**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:

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Este conjunto de datos proporciona información sobre los niveles actuales y una predicción de los niveles de pólenes ambientales, recogidos en las estaciones de aerobiologia de Barcelona, Bellaterra, Girona, Lleida, Manresa, Roquetes, Tarragona, Vielha y Planes de Son. Estos datos están recogidos en la web del Punto de Información Aerobiológica de Cataluña

Esta web ha sido creada por la Red de Aerobiológica de Cataluña, cuyo principal objetivo es publicar y divulgar el estado actual y la predicción de futuro de los niveles de pólenes alergógenos en diferentes puntos de Cataluña.

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

Niveles de polen en Cataluña

3. 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).

Recolección de los niveles ambientales actuales y predicción de polinización para los principales pólenes, en las ciudades: Barcelona, Bellaterra, Girona, Lleida, Manresa, Roquetes, Tarragona, Vielha y Planes de Son.

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

dataset visualmente.



Ilustración 1: abeja en una flor de diente de León recogiendo polen

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Gracias al uso de los datos proporcionados en la web: <https://lap.uab.cat/aerobiologia/es/forecast/catalunya>, se ha podido realizar esta base de datos.

En este conjunto de datos se ha recogido la información de la estación de aerobiologia de las ciudades de Barcelona, Bellaterra, Girona, Lleida, Manresa, Roquetes, Tarragona, Vielha y Planes de Son. Los datos están distribuidos en los atributos: nombre de taxon (el origen de la planta), valores niveles ambientales actuales (recogidos en un rango de 0 que se considera nulo al 4 con un nivel máximo), valores de predicción (definidos como aumenta, estable, descenso o excepcional) y fecha de predicción. Se han recogido los datos para estas variables para el día: del 8 al 14 de abril.

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

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

Los datos han sido extraídos de la página web: <https://lap.uab.cat/aerobiologia/es/>, usando técnicas de Web Scraping y el lenguaje de programación Python.

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

preguntas se pretenden responder.

En los últimos años se ha producido un aumento constante en la prevalencia de enfermedades alérgicas a nivel mundial, de forma que entre un 30-40% de la población mundial está afectada por una o más enfermedades alérgicas. Entre los diferentes tipos de alergias, sabemos que las alergias ambientales es una de los tipos más frecuentes y es causa de diferentes problemas de salud como angioedema, reacciones cutáneas, rinitis, conjuntivitis y enfermedades respiratorias como el asma y descompensaciones de enfermedad pulmonar obstructiva crónica.

Diferentes estudios, postulan que este incremento de reacciones alérgicas ambientales es debido al incremento de contaminación y a los cambios climatológicos que produce la contaminación y el cambio climático. Así los datos proporcionados en esta base de datos, combinados con otros datos, puede ayudar a evaluar como los cambios en las condiciones climatológicas y los cambios de polución ambiental afectaran a la polinización, al recuentos de polen ambiental, la presencia de insectos y la presencia de hongos asociados a las enfermedades alérgicas ambientales.

<https://www.worldallergy.org/UserFiles/file/WWBOA_Executive-Summary_Spanish.pdf>

8. Licencia. Seleccione una de estas licencias para su dataset y explique el

motivo de su selección:

La licencia seleccionada es la CC BY-NC-SA 4.0 License, puesto que permite copiar, distribuir, exhibir y representar la obra y hacer obras derivadas si se cumple:

* se cita al autor de la fuente original,
* se usa sin fines comerciales, (nuestra página está destinada a la provisión de información y a la investigación y por tanto no tiene un fin comercial sino meramente informativo y de investigación)
* y se distribuye la obra derivada bajo las mismas licencias que la fuente original.

○ 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

<https://es.wikipedia.org/wiki/Licencias_Creative_Commons#Licencias>

<https://creativecommons.org/licenses/?lang=es_ES>

<https://opendatacommons.org/licenses/odbl/1-0/index.html>

<https://en.wikipedia.org/wiki/Open_Database_License>

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

preferiblemente en Python o, alternativamente, en R.

10. Dataset. Presentar el dataset en formato CSV

**Recursos**

Los siguientes recursos son de utilidad por la realización de la PEC:

● Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

● Masip, D. El lenguaje Python. Editorial UOC. Tipología y ciclo de vida de los datos Práctica 1 **pág 2**

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

● Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015).

Automated Data Collection with R: A Practical Guide to Web Scraping and Text

Mining. John Wiley & Sons.

● Tutorial de Github https://guides.github.com/activities/hello-world .