Santiago Andres Cobos 202122212
Daniela Herrera 202113704
Juan Sebastian Sierra 202123725

Reporte

- **Decisiones de procesamiento de datos:**

Marco General

Lo primero que hacemos es entender tenemos dos bases de datos “adult.data” (lo vamos a usar como train) y “adult.test (lo vamos a usar para validación en la prueba final). Para garantizar que ambos tuvieran los mismos nombres de columnas, primero extrajimos la lista oficial desde adult.names y la aplicamos a ambos; además en el test cambiamos el nombre de '>50K.' por '>50K' para que el train y test tengan las mismas columnas y el mismo formato de la etiqueta antes de cualquier transformación. Con los dataframes ya uniformes, tomamos el test original y lo dividimos 50/50 con estratificación por ingreso (dividir los datos manteniendo la misma proporción de las clases) para construir la validación y la prueba final: así obtuvimos alrededor de 8140 filas en cada uno y conservamos proporción de la clase con objetivo (en train quedó <= alredor del 75.9% y >50k alrededor de 24.1% y en val/test se mantuvo prácticamente igual). La estratificación (división manteniendo las proporciones de las clases en train, validación y en test) la hacemos con el objetivo de tener una medida de comparación justa entre experimentos, es decir que las medidas de comparación no se vean afectadas por las proporciones.

Valores Faltantes:

Después del Split estratificado, en cada partición (train, validación y test), encontramos que existen valores nulos que se concentraban en 3 categorías (workclass, occupation y native-country). En lugar de borrar filas (habría reducido la muestra y sesgado el dataset), remplazamos cada”?”/NaN por la categoría “UnKnown”. Esto mantiene toda la información disponible y deja que el modelo aprenda explícitamente que un valor desconocido es una posibilidad real dentro de esas variables. Además, tratar los faltantes con una categoría fija tras el split tiene varias ventajas prácticas: (i) evita data leakage, porque el modelo no aprende de información de validación o de test (como medias o modas) que puede tender a generar métricas demasiado optimistas; (ii) preserva el tamaño muestral y, por tanto, el poder estadístico del modelo, al no descartar filas.

Variable numérica:

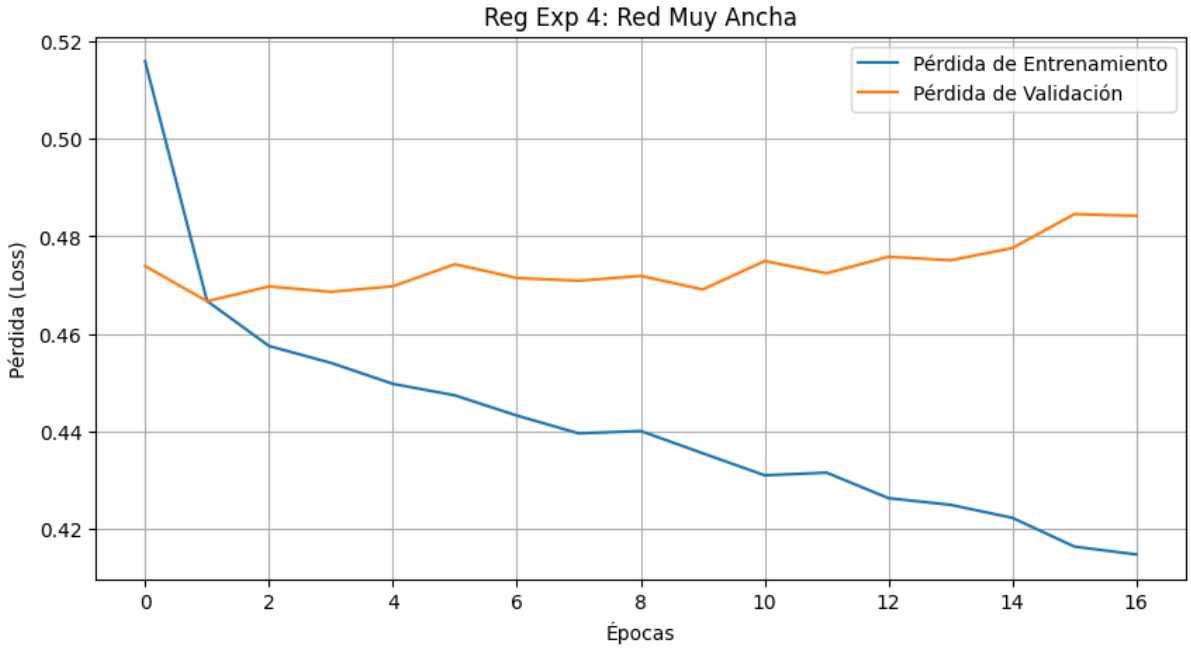
Para las variables numéricas dentro de la base de datos calculamos la media y las desviaciones estándar de cada variable y estandarizamos los valores cada variables . Puesto que, que queremos todas las variables numéricas en la misma escala (media 0 y varianza 1). Eso ayuda a que la red neuronal entrene con gradientes más estables y que tanto la regresión logística como el MLP no se sesguen hacia variables grandes solo por tener valores enormes.

