháng 8, tại Vietnam là từ tháng 7 đến tháng 11.
- Các sự kiện sạt lở diễn ra vào mùa mưa bão, có bão tuyết.

Kết luận ảnh hưởng của các thuộc tính đến quy mô sạt lở:

- Nhìn chung, 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.
- Những khu vực có chỉ số nhiệt độ '*temp*' và '*feelikes*' cao, chỉ số *'precipcover'* thấp 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.
- Có sự khác nhau chưa rõ rệt về quy mô sạt lở giữa các châu lục.

Các kết luận trên chỉ là những cơ sở ban đầu để lựa chọn các thuộc tính quan trọng Việc lựa chọn các thuộc tính cuối cùng để đưa vào mô hình phân lớp sẽ sử dụng các kết quả đạt được ở phân tích thăm dò kết hợp thêm với phương pháp thống kê ở mục 2.4.

*(Hình ảnh từ các phân tích thăm dò bổ sung tại **phụ lục B**)*

2.4. Feature selection

Vì biến mục tiêu *'landslide_size'* là biến định tính, nên chúng tôi áp dụng hai phương pháp để lựa chọn đặc trưng là: *One Way ANOVA* và *Chi-Square Test*.

2.4.1. Lựa chọn biến dạng số (thuộc tính định lượng)

One Way ANOVA là phương pháp thống kê đo lường mối quan hệ tương quan *giữa biến định lượng và biến định tính*, bằng cách so sánh phương sai của biến định lượng theo từng lớp của biến định tính.

Khi áp dụng *One Way ANOVA* để tìm mối quan hệ tương quan giữa các thuộc tính định lượng so với biến mục tiêu *'landslide_size'*, kết quả trả về là 2 giá trị F-score và P-value cho mỗi thuộc tính. *F-score càng cao* và *P-value càng thấp* thì thuộc tính đó *càng quan trọng.* [2]

2.4.2. Lựa chọn biến phân loại (thuộc tính định tính)

Chi-Square Test là phương pháp thống kê đo lường mối quan hệ tương quan *giữa hai biến định lượng*, bằng cách tính phân phối giữa các giá trị trong hai biến.

Khi áp dụng *Chi-Square Test* để tìm mối quan hệ tương quan giữa các thuộc tính định tính so với biến mục tiêu *'landslide_size'*, kết quả trả về là 2 giá trị Chi-square và P-value cho mỗi thuộc tính. *Chi-square càng cao* và *P-value càng thấp* thì thuộc tính đó *càng quan trọng.* [3]

	Column name	F-Scores	P-values
0	longitude	517.812633	0.000000e+00
1	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

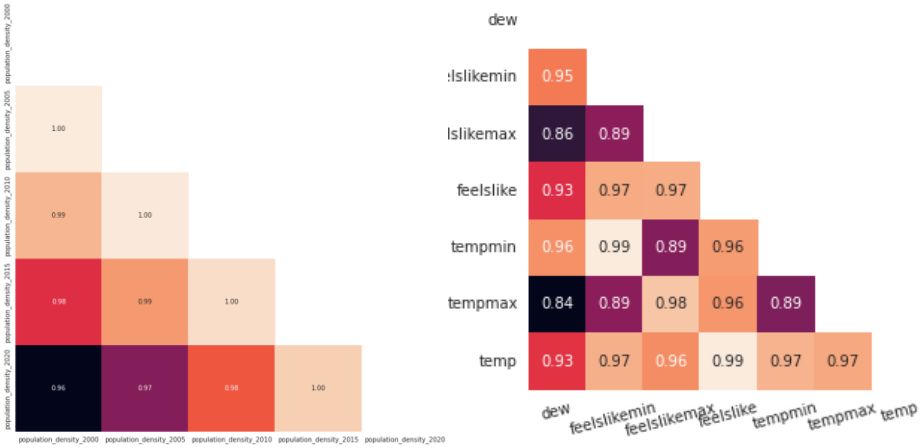
Hình 2. Xếp hạng các thuộc tính theo mức độ quan trọng của biến định lượng (trái) và biến định tính (phải).

Sau khi có được bảng xếp hạng mức độ quan trọng của các thuộc tính, chúng tôi chọn ngưỡng để lựa chọn thuộc tính đưa vào mô hình bằng cách sử dụng những thuộc tính quan trọng đã được tìm thấy sau khi phân tích thăm dò, kết hợp với việc thử nghiệm trên nhiều ngưỡng và đánh giá, lặp lại quá trình này cho đến khi chọn được một ngưỡng phù hợp nhất (cho ta *macro F1-score* cao nhất), chính là 16 biến định lượng (từ *'population_density_avg'* trở lên) và 8 biến định tính (từ '*continent'* trở lên) trong hai bảng trên.

