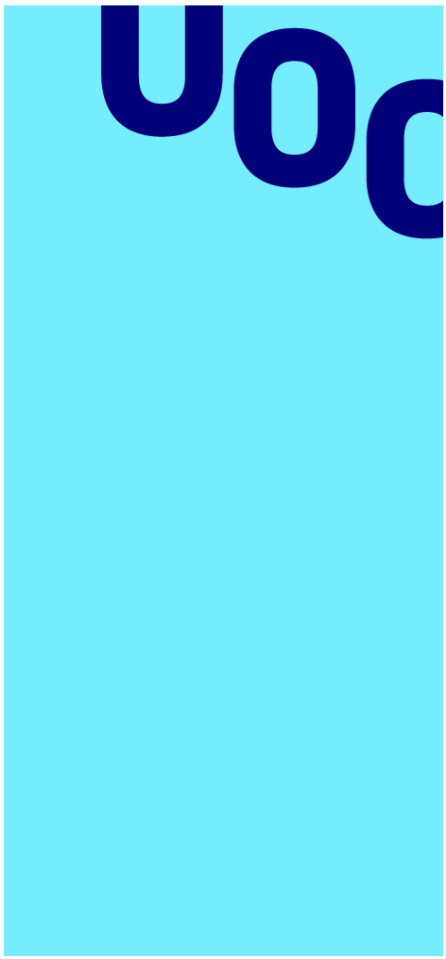
# Tipología y fuentes de datos

## PEC 4



Universitat Oberta de Catalunya

Juan Luis Acebal Rico

# Tipología y fuentes de datos
PEC 4

## Enunciado 1 (1.5 puntos)

Lee el artículo y visualiza el vídeo sobre datos sintéticos del enlace que se incluye a continuación y responde a las siguientes cuestiones:

**https://datos.gob.es/en/documentacion/synthetic-data-what-are-they-and-what-are-they-used**

### 1. ¿Qué son los datos sintéticos? ¿En qué campos son utilizados y para qué finalidad? (0.5 ptos)

**Son datos elaborados artificialmente, que se utilizan para entrenar modelos, para el aprendizaje, o para hacer representaciones parecidas a datos reales. El objetivo puede haber muchos, desde que no existan datos limpios, datos que no estén organizados, por privacidad, por costo, etc.**

**Un ejemplo, para entrenar un modelo de detección de diabetes, no tiene que ser necesariamente datos reales, pueden ser datos sintéticos que han sido extraídos de una muestra representativa de la población, con eso se evita tener que recolectarlos en un lugar concreto, se evitan sesgos, por ejemplo de una población que tenga más predisposición genética, y se dificulta menos la creación de un dataset.**

### 2. ¿Cuáles son sus ventajas? (0.5 ptos)

**Privacidad, ya que permite el uso sin comprometer la informacion de las persones al no contener informacion personal.**

**Pueden ser creados en grandes cantidades, sin limitaciones como los datos reales.**

**Además no tienen sesgos, o se pueden crear con menos sesgos, es decir, puedes crear una muestra representativa de algo para que la inteligencia artificial no tenga un sesgo. Por ejemplo, una foto de personas de la cárcel podría crear un sesgo racista, es por eso que los datos sintéticos pueden ser muy interesantes para entrenar modelos. Por último es más barato que recolectarlo.**

## 3. Describe las técnicas principales utilizadas para desarrollar datos sintéticos (0.5 ptos)

**Técnicas de remuestreo, conlleva la selección y combinación de datos reales, con modelos probabilísticos, que conlleva el uso de modelos de predicción que sigan distribuciones y características que datos reales, y por último, y generativos y métodos de perturbación y enmascaramiento, que modifican datos reales, introducen cambios para producir datos sintéticos conservando algunas de sus propiedades originales.**

# Enunciado 2 (1.5 punto)

Parte 1. Lee el artículo Herramientas de procesado y visualización de datos publicado en la web de datos.gob.es.

https://datos.gob.es/sites/default/files/doc/file/herramientas_de_procesado_y_visualizacion_de_datos.pdf

Menciona brevemente qué te ha llamado más la atención del artículo. ¿Qué herramienta mencionada en el artículo has descubierto y te ha resultado más interesante y por qué?