Santiago Andres Cobos 202122212
Daniela Herrera 202113704
Juan Sebastian Sierra 202123725

Resultados y análisis del MLP:
Primero entrené cinco MLP sin regularización y luego repetí las mismas ideas con regularización. En las redes sin regularización el patrón fue muy claro en todas las gráficas: la pérdida de entrenamiento cae de forma continua, pero la pérdida de validación baja al inicio y luego empieza a subir; eso es overfitting.

- **Hiperparámetros del mejor experimento de MLP**

El modelo con el mejor desempeño, al obtener la menor pérdida de validación final (0.4667), fue el "Reg Exp 4: Red Muy Ancha".

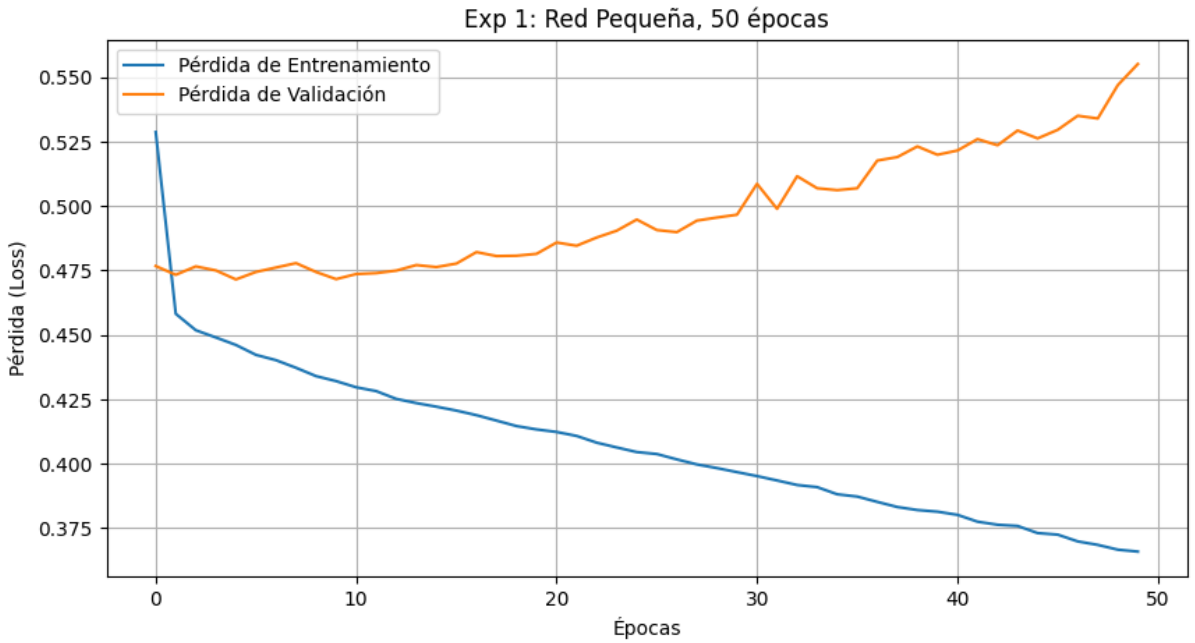


- **Arquitectura**: Una red con 3 capas ocultas de `[512, 256, 128]` neuronas.
- **Función de Activación**: `ReLU`.
- **Optimizador**: `Adam`.
- **Learning rate:** `7e-4` (0.0007).
- **Función de Pérdida**: `BCEWithLogitsLoss` con `pos_weight`.
- **Técnicas de Regularización**:
  - **Dropout**: Tasas de `[0.5, 0.4, 0.3]` después de cada capa oculta.
  - **Early Stopping**: Con una paciencia de 15 épocas.
  - **Regularización L2 (Weight Decay)**: `1e-5`

Comparar los mejores MLP sin regularización y con regularización

Santiago Andres Cobos 202122212
Daniela Herrera 202113704
Juan Sebastian Sierra 202123725

Gráfica del mejor MLP sin regularización:



— MÉTRICAS FINALES DEL MEJOR MODELO SIN REGULARIZACIÓN —

split	loss	accuracy	precision	recall	f1	roc_auc
Prueba (Test)	0.5211	0.829	0.610	0.771	0.681	0.899

Valores mostrados con 3 decimales (salvo *loss* con 4).