2.5. Feature engineering

2.5.1. Kết hợp các nhóm thuộc tính có tương quan cao

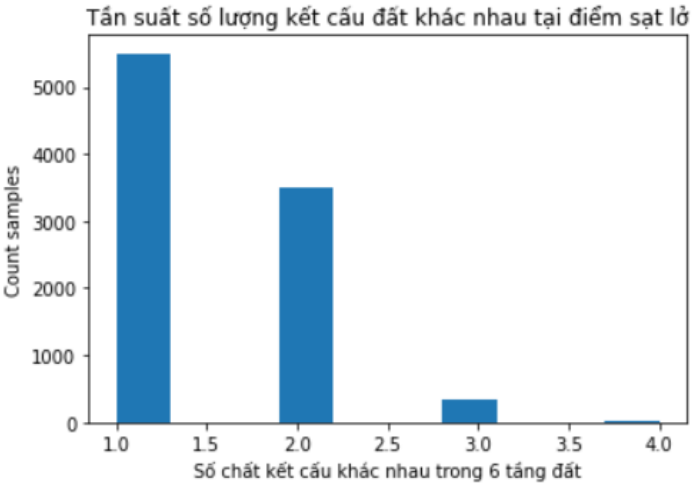
a) Biến dạng số (thuộc tính định lượng)



Dựa trên ma trận tương quan, ta thấy có 2 nhóm thuộc tính có tương quan rất mạnh lẫn nhau là:

- Nhóm thuộc tính về mật độ dân số: *'population_density_2000'*, đến *'population_density_2020'*.
- Nhóm thuộc tính về chỉ số thời tiết: *'dew', 'feelslikemin', 'feelslikemax', 'feelslike', 'tempmin', 'tempmax', 'temp'*.

Nên chúng tôi kết hợp hai nhóm bằng cách lấy *mean(),* thu được hai thuộc tính mới thay thế là *'population_density_avg'* và *'dew_temp_avg'*.



b) Biến dạng phân loại (thuộc tính định tính)

Theo phân tích thăm dò 6 cột về kết cấu của đất theo độ sâu (*'soil_texture_0', 'soil_texture_10', 'soil_texture_30', 'soil_texture_60', 'soil_texture_100', 'soiltexture_200'*), chúng tôi nhận thấy đa số các sample chỉ có từ 1 đến 2 loại đất ở cả 6 tầng đất, nên chúng tôi sử dụng *mode()* để kết hợp 6 cột trở thành một cột duy nhất là *'soil_texture'*: chứa loại đất xuất hiện nhiều nhất trong 6 tầng.

2.5.2. Kết hợp và mã hóa các cột chứa thông tin mô tả (text description)

- Các cột có chứa thông tin mô tả bao gồm: *'event_description', 'event_title'*, và *'location_description'* .
- Xử lý: kết hợp thành cột duy nhất là '*description'*, sử dụng *TF-IDF* để mã hóa dữ liệu văn bản về sparse vector theo unigram và bigram.

2.5.3. Mã hóa One-hot cho các thuộc tính dạng biến phân loại

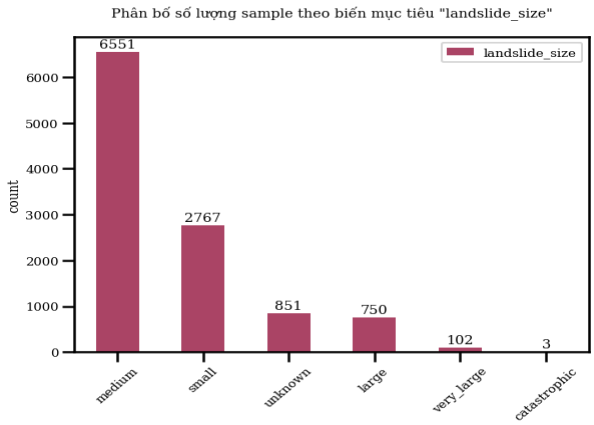
Các cột thuộc tính dạng biến phân loại được chuyển đổi về dạng 0, 1 bằng cách sử dụng hàm *get_dummies()* trong Pandas.

2.5.4. Feature scaling

- Áp dụng *MaxAbsScaler()* để chuẩn hóa miền giá trị của tất cả các thuộc tính về phạm vi [0, 1].
- Chúng tôi chọn *MaxAbsScaler()* do phương pháp này hoạt động tương tự như *MinMaxScaler()* nhưng phù hợp với dữ liệu thưa (sparse data) thu được sau khi mã hóa TF-IDF dữ liệu dạng văn bản hơn. Ngoài ra, nó không làm thay đổi hình dạng của dữ liệu.

