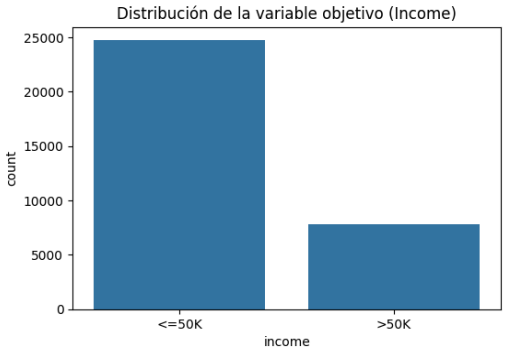
**Reporte parcial 2**

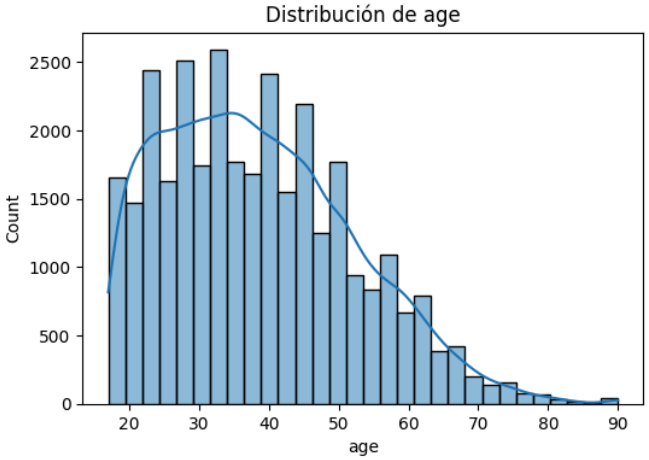
**Julian Fandiño (202021070)**

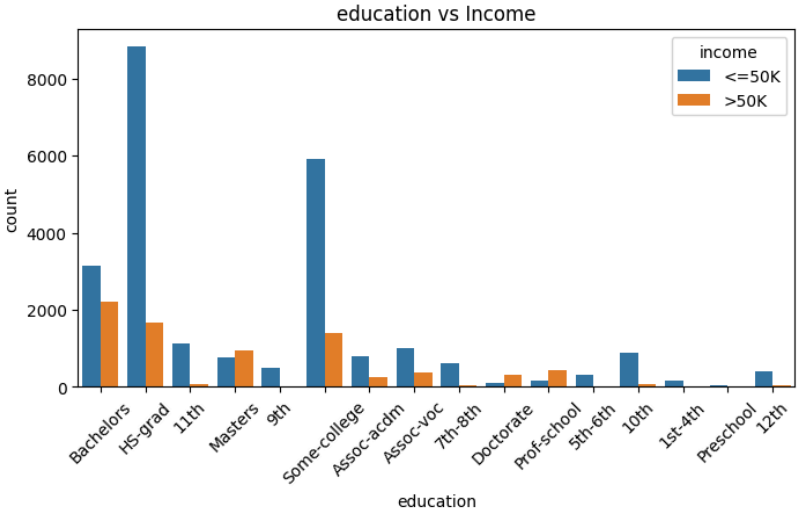
**Sofia Villamizar (202123964)**

1. **Decisiones procesamiento de datos**

   Para realizar el procesamiento de los datos, en primer lugar se usaron todas las variables originales disponibles en el data.set. Luego, se diferenciaron las variables numéricas de las categóricas para poder hacer las transformaciones correspondientes. Ahora, en cuanto a las transformaciones realizadas, para el caso de las variables numéricas se estandarizaron con media cero y desviación estándar uno, para poner todas las variables en la misma escala y evitar que variables con valores grandes dominaran sobre variables con valores pequeños en el entrenamiento del modelo. Para el caso de las categorías, se aplicó la transformación para convertir cada categoría en una columna binaria y se pusieron ceros si aparecían categorías nuevas en validación/test que no estaban en entrenamiento. Al final, cada categoría se presentó como una columna binaria. Además, se hizo un procesamiento de la variable objetivo, también convirtiéndola en una variable binaria donde se asignaba 1 si la persona gana más de 50.000 y 0 si gana menos o igual a 50.000. para evitar el DataLeakage se entrenó el preprocesador solo con los datos de entrenamiento y luego con lo que ya había aprendido se a aplicar el preprocesador a los tres conjuntos, de train, validación y test. A continuación se muestran las gráficas correspondientes al exploratory data analysis que ayudaron a guiar el procesamiento de los datos :



