Práctica 1. Tipología y ciclo de vida de los datos.

Nombre y apellidos: Francisco Javier Fortea

Usuario UOC: javierfortea

Estudios que cursa (Máster o Grado): Máster Universitario en Ciencia de Datos

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

El dataset creado se ha realizado a partir de la información de la web https://airportdatabase.net/, que contiene información de aeropuertos y aerolíneas (en este caso, solo hemos recolectado los datos de las aerolíneas)

Este sitio web contiene información de más de cuarenta mil aeropuertos, con un conjunto de datos importante de cada uno de ellos.

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

El título para el dataset es “Global airports dataset”, ya que se trata de un dataset con información de más de cuarenta mil aeropuertos de todo el mundo.

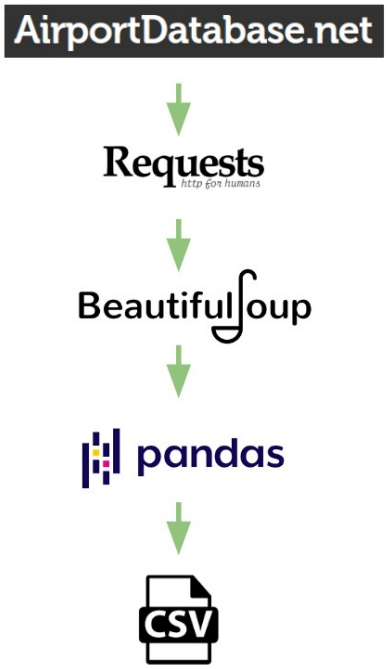
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El dataset incluye información de más de cuarenta mil aeropuertos de diferente tipo (desde pequeños a grandes aeropuertos, pasando por helipuertos). Para cada uno de ellos tenemos su nombre, información geográfica (municipio/región/país/continente/… así como latitud/longitud y elevación), tipo de aeropuerto, código IATA, frecuencia a la que operan el APP/ATIS/GND/TWR, etc.

4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido

El siguiente diagrama representa el proceso de obtención de los datos de https://airportdatabase.net/, usando primero la librería “requests” para realizar las

peticciones HTTP y, a continuación, la librería “BeautifulSoup” para parsear e interpretar el HTML. Finalmente, con “pandas” creamos el dataset final, que es guardado en formato CSV.



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

Los campos que incluye el dataset son los siguientes:

- **airport**_name: nombre del aeropuerto.
- **Ident**: identificador del aeropuerto.
- **type**: tipo de aeropuerto. Los tipos posibles son: “Small Airport”, “Medium Airport”, “Large Airport” y “Heliport”.
- **latitude**: latitud donde se encuentra el aeropuerto.
- **longitude**: longitud donde se encuentra el aeropuerto.
- **elevation**: altitud del aeropuerto.
- **continent**: continente donde se encuentra el aeropuerto.
- **iso country**: país donde se encuentra el aeropuerto.
- **ISO Region**: región donde se encuentra el aeropuerto.
- **ISO Region link**: link con información de la región donde se encuentra el aeropuerto.

- **Municipality**: municipio donde se encuentra el aeropuerto.
- **Scheduled service**: indica si el aeropuerto tiene servicios programados.
- **GPS Code**: código GPS identificador del aeropuerto.
- **IATA Code**: código IATA (International Air Transport Association) del aeropuerto.
- **wikipedia link**: link al artículo sobre el aeropuerto en la Wikipedia.
- **APP**: frecuencia a la que emite el “approach control” del aeropuerto.
- **ATIS**: frecuencia a la que emite el “Automatic terminal information service” del aeropuerto.
- **GND**: frecuencia a la que emiten los “Ground controllers” del aeropuerto.
- **TWR**: frecuencia a la que emite el “traffic control tower” del aeropuerto.
- **Website**: web del aeropuerto.
- **keywords**: listado de palabras claves asociadas al aeropuerto.

Se puede encontrar más información sobre el significado de las columnas APP, ATIS, GND y TWR en el siguiente enlace: https://www.skyradar.com/blog/air-traffic-control-training-what-is-acc-app-twr-gnd-and-smgcs-and-how-is-it-trained-in-an-aerodrome-simulator

Los datos se han recogido recorriendo las diferentes páginas en las que se listan los aeropuertos, y accediendo a cada ficha individual, para capturar los datos de todos los aeropuertos. Para no sobrecargar el servidor, se han añadido esperas entre las peticiones, alargando el tiempo de ejecución del programa.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

Los datos han sido recopilados desde la base de datos online de aeropuertos “AirportDatabase.net” (https://airportdatabase.net/airports), que han realizado una gran labor de documentación de todos los aeropuertos.

He usado python y la libería “BeautifulSoup4” para parsear más fácilmente la información contenida en el HTML de la web, además de la librería “requests” para realizar las peticiones al servidor web.

Hay algunas bases de datos de aeropuertos, como la “World Airport Database” (http://www.world-airport-database.com/database.html), pero que es de pago y contiene menos aeropuertos.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Este conjunto de datos es útil tanto para analizar los datos por sí mismos como para usarlos en conjunción con otros dataset (por ejemplo, de accidentes o tráfico por aeropuerto). Hay bastantes datasets con los aeropuertos de países concretos, pero es más difícil encontrar uno con aeropuertos de todo el mundo y con la información tan completa.

Estos datos se podrían utilizar en conjunción con algunos de los siguientes datasets:

- Busiest Airports by Passenger Traffic (https://www.kaggle.com/jonahmary17/airports)
- Air Traffic Data (https://www.kaggle.com/rohanshetty678/air-traffic-data)

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

La licencia que he escogido es la "**CC BY-SA 4.0 License**", ello se debe a que es una licencia "Creative Commons" que permite la distribución libre con las siguientes características:

- **BY**: esta parte del nombre de la licencia se refiere al derecho de "**atribución**". Se puede copiar, distribuir, derivar el trabajo, etc. siempre y cuando se dé crédito al autor.
- **SA**: esta parte del nombre de la licencia significa "**share-alike**", es decir, se puede distribuir trabajo derivado siempre y cuando se haga con una licencia idéntica (no más restrictiva) que la licencia del trabajo original.
- **Comercial**: usamos la licencia "CC BY-SA 4.0", no la "CC BY-NC-SA 4.0" que incluye el término "non-commercial". Por tanto, estamos usando una licencia que permite el uso comercial.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El dataset y el código en Python, junto a los archivos Pipfile y Pipfile.lock (con la definición de las dependencias usadas), se incluye en este mismo repositorio.

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

El DOI del dataset en Zenodo es “10.5281/zenodo.4679452” y se puede encontrar en el siguiente enlace: https://zenodo.org/record/4679452

Contribuciones	**Firma**
Investigación previa	Javier Fortea
Redacción de las respuestas	Javier Fortea
Desarrollo código	Javier Fortea