Práctica 2 — Análisis del dataset Adult Income (Python)

Integrantes
- NOMBRE APELLIDO (Integrante 1)
- NOMBRE APELLIDO (Integrante 2)

Repositorio: PENDIENTE

Vídeo: PENDIENTE

Fecha de generación: **2025-12-19**

1. Descripción del dataset

El dataset integrado contiene **48,842** registros. La variable objetivo es `income`.
Distribución de clases: `<=50K` = **0** (76.07%), `>50K` = **0** (23.93%).
Se observa un desbalance aproximado de 3:1, por lo que se interpretan métricas por clase además de la accuracy.

2. Integración y selección de los datos

Se integran los conjuntos train y test del Adult Income y se mantienen variables estándar del dominio (edad, educación, horas, ganancias/pérdidas de capital y categóricas de contexto).

3. Limpieza de los datos

3.1 Faltantes y/o valores perdidos

Principales faltantes antes de limpieza (top 5, por columna):

index	missing_count	missing_pct
age	0	0
workclass	0	0
fnlwgt	0	0
education	0	0
education_num	0	0

Las variables categóricas imputan faltantes a la categoría `Unknown`. Las variables numéricas imputan con mediana.

3.2 Tipos de variables y transformaciones

Se normalizan categóricas (strip) y se tipifican numéricas con coerción segura (valores inválidos pasan a NA y se imputan).

3.3 Tratamiento de valores extremos

Para `capital_gain` y `capital_loss` se aplica winsorización al percentil 99.5% para limitar el impacto de colas extremas.

col	cap	n_capped
capital_gain	41310	244
capital_loss	2258	237

3.4 Consideraciones adicionales

Se preserva el tamaño muestral evitando eliminar filas con faltantes, lo que reduce riesgo de sesgo por eliminación.

4. Análisis y métricas

4.1 Supervisado y no supervisado

Modelo supervisado (Regresión logística): ROC-AUC = **0.9048**, Accuracy = **0.8529**.
Para la clase `>50K` (positiva): Precision = **0.736**, Recall = **0.601**, F1 = **0.662**.
Interpretación: el AUC alto indica buena capacidad discriminativa, pero el recall moderado sugiere dificultad para capturar todos los casos `>50K`, consistente con el desbalance.

No supervisado (PCA+KMeans): muestra n = **800**, k = **2**, silhouette = **0.4118**.
Interpretación: el clustering es exploratorio y depende del muestreo; no se extraen conclusiones predictivas fuertes.

4.2 Contraste de hipótesis

Contraste entre grupos de `income` sobre `hours_per_week` usando **Mann–Whitney U**.
Medias: <=50K = **38.84**, >50K = **45.45**. Medianas: <=50K = **40.00**, >50K = **40.00**.
p-value = **0**.
Interpretación: una p-value muy pequeña indica diferencias estadísticamente significativas; esto refleja asociación, no causalidad.

5. Representación de resultados

Figuras generadas en `reports/figures/`: `roc_curve.png` y `confusion_matrix.png`. Tablas en `reports/tables/`.

6. Conclusiones

El dataset permite construir un clasificador con buen desempeño (AUC alto), aunque la recuperación de la clase `>50K` es moderada por el desbalance. El contraste sugiere diferencias consistentes en horas trabajadas entre grupos. El análisis no supervisado es exploratorio y no muestra separaciones nítidas sin supervisión.

7. Código

El código fuente se encuentra en `src/`. Para ejecutar el pipeline: `python -m src.run_all`.

8. Vídeo

Enlace al vídeo (Google Drive UOC): PENDIENTE

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	AA, BB
Redacción de las respuestas	AA, BB
Desarrollo del código	AA, BB
Participación en el vídeo	AA, BB