Universidad de los Andes
Facultad de Economía
HE2 : Inteligencia Artificial Aplicada
Andres Felipe Rosas

Predicción de ingresos individuales mediante modelos de clasificación supervisada: aplicación al conjunto de datos Adult Census

Contenido

Introducción

En este trabajo se aborda la construcción y evaluación de modelos de clasificación binaria a partir del conjunto de datos Adult Census Income, utilizado comúnmente como referencia en problemas de aprendizaje supervisado. El objetivo fundamental consiste en predecir si un individuo tiene ingresos superiores a 50.000 dólares anuales a partir de un conjunto de variables demográficas y socioeconómicas tales como edad, nivel educativo, ocupación, número de horas trabajadas por semana, estado civil, raza, género y país de origen.

La motivación de este ejercicio radica en la necesidad de comprender los factores asociados a la distribución de ingresos y en la importancia de desarrollar modelos que

permitan anticipar patrones de desigualdad económica. Desde la perspectiva del aprendizaje automático, este problema constituye un caso de estudio interesante por el carácter mixto de las variables disponibles, que combinan información numérica y categórica, y por la presencia de un cierto desbalance en las clases que dificulta la tarea de predicción.

En este contexto, el informe se centra en dos etapas principales. En primer lugar, se realiza el preprocesamiento de los datos con el fin de garantizar que los algoritmos de aprendizaje trabajen únicamente con representaciones numéricas estandarizadas, evitando así problemas de fuga de información entre los conjuntos de entrenamiento, validación y prueba. En segundo lugar, se implementan y comparan distintos modelos de referencia, desde un clasificador lineal basado en regresión logística hasta redes neuronales multicapa con regularización, para evaluar su capacidad predictiva bajo métricas propias de la clasificación binaria como exactitud, precisión, exhaustividad, F1 y área bajo la curva ROC.

Metodología

2.1 División de los datos

El conjunto de datos Adult Census se conforma por 32.561 observaciones en el subconjunto de entrenamiento y 16.280 observaciones en el subconjunto de prueba original. Para la correcta evaluación de los modelos se dividió el conjunto de prueba en dos partes equivalentes, generando un subconjunto de validación con 8.140 observaciones y un subconjunto de prueba final con 8.140 observaciones. De esta forma, el procedimiento experimental se llevó a cabo sobre tres particiones: entrenamiento, validación y prueba. La elección de esta estrategia asegura la disponibilidad de un conjunto independiente para el ajuste de hiperparámetros (validación) y otro exclusivo para la evaluación final del modelo (prueba), lo cual previene sesgos optimistas en la estimación del desempeño.

2.2 Análisis exploratorio de los datos (EDA)

El análisis exploratorio permitió identificar patrones de distribución y posibles irregularidades en las variables. La edad presenta una distribución concentrada entre los 20 y 50 años, con un decrecimiento progresivo hacia edades mayores. Las horas trabajadas por semana muestran un comportamiento atípico con un pico muy marcado en 40 horas, lo que refleja la jornada laboral estándar. En cuanto a las variables de capital, tanto la ganancia como la pérdida presentan distribuciones altamente sesgadas hacia cero, con una mayoría de individuos sin registros de ganancias o pérdidas de capital.

El nivel educativo, representado por la variable ordinal education-num, concentra su distribución en valores intermedios correspondientes a educación secundaria y universitaria inicial, lo cual se corresponde con los histogramas de la variable categórica education. Adicionalmente, los boxplots revelan diferencias en edad, horas trabajadas y nivel educativo según la clase de ingreso: los individuos con ingresos superiores a 50.000 dólares tienden a ser de mayor edad, trabajar más horas por semana y presentar un mayor nivel educativo. El análisis de las variables categóricas indica un predominio del sector privado en workclass y del nivel de secundaria completa en education.

2.3 Preprocesamiento y prevención de fuga de datos

El preprocesamiento consistió en varias etapas con el fin de transformar las variables originales en representaciones adecuadas para los algoritmos de aprendizaje supervisado. En primer lugar, se imputaron los valores faltantes de las variables categóricas workclass, occupation y native-country utilizando la moda, de modo que se mantuviera la consistencia de las categorías. Posteriormente, se aplicó codificación one-hot a todas las variables categóricas, incluyendo trabajo, estado civil, ocupación, relación familiar, raza, sexo y país de origen, con el fin de obtener una representación numérica binaria sin introducir un orden artificial entre categorías.

