LAURA DANIELA DIAZ TORRES
VIVIAN CABANZO FERNÁNDEZ
ZENNETH OLIVERO TAPIAS
CRISTIAN FELIPE MUÑOZ GUERRERO

2025-02

BIG DATA Y MACHINE LEARNING PARA ECONOMIA APLICADA

DATOS

Universidad de los Andes

Descripción de los datos y muestra

- **Fuente:** Gran Encuesta Integrada de Hogares (GEIH) – DANE
- **Año de análisis:** 2018
- **Muestra de estudio:** Ocupados mayores a 18 años con ingresos positivos reportados.
- **Cobertura:** Bogotá.
- **Obtención de datos:** Información tomada del portal de GitHub del profesor Ignacio Sarmiento (Uniandes).

Variables y creación de variable de interés

Variables base de datos:

- **Impa:** ingreso monetario primera actividad
- **Hours_work_usual:** horas trabajadas usualmente por semana (primera actividad).
- **Max_educ_level:** máximo nivel educativo alcanzado
- **Relab:** Tipo de ocupación
- **Sex:** sexo.
- **Age:** edad
- **Mes:** mes en la que se realizó la encuesta
- **Size_firm:** tamaño de la firma donde trabaja
- **Directorio, secuencia_p, orden:** identificación del individuo

Variable creada:

$$\text{Ingreso por hora} = \frac{impa}{hours\_work\_usual \times 4{,}33}$$

(4,33 es el promedio de semanas en un mes)

Universidad de los Andes

Manipulación de datos

- Exclusión de ingresos = 0 o faltantes
- Imputación de valor faltante en la variable max_educ_level (moda)
- Transformación: logaritmo natural del ingreso/hora
  - Justificación: reducir peso de valores extremos (outliers)

Descriptivos

- Tamaño de empresa: 30% en firmas >15 empleados
- Edad promedio: 38 (mujeres y hombres)
- Horas trabajadas: 44.3 (M) / 49.9 (H)
- Ingreso hora promedio: $8.402 (M) / $8.849 (H)
- Educación: más mujeres con nivel terciario
- Formalidad: 60% en ambos sexos