2.6. Phát triển mô hình

2.6.1. Xử lý biến mục tiêu '*landslide_size'*



Dựa trên số lượng các mẫu theo từng lớp của *'landslide_size'*, chúng tôi nhận thấy đây là bài toán với dữ liệu mất cân bằng, đa số các điểm dữ liệu thuộc lớp '*medium'*, chỉ có một số ít điểm thuộc lớp *'very_large'* và *'catastrophic'*. Để mô hình phân lớp có thể hoạt động tốt, chúng tôi cần phải giải quyết vấn đề mất cân bằng trước. Chúng tôi đã giải quyết vấn đề này qua 2 bước:

- Gộp lớp *'catastrophic'* (thảm khốc) chung với lớp *'very_large'*

  Vì 2 lớp này mang ý nghĩa tương tự nhau, và số lượng mẫu thuộc lớp *'catastrophic'* và *'very_large'* rất ít (3 và 81 trên tổng 9345 điểm dữ liệu.
- Oversampling cho các lớp có ít điểm dữ liệu hơn

  Chúng tôi áp dụng kỹ thuật *ADASYN* để oversampling cho những lớp có ít điểm dữ liệu hơn. ADASYN là một kỹ thuật có thể bổ sung nhiều điểm dữ liệu hơn cho các lớp thiểu số (chẳng hạn như *"very_large"* và *"catastrophic"*), giúp chúng cân bằng hơn với lớp đa số ("*medium*") mà không làm thay đổi phân phối hay hình dạng của dữ liệu. [5]

2.6.2. Lựa chọn mô hình và độ đo đánh giá

- 4 mô hình phân lớp: *Logistic Regression, Support Vector Machine, Random Forest và PassiveAggressiveClassifier.*
- Độ đo đánh giá: *macro F1-score* và *accuracy*.

Trong đó chúng tôi tập trung vào *macro F1-score* để xếp hạng các mô hình bởi vì đây là độ đo không nhạy cảm sự mất cân bằng của các lớp dữ liệu, phù hợp với bài toán mất cân bằng mà chúng tôi đang giải quyết.

2.6.3. Tinh chỉnh mô hình

Sau khi thử nghiệm trên 4 loại mô hình, chúng tôi chọn ra mô hình tốt nhất để tiếp tục tối đa hóa kết quả bằng cách sử dụng *GridSearchCV* trong *Scikit Learn* để tinh chỉnh các hyperparameter cho mô hình tốt nhất.

2.7. Kết quả và đánh giá

Phương pháp	Mô hình	Macro F1-score	Accuracy
Trước khi xử lý mất cân bằng	Random Forest	0.481892	0.713623
	PassiveAggressiveClassifier	**0.519439**	0.715407
	Logistic Regression	0.511298	0.718973
	SVM	0.472977	0.717190
Sau khi xử lý mất cân bằng	Random Forest	0.498321	0.716476
	PassiveAggressiveClassifier	0.520779	0.711484
	Logistic Regression	**0.521931**	0.715050
	SVM	0.477818	0.706847
GridSearch	**Logistic Regression**	**0.522588**	**0.717189**

3. KẾT LUẬN

Sau khi quá trình thu thập, tiền xử lý và phân tích thăm dò, kết quả thu được cho thấy những thuộc tính chúng tôi thu thập thêm (*'dew_temp_avg', 'fatality_count', 'injury_count', 'precipcover', 'elevation', 'treecover2000', 'population_density_avg'*) có ảnh hưởng đến biến mục tiêu *'landslide_size'* và góp phần quan trọng trong việc phân tích cũng như phát triển mô hình phân lớp.

Sau khi kết hợp những thuộc tính được lựa chọn ra sau khi phân tích thăm dò được kết hợp với những thuộc tính quan trọng khi kiểm thử bằng *One Way ANOVA* và *Chi-Square*, chúng tôi đã lựa chọn ra được 24 thuộc tính quan trọng từ 67 thuộc tính ban đầu.

Việc áp dụng *ADASYN* để giải quyết vấn đề mất cân bằng dữ liệu đã giúp cho 4 mô hình đều đạt được độ chính xác cao hơn. Cùng với đó, chúng tôi sử dụng *Grid Search* để tinh chỉnh mô hình tốt nhất là *Logistic Regression* nhằm nâng cao hiệu suất mô hình, với kết quả cuối cùng là *marco F1-score* = 0.522588 và *accuracy* = 0.717189, cao hơn kết quả của mô hình trước khi xử lý mất cân bằng và trước khi tinh chỉnh mô hình.

TÀI LIỆU THAM KHẢO

[1] The Landslide Blog | The Kedarnath debris flow disaster in Uttarakhand https://blogs.agu.org/landslideblog/2013/06/21/the-kedarnath-debris-flow-disaster-in-uttarakhand/

[2] Toward Data Science | ANOVA for Feature Selection in Machine Learning https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476

[3] Toward Data Science | Using the Chi-Squared test for feature selection with implementation https://towardsdatascience.com/using-the-chi-squared-test-for-feature-selection-with-implementation-b15a4dad93f1

[4] Geographic Data with Basemap | Python Data Science Handbook

[5] Toward Data Science | ADASYN: Adaptive Synthetic Sampling method for imbalanced data https://towardsdatascience.com/adasyn-adaptive-synthetic-sampling-method-for-imbalanced-data-602a3673ba16