Las variables numéricas, como edad, horas trabajadas, ganancias y pérdidas de capital, fueron escaladas mediante estandarización a media cero y varianza unitaria, lo que permite mejorar la estabilidad numérica y la eficiencia en el entrenamiento de modelos. La variable objetivo income fue transformada en binaria, asignando el valor 0 a ingresos menores o iguales a 50.000 dólares y el valor 1 a ingresos superiores.

Para evitar la fuga de información entre conjuntos, el ajuste del preprocesador se realizó exclusivamente sobre el conjunto de entrenamiento, aplicándose posteriormente de manera consistente a validación y prueba. Esta decisión metodológica asegura que la información contenida en los conjuntos de validación y prueba no influya en el ajuste de parámetros ni en la estimación de la distribución de las variables.

Resultados

En esta sección se reportan y analizan los cinco experimentos realizados: una regresión logística como línea base y cuatro redes neuronales multicapa (MLP) con diferentes arquitecturas y niveles de regularización. Todas las conclusiones se apoyan en métricas de clasificación binaria calculadas en validación y prueba bajo dos umbrales de decisión: el estándar de 0,5 y el umbral óptimo determinado por maximización del F1 en validación. Para el mejor MLP con regularización se documentan además las métricas en

entrenamiento, validación y prueba, con y sin calibración de umbral. La información comparativa se sintetiza en la Tabla 1, y el detalle por partición del mejor MLP en la Tabla 2.

3.1 Metodología de validación y criterio de selección

La etapa de validación cumple dos funciones centrales. Primero, estimar de manera imparcial el desempeño de cada configuración para decidir cuál generaliza mejor, evitando la fuga de información al mantener la prueba como conjunto estrictamente final. Segundo, calibrar el umbral de decisión que optimiza el F1 en validación. Este procedimiento es particularmente relevante en un problema con desbalance de clases, pues el umbral por defecto 0,5 tiende a privilegiar la precisión a costa del recall o viceversa, mientras que la selección basada en F1 ofrece un compromiso operativo más equilibrado. El umbral óptimo determinado en validación se traslada después a la evaluación en prueba, sin reoptimizarlo, lo que permite medir su capacidad de generalización.

3.2 Regresión logística

La regresión logística con regularización L2 y ponderación de clases logró en validación un AUC de 0,9007. Con umbral 0,5 alcanzó una precisión de 0,559, un recall de 0,834 y un F1 de 0,6689. Al aplicar el umbral óptimo de 0,636, la precisión subió a 0,6393 y el recall se ajustó a 0,7254, con un F1 de 0,6797. En prueba, el AUC fue 0,9064; el F1 pasó de 0,6734 con umbral 0,5 a 0,6851 con el umbral 0,636. La validación indica que la calibración del umbral aporta una ganancia consistente en F1, producto de un mejor balance precisión–recall, patrón que se replica en prueba. En términos operativos, el modelo lineal ofrece una base sólida, fácilmente interpretable y con brechas reducidas entre validación y prueba.

3.3 MLP [64] sin dropout

El MLP de una sola capa con 64 neuronas, sin dropout, alcanzó en validación AUC 0,9071. Con umbral 0,5 obtuvo F1 de 0,6761; con umbral óptimo 0,574, el F1 mejoró a 0,6884. En prueba, el AUC fue 0,9136; el F1 pasó de 0,6874 (umbral 0,5) a 0,6992 (umbral 0,574). El resultado más alto de F1 en prueba entre todos los experimentos muestra que una arquitectura sencilla, con adecuada calibración de umbral, captura no linealidades útiles sin incurrir en sobreparametrización. La etapa de validación es decisiva para fijar un umbral menos conservador que 0,5, lo que mejora el compromiso precisión–recall y se traduce en una mejoría estable en prueba.

3.4 MLP [128,64] con dropout 0,2

La red de dos capas (128 y 64 neuronas) con dropout 0,2 presentó en validación AUC 0,9082. El F1 pasó de 0,6795 con umbral 0,5 a 0,6932 con umbral óptimo 0,604. En prueba, AUC 0,9118 y F1 de 0,6790 (umbral 0,5) frente a 0,6942 (umbral 0,604). El

desempeño es muy estable entre validación y prueba y ligeramente superior al modelo lineal cuando se emplea el umbral calibrado. La validación evidencia que pequeñas variaciones en el umbral producen ganancias marginales pero consistentes en F1, sin deteriorar la discriminación global.

