Clase 7: Ingeniería de características

FM849 - Programación Científica para Proyectos de Inteligencia Artificial (IA)

8 de enero de 2026

Motivación

En general, los datos reales vienen con problemas como:

- Datos faltantes.
- Datos desordenados.

Es importante entregar datos no nulos y no repetidos a modelos, ya que estas situaciones pueden crear un sesgo. En esta clase, seguiremos aprendiendo algunas funciones útiles en `pandas` que nos permitirán modificar *DataFrames* y realizar un análisis exploratorio sobre datos tabulares.

Continuación del ejemplo práctico

Vamos a seguir explorando el conjunto de datos que tiene información sobre Pokémon en Google Colab.

Limpieza y preparación de datos

- Tratamiento de valores faltantes.
  - Detectar valores NA → `isna()`, `notna()`.
  - Eliminar valores NA → `dropna()`, `fillna()`.

  `NA` se refiere a valores nulos/no disponibles.
- Corrección de tipos de datos.
  - Cambiar el tipo → `astype()`.
  - Crear fechas → `to_datetime()`.
- Limpieza básica.
  - Reemplazar valores → `replace()`.
  - Identificar y remover duplicados → `duplicated()`, `drop_duplicates()`.

Agregación de datos

- Cálculo de estadísticas que resumen los datos.
  - `mean()`, `sum()`, `count()`, `min()`, `max()`, `std()`.
- Agregación por grupos.
  - `agg()`.
- Transformaciones agregadas.
  - Transformar datos $\rightarrow$ `transform()`.

Agrupamiento de datos

- Agrupamiento por variables categóricas.
  - `groupby()`.
- Agrupamiento por intervalos.
  - `cut()`.
  - `qcut()`.
- Reestructuración de datos.
  - `pivot()`, `pivot_table()`.
  - `melt()`.

Referencias

Wes McKinney. *Python for Data Analysis*. O'Reilly Media, 3 edition, 2022.