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Thị Minh Phương	- Lựa chọn thuộc tính - Phát triển mô hình - Viết báo cáo và làm slide
2	Chu Hà Thảo Ngân	- Tiền xử lý dữ liệu - Phân tích thăm dò - Viết báo cáo và làm slide
3	Thái Minh Triết	- Thu thập dữ liệu - Phân tích thăm dò - Viết báo cáo và làm slide

PHỤ LỤC

Link github của đồ án: https://github.com/minhphuongzzz/DS105-final-project

A. Mô tả thuộc tính của bộ dữ liệu thu thập thêm

1. Weather

STT	Tên thuộc tính	Kiểu dữ liệu	Mô tả	Miền giá trị
32	tempmax	float64	Nhiệt độ cao nhất trong ngày tại địa điểm sạt lở	-22.1 to 45.1 (ºC)
33	tempmin	float64	Nhiệt độ thấp nhất trong ngày tại địa điểm sạt lở	-27.1 to 32.1 (ºC)
34	temp	float64	Nhiệt độ tại địa điểm sạt lở	-24.8 to 38.4 (ºC)
35	feelslikemax	float64	Nhiệt độ cảm thấy cao nhất	-22.1 to51.9 (ºC)
36	feelslikemin	float64	Nhiệt độ cảm thấy thấp nhất	-32.3 to 41.0 (ºC)
37	feelslike	float64	Nhiệt độ cảm thấy	-28.0 to 43.6 (ºC)
38	dew	float64	Điểm sương (nhiệt độ hóa sương)	-27.9 to 175.5 (ºC)
39	humidity	float64	Độ ẩm tương đối	14.69 to 100.0 (%)
40	precip	float64	Lượng mưa tại thời điểm "datetime"	0.0 to 498.0 (mm)
41	precipcover	float64	Tỉ lệ số giờ trong ngày có lượng mưa khác 0	0.0 to 100.0 (%)
42	windgust	float64	Gió giật	0.0 to 316.8 (m/s)
43	windspeed	float64	Tốc độ gió trung bình trên 1 phút	0.0 to 222.1 (m/s)
44	winddir	float64	Hướng gió	0.0 to 360.0 (m/s)

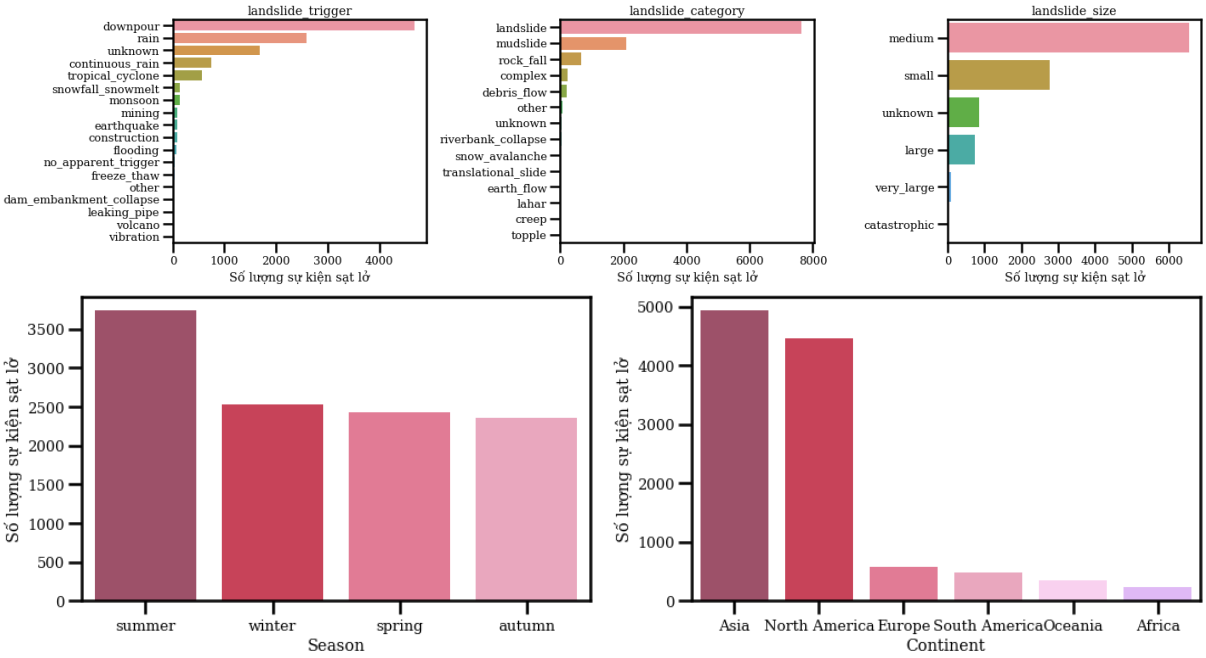
45	pressure	float64	Áp suất khí quyển theo mực nước biển hoặc khí áp	951.8 to 1084.0 (Pa)
46	cloudcover	float64	Tỷ lệ bầu trời bị che phủ bởi mây	0.0 to 100.0 (%)
47	visibility	float64	Tầm nhìn xa	0.0 to 1030.1 (km)
48	moonphase	float64	Tỉ lệ hình dạng của mặt trăng theo chu kỳ	0.0 to 1.0
49	conditions	object	Điều kiện thời tiết	Partially cloudy; Rain, Overcast; Snow;…
50	stations	object	Trạm thời tiết	

