Nuria García Fuentes y Alicia Perdices Guerra

Memoria web scraping

Práctica 1

**Índice**

|  |  |
| --- | --- |
| CONTENIDO | Págs. |
| Contexto | 2 |
| Título dataset | 2 |
| Descripción dataset | 2-3 |
| Representación gráfica dataset | 3 |
| Contenido dataset | 3-6 |
| Agradecimientos propietarios | 6-7 |
| Inspiración. Preguntas que responde | 7-8 |
| Licencia | 8-9 |
| Código Python | 9-10 |
| Publicación Zenodo | 10-11 |
| Entrega GitHub | 11 |
| Referencias | 12-13 |

**Descripción de la Práctica a realizar**

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

* Contexto. Explicar en qué contexto se ha recolectado la información. Explique

por qué el sitio web elegido proporciona dicha información.

En el contexto de la práctica propuesta, y motivadas sobre todo por las ventajas y beneficios que el estudio de datos procedentes de enfermedades y de desastres naturales puede aportar a la comunidad científica, no solo para la prevención, sino para establecer protocolos, hemos decidido estudiar y hacer web scraping de los siguientes recursos:

[https://data.europa.eu/euodp/es/data/publisher/ecdc](https://www.google.com/url?q=https://data.europa.eu/euodp/es/data/publisher/ecdc&sa=D&source=hangouts&ust=1584813221920000&usg=AFQjCNFXaBDxR7QEoA9NpKMW7cIcEjJQbQ)

<https://reliefweb.int/disasters>

Por un lado, se ha elegido **El portal de datos abiertos de la UE**, creado en 2012, ya que ofrece una gran variedad de datos (geográficos, geopolíticos, financieros, estadísticos, resultados electorales, salud, jurídicos, medio ambiente, transporte e investigación científica) de libre acceso, incluidos en un catálogo normalizado, APIs que facilitan el acceso a éstos y una lista de herramientas y aplicaciones que permiten reutilizarlos.

Por otro lado, **ReliefWeb**, como el portal de información humanitaria más grande del mundo, ofrece información independiente y a medida que se desarrollan crisis humanitarias, además de poner el foco sobre las que sucedieron en el pasado proporcionando información muy útil en los protocolos de acción de las actuales.

Se han elegido ambos recursos por ser una fuente fiable y variada de información.

* Definir un título para el dataset. Elegir un título que sea descriptivo.

El nombre elegido para el dataset es: Mosquito\_Deseases\_2013\_2018.csv.

Nos parecía interesante, dentro del contexto de enfermedades infecciosas que nos ha tocado vivir, buscar patrones en incidencia y desarrollo de enfermedades causadas por mosquitos, teniendo en cuenta el tipo de población infectada y región. Además, sería interesante descubrir qué hábitos turísticos o qué características de la fauna y flora local pudieran explicar según qué tipo de enfermedad.

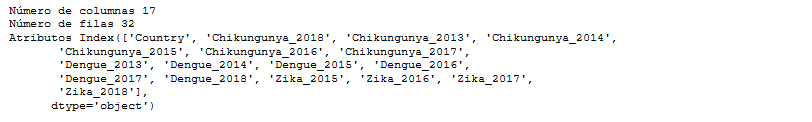
* Descripción del dataset. Desarrollar una descripción breve del conjunto de

datos que se ha extraído (es necesario que esta descripción tenga sentido con

el título elegido).

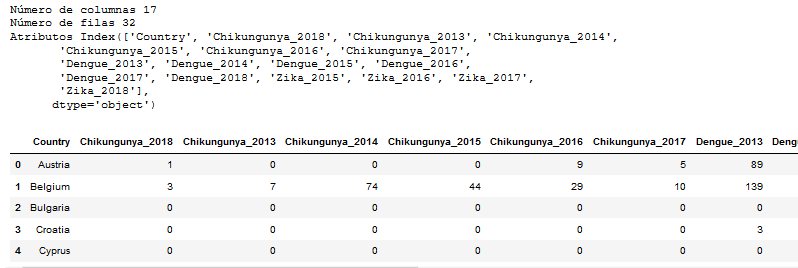
Se ha procedido a crear un dataset único partiendo de los descargados, a través de scraping, del sitio web para proceder con la investigación que nos hemos propuesto.

