Reporte HE2 AI Parcial 2

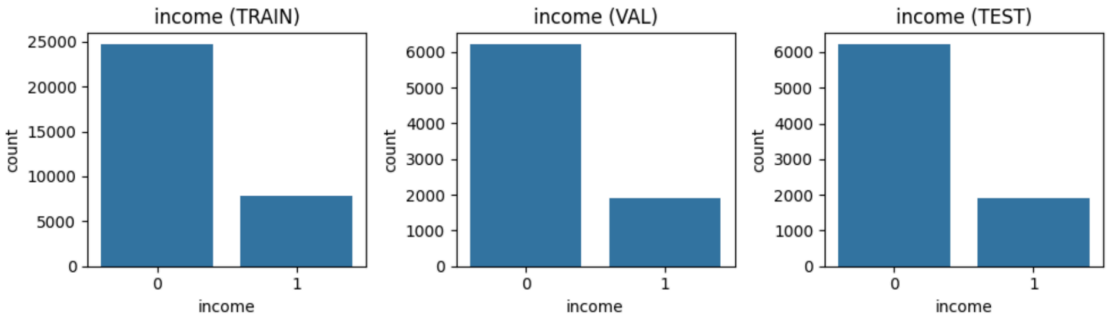
Por: Juan Felipe Benites y Sofia Angulo
Septiembre 2025
HE 2 IA en la Economía

1) Introducción y resumen ejecutivo

Este proyecto aborda la predicción de ingresos superiores a $50K utilizando el dataset Adult Census (UCI, 1994) e implementación en PyTorch. El flujo completo incluye: EDA sin fuga de información, preprocesamiento , un baseline de Regresión Logística y redes neuronales MLP en dos variantes: sin regularización y con regularización
Se definieron tres particiones: TRAIN (32,561), VALIDACIÓN (8,140) y PRUEBA (8,141), manteniendo una proporción de clases estable de ≈76% (≤$50K) y ≈24% (>$50K) en los tres subconjuntos. Despues de one-hot, cada muestra queda con 100 características.

Resultados:

- **Regresión Logística** (umbral optimizado t*=0.53): desempeño sólido y estable con AUC=0.9106, Accuracy=0.8216 y F1=0.6852 en TEST. El ajuste del umbral mejora el equilibrio precisión/recuperación sin afectar el AUC (invariante al umbral).
- **MLP sin regularización:** muestra sobreajuste (pérdida de validación creciente y oscilación de accuracy ~0.79–0.82) pese a una accuracy de entrenamiento muy alta (>0.90).
- **MLP con regularización:** la validación se estabiliza (pico ValAcc≈0.809, ValLoss≈0.576–0.582) y queda cerca del baseline; en los artefactos revisados no hay evidencia de superarlo en TEST.



2) Decisiones de procesamiento de datos

El objetivo fue garantizar comparabilidad entre modelos y evitar leakage. Se utilizaron las 14 columnas originales (numéricas y categóricas) sin descartar categorías raras para preservar señal.

- Imputación y escalado.
    - Numéricas → mediana; categóricas → moda.
    - Numéricas → StandardScaler (media 0, var 1).
- Codificación categórica.

    - One-Hot Encoding con drop='first' para reducir colinealidad y handle_unknown='ignore' para categorías no vistas.

- Encaje exclusivo en TRAIN.
    - Todas las estadísticas (medianas, modas, escalador, esquema OHE) se ajustaron solo con TRAIN; VALIDACIÓN/PRUEBA se transforman con el mismo objeto.