2. Các features khác

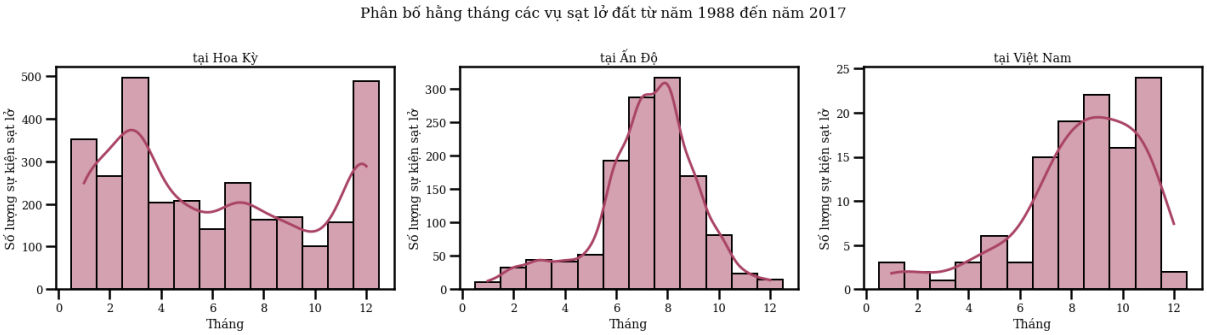
STT	Tên thuộc tính	Kiểu dữ liệu	Mô tả	Miền giá trị
51	elevation	int64	Độ cao so với mực nước biển (mét)	-407.0 to 6916.0 (m)
52	continent	object	Châu lục	Asia, North America, South America, Africa, Europe, Oceania
53	season	object	Mùa	spring, summer, autumn, winter
54	treecover2000	int64	Phần trăm độ phủ tán cây tại khu vực trong năm 2000 (áp dụng với thảm thực vật cao trên 5 mét)	0 to 100 (%)
55	loss	int64	Sự thay đổi từ forest sang non-forest trong khoảng thời gian từ năm 2000 - 2012	0 hoặc 1

56	gain	int64	Sự thay đổi toàn bộ từ non-forest sang forest trong khoảng thời gian từ năm 2000 - 2012	0 hoặc 1
57	soil_texture_0	object	Loại kết cấu đất ở bề mặt	- Lo - ClLo - SaClLo - SaLo - SaCl - SiLo - Cl - LoSa - SiClLo - Sa
58	soil_texture_10	object	Loại kết cấu đất ở độ sâu 10 cm	
59	soil_texture_30	object	Loại kết cấu đất ở độ sâu 30 cm	
60	soil_texture_60	object	Loại kết cấu đất ở độ sâu 60 cm	
61	soil_texture_100	object	Loại kết cấu đất ở độ sâu 100 cm	
62	soil_texture_200	object	Loại kết cấu đất ở độ sâu 200 cm	
63	population_density_2000	float64	Mật độ dân số tại khu vực vào năm 2000	0.0 to 110884.89 (người/km^2)
64	population_density_2005	float64	Mật độ dân số tại khu vực vào năm 2005	0.0 to 116474.90 (người/km^2)
65	population_density_2010	float64	Mật độ dân số tại khu vực vào năm 2010	0.0 to 121829.19 (người/km^2)
66	population_density_2015	float64	Mật độ dân số tại khu vực vào năm 2015	0.0 to 126877.76 (người/km^2)
67	population_density_2020	float64	Mật độ dân số tại khu vực vào năm 2020	0.0 to 130805.85 (người/km^2)