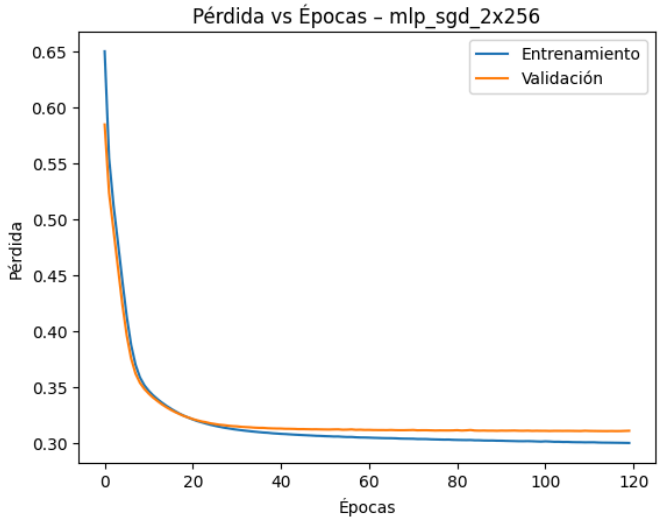



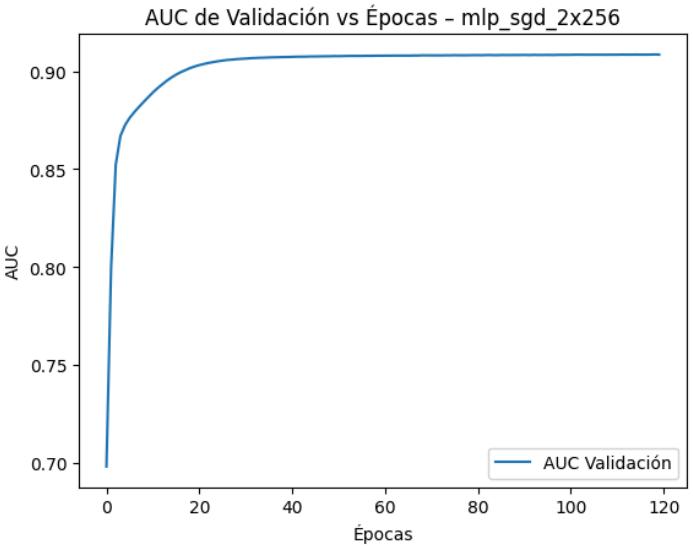



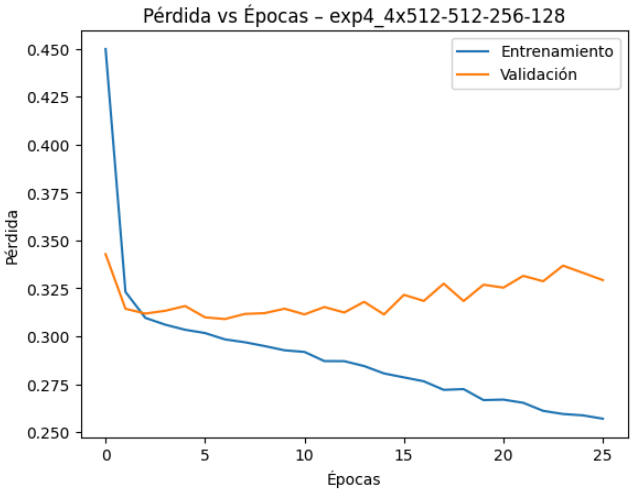

2. **Hiperparámetros del mejor experimento de MLP :**
    - Sin regularización : 2 capas ocultas y 256 neuronas en ambas capas con SGD. Con learning rate de 0.005 (SGD). Con un AUC de 0.91
    - Con regularización : 4 capas ocultas y 512 neuronas en la primera capa, 512 neuronas en la segunda capa , 256 en la tercera capa y en la cuarta 128 neuronas, con un dropout de 0.3. Con AUC de 0.91