El dataset consta de 33 filas correspondientes a la información de **33 países diferentes** de la UE. Tiene, además, **17 atributos**, proporcionando número de casos de las tres enfermedades tratadas en este estudio (**Chikungunya, Dengue y Zika**), entre los años 2013-2018.



* Representación gráfica. Presentar una imagen o esquema que identifique el

dataset visualmente.



* Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de

los datos y como se ha recogido.

El dataset creado parte de la información originada por los siguientes dataset:

* **Zika-annual-epidemiological-report-2018**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported\_cases\_2015 | Casos de zika reportados en el año 2015 | String |
| Reported\_cases\_2016 | Casos de zika reportados en el año 2016 | String |
| Reported\_cases\_2017 | Casos de zika reportados en el año 2017 | String |
| Reported\_cases\_2018 | Casos de zika reportados en el año 2018 | String |

* **AER\_for\_2017-Zika-virus-disease**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| 2015 Reported cases\_2015 | Casos de zika reportados en el año 2015 | String |
| 2015 Reported cases\_2016 | Casos de zika reportados en el año 2016 | String |
| 2015 Reported cases\_2017 | Casos de zika reportados en el año 2017 | String |

* **dengue-annual-epidemiological-report-2017**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported\_cases\_2013 | Casos de dengue reportados en 2013 | String |
| Rate\_2013 | Ratio dengue/ 100.000 hab en 2013 | String |
| Reported\_cases\_2014 | Casos de dengue reportados en 2014 | String |
| Rate\_2014 | Ratio dengue/ 100.000 hab en 2014 | String |
| Reported\_cases\_2015 | Casos de dengue reportados en 2015 | String |
| Rate\_2015 | Ratio dengue/ 100.000 hab en 2015 | String |
| Reported\_cases\_2016 | Casos de dengue reportados en 2016 | String |
| Rate\_2016 | Ratio dengue/ 100.000 hab en 2016 | String |
| Reported\_cases\_2017 | Casos de dengue reportados en 2017 | String |
| Rate\_2017 | Ratio dengue/ 100.000 hab en 2017 | String |
| Confirmed\_cases | Casos confirmados en 2017 | String |

* **dengue-annual-epidemiological-report-2018**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported \_2014 | Casos de dengue reportados en 2014 | String |
| Rate\_2014 | Ratio dengue/ 100.000 hab en 2014 | String |
| Reported \_2015 | Casos de dengue reportados en 2015 | String |
| Rate\_2015 | Ratio dengue/ 100.000 hab en 2015 | String |
| Reported \_2016 | Casos de dengue reportados en 2016 | String |
| Rate\_2016 | Ratio dengue/ 100.000 hab en 2016 | String |
| Reported \_2017 | Casos de dengue reportados en 2017 | String |
| Rate\_2017 | Ratio dengue/ 100.000 hab en 2017 | String |
| Reported \_2018 | Casos de dengue reportados en 2018 | String |
| Rate\_2018 | Ratio dengue/ 100.000 hab en 2018 | String |
| ASR\_2018 | Age standarized rate | String |
| Confirmed\_cases | Casos confirmados en 2018 | String |

* **chikungunya-virus-disease-annual-epidemiological-report-2017**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported \_2013 | Chikungunyas reportados en 2013 | String |
| Rate\_2013 | Chikungunyas / 100.000 hab en 2013 | String |
| Reported \_2014 | Chikungunyas reportados en 2014 | String |
| Rate\_2014 | Chikungunyas / 100.000 hab en 2014 | String |
| Reported \_2015 | Chikungunyas reportados en 2015 | String |
| Rate\_2015 | Chikungunyas / 100.000 hab en 2015 | String |
| Reported \_2016 | Chikungunyas reportados en 2016 | String |
| Rate\_2016 | Chikungunyas / 100.000 hab en 2016 | String |
| Reported \_2017 | Chikungunyas reportados en 2017 | String |
| Rate\_2017 | Chikungunyas / 100.000 hab en 2017 | String |
| Confirmed | Casos confirmados en 2017 | String |

