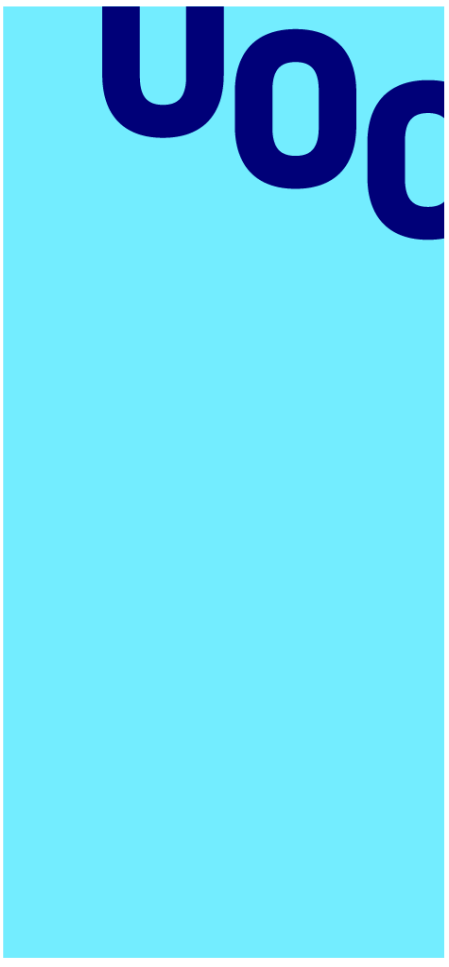
Tipología y fuentes de datos

PEC 1



Universitat Oberta de Catalunya

Fecha de entrega :
12 de octubre de 2021

Tipología y fuentes de datos
PEC1

Enunciado 1 (1,5 puntos)

Indique si son ciertas o falsas las siguientes afirmaciones:

1. Un modelo de datos permite describir la estructura de los datos y también las restricciones que éstos deben cumplir.

2. En la pirámide DIKW, el nivel más bajo trataría es explicar el cómo se conectan unos datos con otros.

3. Una imagen en formato png tiene un formato fijo de puntos repartidos en filas y columnas, aun así no son datos estructurados.

4. La suma de dos datos categóricos ordinales es otro dato categórico ordinal.

5. Una restricción de unicidad es que los resultados de la evaluación de una asignatura en la UOC sean [N, D, C-,C+, B, A].

6. En las tablas de una base de datos relacional no se puede almacenar un modelo en red

7. Las imágenes pueden contener la localización geográfica de donde fueron tomadas. Si estos metadatos se borran, la imagen no se podría visualizar correctamente.

8. Dublin Core, o más formal Dublin Core Metadata Element Set, es un estándar ISO 15836-2003.

9. La exhaustividad (recall) es la mejor medida cuando los falsos positivos son más importantes que los falsos negativos.

10. El formato JSON es más versátil que CSV

Enunciado 2 (3,5 puntos)

Para resolver este enunciado hay que cargar el jupyter notebook en Google Colab (ver Anexo) y cargar el dataset del Titanic. Leed con atención el ejercicio y resolver los fragmentos de código marcados con `# TODO` para obtener las respuestas.

Responda a las siguientes preguntas:

Q1 - ¿Cuáles de las variables del conjunto de datos del Titanic son categóricas?
Q2 - ¿Alguna de estas variables categóricas es ordinal?
Q3 - ¿Y alguna de estas variables categóricas es binaria?
Q4 - ¿Qué tipos de datos utiliza Python para codificar estas variables categóricas?
Q5 - Y cuáles son numéricas?
Q6 - Gestionar el error cuando se selecciona una columna que no existe
Q7 - Determinar el porcentaje de personas que sobrevivieron
Q8 - Qué género tuvo más oportunidad de sobrevivir
Q9 - Qué clase social tuvo más posibilidades de sobrevivir
Q10 - Qué grupo de edad tuvo más posibilidades de sobrevivir

Enunciado 3 (2.5 puntos)

Los protocolos Open Graph y Twitter Cards controlan cómo se visualizan las URLs cuando se comparten por redes sociales. Para ello utilizan metadatos que se encuentran en la página web. En esta parte veremos cómo extraerlos y utilizarlos mediante una librería de Python denominada 'BeautifulSoup' y que será de mucha utilidad para hacer *scraping* de la web.

