# **🛳️ Titanic Dataset – Descripción**

El dataset del Titanic proviene de Kaggle y está basado en los registros reales de los pasajeros del barco que naufragó en 1912.  
 El objetivo es predecir si un pasajero sobrevive (Survived = 1) o muere (Survived = 0) en base a sus características personales y del viaje.

📂 Columnas del dataset train.csv

* PassengerId: identificador único de cada pasajero. No aporta información para el modelo → se suele eliminar.
* Survived: variable objetivo (0 = no sobrevivió, 1 = sobrevivió).
* Pclass: clase del billete del pasajero (1ª, 2ª o 3ª). Relacionada con estatus socioeconómico.
* Name: nombre completo del pasajero. Contiene títulos (*Mr, Mrs, Miss, Master*) que pueden extraerse como *feature engineering*.
* Sex: género del pasajero (male/female). Una de las variables más influyentes.
* Age: edad del pasajero (en años). Tiene valores nulos.
* SibSp: número de hermanos/esposos/as a bordo.
* Parch: número de padres/hijos a bordo.
* Ticket: número del billete. Poco útil directamente; a veces se analiza el prefijo.
* Fare: tarifa pagada por el pasajero. Variable numérica, con algunos outliers.
* Cabin: número de cabina. Tiene muchos valores nulos; en algunos análisis se usa solo la letra (zona del barco).
* Embarked: puerto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton). Tiene algunos nulos.

🔹 Posibles problemas y decisiones de limpieza

* Duplicados: revisar si los hay.
* Nulos: en Age, Embarked y especialmente en Cabin.
* Outliers: tarifas (Fare) con valores muy altos.
* Columnas poco útiles: PassengerId, Ticket, Cabin (en muchos casos se eliminan).

**Objetivo en Machine Learning**

El reto clásico es predecir la variable Survived a partir del resto de características.  
 Se suele usar:

* Modelos de clasificación binaria (Logistic Regression, Random Forest, XGBoost, etc.).
* Features creadas como:
  + FamilySize = SibSp + Parch + 1.
  + IsAlone (viaja solo o acompañado).
  + Títulos extraídos de Name.

## **Preguntas / Instrucciones para el Titanic**

1. Carga el dataset y revisa las primeras filas (head()), la información (info()) y las estadísticas básicas (describe()).
2. ¿Cuántos valores nulos hay en cada columna? (isnull().sum()).
3. Elimina las columnas que no aportan información al modelo (ej: PassengerId, Name, Ticket). ¿Por qué?
4. ¿Qué harías con la columna Cabin? ¿La mantienes o la eliminas?
5. Imputa los valores nulos en Age. ¿Prefieres media, mediana o moda? Justifica tu elección.
6. Imputa los valores nulos en Embarked. ¿Cuál sería la estrategia más razonable?
7. Revisa la variable Fare. ¿Hay outliers? ¿Qué estrategias aplicarías?
8. Convierte la variable Sex en numérica. (ej: 0 = male, 1 = female).
9. Convierte la variable Embarked en variables dummy (one-hot encoding).
10. Agrupa la variable Age en intervalos (ej: niño, joven, adulto, mayor). ¿Cuántos pasajeros hay en cada grupo?
11. ¿Qué porcentaje de mujeres sobrevivió frente a hombres?
12. ¿Qué porcentaje de pasajeros de primera clase sobrevivió frente a los de tercera clase?
13. Crea una nueva variable FamilySize = SibSp + Parch + 1. ¿Influye en la supervivencia?
14. Crea una nueva variable binaria IsAlone (1 si FamilySize = 1, 0 en otro caso).
15. Revisa la correlación entre las variables numéricas y el target (Survived). ¿Qué observas?
16. Haz un conteo de pasajeros por clase (Pclass) y compara con la tasa de supervivencia.
17. ¿Qué columnas dejarías en tu dataset final para modelar? ¿Cuáles eliminarías y por qué?
18. Normaliza o estandariza las variables numéricas (Age, Fare) si fuera necesario.
19. ¿Qué variables derivadas se te ocurren que podrían mejorar un modelo?
20. Guarda el dataset limpio en un CSV final para usarlo en un modelo de ML.