Tipología y ciclo de vida de los datos

# Práctica 2: Limpieza y validación de los datos

Paula León Gil-Gibernau (PLG): aula 1

Silvina Guijarro Domingo (SGD): aula 2

Contenido

[Práctica 2: Limpieza y validación de los datos 1](#_Toc9157801)

[1. Presentación 3](#_Toc9157802)

[2. Descripción del dataset 3](#_Toc9157803)

[2.1. ¿Por qué es importante y qué pregunta/problema pretende responder? 3](#_Toc9157804)

[3. Integración y selección de los datos de interés a analizar 3](#_Toc9157805)

[4. Limpieza de los datos 3](#_Toc9157806)

[4.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 3](#_Toc9157807)

[4.2. Identificación y tratamiento de valores extremos. 3](#_Toc9157808)

[5. Análisis de los datos 3](#_Toc9157809)

[5.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). 3](#_Toc9157810)

[5.2. Comprobación de la normalidad y homogeneidad de la varianza. 3](#_Toc9157811)

[5.3. 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](#_Toc9157812)

[6. Representación de los resultados a partir de tablas y gráficas. 4](#_Toc9157813)

[7. Resolución del problema 4](#_Toc9157814)

[7.1. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? 4](#_Toc9157815)

[7.2. ¿Los resultados permiten responder al problema? 4](#_Toc9157816)

## Presentación

En esta práctica se debe elaborar un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

El objetivo de esta actividad será el tratamiento del dataset del Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic ).

## Descripción del dataset

El conjunto de datos objeto de análisis se ha extraído de <https://www.kaggle.com/c/titanic>. Este dataset está constituido por dos bases de datos, uno de entrenamiento y otro de test de validación. Para poder predecir las variables asociadas a la supervivencias, nos hemos centrado en el uso del dataset de entrenamiento, que está configurado por un total de 12 variables (columnas) de una muestra de 891 sujetos (filas o registros). Para comprobar nuestras predicciones nos centraremos en el uso del dataset test que esta formado por 11 variables de una muestra 418 sujetos. Los campos de este conjunto de datos encontramos las siguientes variables:

1. PassengerId: Identificacion del pasajero. Tipo: integer.

2. Survived: Sobrevivieron: 0=no, 1=sí. Tipo: Factor.

3. Pclass: Clase de billete: 1=primera, 2=segunda, 3=tercera. Tipo: Factor. (Es un indicador de clase socioeconomica)

4. Name: Nombre. Tipo: Factor.

5. Sex: Sexo. Tipo: Factor.

6. Age: Edad en años. Tipo: Num.

7. SibSp: Número de parientes como hermanos, hermanas, esposo o esposa.. Tipo: Integer.

8. Parch: Número de parientes como padre, madre, hijo, hija.. Tipo: Integer.

9. Ticket: Número de ticket. Tipo: Factor.

10. Fare: Tarifa de pasajero. Tipo: Num.

11. Cabin: Número de cabina. Tipo: Factor.

12. Embarked: Puerto de embarque: C=Cherbourg, Q=Queenstown, S=Southampton. Tipo: Factor.

### ¿Por qué es importante y qué pregunta/problema pretende responder?

El hundimiento del Titanic ha sido uno de los naufragios más famosos de la historia. La noche del 14 al 15 de abril de 1912, en su viaje inaugural, el Titanic se hundió después de chocar con un iceberg en el océano Atlántico. En dicho naufragio fallecieron 1502 personas, de los 2224 personas contabilizados entre pasajeros y tripulación. Esta desgracia conmocionó a la comunidad internacional, tras su hundimiento se realizaron varias comisiones de investigación de la tragedia que sirvieron para realizar mejoras en la regulación de seguridad marítima de transatlánticos.

Una de las razones por las que el naufragio produjo tantas pérdidas fue porque no había suficientes botes salvavidas para los pasajeros y la tripulación. Aunque es probable que hubiera algún factor de azar involucrado en la supervivencia al hundimiento, algunos grupos de personas como las mujeres, los niños y los pasajeros de clase alta tuvieron más probabilidades de sobrevivir que otros.

El siguiente análisis tiene como objetivo poder predecir qué características poseían las personas que tuvieron más probabilidades de sobrevivir.

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

## Limpieza de los datos

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

### Identificación y tratamiento de valores extremos.

## Análisis de los datos

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

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

### 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.

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

## Resolución del problema

### A partir de los resultados obtenidos, ¿cuáles son las conclusiones?

### ¿Los resultados permiten responder al problema?