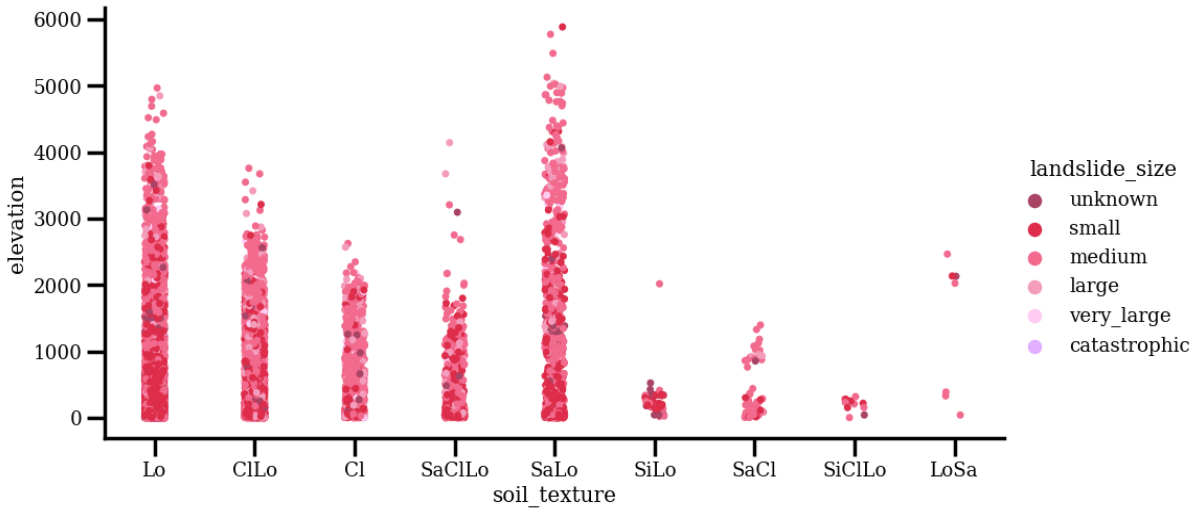
B. Hình ảnh từ phân tích thăm dò



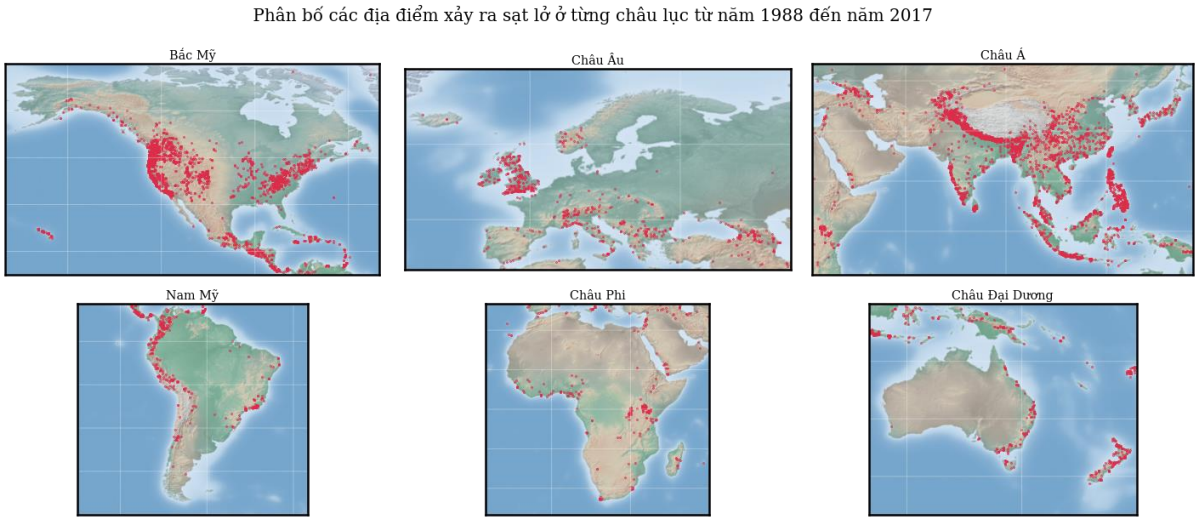
Phân bố số lượng sample của thuộc tính: landslide_trigger, landslide_category, landslide_size, season, continent



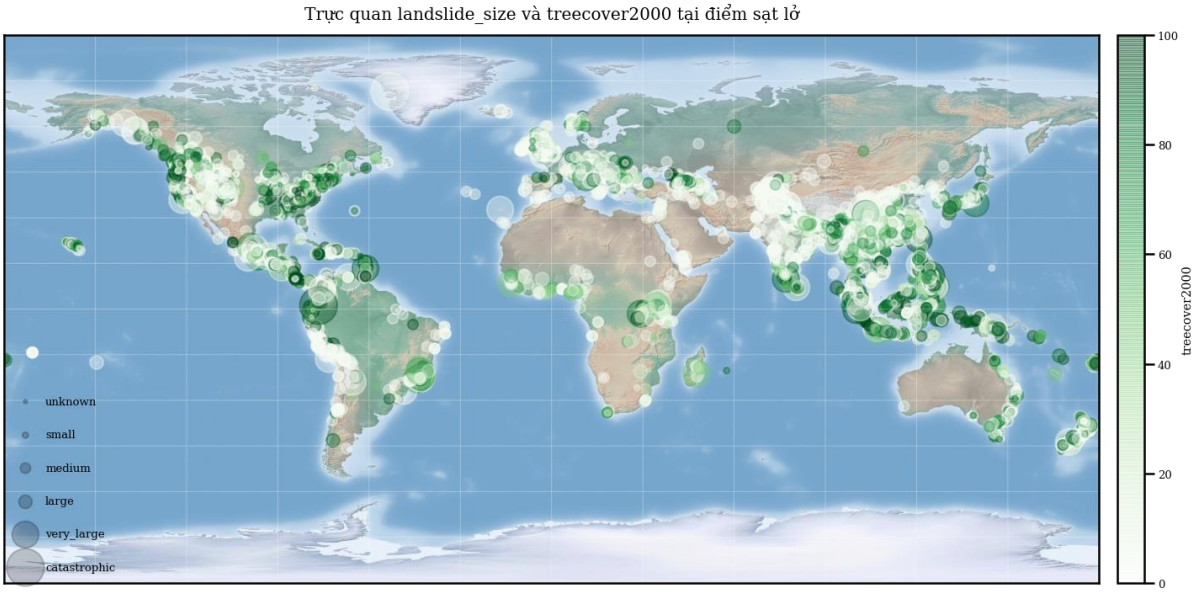
Phân bố hằng tháng các vụ sạt lở đất từ năm 1988 đến năm 2017 tại 3 quốc gia: United States, India, Vietnam. Những vụ sạt lở đất diễn ra vào mùa mưa bão, bão tuyết, mưa lớn kéo dài.



