

# DATASET: TSUNAMIS OCURRIDOS A NIVEL MUNDIAL

Marta Martin Llambes

## 1. SUBTÍTULO

El *dataset* recoge datos de tsunamis ocurridos a nivel mundial desde el año 2000 antes de Cristo, hasta el año 2018. Entre dichos datos se encuentran por ejemplo la fecha, la localización, la altura máxima, la causa del tsunami, etc.

## 2. IMAGEN

La imagen identificativa escogida para el dataset, ha sido la siguiente:



<https://pixabay.com/es/st-peter-ording-tsunami-maremoto-1897303/>

### 3. CONTEXTO

Los datos extraídos corresponden a las informaciones que se tienen de los diferentes tsunamis que han ocurrido en el mundo durante la historia de la humanidad. Los datos proceden de todo tipo de fuentes, científicas y no científicas, ya que recopilan informaciones desde tiempos anteriores al nacimiento de Cristo.

### 4. CONTENIDO

El *dataset* recoge una colección de datos sobre 2627 tsunamis (resultado a día 25 de octubre), ocurridos entre el año 2000 a.C y el año 2018 alrededor del mundo, y se encuentran ordenados de más antiguos a más recientes.

Para cada tsunami, no se han extraído todos los datos, sino los que parecían más interesantes:

- **year:** año en el que sucedió el tsunami.
- **month:** mes en el que sucedió el tsunami.
- **day:** día en el que sucedió el tsunami.
- **validity:** con un valor en la escala de -1 a 4, define cuán probable es que el evento registrado fuera realmente un tsunami.

Valor	Significado
4	Definitivamente fue un tsunami.
3	Probablemente fue un tsunami.
2	Dudablemente fue un tsunami.
1	Muy dudablemente fue un tsunami.
0	Evento que solo causó un seiche o perturbación en un río de interior.
-1	Entrada errónea.

- **causeCode:** con un valor del 0 al 11 define la causa del tsunami.

Valor	Causa del tsunami
0	Desconocida.
1	Terremoto.
2	Terremoto cuestionable.
3	Terremoto y deslizamiento de tierra.
4	Volcán y terremoto.
5	Volcán, terremoto y deslizamiento de tierra.
6	Volcán.
7	Volcán y deslizamiento de tierra.
8	Deslizamiento de tierra.

<b>9</b>	Meteorológica.
<b>10</b>	Explosión.
<b>11</b>	Marea astronómica.

- **earthquakeMagnitude**: si la causa del tsunami fue un terremoto, especifica de qué magnitud fue éste.
- **locationCountry**: qué país sufrió el tsunami.
- **locationName**: localización aproximada donde sucedió el tsunami.
- **locationLat**: coordenada geográfica de latitud donde sucedió el tsunami.
- **locationLong**: coordenada geográfica de longitud donde sucedió el tsunami.
- **maxWaterHeight**: altura máxima que alcanzaron las olas del tsunami.
- **runups**: número de localizaciones afectadas por olas del mismo tsunami.
- **numDeaths**: cuántas personas murieron a causa del tsunami.
- **numInjuries**: cuántos heridos hubo a causa del tsunami.
- **descrDamage**: escala que evalúa los daños económicos aproximados causados por el tsunami.

Valor	Daños
<b>0</b>	Ninguno
<b>1</b>	Limitados (menos de 1 millón de dólares)
<b>2</b>	Moderados (de 1 a 5 millones de dólares)
<b>3</b>	Severos (de 5 a 24 millones de dólares)
<b>4</b>	Extremos (más de 25 millones de dólares)

- **numDestroyedHouses**: número de casas destruidas por el tsunami.

Algunos de los anteriores atributos aparecen frecuentemente vacíos, pero será tarea de quien quiera usar el *dataset* decidir cuáles utilizar y cuáles no.

