Copilot

La limpieza de datos, también conocida como “Data Cleaning” o “Data Cleansing”, es un proceso esencial en Data Science y Machine Learning1. Consiste en resolver anomalías en conjuntos de datos (Datasets), para poder explotarlos después1.

Este proceso engloba varias técnicas destinadas a mejorar la calidad de los datos1. En general, esto significa identificar y sustituir los datos o registros incompletos, inexactos, corruptos o irrelevantes1. Después de una limpieza de datos correctamente realizada, todos los conjuntos de datos deben ser coherentes y estar libres de errores1.

La limpieza de datos es crucial por varias razones:

- **Calidad de los datos**: Es muy importante garantizar la calidad de los datos, ya que son el combustible de tecnologías como la ciencia de los datos, la inteligencia artificial y el machine learning1.
- **Análisis Precisos**: Sin la limpieza, es probable que los resultados de los análisis estén distorsionados1.
- **Rendimiento de los Modelos**: Un modelo de machine learning o de IA entrenado con datos erróneos puede estar sesgado o ofrecer un rendimiento deficiente1.
- **Costos**: Según un estudio de IBM, la mala calidad de los datos cuesta a Estados Unidos 3,1 billones de dólares al año1. La prevención a través del Data Cleaning es relativamente asequible, pero arreglar los problemas existentes puede costar diez veces más1.

Por lo tanto, la limpieza de datos es una etapa crucial en cualquier proyecto de Data Science para asegurar la precisión y la eficacia de los análisis y modelos1.