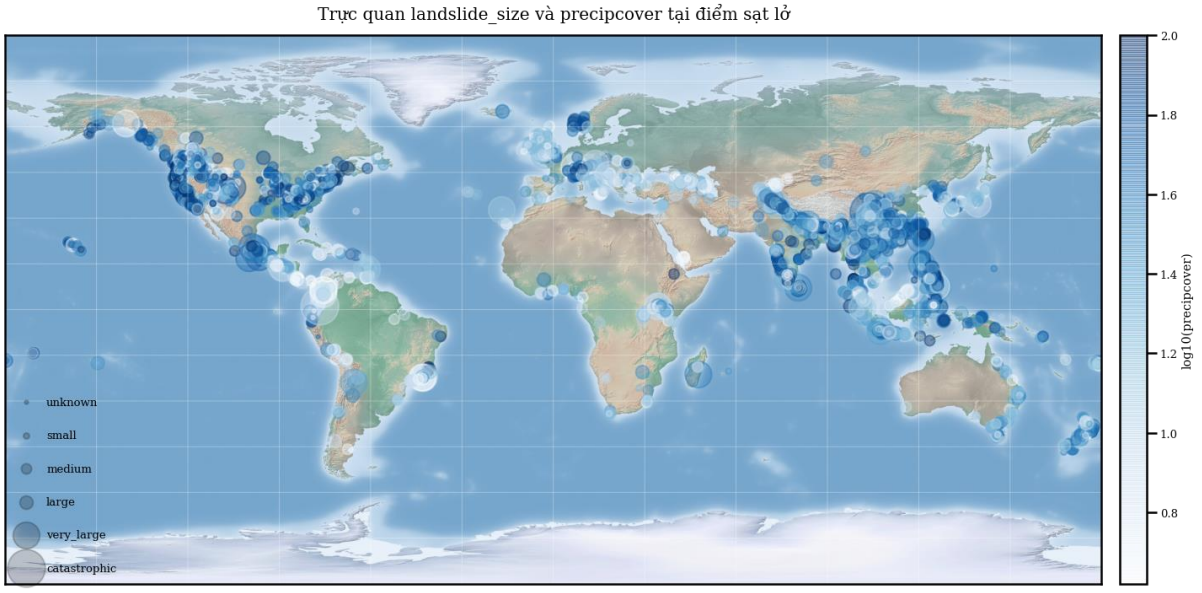
Độ cao địa hình và quy mô sạt lở đất theo từng loại kết cấu đất



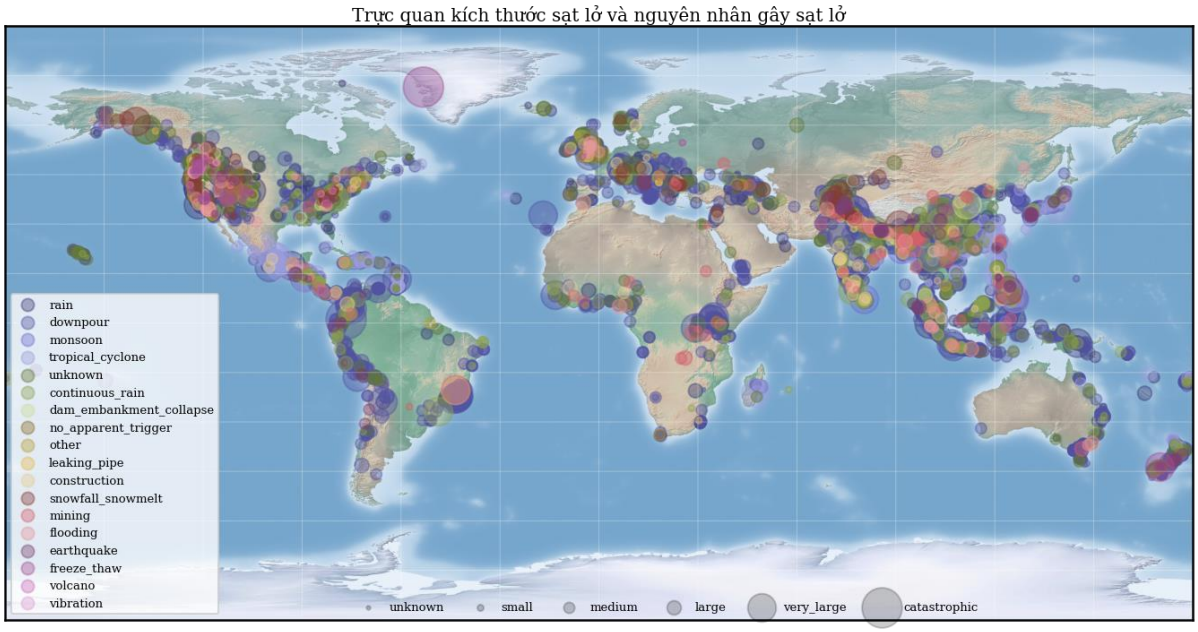
Phân bố các địa điểm xảy ra sạt lở ở từng châu lục từ năm 1988 đến năm 2017



