

## Práctica 1 - Preparación de datos

Sistemas de Aprendizaje Automático - C.E. Inteligencia Artificial y Big Data.

I.E.S. Camas - Curso 2025-26

---

Realiza con Python un programa que cumpla con los requisitos que se muestran a continuación. Entrégalo en un cuaderno de Júpiter de Google Colab.

En este ejercicio vamos a preparar un conjunto de datos para un posterior estudio. Utilizaremos el dataset del Titanic, que almacena una serie de atributos de 891 pasajeros que viajaron en el Titanic y la etiqueta objetivo de si sobrevivieron o no.

La descripción del dataset es la siguiente:

1. **Survived** (Sobrevivió): Variable binaria que indica si el pasajero sobrevivió al hundimiento (1) o no (0).
2. **Pclass** (Clase): Clase del billete, que puede ser 1, 2 o 3. La clase 1 es la más alta.
3. **Sex** (Sexo): Género del pasajero, ya sea masculino o femenino.
4. **Age** (Edad): Edad del pasajero.
5. **SibSp**: Número de hermanos/cónyuges a bordo del Titanic.
6. **Parch**: Número de padres/hijos a bordo del Titanic.
7. **Fare** (Tarifa): Tarifa pagada por el pasajero, en libras.
8. **Embarked** (Embarcado): Puerto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton).
9. **Class** (Clase): Clase en la que viajaba el pasajero.
10. **Who**: Indica el sexo en palabras (man/woman).
11. **Adult\_male**: Indica si es un hombre adulto o no.
12. **Deck** (cubierta): Cubierta en la que viajaba el pasajero, de la A a la G.
13. **Embark\_town** (Ciudad de Embarque): Cherbourg, Queenstown, o Southampton.
14. **Alive** (vivo): Indica si sobrevivió al naufragio.
15. **Alone** (solo): Si viajaba solo o no.

Realiza las siguientes tareas:

1. **Tarea 1: Importación de librerías, carga del dataset y eliminación de columnas redundantes.**
  - 1.1. Carga el dataset en un dataframe. Usa el fichero de GitHub ubicado este enlace:
  - 1.2. <https://raw.githubusercontent.com/mwaskom/seaborn-data/master/titanic.csv>

Observa un ejemplo de carga de un fichero desde una URL en un Dataframe en Google Colab:

```
url = "https://raw.githubusercontent.com/mwaskom/seaborn-data/master/titanic.csv"  
midataframe = pd.read_csv(url)
```

- 1.3. Se eliminarán las columnas *Class*, *Who*, *Adult\_male*, *Embark\_town* y *Alive*.

2. **Tarea 2: Codificación de columnas de texto:**
  - 2.1. La columna *sex* debe codificarse de forma numérica como entero, siendo 0 para hombre y 1 para mujer.

## Práctica 1 - Preparación de datos

Sistemas de Aprendizaje Automático - C.E. Inteligencia Artificial y Big Data.

I.E.S. Camas - Curso 2025-26

- 2.2. La columna *Embarked* debe codificarse de forma numérica como entero, siendo 0 para Cherbourg, 1 para Queenstown y 2 para Southampton.
- 2.3. La columna *Deck* debe codificarse de forma numérica como entero, siendo 0 para A, 1 para B, ... y 6 para G.
- 2.4. La columna *Alone* debe codificarse de forma numérica como entero, siendo 0 si no viajaba solo y 1 si viajaba solo.

### 3. Tarea 3: Limpieza de datos vacíos y de datos fuera de rango

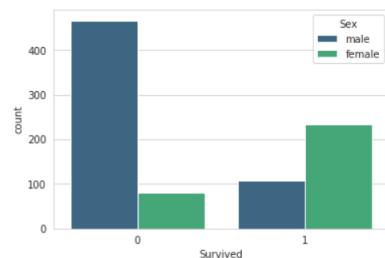
Considera el siguiente rango como el rango correcto de los atributos:

- Survived: {0, 1}
- Pclass: {1, 2, 3}
- Sex: {0, 1}
- Age: [0-100]
- SibSp: [0-10]
- Parch: [0-10]
- Fare: >0
- Embarked {0, 1,2}
- Deck: {0, 1, 2, 3, 4, 5, 6}
- Alone: {0, 1}

- 3.1. Se consideran datos inválidos aquellos que están vacíos, o que tienen un valor fuera del rango indicado. Para cada uno de los atributos, calcula el número de elementos inválidos (puedes usar el método *pandas.DataFrame.info*).
  - Si en alguna de las columnas el porcentaje de datos inválidos es superior al 25%, elimina esa columna completa.
  - Después de eso, en el resto, elimina las filas completas que tengan algún atributo inválido.

### 4. Tarea 4: Cálculos estadísticos básicos.

- 4.1. Haz un análisis estadístico relacionando el atributo *Sex* con el target (*Survived*). Muestra datos numéricos y un gráfico similar a este:



- 4.2. Haz un análisis estadístico relacionando el atributo *Pclass* con el target (*Survived*). Muestra datos numéricos y un gráfico similar al anterior.
- 4.3. Haz un análisis estadístico de la edad de los pasajeros que contenga, al menos, la media, la mediana, la desviación típica, el mínimo, el máximo, y los cuartiles Q1, Q2 y Q3. Representa la distribución de la edad de forma gráfica.

## Práctica 1 - Preparación de datos

Sistemas de Aprendizaje Automático - C.E. Inteligencia Artificial y Big Data.

I.E.S. Camas - Curso 2025-26

---

### 5. Tarea 5: Escalamiento de los datos

- 5.1. Escala TODOS los atributos por el método de estandarización por rangos (MinMaxScaler).

### 6. Tarea 6: Guarda el dataframe modificado en un fichero csv

- 6.1. Guarda el dataframe con todas las modificaciones en tu disco local, con el nombre *titanic\_nombre\_apellido1\_apellido2.csv*.

Observa el siguiente ejemplo para exportar un dataframe a csv y descargarlo al disco local de tu PC:

```
from google.colab import files  
mi_dataframe.to_csv('mi_csv.csv')  
files.download('mi_csv.csv')
```

---

#### OBSERVACIONES A TENER EN CUENTA:

- ✓ El código debe estar comentado.
  - ✓ Se valorará la simplicidad y legibilidad del código.
  - ✓ Se valorará el añadido de bloques de tipo texto entre los bloques de código, en los que se explique el proceso y se analicen y valoren los resultados que se vayan obteniendo.
- 

#### ENTREGA DE LA PRÁCTICA:

En la tarea de la plataforma Moodle, debes entregar lo siguiente:

1. El código de la práctica, que debe ser un archivo de Google Colab, con el nombre *Práctica1\_nombre\_apellido1\_apellido2.ipynb*.
  2. El fichero csv resultante, con el nombre *titanic\_nombre\_apellido1\_apellido2.csv*.
- 

#### RÚBRICA DE CORRECCIÓN:

ÍTEMES EVALUABLES		PESO
ÍTEM 1	Tarea 1: Importación de librerías, carga del dataset y eliminación de columnas redundantes	10%
ÍTEM 2	Tarea 2: Normalización/codificación de columnas	20%
ÍTEM 3	Tarea 3: Limpieza de datos vacíos y de datos fuera de rango	20%
ÍTEM 4	Tarea 4: Cálculos estadísticos básicos.	20%
ÍTEM 5	Tarea 5: Escalamiento de los datos	15%
ÍTEM 6	Tarea 6: Guarda el dataframe modificado en un fichero csv	5%
ÍTEM 7	Comentarios del código y explicación del proceso en bloques de texto.	10%