UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Python para Ciencia de Datos

Trabajo de asignatura

Análisis de los Juegos Olímpicos con Python y Pandas

Máster F.P. en IA & Big Data Analytics
Curso 2025-2026

UPV

Objetivos

- Este trabajo propone explorar datos de atletas y medallas en los Juegos Olímpicos
- Los datos de entrada están en dos tablas: Atletas y Medallas
- Se busca combinar análisis con pandas y programación en Python
- Se incluye un análisis de determinadas características y tendencias de los datos originales, y en su caso, su segregación por género, juegos de invierno o verano, país de origen, etc.
- El trabajo incorpora ciertos retos cerrados con otros abiertos a elección del estudiantado

Retos a realizar

- El trabajo incluye **4 retos** diferenciados y para cada uno, propone una serie de **tareas** de procesamiento, manipulación y visualización de datos, y una **función** en Python relacionada
- Los retos son:
    - **RETO 1**: Relacionar los datos de entrada y obtener una visión completa de los/las atletas
    - **RETO 2**: Analizar perfiles físicos y demográficos
    - **RETO 3**: Estudiar las medallas obtenidas y el rendimiento
    - **RETO 4**: Formular un objetivo abierto y resolverlo
- El trabajo puede realizarse por **parejas** o de manera **individual**

Formato de entrega

- El formato de entrega será un único fichero **Notebook de Jupyter** autocontenido
- Este fichero deberá tener una **sección** para cada uno de los retos que propone el trabajo
- En cada sección se deberá incluir la secuencia de **tareas** propuestas en **celdas** distintas y, debajo del enunciado de cada tarea, una o varias celdas con la resolución de dicha tarea
- Se valorará que el fichero sea **visualmente atractivo** y **fácil de seguir**: estructurado por retos, y para cada uno, por tareas, y para cada tarea, su enunciado, procesamiento, gráficas, comentarios, etc.

RETO 1: Relación entre los datos de entrada (tablas)

- Explica brevemente (en 5–8 líneas) qué aporta cada tabla (atletas y medallas) y cómo se relacionan entre sí
- Muestra un listado de ejemplo (5–10 filas) con: nombre del atleta, país, tipo de juegos, años en los que participó y medallas obtenidas en cada año
- Indica 3 preguntas que no se podrían responder usando solo una de las tablas
- Haz un gráfico de barras apiladas con el nº de participaciones por país para los 10 países con más participaciones, separando Verano/Invierno. Justifica por qué barras apiladas es una buena elección (o propón otro tipo de gráfica y justifícala)

RETO 1: Relación entre los datos de entrada (tablas)

- Implementa la **función** `resumen_participacion(id_atleta, df_medallas)` que resuma los **años de participación** de un atleta concreto en intervalos consecutivos

- **Salida**: diccionario con número de participaciones y una lista de intervalos

- Por ejemplo, ante una participación como esta:

  `[1992, 1996, 2000, 2008, 2012, 2016]`

  la salida sería:

  ```
  {"participaciones": 6, "intervalos": ["1992–2000", "2008–2016"]}
  ```

RETO 1: Relación entre los datos de entrada (tablas)

- **Notas**:
  - Ordenar los años, y cuidado con posibles duplicados
  - Construir intervalos lo más largos posibles de participaciones consecutivas. Son consecutivos aquellos años que siguen el salto olímpico esperado (habitualmente cada 4 años en Juegos de Verano o Invierno)

- **Caso de prueba** (*Michael Phelps*):

```
resumen_participacion(id_atleta=94406, df_medallas=df_medallas)
```

Resultado:

```
{"participaciones": 5, "intervalos": ["2000-2016"]}
```

RETO 2: Perfil de las/los atletas

- Obtén la edad media de los atletas diferenciando entre juegos de Verano e Invierno (muestra ambas y compáralas en 3–5 líneas)
- Muestra un histograma de edades de los 5 países con más atletas participantes (un *FacetGrid* de histogramas por país)
- Muestra la distribución porcentual segregando por Hombre/Mujer y Verano/Invierno, en una gráfica adecuada (en porcentajes, no conteos). Justifica el tipo de gráfica por el que has optado.
- Calcula la altura y peso medio de las mujeres en los 5 deportes en los que hay más mujeres que hombres, y muestra una visualización que permita comparar esa distribución (p. ej., *boxplots* por deporte). Añade al menos 2 comentarios sobre qué muestran los resultados obtenidos