**Open Refine. Me parece interesante para procesos ETL, depuración y operaciones de enriquecer y depurar datos. Nunca la había escuchado**

**Del resto hay varias interesantes, especialmente Grafana, que ya había escuchado, y que por el tipo de visualizaciones posibles haciendo dashboards, debe ser bastante interesante.**

¿Podrías proponer alguna herramienta o librería no incluida en el artículo que consideras importante? (0.5 ptos).

**Yo diría alguna herramienta concreta de Python, del estilo seaborn**

Parte 2. Supongamos que estás liderando un proyecto de análisis de datos que incluye el análisis de datos estructurados (por ejemplo, datos de ventas en una base de datos relacional) y datos no estructurados (por ejemplo, comentarios de clientes en forma de texto). Tu objetivo es seleccionar las herramientas apropiadas para procesar, analizar y visualizar estos datos de manera efectiva. Puedes suponer que ya tienes los datos exportados en ficheros. (1 pto)

Elabora un pequeño informe abarcando las siguientes cuestiones:

1. **Procesamiento de Datos Estructurados:**
   - Identifica una herramienta **específica de Python** adecuada para procesar los datos estructurados, como los datos de ventas en una base de datos relacional. Justifica tu elección.

   **Yo recomendaría pandas, es muy cómodo de tratar "tablas" o datos de más de 1 dimensión, además podríamos usarlo en muchos contextos con datos. Además Orange que es muy potente tanto en su versión librería como su aplicación.**

2. **Procesamiento de Datos No Estructurados:**

- Elige una herramienta **especifica de Python** para manejar los datos no estructurados, como los comentarios de clientes en forma de texto. Explica por qué esta herramienta es apropiada para esta tarea.

**NLTK es la mejor opción para procesar texto, podemos hacer tokens con nuestro texto y luego hacer clasificaciones.**

3. **Análisis de Datos:**
   - ¿Qué librerías de Python usarías para realizar análisis comparativos y extraer patrones tanto de datos estructurados como no estructurados?
   - **Scikit para aprendizaje automatico. Quizás estaría bien pandas (como en cualquier uso de datos en Python, casi imposible no utilizarla), y SpaCy tanto en análisis de datos como en procesamiento de datos no estructurados. StatsModels en algún momento es también de uso obligado por sus métricas para modelos de IA**
4. **Visualización de Datos:**
   - Finalmente, selecciona una herramienta de visualización en Python que te permita comunicar los resultados del análisis de manera efectiva.
   - **Matplotlib, seaborn, ambas son importantes e imprescindibles en el dia a dia en python**

Nota: preferiblemente se deben seleccionar herramientas y librerías de Python que estén incluidas en el artículo de la parte 1.