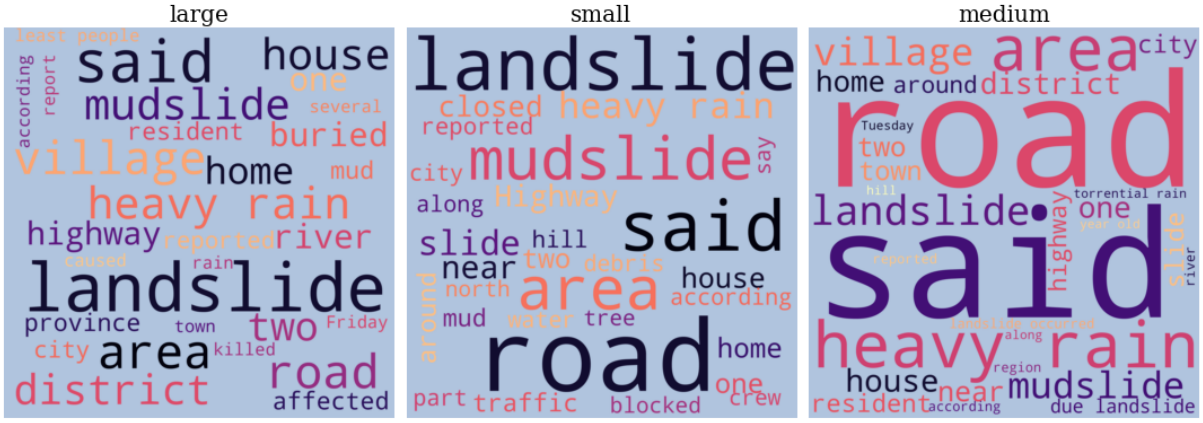
Trực quan mức độ sạt lở và độ che phủ rừng tại điểm sạt lở

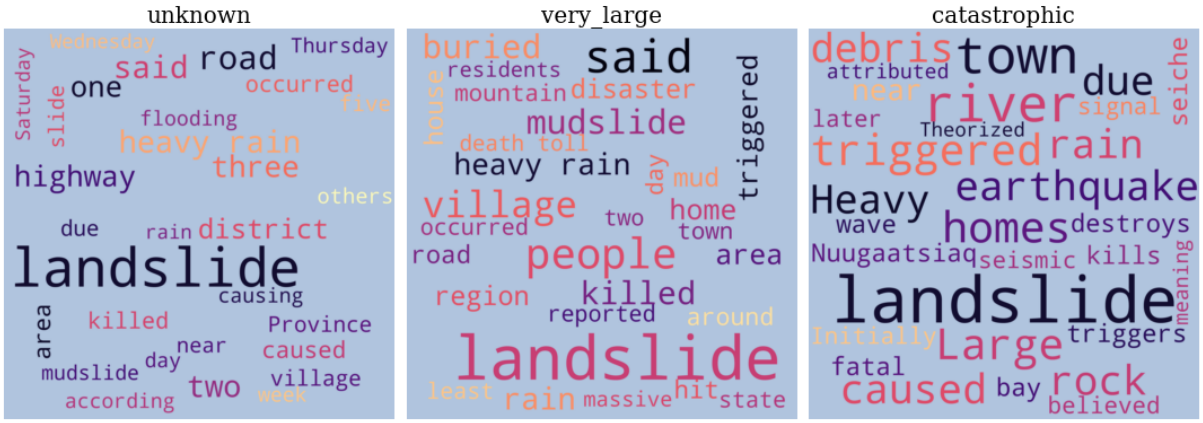


Trực quan mức độ sạt lở và lượng mưa tại điểm sạt lở



Trực quan mức độ sạt lở và nguyên nhân gây sạt lở





Wordcloud những từ xuất hiện nhiều ở các nhãn mức độ sạt lở đất (target) dựa trên 'event_description'