* **chikungunya-annual-epidemiological-report-2018**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported \_2014 | Chikungunyas reportados en 2014 | String |
| Rate\_2014 | Chikungunyas / 100.000 hab en 2014 | String |
| Reported \_2015 | Chikungunyas reportados en 2015 | String |
| Rate\_2015 | Chikungunyas / 100.000 hab en 2015 | String |
| Reported \_2016 | Chikungunyas reportados en 2016 | String |
| Rate\_2016 | Chikungunyas / 100.000 hab en 2016 | String |
| Reported \_2017 | Chikungunyas reportados en 2017 | String |
| Rate\_2017 | Chikungunyas / 100.000 hab en 2017 | String |
| Reported \_2018 | Chikungunyas reportados en 2018 | String |
| Rate\_2018 | Chikungunyas / 100.000 hab en 2018 | String |
| ASR\_2018 | Age standarized rate | String |
| Confirmed | Casos confirmados en 2018 | String |

Nuestro dataset en cuestión consta de los siguientes campos:

* Mosquito\_Deseases\_2013\_2018.csv

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Chikungunya\_2018 | Chikungunyas reportados en 2018 | String |
| Chikungunya\_2013 | Chikungunyas reportados en 2013 | String |
| Chikungunya\_2014 | Chikungunyas reportados en 2014 | String |
| Chikungunya\_2015 | Chikungunyas reportados en 2015 | String |
| Chikungunya\_2016 | Chikungunyas reportados en 2016 | String |
| Chikungunya\_2017 | Chikungunyas reportados en 2017 | String |
| Dengue\_2013 | Dengues reportados en 2013 | String |
| Dengue\_2014 | Dengues reportados en 2014 | String |
| Dengue\_2015 | Dengues reportados en 2015 | String |
| Dengue\_2016 | Dengues reportados en 2016 | String |
| Dengue\_2017 | Dengues reportados en 2017 | String |
| Dengue\_2018 | Dengues reportados en 2018 | String |
| Zika\_2015 | Zikas reportados en 2015 | String |
| Zika\_2016 | Zikas reportados en 2016 | String |
| Zika\_2017 | Zikas reportados en 2017 | String |
| Zika\_2018 | Zikas reportados en 2018 | String |

Los 6 datasets originales fueron extraídos a través del proceso descrito en ‘enfermedades\_scraping\_Garcia\_Perdices.ipynb’ y explicado en Web\_Scraping\_NuriaGarcia\_AliciaPerdices\_Prac1\_v4.docx.

A partir de dichos datasets se crea el dataset estudiado ‘Mosquito\_Deseases\_2013\_2018.csv’ donde se realiza una recopilación de los datos para poder estudiar las 3 enfermedades en distintos países y años. Para obtener dicho dataset sería necesario ejecutar el código que se puede encontrar en ‘Dataset\_unico\_Enfermedades\_Garcia\_Perdices\_Prac1\_v2.ipynb’.

Como su nombre indica, el dataset contiene datos desde 2013 a 2018 de las diferentes enfermedades mencionadas y países.

* Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario

incluir citas de investigación o análisis anteriores (si los hay).

Agradecemos la obtención y tratamiento de dichos datos por parte de **Open Data Portal** (Website: http://data.europa.eu/euodp).

El portal de datos abiertos de la Unión Europea proporciona acceso a una creciente gama de datos de las instituciones y organismos de la UE. Estos datos se pueden utilizar y reutilizar con fines comerciales o no comerciales.

Según indican en su web el objetivo de proporcionar un acceso fácil y gratuito a los datos es ayudar a los ciudadanos a utilizarlos de forma innovadora y a aprovechar su potencial económico. El portal se ha concebido también para hacer más transparentes y responsables las instituciones y organismos de la UE.