# Enunciados 3, 4, 5

Ver en el notebook asociado a la PEC4

# Anexo

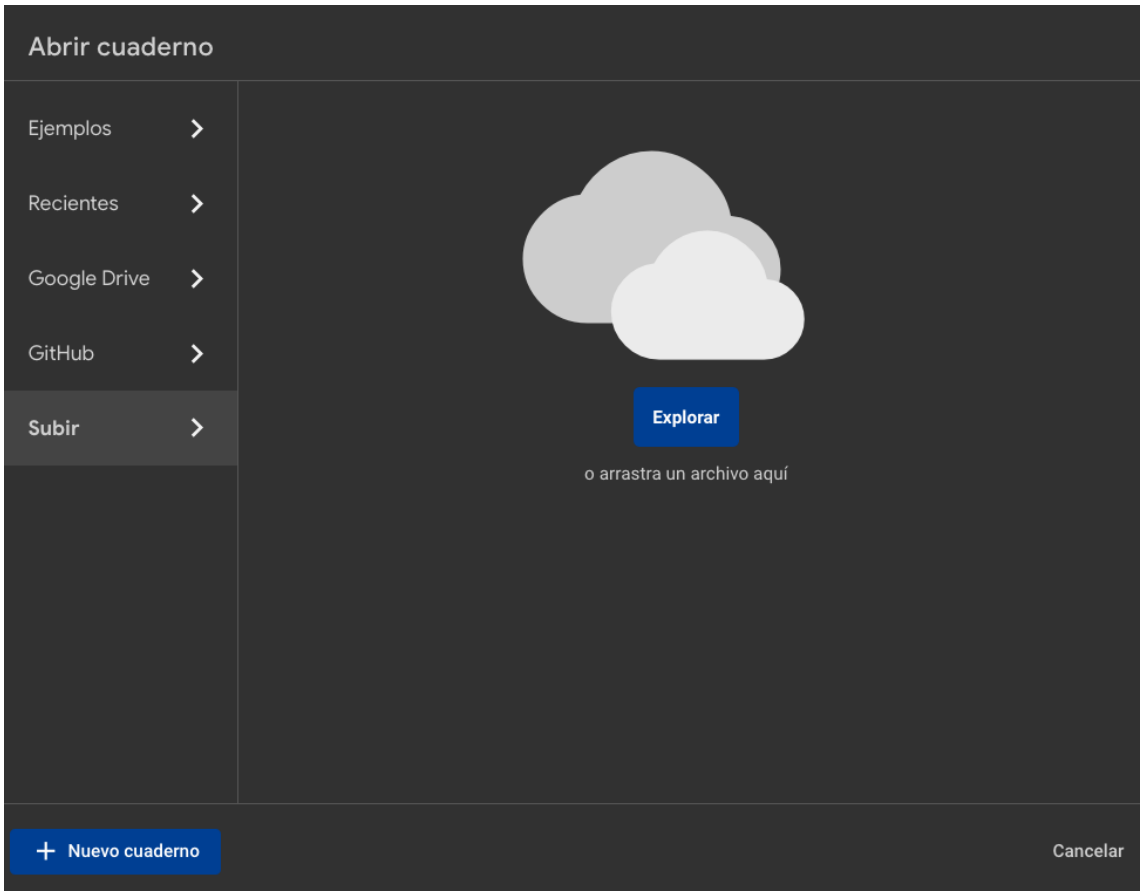
Descarga en tu equipo el notebook TyFdD_PEC4_2324_S2.ipynb. El notebook está pensado para que lo ejecutes en Google Colab.
**Google Colab** es un servicio en la nube que permite ejecutar Jupyter Notebooks accediendo con un navegador web. Tiene además las siguientes ventajas:

- Posibilidad de ejecución mediante GPUs
- Basado en jupyter notebook pudiendo crear y ejecutar libros en Python 2 o 3

- Tiene preinstaladas las librerías comunes usadas en ciencia de datos y la posibilidad de instalar otros.
- Enlaza con cuentas de Google Drive y desde github

Primero es necesario entrar en sesión ( login ) con una cuenta de Google (la de la uoc debería funcionar).

Ahora ya se puede subir el notebook de la PEC a Colab :



La ejecución del libro es exactamente igual que en cualquier jupyter notebook . Hay que pulsar Shift + Enter para que el código ( python ) se ejecute.

A partir de ahí, debes completar el código python que falta y que está marcado con `#TODO` y contestar a las cuestiones planteadas escribiendo tanto las respuestas como el código con el que obtienes las respuestapes.
Una vez finalizado, debes descargar el archivo . ipynb para poder realizar la entrega.

**Criterios de valoración**
Cada uno de los apartados tiene un peso asignado en el total de la PEC. Se valorará, por cada apartado, la validez de la solución y claridad de la argumentación.

**Formato y fecha de entrega**
Debe enviar la PEC al buzón de entrega y registro del AC disponible en el aula (apartado de Evaluación). El formato del archivo que contenga su solución puede ser pdf ,. odt , .doc y . docx . Para otras opciones, por favor, contactar previamente con su profesor colaborador. El número del fichero debe contener el código de la asignatura, su apellido y su nombre, así como el número de la actividad (PEC4). Por ejemplo apellido1_nombre_tifdd_pecX.pdf

**La autoría de la PECdebe ser propia e individual.**

**Propiedad intelectual**
Al presentar una práctica o PEC que haga uso de recursos ajenos, debe presentarse junto con ella un documento donde se detallan todos ellos,
especificando el nombre de cada recurso, su autor, el lugar donde se obtenga y su estatus legal: si la obra está protegida por el copyright o se acoge a alguna otra licencia de uso ( Creative Commons , licencia GNU, GPL etc.). El estudiante deberá asegurarse de que la licencia que sea no impida específicamente su uso en el marco de la práctica o PEC. En caso de no encontrar la información correspondiente, se tendrá que asumir que la obra está protegida por el copyright.

Será necesario, además, adjuntar los archivos originales cuando las obras
utilizadas sean digitales, y su código fuente, si así corresponde.