**BÁO CÁO TIẾN ĐỘ ĐỒ ÁN MÔN HỌC**

**THÔNG TIN NHÓM**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ và tên** | **Vai trò** |
| 1 | 19522065 | Nguyễn Thị Minh Phương | Nhóm trưởng |
| 2 | 19521882 | Chu Hà Thảo Ngân |  |
| 3 | 19522397 | Thái Minh Triết |  |

**TÊN ĐỀ TÀI:** PHÂN TÍCH VÀ DỰ ĐOÁN ẢNH HƯỞNG CỦA CÁC YẾU TỐ TỰ NHIÊN ĐẾN TÌNH TRẠNG SẠT LỞ ĐẤT.

**I. Nguồn dữ liệu**

**1. Bộ dữ liệu chính**

- Tên bộ dữ liệu: *Global Landslide Catalog (GLC).*

- Nguồn dữ liệu:<https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog-Export/dd9e-wu2v>

- 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.

**2. Dữ liệu thu thập thêm**

**2.1. Weather**

- Nguồn dữ liệu: <https://www.visualcrossing.com/weather-api>

- Mô tả dữ liệu: dữ liệu được thu thập thêm từ Weather API của *VisualCrossing*. Bộ dữ liệu có chứa các thông tin liên quan đến thời tiết tại thời điểm và địa điểm xảy ra sự kiện sạt lở đất trong bộ *GLC*.

**2.2. Elevation**

- Nguồn dữ liệu: https://developers.airmap.com/docs/elevation-api

- Mô tả dữ liệu: dữ liệu được thu thập từ Eleavation API của trang web Airmap. Chứa thông tin về độ cao so với mực nước biển cho hầu hết các vị trí địa lý trên Trái Đất. Độ đo được sử dụng là mét. Độ phân giải không gian: 1 arc-second (khoảng 30 mét) (arc-second cái này tìm hiểu thêm cho chắc ăn nha :>)

**2.3. Continent**

- Nguồn dữ liệu: package pycountry\_convert

- Mô tả dữ liệu: package sử dụng dữ liệu từ Wikipedia, cho phép thực hiện việc chuyển đổi giữa tên quốc gia (chuẩn ISO) sang mã quốc gia và châu lục. Các giá trị của châu lục thu được từ bộ dữ liệu gồm có:

* Asia: Châu Á
* North America: Bắc Mĩ
* South America: Nam Mĩ
* Europe: Châu Âu
* Africa: Châu Phi
* Oceania: Châu Đại Dương

**2.4. Season**

- Nguồn dữ liệu: https://www.nationalgeographic.org/encyclopedia/season/

- Mô tả dữ liệu: thu thập dựa vào thời điểm và vị trí địa lý so với đường xích đạo tại nơi xảy ra sự kiện sạt lở.

* Tại bắc bán cầu:

|  |  |
| --- | --- |
| **Season** | **Thời gian** |
| Spring | 21/03 – 20/06 |
| Summer | 21/06 – 22/09 |
| Autumn | 23/09 – 20/12 |
| Winter | 21/12 – 20/03 năm sau |

* Tại nam bán cầu

|  |  |
| --- | --- |
| **Season** | **Thời gian** |
| Autumn | 21/03 – 20/06 |
| Winter | 21/06 – 22/09 |
| Spring | 23/09 – 20/12 |
| Summer | 21/12 – 20/03 năm sau |

