

Tarea: Análisis Exploratorio de Datos – Adult Income Dataset

1. Introducción

El objetivo de esta tarea es realizar un análisis exploratorio de datos completo sobre el dataset *Adult Income*. Se deberá identificar patrones, relaciones y factores que influyen en el ingreso de las personas (>50K), y presentar sus hallazgos en un notebook y un PDF resumen.

2. Descripción del Dataset

El dataset Adult Income contiene información de censos de individuos incluyendo edad, educación, sexo, horas trabajadas, ocupación, ganancias y pérdidas de capital, entre otras variables. La variable objetivo es 'income' (<=50K, >50K).

Variable	Tipo	Descripción
age	Numérica	Edad del individuo en años
workclass	Categórica	Tipo de trabajo (ej. Private, Self-emp-not-inc, Govt)
fnlwgt	Numérica	Peso final de la muestra (representa la cantidad de personas que representa cada fila)
education	Categórica	Nivel educativo (ej. Bachelors, HS-grad, Masters, etc.)
education-num	Numérica	Nivel educativo en número (ej. 1-16, equivalente a años de educación)
marital-status	Categórica	Estado civil (ej. Married-civ-spouse, Never-married, Divorced)
occupation	Categórica	Tipo de ocupación laboral (ej. Tech-support, Craft-repair, Exec-managerial)
relationship	Categórica	Relación familiar (ej. Wife, Husband, Own-child, Unmarried)
race	Categórica	Raza (ej. White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other)
gender	Categórica	Género (Male/Female)
capital-gain	Numérica	Ganancia de capital obtenida durante el año
capital-loss	Numérica	Pérdida de capital durante el año
hours-per-week	Numérica	Número de horas trabajadas por semana
native-country	Categórica	País de origen (ej. United-States, Mexico, Philippines, etc.)
income	Categórica	Variable objetivo: ingreso anual (<=50K o >50K)

3. Objetivos de la tarea

- Cargar y preparar los datos.
- Realizar estadísticas descriptivas de variables numéricas y categóricas.
- Analizar distribuciones univariadas y detectar outliers.
- Explorar relaciones bivariadas entre variables y la variable objetivo.
- Generar gráficos tipo subplot, boxplots, countplots e histogramas.
- Calcular la matriz de correlación y visualizarla con heatmap.
- Identificar hallazgos y conclusiones relevantes.
- Presentar el trabajo en un notebook y un PDF resumen con gráficos y conclusiones.

4. Requerimientos de entrega

La tarea debe entregarse en dos formatos:

- a. Notebook (.ipynb) con todas las secciones de análisis y gráficos.
- b. PDF resumen con los principales gráficos y conclusiones.

Se recomienda organizar las gráficas en subplots y usar títulos y ejes en negrita para mayor claridad.

5. Secciones sugeridas para el notebook

Secciones sugeridas para el notebook: - Introducción - Carga y preparación de datos – Estadísticas descriptivas - Distribuciones univariadas - Análisis bivariado - Correlaciones y heatmaps - Hallazgos y conclusiones