Generalización y Sobreajuste (Loss y ROC AUC): La pérdida (Loss) en el conjunto de prueba es significativamente menor en el modelo regularizado. Esto, junto con un ROC AUC más alto, es la evidencia más clara de que la regularización combatió con éxito el sobreajuste. El modelo sin regularización aprendió muy bien los datos de entrenamiento (como vimos en sus gráficas), pero no pudo generalizar ese conocimiento a datos nuevos. El modelo regularizado, en cambio, aprendió patrones más robustos y universales, lo que resultó en un rendimiento mucho mejor en el mundo real (el conjunto de prueba).

Mejor Rendimiento General (Accuracy y F1-Score): El modelo regularizado tiene una mayor exactitud y un F1-Score superior, lo que indica un mejor equilibrio entre Precisión y Recall.

Mejor en la Tarea Clave (Recall): Incluso en el Recall, donde el modelo sin regularización ya era bueno (0.771), el modelo con regularización lo supera (0.803). Esto significa que es aún más eficaz para encontrar a la mayor cantidad posible de personas con ingresos altos.

Santiago Andres Cobos 202122212
Daniela Herrera 202113704
Juan Sebastian Sierra 202123725

El modelo con regularización es superior al modelo sin regularización en todas las métricas evaluadas.

- **Compare el mejor MLP y la regresión lineal a partir de sus métricas. Interprete los resultados.**

Métricas regresión lineal:

split	accuracy	precision	recall	f1	roc_auc
train	0.854	0.738	0.605	0.665	0.909
val	0.848	0.717	0.587	0.646	0.899
test	0.854	0.743	0.612	0.671	0.911

Métricas mejor MPL:

split	loss	accuracy	precision	recall	f1	roc_auc
Entrenamiento (Train)	0.4422	0.845	0.645	0.795	0.713	0.920
Validación (Validation)	0.4689	0.832	0.617	0.761	0.682	0.906
Prueba (Test)	0.4443	0.847	0.645	0.785	0.708	0.917

La Regresión Logística tiene una exactitud ligeramente mayor (85.8% vs. 84.4%). Sin embargo, en un conjunto de datos desbalanceado como este (donde hay muchos más individuos con ingresos <=50K que >50K), la exactitud puede ser una métrica engañosa. Un modelo podría obtener una alta exactitud simplemente prediciendo la clase mayoritaria la mayoría de las veces.

Regresión Logística (Ganador en Precisión): Tiene una precisión de 0.743. Esto significa que cuando este modelo predice que una persona gana >50K, tiene razón el 74.3% de las veces. Es un modelo más "cauteloso" y confiable en sus predicciones positivas, minimizando los falsos positivos.

MLP (Ganador Absoluto en Recall): Tiene un recall de 0.785, muy superior al 0.612 de la Regresión Logística. Esto significa que el MLP es capaz de encontrar al 80.3% de todas las personas que realmente ganan >50K en el conjunto de datos. La Regresión Logística, en cambio, solo encuentra al 61.2%. El MLP es un modelo mucho más "exhaustivo" y eficaz para no dejar pasar a los individuos de la clase de interés.

El F1-Score (la media armónica de precisión y recall) es mayor para el MLP (0.709 vs 0.671). Esto confirma que, a pesar de su menor precisión, el MLP logra un mejor equilibrio general entre ser correcto y ser exhaustivo.

Santiago Andres Cobos 202122212
Daniela Herrera 202113704
Juan Sebastian Sierra 202123725

El ROC AUC es también ligeramente superior en el MLP (0.917 vs 0.911). Esta métrica indica la capacidad del modelo para distinguir entre las dos clases. Un valor más alto significa que el MLP es, en general, un mejor clasificador a la hora de separar a los que ganan más de 50K de los que no.