Respecto a análisis o estudios en dicho dichas enfermedades hemos encontrado artículos:

‘Mosquito-borne diseases’ by World Health Organization <https://www.who.int/neglected_diseases/vector_ecology/mosquito-borne-diseases/en/>

‘How Dengue, a Deadly Mosquito-Borne Disease, Could Spread in a Warming World’ By KENDRA PIERRE-LOUIS and NADJA POPOVICH JUNE 10, 2019

<https://www.nytimes.com/interactive/2019/06/10/climate/dengue-mosquito-spread-map.html>

‘Mosquito-Borne Disease Could Threaten Half the Globe by 2050’ By Chelsea Harvey, E&E News on March 7, 2019

<https://www.scientificamerican.com/article/mosquito-borne-disease-could-threaten-half-the-globe-by-2050/>

De hecho, en este artículo comentan: *‘Para 2050, la mitad de la población mundial podría estar en riesgo de contraer enfermedades transmitidas por mosquitos como el dengue o el virus del Zika, sugiere una investigación reciente.’*

Se ha encontrado un dataset donde se encuentran unos tipos de mosquito que son capaces de transmitir más rápidamente estas enfermedades: Zika, Dengue Chicunbunya y otros virus. El dataset de los mosquitos Aedes aegypti se puede encontrar en:

<https://data.world/zika-virus-data/mosquito-data>

De hecho, en los pdf’s de donde se ha extraído los datasets se puede ver un estudio de cada enfermedad por separado. Los pdfs se pueden encontrar buscando cada enfermedad en <http://data.europa.eu/euodp> (también los podéis ver listados en ‘enfermedades\_scraping\_Garcia\_Perdices.ipynb’)

* Inspiración. Explique por qué es interesante este conjunto de datos y que

preguntas se pretenden responder.

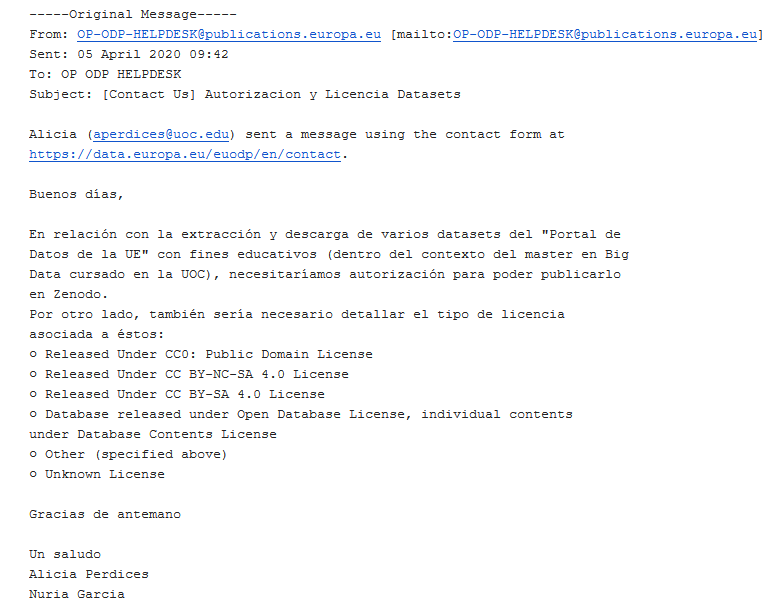
Tal y como hemos indicado anteriormente, en el contexto de la práctica propuesta, y motivadas sobre todo por las **ventajas y beneficios** que el estudio de datos procedentes de enfermedades puede aportar a la **comunidad científica**, no solo para la **prevención**, sino también para establecer **protocolos de acción ante epidemias,** hemos decidido profundizar en el estudio de los datos de la población que ha desarrollado los tres tipos de enfermedades especificados: **Chikungunya, Dengue y Zika**, con el fin de dar respuesta a las siguientes preguntas:

