

# DESCRIPCION DATASET- PARCIAL –APRENDIZAJE AUTOMATICO

## Segunda entrega:

### Descripción del dataset

#### ❖ Origen del dataset

Los datos utilizados en este proyecto provienen de la Encuesta Permanente de Hogares (EPH), publicada por el INDEC (Instituto Nacional de Estadística y Censos de Argentina).

Se utilizó el archivo correspondiente al primer trimestre del año 2023 (T423), específicamente el módulo de personas (usu\_individual\_T423.txt), descargado desde el sitio oficial en mayo de 2025.

#### ❖ Proceso de recopilación y preprocesamiento

- Se cargó el archivo con separador (:) y codificación (latin1).
- Se filtraron únicamente las personas en condición de actividad económica: ocupados (estado = 1) y desocupados (estado = 2).
- Se eliminaron columnas irrelevantes y se seleccionaron solo las variables necesarias para el análisis.
- Se creó una nueva variable target, donde:
  - \* 1 indica persona ocupada
  - \* 0 indica persona desocupada

#### ❖ Características del dataset final

- Cantidad de instancias (filas): 22.367
- Cantidad de variables (columnas): 5
- Variables incluidas:

Variable	Tipo	Descripción
sexo	numérica (int64)	Sexo biológico (1 = varón, 2 = mujer)
edad	numérica (int64)	Edad en años
nivel_educativo	numérica (int64)	Nivel educativo alcanzado (codificado por EPH)
categoria_ocupacional	numérica (int64)	Tipo de ocupación principal
target	numérica (int64)	Variable objetivo (1 = ocupado, 0 = desocupado)

Este dataset procesado puede ser consultado y utilizado dentro de la carpeta data/processed/ del repositorio del proyecto.