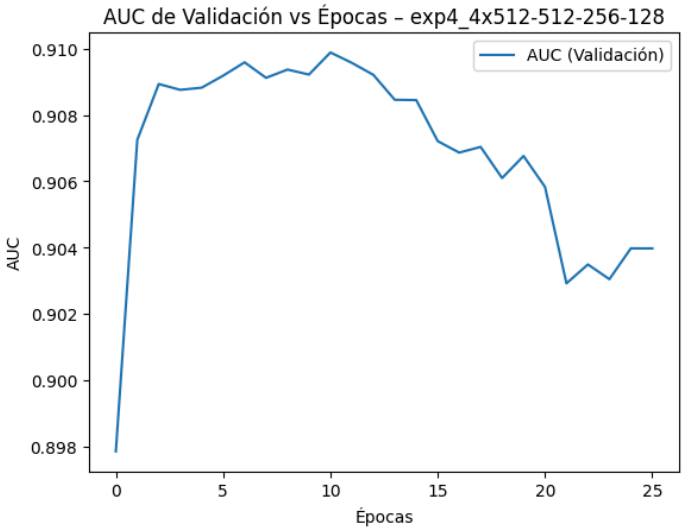
**Comparación de mejores MLP sin regularización y con regularización :**

En MLP sin regularización,el mejor modelo fue el de 2 capas ocultas con 256 neuronas en cada capa y learning rate de 0.005, alcanzando un AUC de 0.91 en validación. Por su parte, en los modelos más grandes, después de cierto número de épocas el error en validación empezó a crecer, lo que indicaba que el modelo no estaba generalizando bien a los datos nuevos sino memorizando del conjunto de entrenamiento (overfitting) y también su AUC comenzaba a bajar. Así que los mejores MLP sin regularización fueron aquellos con modelos simples, es decir con pocas capas ocultas y pocas neuronas en cada una de estas. Por su parte, en MLP con regularización, el mejor modelo fue uno más complejo que obtuvo un AUC de aproximadamente 0.91 también. Sin embargo, a pesar de que los AUC son similares, la regularización permitió que los modelos más grandes mantuvieran un desempeño estable, evitando el aumento del error de validación y reduciendo el overfitting gracias al dropout. En este caso, los mejores MLP con regularización fueron aquellos con arquitecturas más complejas, es decir con más capas. En cuanto al análisis de sus métricas, el MLP sin regularización obtuvo la mejor precisión y AUC en el conjunto de test, mientras que el MLP con regularización tuvo un recall superior, mostrando mayor capacidad de identificar positivos sin overfitting. Además, se destaca que ambos tienen un valor de AUC aproximado similar, lo que muestra que los dos son buenos clasificadores globales, superiores a un clasificador aleatorio.




AUC de Validación vs Épocas – mlp_sgd_2x256
AUC
Épocas
AUC Validación
0.90
0.85
0.80
0.75
0.70
0
20
40
60
80
100
120


Pérdida vs Épocas – exp4_4x512-512-256-128
Pérdida
Épocas
Entrenamiento
Validación
0.450
0.425
0.400
0.375
0.350
0.325
0.300
0.275
0.250
0
5
10
15
20
25


AUC de Validación vs Épocas – exp4_4x512-512-256-128
AUC
Épocas
AUC (Validación)
0.910
0.908
0.906
0.904
0.902
0.900
0.898
0
5
10
15
20
25

3. **Compare el mejor MLP y la regresión lineal a partir de sus métricas e interprete los resultados**

En cuanto al accuracy, la regresión obtuvo un valor de 0.86 mientras que el mejor MLP un valor de 0.84, esto muestra que en cuanto al valor de predicciones correctas sobre el total de predicciones, la regresión tuvo un mejor desempeño. Para la precisión de ambos modelos, la regresión obtuvo una precisión de 0.75 y la mejor MLP una precisión de 0.64, esto muestra que la regresión tiene una menor incidencia en predecir falsos positivos. Para el recall, la regresión tuvo un valor de 0.61 mientras que el mejor MLP un valor de 0.76, en este caso se muestra que MLP es mejor para detectar más positivos reales y evitar menos falsos negativos. El f1-score de la regresión fue 0.67 y el del mejor MLP fue 0.69, esto muestra que el MLP tiene un mejor balance entre precisión y recall, mientras que la regresión , aunque fue más precisa, perdió sensibilidad (recall), lo que redujo su F1-score. Por último para las métricas de AUC y ROC, la regresion tuvo un valor de 0.90 mientras que el mejor MLP, tuvo un valor de 0.91, esto muestra que el MLP tiene una ventaja en la capacidad de discriminacion entre clases, sin embargo ambos son buenos ya que están bastante cerca de 1. En términos generales, el MLP puede mostrar un ligera mejora respecto a la regresión, pues logra un mejor AUC, F1-score y recall , la regresión sigue siendo un modelo competitivo con mayor precisión y un accuracy similar al de la mejor MLP.