Enunciado 4 (2.5 puntos)

Suponer que partimos de unos vectores de documentos de 0s y 1s así como una consulta. Vamos a calcular las similitud entre éstos utilizando la métrica del coseno.

De manera similar, vamos a obtener unos objetivos y unas predicciones para obtener la matriz de confusión así como unas métricas (precisión, exhaustividad , F1, y exactitud) desarrollando las funciones de cálculo en Python.

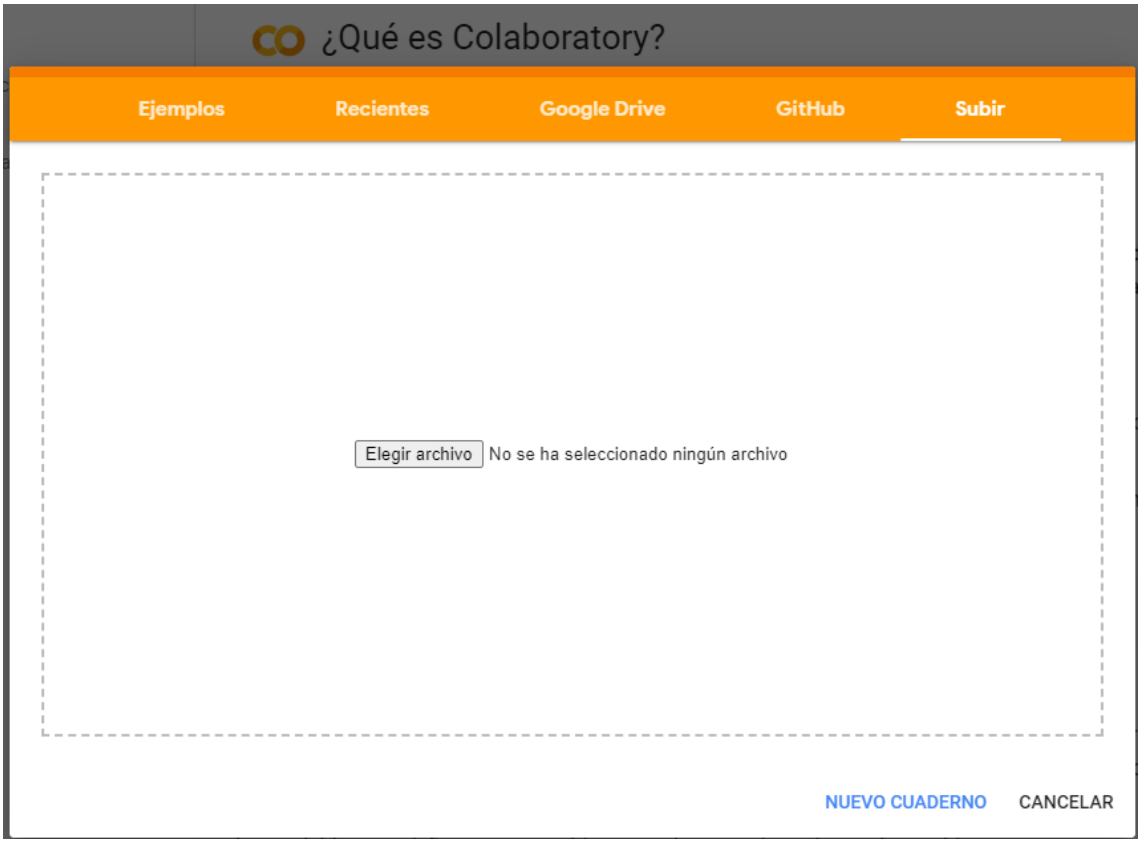
Anexo

Descarga en tu equipo el notebook `TyFdD-PEC1-2021.ipynb` así como los archivos de datos en csv. 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 otras.
- Enlaza con cuentas de Google Drive y desde github

