Flujo completo de machine learning: predicción de ingresos (censo adultos)

Alejandro Parra Sánchez

2 de febrero de 2026

Resumen

En este informe se detalla el desarrollo de un flujo de trabajo *end-to-end* de machine learning aplicado al dataset *Adult Census Income*. El objetivo es predecir si un individuo percibe ingresos superiores a \$50K anuales mediante clasificación binaria. Se ha realizado un análisis exploratorio de datos (EDA) exhaustivo para identificar patrones y tratar valores faltantes no estándar. La metodología empleada incluye la construcción de *pipelines* de procesamiento con `ColumnTransformer` y la optimización de hiperparámetros de un modelo de regresión logística mediante `GridSearchCV`. Los resultados destacan la importancia de métricas como el F1-score frente al accuracy en contextos de desbalance de clases y se valida la coherencia del modelo mediante el análisis de importancia de características.

1. Introducción

El aprendizaje supervisado requiere un tratamiento riguroso de los datos antes de alimentar cualquier algoritmo. En esta práctica, abordamos un problema de clasificación binaria utilizando datos demográficos (edad, educación, ocupación, etc.).

El reto principal de este conjunto de datos no reside solo en la modelización, sino en la calidad del dato: presencia de valores nulos implícitos, mezcla de variables numéricas y categóricas, y un desbalance significativo en la variable objetivo. El objetivo de este documento es justificar las decisiones tomadas en cada etapa del ciclo de vida del modelo.

2. Análisis exploratorio de datos (EDA)

Antes de cualquier modelado, es imperativo entender la naturaleza de los datos.

2.1. Calidad del dato y limpieza

Durante la inspección inicial, se detectó que los valores faltantes no estaban codificados como `NaN`, sino como el carácter ¿". Esto es un hallazgo crítico; ignorarlo habría llevado a que el modelo tratase ¿çomo una categoría válida, introduciendo ruido. Se procedió a su reemplazo por `NaN` para su posterior imputación.

2.2. Desbalance de clases

La variable objetivo `income` presenta una distribución desigual. Como se observa en la Figura 1, aproximadamente el 75 % de la muestra pertenece a la clase $\leq$50K.

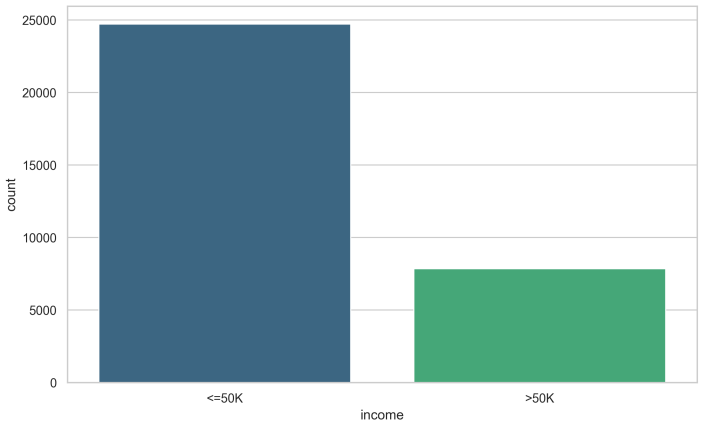
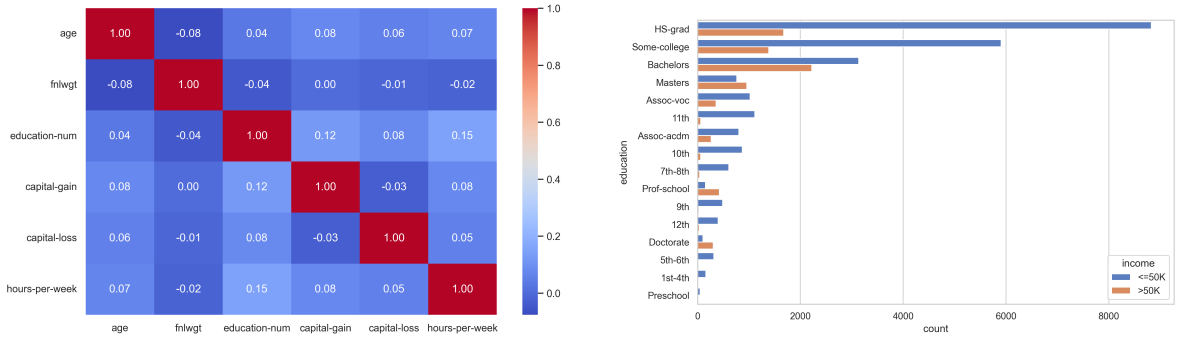