**II. Mô tả chi tiết từng feature**

**1. *Global Landslide Catalog (GLC)***

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Name** | **Datatype** | **Description** |
| 1 | source\_name | object | Tên báo đưa tin |
| 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 |
| 5 | event\_time | float64 | Giờ diễn ra sạt lở đất |
| 6 | event\_title | object | Tiêu đề tin tức sạt lở đất |
| 7 | location\_description | object | Mô tả thông tin vị trí sạt lở |
| 8 | location\_accuracy | object | Khoảng cách chênh lệch giữa vị trí ghi nhận so với vị trí thực tế |
| 9 | event\_description | object | Mô tả sự kiện sạt lở đất |
| 10 | landslide\_category | object | Loại sạt lở đất |
| 11 | landslide\_trigger | object | Nguyên nhân gây ra sạt lở đất |
| 12 | landslide\_size | object | Mức độ sạt lở đất |
| 13 | landslide\_setting | object | Môi trường xung quanh vị trí sạt lở đất |
| 14 | fatality\_count | float64 | Số lượng người tử vong |
| 15 | injury\_count | float64 | Số lượng người thương vong |
| 16 | storm\_name | object | Tên cơn bão xảy ra trước khi sạt lở |
| 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 |
| 20 | event\_import\_id | float64 | Mã cung cấp sự kiện sạt lở |
| 21 | country\_name | object | Tên quốc gia nơi xảy ra sự kiện |
| 22 | country\_code | object | Mã quốc gia |
| 23 | admin\_division\_name | object | Tên đơn vị hành chính |
| 24 | admin\_division\_population | float64 | Dân số của đơn vị hành chính |
| 25 | gazeteer\_closest\_point | object | Vị trí trên bản đồ gần nơi xảy ra sạt lở nhất |
| 26 | gazeteer\_distance | float64 | Khoảng cách từ "gazeteer\_closest\_point" tới nơi xảy ra sạt lở |
| 27 | submitted\_date | object | Ngày nộp/hoàn thành sample trên dataset |
| 28 | created\_date | object | Ngày tạo sample trên dataset |
| 29 | last\_edited\_date | object | Ngày cuối cùng chỉnh sửa sample trên dataset |
| 30 | latitude | float64 | Vĩ độ nơi xảy ra sạt lở |
| 31 | longitude | float64 | Kinh độ nơi xảy ra sạt lở |

**2. Weather**

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Name** | **Datatype** | **Description** |
| 1 | event\_id | int64 | Mã sự kiện sạt lở đất |
| 2 | datetime | object | Ngày/giờ của thời tiết được yêu cầu |
| 3 | datetimeEpoch | int64 | Số giây kể từ 01/01/1970 đến “datetime” |
| 4 | tempmax | float64 | Nhiệt độ cao nhất trong ngày tại địa điểm sạt lở |
| 5 | tempmin | float64 | Nhiệt độ thấp nhất trong ngày tại địa điểm sạt lở |
| 6 | temp | float64 | Nhiệt độ tại địa điểm sạt lở |
| 7 | feelslikemax | float64 | Nhiệt độ cảm thấy cao nhất (ºC) |
| 8 | feelslikemin | float64 | Nhiệt độ cảm thấy thấp nhất (ºC) |
| 9 | feelslike | float64 | Nhiệt độ cảm thấy (ºC) |
| 10 | dew | float64 | Điểm sương (nhiệt độ hóa sương) |
| 11 | humidity | float64 | Độ ẩm tương đối (tính theo %) |
| 12 | precip | float64 | Lượng mưa tại thời điểm “datetime” |
| 13 | precipprob | float64 | (forecase only) |
| 14 | precipcover | float64 | Tỉ lệ số giờ trong ngày có lượng mưa khác 0 |
| 15 | preciptype | float64 | (forecase only) |
| 16 | snow | float64 | Lượng tuyết rơi |
| 17 | snowdepth | float64 | Độ dày của lớp tuyết trên mặt đất |
| 18 | windgust | float64 | Gió giật |
| 19 | windspeed | float64 | Tốc độ gió trung bình trên 1 phút |
| 20 | winddir | float64 | Hướng gió |
| 21 | pressure | float64 | Áp suất khí quyển theo mực nước biển hoặc khí áp (tính theo đơn vị millibars hay hectopascals) |
| 22 | cloudcover | float64 | Tỉ lệ bầu trời bị che phủ bởi mây (%) |
| 23 | visibility | float64 | Tầm nhìn xa (km) |
| 24 | solarradiation | float64 | Bức xạ mặt trời (W/m^2) |
| 25 | solarenergy | float64 | Tổng năng lượng mặt trời tích tụ (MJ hoặc kWh) |
| 26 | uvidex | float64 | Mức độ của tia UV (giá trị từ 0 đến 10 thể hiện mức độ từ thấp đến cao) |
| 27 | sunrise | object | Thời gian mặt trời mọc |
| 28 | sunriseEpoch | int64 | Thời gian mặt trời mọc tính bằng số giây kể từ 01/01/1970 |
| 29 | sunset | object | Thời gian mặt trời lặn |
| 30 | sunsetEpoch | int64 | Thời gian mặt trời lặn tính bằng số giây kể từ 01/01/1970 |
| 31 | moonphase | float64 | Tỉ lệ hình dạng của mặt trăng theo chu kỳ |
| 32 | conditions | object | Điều kiện thời tiết |
| 33 | description | object | Mô tả chi tiết điều kiện thời tiết |
| 34 | icon | object | Dùng để hiển thị biểu tượng |
| 35 | stations | object | Trạm thời tiết |

