Entrega práctica – Bloque 2

Peso en la nota final: 35%
Formato de entrega:

- Jupyter Notebook *(.ipynb)* o script Python *(.py)*.
- El **código es obligatorio**; si alguna parte no se ejecuta, **la calificación será 0**.
- Podéis incluir comentarios en el script o si preferís podéis adjuntar un PDF adicional con el análisis, pero el código debe ser entregado.
- Trabajo en **parejas**: basta con que uno entregue, indicando **ambos nombres y apellidos** al inicio del código. No se permite trabajar con la misma pareja que en la primera entrega.
- Trabajos idénticos tendrán una calificación de 0 para ambos grupos.

Contexto del problema

En esta segunda entrega partiremos del mismo contexto general que en la primera práctica:
el análisis de los factores que pudieron influir en la **supervivencia de los pasajeros del Titanic** en su hundimiento el 15 de abril de 1912.

Sin embargo, en esta ocasión **trabajaremos con un dataset ligeramente adaptado** respecto al anterior:

- Contiene más registros y algunas columnas adicionales.
- Además, dispondrás de una base de datos complementaria con información meteorológica del día del suceso, que incluye variables como el clima y la temperatura media.

Conjunto de datos

Archivo: *Titanic__2.csv*

Descripción de algunas variables:

- *Survived:* Supervivencia (*0 = No, 1 = Sí*).
- *Pclass:* Clase del billete (*1 = 1ª clase, 2 = 2ª clase, 3 = 3ª clase*)
- *SibSp:* Número de hermanos/esposos a bordo
- *Parch:* Número de padres/hijos (algunos viajaban solo con niñera → *Parch == 0*)
- *Ticket:* Número del billete
- *Fare:* Tarifa pagada
- *Embarked:* Puerto de embarque (*C = Cherbourg, Q = Queenstown, S = Southampton*)
- *Barco:* Nombre del barco en el que viajaban los pasajeros.

Indicaciones generales

- Debéis ir completando los ejercicios propuestos en orden.
- **Lo más importante no será únicamente obtener los resultados**, sino **el análisis crítico y la justificación de las conclusiones** en cada apartado.
- Puedes reutilizar variables creadas en la primera entrega si las consideras útiles.
- Usa *random_state* fijo si aplicas métodos con componente aleatoria.

EJERCICIOS:

1.- Variables poco relevantes (0.5 puntos)

- Examina todas las variables del dataset e indica cuáles crees que inicialmente no van a aportan valor predictivo o son redundantes. **Justifica** tu elección.

2.- Valores perdidos (2 puntos)

- Calcula el porcentaje de valores perdidos por variable.
- Evalúa si sería conveniente eliminar TODOS los registros que contengan algún valor NaN en alguna de las variables:
    - ¿Qué porcentaje de registros se perdería?
    - ¿Crees que en este caso es una opción razonable? **Justifica** tu decisión y aplica los cambios que consideres oportunos.
- Imputa los valores faltantes aplicando dos métodos distintos (uno basado es estadísticos y otro basado en un método iterativo). **Justifica** tu elección.
    - ¿En este caso, que método consideras más adecuado?¿Por qué?
- Comprueba que no queden valores NaN y analiza si las distribuciones cambian significativamente tras imputar.

3.- Correcciones ortográficas (1 punto)

- Selecciona las variables categóricas o de tipo texto.
- Revisa que no existan errores tipográficos. En caso de detectar errores, corrígelos usando algún método (no sirve de forma manual).

4.- Valores extremos (outliers) (1.5 puntos)

- Selecciona las variables numéricas.

- ¿Qué gráfico consideras que es el más adecuado para detectar valores extremos? Realiza dicho gráfico para las variables numéricas seleccionadas. Dame una descripción **detallada** de cada uno de ellos.
- Analiza los gráficos junto con la descripción inicial de los datos e indica:
    - ¿Qué variables presentan valores extremos?
    - ¿Son reales o parecen errores de registro? **Justifica** tu respuesta.
- Aplica el tratamiento que consideres más apropiado y **justifica** tu elección.

5.- Filtrado y agregaciones (3 puntos)

- Comprueba si existen registros duplicados. Aplica el tratamiento que consideres adecuado **justificando** tu elección.
- Crea dos subconjuntos:
    - *df_male*: registros con *Sex == "male"*
    - *df_female*: registros con *Sex=="female"*

  Para cada subconjunto (utilizando funciones de agregación):
    - Agrupa por *Pclass* y calcula la edad media, mínima y máxima.
    - Agrupa por *Pclass* y calcula la tarifa media y desviación estándar (*Fare*).
    - Interpreta los resultados: ¿hay diferencias relevantes entre clases o sexos?.
- Une ambos subconjuntos en un solo DataFrame (*df_total*) utilizando la función adecuada.
- Agrupa *df_total* por *Pclass* y *Sex* (utilizando funciones de agregación), y calcula:
    - El total de pasajeros en cada grupo.
    - El número de sobrevivientes.
    - El porcentaje de supervivencia.
- Presenta los resultados en un DataFrame resumen con las columnas:
  1. *Total pasajeros*
  2. *Sobrevivientes*
  3. *% Supervivencia*
- Interpreta los resultados:
    - ¿Qué diferencias observas entre hombres y mujeres dentro de cada clase?
    - ¿Cómo cambia la probabilidad de supervivencia según la clase?
    - ¿Qué conclusiones puedes sacar sobre la relación entre género, clase y supervivencia?

- Realiza otro filtro y función de agregación que consideres interesante para el análisis. Puedes elegir cualquier variable o combinación de variables. Justifica por qué has elegido esa combinación y qué información nueva aporta.

6.- Uniones (1 punto)

- Une tu dataset principal con *condiciones_metereologicas.csv* usando la clave adecuada. Añade unicamente las variables *TemperaturaMedia* y *Clima*. Comprueba que la unión se haya realizado correctamente.
- Analiza si estas dos variables podrían aportar valor a tu modelo predictivo. **Justifica** tu respuesta.

7.- Conclusión final (1 punto)

- Guarda la base de datos final en un archivo *.csv.*
- Elabora una breve reflexión (máx. de 15 líneas) que incluya:
    - Principales problemas de calidad de datos detectados.
    - Decisiones tomadas para resolverlos y su justificación.
    - Cómo estas decisiones podrían afectar el rendimiento de un modelo predictivo posterior.