ANÁLISIS EXPLORATORIO DE DATOS (AED)

1 ¿QUÉ ES?

Conjunto de **técnicas estadísticas** dirigidas a explorar, describir y resumir la información que contienen los datos, maximizando su comprensión.

Gracias a ello puedes:

- Realizar un análisis descriptivo
- Identificar posibles errores
- Revelar la presencia de datos atípicos
- Comprobar la relación entre las variables

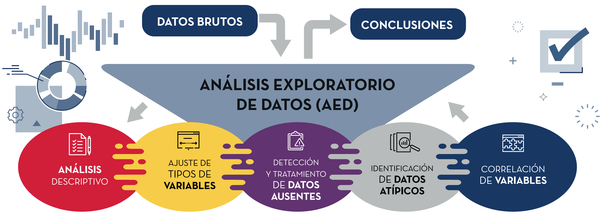
¿POR QUÉ ES IMPORTANTE?

- Las técnicas estadísticas de **análisis de datos y el machine learning** presuponen el cumplimiento de unas condiciones previas para garantizar la **objetividad e interoperabilidad** de los datos.
- El **AED** es esencial para garantizar que los resultados de cualquier análisis estadístico sean **consistentes y veraces**.
- Además, permite comprender exhaustivamente los **datos** antes de analizarlos, caracterizar sus **atributos** principales y descubrir las **interrelaciones entre variables**.

2 ¿CUÁLES SON LOS PASOS A SEGUIR?

Fuente: se han tomado como referencia dos libros:

- **R for Data Science de Wickman y Grolemund (2017)**
- **Python Data Science Handbook" de Jake VanderPlas (segunda edición 2023)**



1 ANÁLISIS DESCRIPTIVO

- **¿Qué es?** Síntesis de la información que proporciona el conjunto de datos, extrayendo sus características más representativas.
- **¿Por qué es necesario?** Permite conocer los tipos de datos, descubrir patrones y preparar los datos para futuros análisis.
- **Tratamiento:** Aplicar funciones de estadística descriptiva para explorar la estructura del conjunto de datos, examinar los datos y las variables que presenta.

2 RE-AJUSTE DE LOS TIPOS DE VARIABLES

- **¿Qué es?** Verificar que las variables se han almacenado con el tipo de valor correspondiente.
- **¿Por qué es necesario?** Una mala codificación de las variables puede influir negativamente en la agrupación de los datos o los resultados de los análisis.
- **Tratamiento:** Aplicar la codificación apropiada para cada una de las variables.

3 DETECCIÓN Y TRATAMIENTO DE DATOS AUSENTES

- **¿Qué es?** Identificar la falta de algunos de los datos en la variable.
- **¿Por qué es necesario?** Los datos ausentes pueden generar problemas a la hora de aplicar técnicas de machine learning, elaborar modelos predictivos, realizar análisis estadísticos o generar representaciones gráficas.
- **Tratamiento:** Existen varias maneras de tratar los valores ausentes, como por ejemplo sustituirlos por la media o la mediana, o completar los valores faltantes con el valor anterior o posterior de la columna.

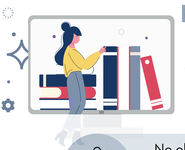
4 DETECCIÓN Y TRATAMIENTO DE DATOS ATÍPICOS

- **¿Qué es?** Identificar datos con valores significativamente distintos a los que presenta la variable.
- **¿Por qué es necesario su tratamiento?** Pueden modificar los resultados y restar potencia a los análisis estadísticos o técnicas de machine learning aplicadas.
- **Tratamiento:** Disminuir su influencia en análisis posteriores o, en casos muy extremos, eliminarlos del conjunto de datos.

5 ANÁLISIS DE CORRELACIÓN DE VARIABLES

- **¿Qué es?** Analizar la relación entre dos o más variables.
- **¿Por qué es necesario?** Entre otras razones, para descartar posibles variables que aporten información redundante en el conjunto de datos, ocasionando ruido en los análisis.
- **Tratamiento:** Calcular los coeficientes de correlación para las variables para detectar coeficientes cercanos a 1 ó -1.

3 TENDENCIA EMERGENTE
ANÁLISIS EXPLORATORIO DE DATOS AUTOMATIZADO



Actualmente existen bibliotecas que ofrecen soluciones eficientes para **generar informes y visualizaciones de AED de manera automática**.

Ejemplo

YData Profiling

No obstante, el AED automatizado tiene **limitaciones**. El científico de datos debe interpretar los resultados con criterio y conocimiento del contexto.

VENTAJAS

- **EFICIENCIA Y RAPIDEZ.**
- **VISIÓN GENERAL COMPLETA.**
- **DETECCIÓN TEMPRANA DE PROBLEMAS.**

4 ¿QUIERES SABER MÁS?

¡Aprende practicando las técnicas anteriores!

Descubre nuestras **guías prácticas**, donde se explica cómo realizar un AED paso a paso utilizando un conjunto de datos real

- Introducción al Análisis Exploratorio con R
- Introducción al Análisis Exploratorio con Python

* Ambas guías utilizan el mismo **conjunto de datos** para poder comparar: registro de la calidad del aire en Castilla y León.

Todos los materiales están disponibles en **Github** para que puedas replicar el ejercicio: laboratorio-de-Datos/Data Science at main · Admindatosgobes/Laboratorio-de-Datos · GitHub