RETO 2: Perfil de las/los atletas

- Implementa la **función** `extremos_atletas(df_atletas)` que devuelva un diccionario con los/las atletas "extremos" (más joven, más mayor, de mayor altitud, menor altitud, con menor peso y mayor peso), y que aplique ciertas reglas de desempate (descritas a continuación)

- **Salida**: la función debe devolver la siguiente estructura de datos:

```
{
  "mas_joven": {"id": ..., "nombre": ..., "edad": ...},
  "mas_mayor": {"id": ..., "nombre": ..., "edad": ...},
  "mas_alto":  {"id": ..., "nombre": ..., "altura": ...},
  "mas_bajo":  {"id": ..., "nombre": ..., "altura": ...},
  "mas_ligero":{"id": ..., "nombre": ..., "peso": ...},
  "mas_pesado":{"id": ..., "nombre": ..., "peso": ...}
}
```

RETO 2: Perfil de las/los atletas

- **Notas**:
  - Descartar registros sin valor numérico válido (edades/alturas/pesos nulos o “no plausibles”)
  - Documentar y justificar qué umbrales has usado para considerar valores “no plausibles” (p. ej., alturas <120 cm o >250 cm)
  - Cuando haya empates, desempatar por más participaciones olímpicas (si hay información), y si persiste, por orden alfabético de nombre.
  - Para las participaciones puedes utilizar la función desarrollada en la tarea anterior.

RETO 3: Cantidad de medallas y rendimiento

- Dibuja un mapa de calor de país × deporte (en el que la intensidad sea el número total de medallas
- Obtén la clasificación de los 10 países con mayor número total de medallas (de cualquier tipo) y, para cada uno de esos 10 países, el deporte en el que más medallas ha conseguido
- De los 5 países líderes en medallas, calcula la evolución por décadas de su cantidad de medallas (elige y justifica cómo discretizar las décadas). Muestra una serie temporal por décadas y añade una breve conclusión
- Elige un deporte y analiza si hay relación entre la altura/peso de los deportistas y probabilidad de medalla en función del país. Compara todos los países para el deporte elegido, muestra una gráfica adecuada e interpreta los resultados (en 5–8 líneas)

RETO 3: Cantidad de medallas y rendimiento

- Implementa la **función** `ranking_paises_medallas(df_medallas, k)` que devuelva devuelva el *Top-k* de países, siguiendo de manera estricta la siguiente prioridad de desempate:
    1. El de más Oros
    2. Si empatan, el de más Platas
    3. Si empatan, el de más Bronces
    4. Si empatan, el del primer año en que el país logró una medalla (más antiguo gana)
    5. Si persiste el empate, por orden alfabético del país

- **Salida**: lista ordenada de *k* diccionarios como la siguiente:

```
[
  {“orden”: 1, "país": p1, "oro": n1, "plata": n2, "bronce": n3, "primer_año": a1},
  {“orden”: 2, "país": p2, "oro": n4, "plata": n5, "bronce": n6, "primer_año": a2},
  ...
]
```

RETO 4: Objetivo abierto

- Plantea un reto propio de análisis del *dataset* (claro y verificable) y resuélvelo
- Incluye una lista de 3-5 tareas de análisis similares a las que se han propuesto para los retos anteriores
- Debe haber al menos una visualización (gráfico) y se debe justificar el porqué ese tipo de gráfico es adecuado
- Segrega por juegos de Verano/Invierno, si aplica
- Se valorará si para la resolución de las tareas propuestas se implementa alguna función Python de soporte. En ese caso, debes documentar brevemente la función (su perfil, salida, qué hace, casos de prueba,...)
- Finaliza el reto con una conclusión interpretando los resultados observados y patrones o tendencias detectadas, y si procede, posibles limitaciones del *dataset* que afecten a dicha interpretación