Para estudiantes de Educación Básica y Media.
UNIVERSIDAD DE CHILE

Clase 7: Ingeniería de Características.

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

Contenidos de hoy

- Limpieza y Preparacion de Datos.
- Agregacion de Datos.
- Agrupamiento de Datos.

Motivación

En general, los datos reales vienen con problemas:

- 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 utiles en pandas que nos permitirán modificar dataframes y realizar análisis exploratorio sobre datos tabulares.

Vamos a continuar con el ejemplo de la clase anterior

Vamos a continuar explorando el dataset de Pokemons: https://colab.research.google.com/drive/1qzo9PRNI5_31AzmoNyVx1I22hJy5MCDf?usp=sharing.

Figura 1: Pokemones de primera generación.

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()`
  - Remover duplicados: → `duplicated()`, `drop_duplicates()`

Agregación de Datos

- **Cálculo de estadísticas resumen**
  - Calcular estadisticas → `mean()`, `sum()`, `count()`
  - Calcular estadisticas → `min()`, `max()`, `std()`
- **Agregación por grupos**
  - Agregar datos → `agg()`
- **Transformaciones agregadas**
  - Transformar datos → `transform()`

Agrupamiento de Datos

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

Referencias:

- Wes McKinney. (2022). Python for Data Analysis. Third Edition.