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

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

## 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?