* **Número de infecciones totales por enfermedades en un periodo de tiempo** concreto para poder implementar acciones destinadas a paliar, prevenir enfermedades con mayor incidencia.
* **Países de la EU más afectados por estas enfermedades**:
  + Al comprobar que el Dengue es la infección más predominante, obtener una lista de países con mayor incidencia para esta enfermedad, para poder llevar a cabo planes de abastecimiento de medicamentos y material sanitario necesario por regiones afectadas.
  + Lista de países más afectadas por las tres enfermedades para del mismo modo complementar el abastecimiento necesario de forma adecuada.
* Establecer comparaciones entre países afectados para descubrir **qué características de la fauna y flora local, así como patrones turísticos puedan justificar la aparición de este tipo de enfermedades.**
* Licencia. Seleccione una de estas licencias para su dataset y explique el

motivo de su selección:

* Released Under CC0: Public Domain License
* Released Under CC BY-NC-SA 4.0 License
* Released Under CC BY-SA 4.0 License
* Database released under Open Database License, individual contents
* Under Database Contents License
* Other (specified above)
* Unknown License

En relación a la licencia de los datos obtenidos del recurso, la respuesta del administrador ha sido la siguiente:





Sería, por tanto: **Released Under CC0: Public Domain License**

* Código. Adjuntar el código con el que se ha generado el dataset,

preferiblemente en Python o, alternativamente, en R.

Véanse los archivos: **\*.ypynb en GitHub.**

A continuación, detallamos los elementos de este estudio que se pueden encontrar en Github:

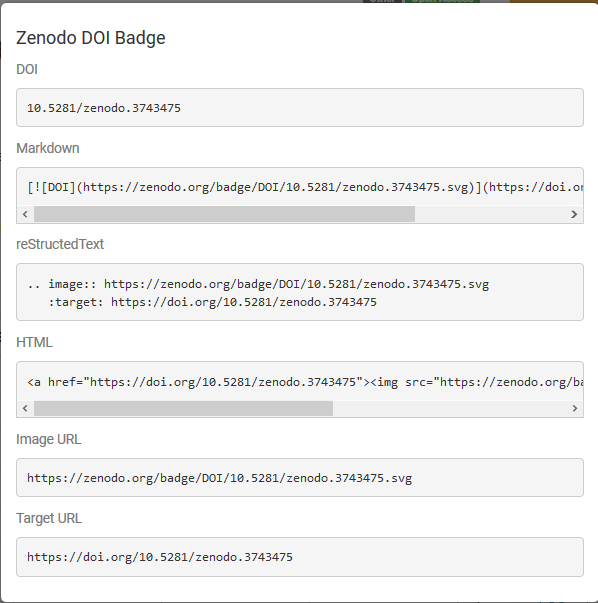
|  |  |
| --- | --- |
| Contribuciones | Firmas |
| Investigación previa:   * Web\_Scraping\_NuriaGarcia\_AliciaPerdices\_Prac1\_v4.docx \*:   (estudio inicial y explicación proceso.)  Carpeta Enfermedades:   * enfermedades\_scraping\_Garcia\_Perdices.ipynb:   (código de scraping para obtener los datasets)   * Dataset\_unico\_Enfermedades\_Garcia\_Perdices\_Prac1.ipynb:   (cómo obtener dataset único con pequeño estudio, diferentes versiones)   * Carpeta datas: podéis encontrar todos los datasets creados (los 6 de scraping y el dataset único) * dengue\_zika-annual-epidemiological-report\_Garcia\_Perdices\_Prac1.ipynb:   (report inicial de los datos Zika, notebook de los presentado en el documento de estudio previo \*)   * Dengue\_zika\_chikungunya\_report\_Garcia\_Perdices\_Prac1\_v2.ipynb:   (análisis de los datos Dengue, Zika, Chicunbunya, estudio enfermedades por separado)   * Carpetas html y pdf (encuentras los .ypbn en el formato que lleva por nombre la carpeta)   Carpeta Desastres:   * web\_scarping\_desastres\_Garcia\_Perdices\_Prac1.ipynb:   (código de scraping para obtener el dataset )   * Desastres\_Analisis\_Garcia\_Perdices\_Prac1.ipynb:   (análisis del dataset obtenido desatres)   * Carpeta data: 500\_registros\_desastres.csv   (dataset obtenido del scraping) | Alicia Perdices  Núria García |
| Redacción de las respuestas:   * Memoria\_Web\_Scraping\_Garcia\_Perdices\_Prac1.docx:   (documento donde se da respuesta a las preguntas planteadas en la práctica) | Alicia Perdices  Núria García |
| Desarrollo código:  Carpeta Práctica 1:  (carpeta creada con la intención de englobar solo la parte de codificación del estudio realizado y contestado en la memoria)   * enfermedades\_scraping\_Garcia\_Perdices.ipynb:   (código de scraping de los datasets enfermedades (6 ficheros))   * Dataset\_unico\_Enfermedades\_Garcia\_Perdices\_Prac1\_v2.ipynb:   (forma de crear el dataset de estudio y estudio del dataset)   * Dengue\_zika\_chikungunya\_report\_Garcia\_Perdices\_Prac1\_v2.ipynb:   (análisis de los datos Dengue, Zika, Chicunbunya, estudio enfermedades por separado) | Alicia Perdices  Núria García |

* Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña

descripción.

El enlace de la publicación es el siguiente:

<https://zenodo.org/record/3743475#.XozOxXJS8dU>



* Entrega. Presentar el trabajo con el DOI del dataset en GitHub.

**Referencias**

**Básicos**

* Subirats, L., Calvo, M. (2019). **Web Scraping**. Editorial UOC.

**Complementarios**

* Lawson, R. (2015). **Web Scraping with Python**. Packt Publishing Ltd. Chapter 2

**Web**

* **Centro Europeo para la prevención y control de las enfermedades**. Recuperado de:

<https://data.europa.eu/euodp/es/data/publisher/ecdc>

* **Disasters**. Recuperado de: <https://reliefweb.int/disasters>
* **Scraping en Python con ejemplos.** Recuperado de:

<https://jarroba.com/scraping-python-beautifulsoup-ejemplos/>

* **Monte Yasur en Vanuatu: El volcán en activo más accesible del mundo.** Recuperado de:

<https://tiempodexplorar.com/monte-yasur-en-vanuatu-el-volcan-en-activo-mas-accesible-del-mundo/>

* **All about the ReliefWeb API :** <https://apidoc.rwlabs.org/#doc-use>
* **Tabula documentation:**

<https://tabula-py.readthedocs.io/en/latest/tabula.html>

* Mar 08 2020**, “Tabula-py”.**

<https://readthedocs.org/projects/tabula-py/downloads/pdf/latest/>

* MedinePlus **“Virus del Zika”.**

<https://medlineplus.gov/spanish/zikavirus.html>

* Stackoverflow, 2017-11-28**. “How to extract table as text from the PDF using Python”**

<https://stackoverflow.com/questions/47533875/how-to-extract-table-as-text-from-the-pdf-using-python>

* Pythonic Excursions, Feb 02, 2019**. “Parse PDF Files While Retaining Structure with Tabula-py”**

<https://aegis4048.github.io/parse-pdf-files-while-retaining-structure-with-tabula-py>

* Datacamp**,** November 2nd, 2018**.** Derrick Mwitti **“JSON Data in Python”**

<https://www.datacamp.com/community/tutorials/json-data-python>

* **RIP tutorial:** <https://riptutorial.com/es/pandas/example/5621/eliminar-una-columna-en-un-dataframe>
* BBC.2 abril 2014, Lizzie Crouch y Paula McGrath. **“Las enfermedades que produce la picadura del mosquito”**

<https://www.bbc.com/mundo/noticias/2014/04/140402_salud_enfermedades_mosquitos_gtg>

* **Beautiful Soup Documentation:**

<https://www.crummy.com/software/BeautifulSoup/bs4/doc/#find-parents-and-find-parent>

* **ReliefWeb**

<https://es.wikipedia.org/wiki/ReliefWeb>