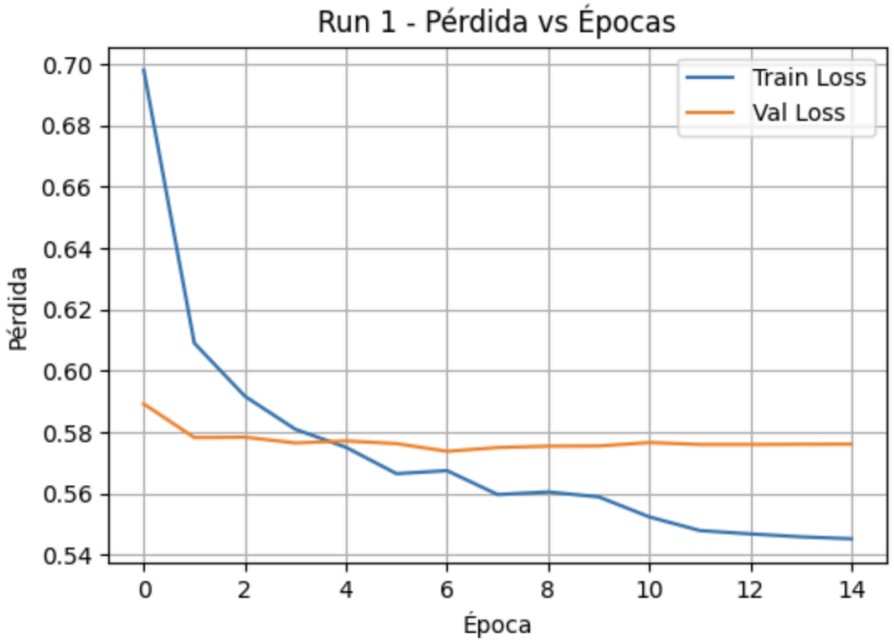
- Dimensionalidad y carga.
    - 100 features post-OHE; tensores float32; DataLoader con batch=128 y shuffle=True solo en TRAIN.

3) MLP: mejores hiperparámetros y comparación sin vs. con regularización

3.1 Mejor configuración de MLP con regularización
El mejor experimento observado corresponde a una red mediana que balancea capacidad y control de varianza

Bloque	Configuración / Valor
Arquitectura	Capas (128, 64) con ReLU (He init)
Normalización	BatchNorm en capas densas
Dropout	0.2
Optimización	AdamW (lr=1e-3, weight_decay=1e-4)
Criterio	BCEWithLogitsLoss
Regularización adicional	Gradient clipping (activo)
EarlyStopping	Paciencia=6; restaura el mejor checkpoint
Batch size	128
Desempeño Validación (mejor punto)	ValLoss≈0.576 *(rango 0.576–0.582)*; ValAcc pico≈0.809

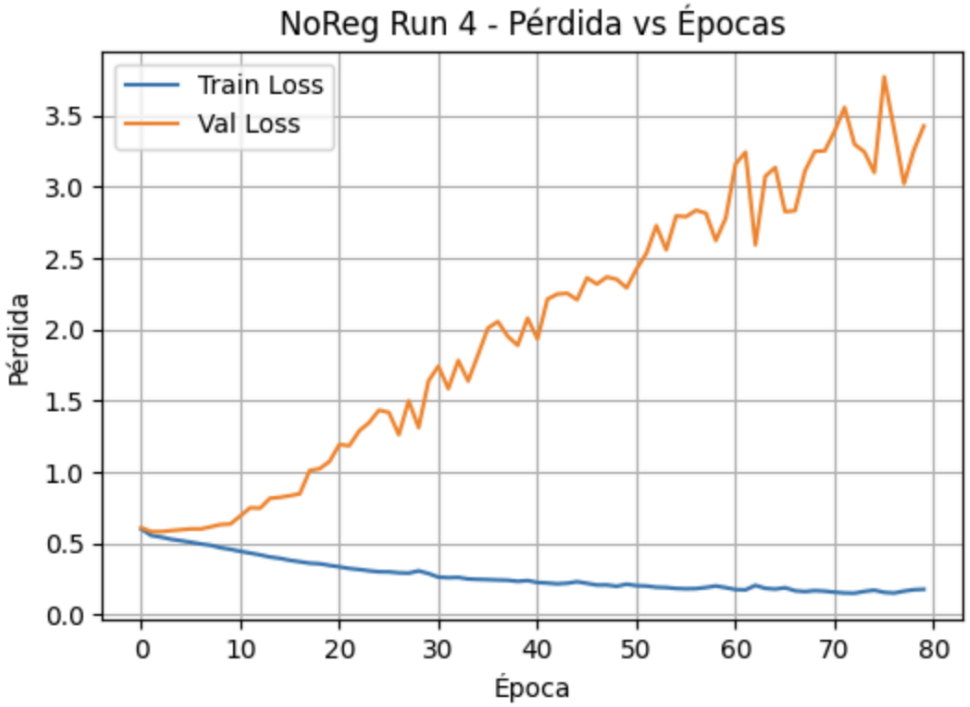
Resultado**:** ValLoss≈0.576 (rango 0.576–0.582) y ValAcc pico≈0.809.
Las curvas por época muestran que, tras unas iteraciones, la pérdida de validación se estabiliza y el EarlyStopping detiene el entrenamiento cerca de la época ~12, evitando que el modelo sobreentrene.