Los datos del *dataset* creado provienen de la base de datos de tsunamis de la organización estadounidense NOAA/DWS. Dichos datos los han ido recogiendo tanto de fuentes científicas como académicas, de catálogos regionales y mundiales, de mareógrafos, de sensores en aguas profundas, de informes de eventos individuales y de trabajos no publicados.

Para crear el *dataset*, se han obtenido de la base de datos mencionada todos los datos de los que disponen sin ningún filtraje, mediante un script en Python y las técnicas de *Web Scraping* para páginas HTML. Finalmente, se han guardado los datos recolectados en un archivo csv.

## 5. AGRADECIMIENTOS

Los datos han sido extraídos de la “NGDC/WDS Tsunami Event Database”, a la cual se puede acceder mediante este enlace:

<https://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=70&d=7>

Dicha base de datos pertenece a la administración estadounidense NOAA (National Oceanic and Atmospheric Administration).

La referencia bibliográfica a dicha base de datos es la siguiente:

*National Geophysical Data Center / World Data Service (NGDC/WDS): Global Historical Tsunami Database. National Geophysical Data Center, NOAA. [doi:10.7289/V5PN93H7](https://doi.org/10.7289/V5PN93H7) [25th October 2018]*

## 6. INSPIRACIÓN

Este conjunto de datos puede ser interesante para diferentes tipos de estudios relacionados con tsunamis. Se pueden llevar a cabo varios análisis descriptivos como, por ejemplo, en qué zonas geográficas se produce un mayor número de tsunamis, cuáles son las causas más comunes que originan tsunamis, etc.

También imagino que se podría intentar llevar a cabo análisis más complejos, como intentar identificar patrones, por ejemplo, relacionando la causa del tsunami o magnitud del terremoto que lo origine con la altura máxima de sus olas, intentar hacer predicciones de futuros tsunamis, o predecir un modelo que nos ayude a avisar cuanto antes a poblaciones que vayan a sufrir un tsunami para que tengan un mayor tiempo de evacuación, etc.

## 7. LICENCIA

La licencia escogida para el *dataset* es la **CC BY-NC-SA 4.0**. Las razones por las que se ha decidido escoger dicha licencia son:

- Permite a otras personas tanto compartir los datos (copiarlos y distribuirlos en cualquier medio o formato), como adaptarlos (transformarlos, limpiarlos...).
- No permite que se utilicen los datos para un uso comercial. Como después de investigar la página web de donde se han obtenido los datos de la base de datos no se ha conseguido averiguar qué política quieren seguir en este ámbito, se ha considerado que prohibir el uso de los datos con fines comerciales sería lo más seguro. De este modo, los datos se seguirán usando en ámbitos más educativos que comerciales.

- Quien quiera usar los datos tiene que referenciarnos, así como indicar los cambios que realizó a los datos.
- Quien modifique y utilice los datos para sus propios estudios, los puede volver a compartir, así como su trabajo realizado, pero solo bajo la misma licencia que hemos escogido. De este modo, se garantiza que se sigan cumpliendo las condiciones que hemos escogido.

## 8. CÓDIGO

El *dataset* se ha generado mediante código Python, y dicho código puede encontrarse en el siguiente enlace:

<https://github.com/mmartinlla/tsunami-scraper/tree/master/python%20code>

## 9. DATASET

El *dataset* obtenido en formato CSV se puede encontrar en el siguiente enlace:

<https://github.com/mmartinlla/tsunami-scraper/tree/master/csv%20dataset>

## 10. RECURSOS

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- National Geophysical Data Center / World Data Service (NGDC/WDS): Global Historical Tsunami Database. National Geophysical Data Center, NOAA. doi:10.7289/V5PN93H7 [25th October 2018]
- <https://creativecommons.org/licenses/>
- <https://creativecommons.org/licenses/by-nc-sa/4.0/>
- <https://pixabay.com/es/st-peter-ording-tsunami-maremoto-1897303/>
- <https://github.com/tteguayco/Web-scraping>
- <https://github.com/rafoelhonrado/foodPriceScraper>