# M2.851 - Tipología y ciclo de vida de los datos

Práctica 2

Alumnos:

Michaelle Estefanía Valenzuela Sangoquiza

Juan Manuel Penalta Rodríguez.

Índice

[M2.851 - Tipología y ciclo de vida de los datos 1](#__RefHeading___Toc1001_781917998)

[1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? 3](#__RefHeading___Toc2709_2179713084)

[2. Integración y selección de los datos de interés a analizar. 3](#__RefHeading___Toc2711_2179713084)

[3. Limpieza de los datos. 3](#__RefHeading___Toc2713_2179713084)

[3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 3](#__RefHeading___Toc2715_2179713084)

[3.2. Identificación y tratamiento de valores extremos. 3](#__RefHeading___Toc2717_2179713084)

[4. Análisis de los datos. 3](#__RefHeading___Toc2719_2179713084)

[4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). 3](#__RefHeading___Toc2721_2179713084)

[4.2. Comprobación de la normalidad y homogeneidad de la varianza. 3](#__RefHeading___Toc2723_2179713084)

[4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes. 3](#__RefHeading___Toc2725_2179713084)

[5. Representación de los resultados a partir de tablas y gráficas. 3](#__RefHeading___Toc2727_2179713084)

[6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? 4](#__RefHeading___Toc2729_2179713084)

## 1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset está compuesto por datos de pasajeros del Titanic, hundido en el oceano atlántico en el año 1912 tras chocar con un iceberg. En su momento fue una de las tragedias náuticas más importantes en tiempos de paz. Murieron 1496 de los 2208 pasajeros.

Los datos de este dataset fueron obtenidos originalmente de la página Kaggle (<https://www.kaggle.com/c/titanic/data>). Los datos con los que se realiza este trabajo se pueden consultar en el siguiente enlace:

[Enlace a github donde están los datos]

Los datos de los que disponemos son los siguientes:

* **Survival**: Variable numérica que indica si el pasajero surpervivió o murió en el hundimiento.
  + 1 = Superviviente
  + 0 = No superviviente
* **pclass**: Variable numérica que indica el tipo de ticket que tenía el pasajero.
  + 1 = Primera Clase
  + 2 = Segunda Clase
  + 3 = Tercera Clase
* **sex**: Variable alfanúmerica que indica el genero del pasajero: male o female
* **age**: Variable numérica que indica la edad en años del pasajero.
* **sibsp**: Variable numérica que indica el número de hermanos o esposos del pasajero que viajan con él.
* **parch**: Variable numérica que indica el número de padres o hijos del pasajero que viajan con él.
* **ticket**: Variable alfanumérica con el número de ticket del pasajero.
* **fare**: Variable numérica que indica la tarifa pagada por el pasajero.
* **cabin**: Variable alfanumérica que indica la cabina o cabinas que ocupaba el pasajero y sus parientes.
* **embarked**: Variable alfanumérica que indica el puerto de embarque del pasajero.
  + C = Cherbourg
  + Q = Queenstown
  + S = Southampton

En total el dataset dispone de 891 entradas.

Con estos datos vamos a intentar analizar si existe alguna relación entre las variables disponibles y la supervivencia de los pasajeros. En concreto intentaremos averiguar en este trabajo:

* Si se cumplió el protocolo de salvamento que rige el proceso de evacuación de la nave, conocido como «mujeres y niños primero». Es decir si existe una relación entre el sexo y la edad y la supervivencia
* Si existió una relación entre la clase social del pasaje y sus posibilidades de supervivencia.

## 2. Integración y selección de los datos de interés a analizar.

## 3. Limpieza de los datos.

### 3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

### 3.2. Identificación y tratamiento de valores extremos.

## 4. Análisis de los datos.

### 4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

### 4.2. Comprobación de la normalidad y homogeneidad de la varianza.

### 4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

## 5. Representación de los resultados a partir de tablas y gráficas.

## 6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

## 7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos.

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa |  |
| Redacción de las respuestas |  |
| Desarrollo código |  |