3.2 MLP sin regularización: diagnóstico de sobreajuste

Los experimentos sin regularización exhiben un patrón consistente de overfitting:

- Accuracy (Train) supera 0.93.
- Accuracy (Validación) se mantiene en ~0.80–0.82.
- ValLoss aumenta con las épocas, aunque la pérdida de entrenamiento desciende.
- Alta varianza entre corridas.




```
[NoReg Run 4] Diagnóstico: overfitting
```

3.3 Comparación sin vs. con regularización

Configuración	Diagnóstico	Accuracy (Train)	Accuracy (Validación)	ValLoss (tendencia)	Observaciones
MLP sin regularización	Overfitting severo	>0.93	0.80–0.82	Crece con las épocas	Generalización deficiente; alta varianza entre corridas
MLP con regularización	Ajuste razonable	≈0.81	0.80–0.82 *(pico ≈0.809)*	Estable (≈0.576–0.582)	EarlyStopping detiene en el valle; métricas consistentes

MLP con regularización queda muy cercano a la logística (≈0.809 vs ≈0.812 en validación), confirmando que la regularización corrige el sobreajuste pero aún no supera el baseline.

4) Comparación entre el mejor MLP y la Regresión Logística

4.1 Resultados del baseline: Regresión Logística (t*=0.53)

- TRAIN: Loss=0.5789, Acc=0.8190, Prec=0.5886, Rec=0.8249, F1=0.6870, AUC=0.9070.
- VALIDACIÓN: Loss=0.5987, Acc=0.8117, Prec=0.5711, Rec=0.8144, F1=0.6714, AUC=0.8984.
- PRUEBA: Loss=0.5654, Acc=0.8216, Prec=0.5876, Rec=0.8216, F1=0.6852, AUC=0.9106.

Accuracy, F1 y AUC en el conjunto de prueba: el modelo logra AUC≈0.91, lo que indica alto poder de discriminación; el F1≈0.685 confirma un buen equilibrio entre precisión y recuperación con el umbral seleccionado.

4.2 Resultados del MLP con regularización - mejor experimento

- ValLoss (mejor): ≈0.576 (rango observado 0.576–0.582).
- ValAcc (pico): ≈0.809.

Las curvas por época evidencian estabilidad y parada temprana alrededor de la época ~12, lo que reduce el riesgo de sobreajuste.

Interpretación de resultado final

- El baseline lineal (con buen preprocesamiento y umbral optimizado) alcanza AUC≈0.91 y F1≈0.685 en PRUEBA, por lo que establece un piso alto de desempeño.
- El MLP con regularización está bien calibrado en validación (ValLoss≈0.576, ValAcc≈0.809) y corrige el sobreajuste observado en las corridas sin regularización; no obstante, con la evidencia disponible no supera el baseline en PRUEBA.
- En datos tabulares, los modelos lineales fuertes son competitivos; para que el MLP supere consistentemente, suele requerirse tuning adicional, ajuste de umbral por F1 y, de ser necesario, calibración de probabilidades.