M2.851 - Tipología y ciclo de vida de los datos

Práctica 1: WebScraping coffee_varieties

Equipo para la práctica

El equipo está formado por:

- Ivan García Jiménez (igarciajimenez0@uoc.edu)
- Itziar Ricondo Iriondo (iricondoi@uoc.edu)

1 Contexto

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

El café es uno de los productos más comercializados y una de las tres bebidas más consumidas del mundo (junto con el agua y el té)[1]. Se usan principalmente dos especies para la preparación su preparación: *Coffea arabica* y *Coffea canephora*. La primera especie abarca casi tres cuartas partes de la producción mundial y se cultiva principalmente desde México hasta Perú.

El cultivo del café está culturalmente ligado a la historia y al progreso de muchos países que lo han producido. En contra de lo que pueda creerse, la producción mundial de café proviene, alrededor de un 70 %, de explotaciones principalmente familiares de superficie inferior a diez hectáreas, incluso generalmente por debajo de cinco hectáreas.

En los últimos años, tres factores han aumentado el riesgo de extinción del cultivo del café, la enfermedad de la roya del café a partir del año 2012, el cambio climático y los precios que recibe el productor[2].

El World Coffee Research (WCR)[3] es un programa colaborativo de investigación y desarrollo sin ánimo de lucro de la industria mundial del café para cultivar, proteger y mejorar el suministro de café de calidad, así como la mejora de calidad de vida de las familias que lo producen.

Este organismo sin ánimo de lucro ha publicado un catálogo de las variedades del café arábiga. El objetivo del catálogo es **ofrecer información a los productores de café, sobre cuál variedad de café se ajusta mejor a sus condiciones, con objeto de tomar la mejor decisión en cuanto a la variable a cultivar.**

El trabajo se ha centrado en adquirir o rastrear la información de este catálogo disponible de forma abierta y generar un conjunto de datos.

[1] https://es.wikipedia.org/wiki/Caf%C3%A9
[2] https://www.sica.int/Iniciativas/cafe
[3] https://worldcoffeeresearch.org/

2 Definir un título para el dataset

Elegir un título que sea descriptivo.

El título elegido para el dataset es coffee_varieties.

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 conjunto de datos proporciona información sobre la apariencia, información agronómica, genética y disponibilidad de cada variedad de café arábica. La información más extensa es la agronómica, datos necesarios para identificar la idoneidad de determinada localización a las condiciones de crecimiento de la variedad de café.

4 Representación gráfica

Presentar una imagen o esquema que identifique el dataset visualmente.

Se adjunta la imagen de variedad Bourbon

5 Contenido

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

Los datos son estáticos, ya que representan propiedades sobre cada variedad de café. Por otro lado, podría darse la circunstancia que la organización propietaria de la fuente de los datos añadiera una nueva variedad, que sería en el conjunto de datos de salida al volver a ejecutar el archivo Python desarrollado para la obtención del conjunto de datos (archivo csv).

El conjunto de datos incluye los siguientes campos:

Variable	Descripción	Tipo	Valores
Name		String	
Description		String	
Stature	Hábito de crecimiento de la planta	Discreto	Tall Dwarf/Compact
Leaf_tip_color	Color de la punta de nuevas hojas.	Discreto	Green Light Bronze Bronze Dark Bronze Green or Bronze
Bean_size	Tamaño de grano de café. Referecia, Caturra = Average, SL28 = Large, and Maragogipe = Very Large	Discreto	Below Average Average Large Very Large
Optimal_Altitude_1	Altitud para máximo rendimiento de calidad y agronómico. Rango latitud: 5°N to 5°S	String	>1600 m
Optimal_Altitude_2	Altitud para máximo rendimiento de calidad y agronómico. Rango latitud: 5–15°N and 5–15°S:	String	1000-1600m
Optimal_Altitude_3	Altitud para máximo rendimiento de calidad y agronómico. Rango latitud: >15°N and >15°S:	String	7000 m
Quality_potential	Calidad potencial si es cultivado a mayores altitudes	Discreto	Very Low Low Good Very Good Exceptional
Yield_potential	Potencial de fruto de árbol de café. Referencia: For reference, Caturra = Good and SL28 = Good	Discreto	Low Medium Good High Very high
Coffee_leaf_rust	Resistencia a roya de cafeto	Discreto	Resistant Tolerant Susceptible
CBD	Resistencia a la antracnosis de la cereza	Discreto	Resistant Tolerant Susceptible
Nematodes	Resistencia a los nematodos	Discreto	Resistant Tolerant Susceptible
Year_First_Production	Años para la primera cosecha	Discreto	Year 2 Year 3 Year 4
Nutrition_Requirement	Requerimientos nutricionales	Discreto	Low Medium

