datos.gob.es

ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Guía Práctica de Introducción al Análisis Exploratorio de Datos

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:

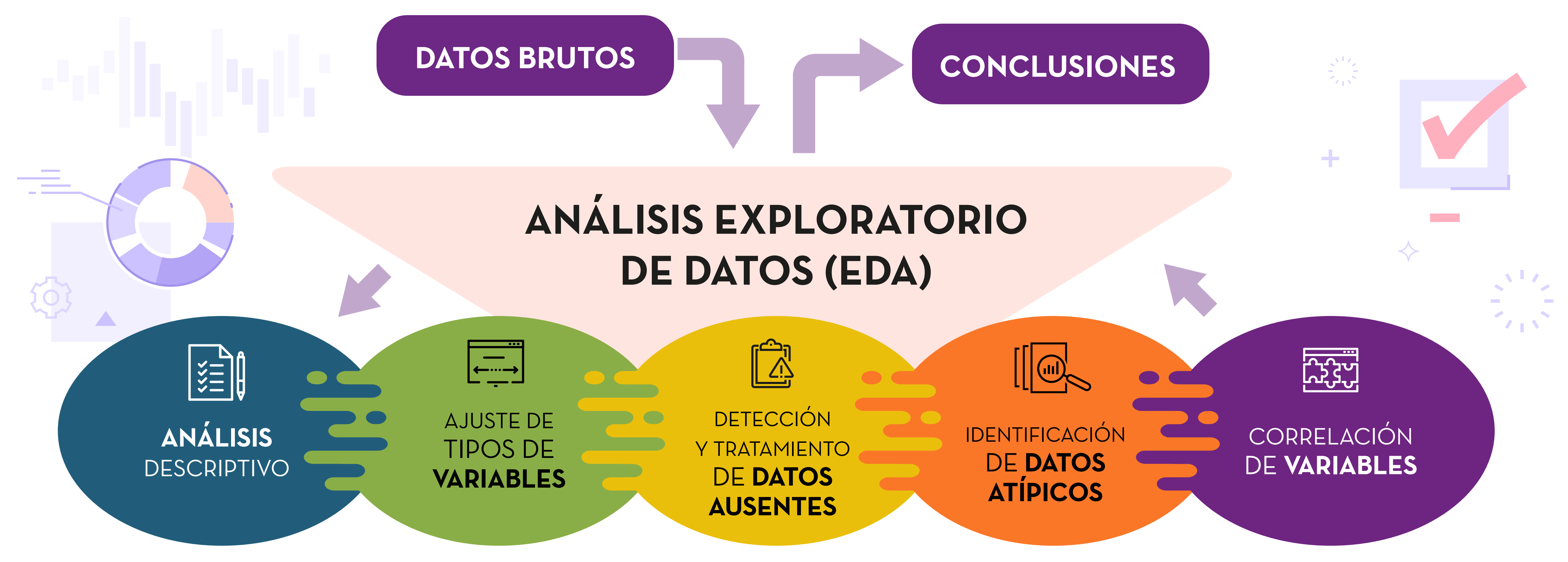
- Realizar un análisis descriptivo
- Identificar posibles errores
- Revelar la presencia de datos atípicos
- Comprobar la relación entre las variables

¿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 **EDA** es esencial para garantizar que los resultados de cualquier análisis estadístico sean **consistentes y veraces**.

2 ¿CUÁLES SON LOS PASOS A SEGUIR?

Fuente: Se ha tomado como referencia el libro **R for Data Science de Wickman y Grolemund (2017)**



DATOS BRUTOS

CONCLUSIONES

ANÁLISIS EXPLORATORIO DE DATOS (EDA)

ANÁLISIS DESCRIPTIVO

AJUSTE DE TIPOS DE **VARIABLES**

DETECCIÓN Y TRATAMIENTO DE **DATOS AUSENTES**

IDENTIFICACIÓN DE **DATOS ATÍPICOS**

CORRELACIÓN DE **VARIABLES**

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 o -1.

3 ¿QUIERES SABER MÁS?

Descubre la "**Guía Práctica de Introducción al Análisis Exploratorio de Datos**"

Aprende las técnicas anteriores mediante el desarrollo de un **caso práctico**.

Metodología para el ejemplo práctico:

- **Conjunto de datos utilizado**: registro de la calidad del aire en Castilla y León.
- **Herramientas utilizadas**: código R en el entorno de programación RStudio

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Iniciativa aporta