Primero hay que 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:

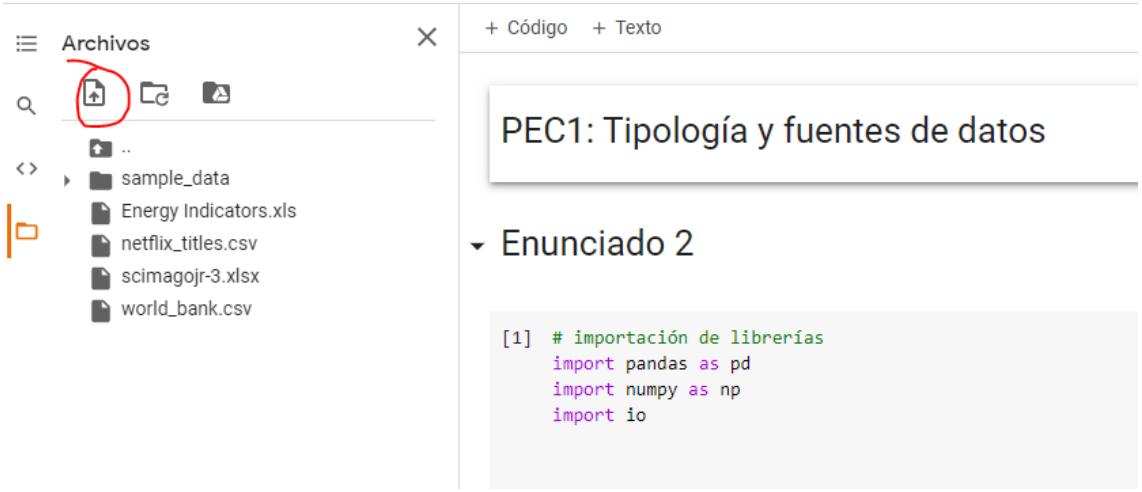


La ejecución del libro es exactamente igual que en cualquier jupyter notebook. Hay que pulsar Shift + Enter para que el código (python) se ejecute. Los conjuntos de datos necesarios para la PEC se cargan mediante código (hay que seleccionar el archivo .csv adjunto):

```
[ ]  # subir todos los archivos de la PEC a colab
     from google.colab import files
     uploaded = files.upload()
```

```
Elegir archivos  Ningún archivo seleccionado  Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving netflix_titles.csv to netflix_titles.csv
Saving Energy Indicators.xls to Energy Indicators.xls
Saving world_bank.csv to world_bank.csv
Saving scimagojr-3.xlsx to scimagojr-3.xlsx
```

Alternativamente también es posible subirlos con la herramienta para subir archivos del propio colab:



No olvides que hay que subir **el archivo csv**.

A partir de ahí debes completar el código python que falta que está marcado con `# TODO` y responder a las cuestiones planteadas escribiendo tanto las respuestas como el código con el que obtienes las respuestas.

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

Formato y fecha de entrega
Tenéis que enviar la PEC al buzón de Entrega y registro de EC disponible en el aula (apartado Evaluación). El formato del archivo que contiene vuestra solución puede ser .pdf,. odt, .doc y .docx. Para otras opciones, por favor, contactar previamente con vuestro profesor colaborador. El nombre del fichero debe contener el código de la asignatura, vuestro apellido y vuestro nombre, así como el número de actividad (PEC2). Por ejemplo *apellido1_nombre_tyfdd_pecX.pdf*. No olvidéis poner vuestro nombre y apellidos en el documento.

La fecha límite para entregar la PEC es **la indicada en la portada.**

Propiedad intelectual
Al presentar una práctica o PEC que haga uso de recursos ajenos, se tiene que presentar junto con ella un documento en que se detallen todos ellos, especificando el nombre de cada recurso, su autor, el lugar donde se obtuvo 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 tendrá que asegurarse que la licencia que sea no impide específicamente su uso en el marco de la práctica o PEC. En caso de no encontrar la información correspondiente tendrá que asumir que la obra está protegida por el copyright.

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