**3. Elevation, continent, season**

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | elevation | int64 |  |
| 2 | continent | object |  |
| 3 | season | object |  |

**III. EDA**

Kích thước dữ liệu: 11033 x 53

Tỉ lệ dữ liệu bị khuyết giá trị: 15.03 %

1. **Landslide**

**Bảng thống kê mô tả giữa số lượng người tử vong, thương vong và dân số địa phương**

|  |  |  |  |
| --- | --- | --- | --- |
|  | fatality\_countfloat64 | injury\_countfloat64 | admin\_division\_populationfloat64 |
| count | 9648 | 5359 | 9471 |
| mean | 3.2194237147595355 | 0.7518193692853145 | 157760.04582409462 |
| std | 59.886178345211086 | 8.458955199140028 | 829734.5446186023 |
| min | 0 | 0 | 0 |
| 25% | 0 | 0 | 1963 |
| 50% | 0 | 0 | 7365 |
| 75% | 1 | 0 | 34021 |
| max | 5000 | 374 | 12691836 |

**Tiền xử lý sơ bộ**

* Bỏ bớt các cột không có giá trị:

'event\_id','source\_name','source\_link', 'photo\_link', 'notes', 'event\_import\_source', 'event\_import\_id', 'country\_code', 'submitted\_date', 'created\_date', 'last\_edited\_date'

* Xử lý missing tên quốc gia bị thiếu bằng cách tìm tên quốc gia dựa trên toạ độ (*Nominatim API*)
* Drop 1 số dòng mà thuộc tính có missing value ('landslide\_category’, 'stations'…)
* Điền giá trị khuyết cho số lượng người thương vong, tử vong, dân số địa phương
* Xử lý một số thuộc tính có outliers (thuộc tính số lượng người thương vong, tử vong, dân số địa phương)

**Target: 'landslide\_trigger' (nguyên nhân gây sạt lở)**

Chart

Description automatically generated

* Số lượng label ban đầu: 18
* Gom một số label về chung 1 label (ví dụ gom các loại mưa về chung label ‘*rain’*)
* Số lượng label sau khi xử lý: 12
* Số lượng label ‘rain’ mất cân bằng => sử dụng phương pháp giải quyết mất cân bằng như SMOTE, ADASYN

**Target: 'landslide\_size’ (mức độ sạt lở):**

Chart, funnel chart

Description automatically generated

* Số lượng label ban đầu: 6
* Gom label ‘*very\_large*’ chung với label ‘*large’*, bỏ label ‘*catastrophic’*
* Số lượng label sau khi xử lý: 4

**Loại sạt lở đất (landslide\_category)**

**Chart, bar chart

Description automatically generated**

**Số lượng những sự kiện sạt lở diễn ra hằng năm**

Chart, bar chart, histogram

Description automatically generated

* Số lượng sự kiện sạt lở bắt đầu tăng từ năm 2007
* Hoặc từ năm 2017 NASA mới bắt đầu thu thập các thông tin sạt lở

**Ma trận tương quan các thuộc tính**

A picture containing text

Description automatically generated

**Ma trận tương quan giữa thương-tử vong và dân số địa phương**

Giữa thương-tử vong và dân số địa phương (đơn vị hành chính) không có tương quan.

Khu vực dân số tuy đông nhưng số thương-tử vong do sạt lở cũng thấp

Chart, waterfall chart, treemap chart

Description automatically generated

Chart

Description automatically generated

**Top 10 quốc gia hay diễn ra sạt lở**

Chart, bar chart

Description automatically generated

**Phân bố hàng tháng các lần xảy ra sạt lở đất ở một số quốc gia**

Tháng diễn ra sạt lở nhiều nhất:

* Tại Mỹ: từ tháng 12 - tháng 3
* Tại Ấn Độ: tháng 6 - tháng 8
* Tại Việt Nam: tháng 7 - tháng 11
* Diễn ra vào mùa mưa bão, bão tuyết

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

1. **Weather Playground**

Kích thước bộ dữ liệu sau khi xử lý là: … 9703 x 42