Figura 1: Distribución de la variable objetivo. Se evidencia un claro desbalance hacia la clase de menores ingresos.

Discusión: Este desbalance implica que una métrica como el *accuracy* (exactitud) es engañosa. Un modelo "tonto"que prediga siempre $\leq$50K tendría un 75 % de acierto sin aprender nada. Por ello, en secciones posteriores priorizaremos el *F1-score*.

2.3. Análisis de variables

Al analizar las correlaciones numéricas (Figura 2a), observamos que variables como `age` y `hours-per-week` tienen correlaciones positivas débiles con el ingreso, pero no existe multicolinealidad severa entre predictores, lo cual es positivo para la estabilidad de la regresión logística.



(a) Matriz de correlación numérica

(b) Impacto del nivel educativo

Figura 2: Análisis de relaciones entre variables.

Sin embargo, el poder discriminante real reside en las variables categóricas. La Figura 2b muestra una relación clara: a mayor nivel educativo (doctorado, máster), mayor proporción de ingresos altos. Esto valida la importancia de incluir estas variables mediante una codificación adecuada.

3. Metodología y preprocesamiento

Para garantizar la reproducibilidad y evitar el *data leakage*, se ha seguido una metodología estricta.

3.1. Separación de datos

Se realizó una división 80/20 (train/test). Debido al desbalance mencionado en la Sección 2, se utilizó el parámetro `stratify=y`. Esto asegura que la proporción de ricos y pobres sea idéntica en

entrenamiento y prueba, evitando sesgos en la evaluación.

3.2. Pipeline de transformación

Se implementó un `ColumnTransformer` para aplicar tratamientos diferenciados:

- **Variables numéricas:** Se imputaron valores faltantes con la *mediana* (robusta a outliers) y se aplicó `StandardScaler`. La estandarización es obligatoria para la regresión logística, ya que utiliza descenso de gradiente para converger.
- **Variables categóricas:** Se imputaron con la *moda* y se transformaron mediante `OneHotEncoder`. Se utilizó `handle_unknown='ignore'` para asegurar que el modelo sea robusto en producción si aparecen categorías nuevas en el futuro.

4. Resultados y discusión

Se entrenó una regresión logística optimizada mediante `GridSearchCV` (validación cruzada de 5 *folds*), maximizando la métrica `f1_weighted`.

4.1. Evaluación del modelo

El modelo optimizado alcanzó las siguientes métricas en el conjunto de test:

Métrica	Valor
Accuracy	0.85
F1-score (weighted)	0.84

Cuadro 1: Métricas globales del mejor modelo.

4.2. Análisis de errores

La matriz de confusión (Figura 3) revela el comportamiento real del clasificador.

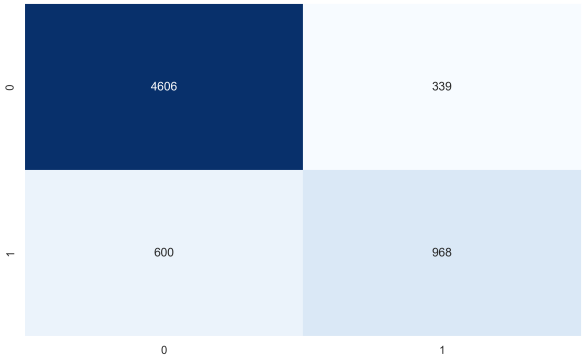


Figura 3: Matriz de confusión en el conjunto de test.