Variable	Descripción	Tipo	Valores
			High Very High
Ripening_of_Fruit	Maduración de la fruta. Referencia: Caturra = Average	Discreto	Early Average Late Very late
CTGB	Rendimiento de cerezas a grano verde. Proporción del grano en relación al grano. Para referencia, Caturra = Average, SL28 = High	Discreto	Low Average High Very High
Planting_Density	Distancia que se debe utilizar para la siembra de esta variedad[4]	Discreto	1000-2000 a/ha (multiple-stem pruning) 2000-3000 a/ha (multiple-stem pruning) Less than 2000 a/ha (multiple-stem pruning) 3000-4000 (single stem pruning) 4000-5000 a/ha (single-stem pruning) 5000-6000 a/ha (single-stem pruning)
Additional	Información adicional	String	
Lineage	Linaje genético	String	
Genetic Description	Grupo genético	Discreto	Ethiopian/landrace Bourbon-Typica group/Typica Bourbon-Typica group/Bourbon Bourbon-Typica group/Typica+Bourbon Introgressed/Catimor Introgressed/Sarchimor F1 hybrid/introgressed F1 hybrid/not introgressed
History	Historia	String	
Breeder	Cultivador	String	
Intellectual_Property_Rights	Propiedad intelectual	String	
Image	Dirección de la imagen de la variedad de café	String	

[4] En Centro América, los árboles generalmente se podan para definirles un solo tallo vertical principal. En Africa, es típico podar los árboles dejándoles múltiples tallos principales verticales por árbol (2-3). Por lo tanto, aunque las densidades de la plantación de árboles son mucho más bajas en África, cada árbol está produciendo relativamente más porque tienen múltiples tallos principales.

6 Agradecimientos

Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

El catálogo “Las variedades del Café Arábica” ha sido publicado por World Coffee Research[5]. Ests organización lleva a cabo investigación básica en la planta de café que asegure el futuro del café y favorecer así la calidad de vida de los agricultores.

El catálogo está disponible en dos formatos — un sitio web interactivo[6] y PDF — y reúne información esencial sobre 55 variedades de arábica de todo el mundo.

El catálogo ha sido desarrollado en colaboración con expertos en café de toda América Central, el Caribe y África.

El catálogo “Las variedades del Café Arábica” por World Coffee Research se distribuye bajo una licencia Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

7 Inspiración

Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El interés del conjunto de datos reside en la información proporcionada para identificar las condiciones idóneas de cultivo de las especies de café. Algunas de las preguntas a las que responde el conjunto de datos son:

- ¿Qué variedades son más resistentes a la enfermedad de la roya?
- ¿Cuál es la altitud en la cual la calidad y el potencial de producción de la variedad es mayor?
- ¿Qué variedades de café producen antes el fruto?
- ¿Qué distancia se debe utilizar para la siembra de una variedad?

8 Licencia

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

El catálogo original está publicado con licencia CC BY-NC-ND 4.0. Por lo tanto, la licencia de nuestro conjunto de datos debe ofrecer las mismas limitaciones. En concreto, esta licencia permite compartir el material original, esto es, **copiar y distribuir el material en cualquier medio o formato**.

Más información sobre el alcance de esta licencia está disponible en Creative Commons[7]

[5] https://worldcoffeeresearch.org/
[6] https://varieties.worldcoffeeresearch.org/es/varieties
[7] https://creativecommons.org/licenses/by-nc-nd/4.0/

9 Código

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

Realizado.

10 Dataset

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

Pendiente.

11 Tabla de contribuciones

Firmas:

- Ivan García, IG
- Itziar Ricondo, IR

Contribuciones	**Firma**
Investigación previa	IG,IR
Redacción de las respuestas	IG,IR
Desarrollo código	IG,IR