datos.gob.es

ANÁLISIS DE DATOS ABIERTOS CON HERRAMIENTAS *OPEN SOURCE* (Parte 1)

¿QUÉ ES?

El análisis de datos es un proceso sistemático que permite **transformar información en bruto en conocimiento valioso**, facilitando:

- Toma de decisiones informadas.
- Identificación de patrones y tendencias.
- Comprensión profunda de conjuntos de datos complejos.

PRIMEROS PASOS DEL PROCESO INTEGRAL DE ANÁLISIS DE DATOS

Depuración de datos
Conversión de datos
Análisis de datos

1. Preparación y depuración de datos

Objetivo: transformar datos en bruto en conjuntos limpios y estructurados.

Ejemplos de herramientas

OpenRefine — COMPLETAMENTE GRATUITA
Limpieza y transformación de datos. Ofrece una interfaz gráfica e intuitiva que es multiplataforma (requiere Java). Detecta:
- Duplicidades
- Datos incompletos
- Inconsistencias estructurales

Talend Open Studio — GRATIS EN SU VERSIÓN BÁSICA
Herramienta ETL (Extraer, Transformar, Cargar).
(!) Requiere conocimientos intermedios de programación.
Permite:
- Programación por componentes
- Integral datos de múltiples fuentes

2. Conversión de datos

Objetivo: adaptar el formato de los datos para facilitar su análisis.

Ejemplos de herramientas

Mr Data Converter — COMPLETAMENTE GRATUITA
- Conversión entre formatos CSV, Excel, JSON, HTM, XLM
- Interfaz web sencilla
- Sin instalación requerida

Pandoc — COMPLETAMENTE GRATUITA
- Conversión universal de documentos
- Soporta más de 20 formatos diferentes
- Línea de comandos potente

Tabula — COMPLETAMENTE GRATUITA
- Extracción de tablas desde PDF
- Convierte documentos en formatos reutilizables
- Útil para informes y documentación oficial

3. Análisis de datos

Objetivo: explorar, procesar y obtener *insights* de los conjuntos de datos.

***Software* de análisis amigable**

WEKA — COMPLETAMENTE GRATUITA
- Aprendizaje automático y minería de datos
- Interfaz gráfica
- Integración con scikitlearn, R y Deeplearning
- Ideal para principiantes en *machine learning*

ORANGE — COMPLETAMENTE GRATUITA
- Paradigma *drag and drop* (arrastrar y soltar)
- Visualizaciones interactivas
- Análisis estadístico accesible

KNIME — GRATIS EN SU VERSIÓN BÁSICA
- Análisis de datos visual
- Flujos de trabajo mediante conexión de nodos
- Amplia biblioteca de componentes

Entornos de desarrollo

Jupyter Notebook — COMPLETAMENTE GRATUITA
- Documentos ejecutables
- Combinación de código, visualizaciones y narrativa
- Soporta múltiples lenguajes
- Ideal para reproducibilidad

RStudio — GRATIS EN SU VERSIÓN BÁSICA
- Entorno completo para lenguaje R
- Integración de consola, editor y visualización
- Herramientas estadísticas avanzadas

Lenguajes de programación para análisis de datos

R — COMPLETAMENTE GRATUITA
Especializado en estadística
- Potente para análisis estadístico y visualización
- Bibliotecas destacadas:
  - » **Tidyverse**
  - » **ggplot2**

Python — COMPLETAMENTE GRATUITA
Lenguaje versátil
- Recomendación: usar Anaconda para gestión de entornos
- Las principales bibliotecas para análisis son:
  - » **Pandas** (manipulación de datos)
  - » **NumPy** (cálculo numérico)
  - » **scikit-learn** (*machine learning*)
  - » **Matplotlib** (visualización)

Herramientas emergentes

Streamlit — COMPLETAMENTE GRATUITA
- Creación rápida de aplicaciones web de datos
- Solo requiere Python
- Prototipado veloz de *dashboards*
  - » En este ejercicio práctico lo utilizamos para crear un chat de datos públicos.

Polars — COMPLETAMENTE GRATUITA
- Alto rendimiento
- Alternativa optimizada de pandas
- Procesamiento paralelo

Apache Spark — COMPLETAMENTE GRATUITA
- Procesamiento distribuido
- Para *Big data*
- APIs en Python, R y Scala

Recomendación
Comienza con herramientas sencillas como Jupyter y Python, y gradualmente explora opciones más avanzadas según tus necesidades de análisis.

Descubre aquí los beneficios y pasos del Análisis Exploratorio de Datos (AED).

GOBIERNO DE ESPAÑA · MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA · SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es · iniciativa aporta · datos.gob.es