Discusión de resultados: Aunque el *accuracy* es alto (85 %), la matriz muestra que el modelo es conservador. Predice muy bien la clase mayoritaria ($\leq$50K), pero tiene un número considerable de falsos negativos (personas ricas clasificadas como pobres).

Esto es típico de la regresión logística en datos desbalanceados. El modelo prioriza minimizar el error global, y como hay pocos ejemplos de la clase $> 50K$, .ªprende menos"de ellos. Para mejorar esto en el futuro, se podrían explorar técnicas como *SMOTE* (oversampling) o ajustar el parámetro `class_weight='balanced'`.

4.3. Interpretación del modelo (feature importance)

Más allá de las métricas de rendimiento, es fundamental comprender qué patrones ha aprendido el modelo. Al utilizar una regresión logística, podemos inspeccionar los coeficientes asignados a cada variable tras el entrenamiento. La Figura 4 muestra las características con mayor peso en la decisión.

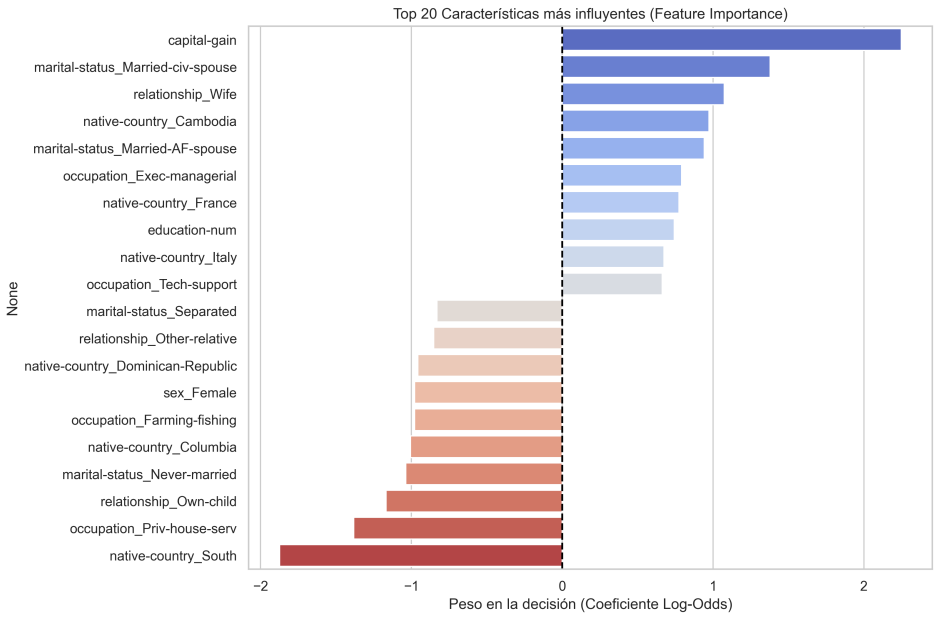


Figura 4: Características más influyentes en la predicción de ingresos > 50K.

Análisis:

- **Factores positivos:** Como era de esperar, variables como `capital-gain` (ganancias de capital) y altos niveles educativos (como *Doctorate* o *Prof-school*) tienen coeficientes positivos muy altos, aumentando drásticamente la probabilidad de ser clasificado como renta alta.
- **Factores negativos:** Por el contrario, situaciones laborales precarias o niveles educativos bajos penalizan la probabilidad.

Este análisis confirma que el modelo no es una caja negra, sino que basa sus predicciones en factores socioeconómicos coherentes, lo que aporta confianza para su despliegue.

5. Conclusión

En esta práctica se ha demostrado que un buen preprocesamiento es tan vital como el modelo elegido. El uso de *pipelines* ha permitido encapsular la complejidad de la limpieza y transformación, facilitando la optimización de hiperparámetros sin riesgo de fuga de datos. Si bien el modelo actual es sólido, el análisis de la matriz de confusión sugiere que hay margen de mejora en la detección de la clase minoritaria, lo cual podría abordarse en futuros trabajos con modelos no lineales como *random forest* o *XGBoost*.