3.5 MLP [256,128] con dropout 0,5

Este MLP más profundo y fuertemente regularizado por dropout alcanzó en validación AUC 0,9083 y un F1 que pasó de 0,6794 (umbral 0,5) a 0,6895 (umbral 0,581). En prueba, el AUC fue 0,9109; el F1 aumentó de 0,6728 (umbral 0,5) a 0,6914 (umbral 0,581). Con umbral 0,5 este modelo mostró el recall más alto en prueba entre todas las configuraciones, a costa de una precisión relativamente baja; la validación permitió identificar un umbral que modera ese sesgo hacia la clase positiva, elevando el F1. La brecha entre validación y prueba se mantiene acotada, lo que sugiere control adecuado de la varianza pese al mayor tamaño del modelo.

3.6 MLP [256,128] con dropout 0,2 y weight decay 1e−3

La combinación de profundidad y regularización L2 explícita obtuvo el mayor AUC en prueba (0,9142). En validación, el F1 pasó de 0,6821 (umbral 0,5) a 0,6923 con el umbral óptimo 0,607. En prueba, el F1 mejoró de 0,6825 (umbral 0,5) a 0,6970 (umbral 0,607). La validación indica que el weight decay ayuda a estabilizar el aprendizaje, y la mejora observada en prueba confirma una discriminación global ligeramente superior sin sacrificar el equilibrio precisión–recall. Este modelo es competitivo en F1 y dominante en AUC, lo que lo vuelve atractivo cuando la prioridad es la calidad discriminante total del clasificador.

3.7 Comparación integrada de los cinco experimentos

La Tabla 1 resume las métricas clave y ordena las configuraciones por F1 en prueba con umbral óptimo. Todas las redes superan a la regresión logística en F1 de prueba, aunque la diferencia es moderada. El mejor F1 corresponde al MLP [64] sin dropout (0,6992), seguido muy de cerca por el MLP [256,128] con dropout 0,2 y weight decay 1e−3 (0,6970). En términos de AUC, el valor más alto se alcanza con esta última configuración (0,9142), con el MLP [64] y el MLP [128,64] muy próximos. La validación cumple un papel crítico en dos frentes: ofrece una estimación imparcial para elegir entre arquitecturas y proporciona el umbral que maximiza F1, mejora que se traslada consistentemente a la prueba. La brecha validación–prueba es pequeña en todos los casos, lo que sugiere que la regularización y la estrategia de particionado han controlado adecuadamente el sobreajuste.

3.8 Mejor MLP regularizado: métricas por partición y efecto del umbral

El mejor MLP regularizado presenta, con el umbral óptimo derivado de validación (0,560), un comportamiento equilibrado en las tres particiones. En entrenamiento, accuracy 0,8363, precisión 0,6165, recall 0,8476, F1 0,7138 y AUC 0,9238; en validación, accuracy 0,8258, precisión 0,5960, recall 0,8154, F1 0,6886 y AUC 0,9091; en prueba, el F1 pasa de 0,6858 con umbral 0,5 a 0,6944 con el umbral óptimo, manteniendo AUC 0,9124. La Tabla 2 recoge estas métricas con sus umbrales. El patrón es coherente: la calibración del umbral eleva sistemáticamente el F1 frente al uso del umbral 0,5, al mejorar el compromiso entre precisión y recall sin perjudicar la discriminación global. Las matrices de confusión asociadas muestran, en todos los casos, que el ajuste del umbral reduce falsos negativos a cambio de un incremento controlado de falsos positivos, lo que resulta adecuado cuando se prioriza recuperar correctamente la clase positiva.

Discusión

Los resultados comparativos muestran que la regresión logística constituye una línea base sólida y difícil de superar ampliamente en este conjunto de datos. En términos de discriminación global, todas las configuraciones alcanzan áreas bajo la curva ROC cercanas o superiores a 0,91, lo que sugiere que una parte sustantiva de la señal es lineal y está bien capturada por el modelo base. No obstante, las redes neuronales multicapa aportan mejoras marginales pero sistemáticas en F1 cuando se emplea un umbral de decisión calibrado en validación. En particular, el MLP de arquitectura mínima [64] sin dropout consigue el mayor F1 en prueba, mientras que la variante [256,128] con dropout 0,2 y weight decay 1e−3 logra el AUC más alto. Esta divergencia entre la métrica de ranking (AUC) y la métrica de punto de operación (F1) subraya que la superioridad de un modelo depende del criterio operativo con el que se evalúe.

La etapa de validación desempeña un papel doble y decisivo. Por un lado, permite elegir entre arquitecturas con desempeños muy próximos pero no idénticos, preservando la prueba como estimación final no sesgada. Por otro, habilita la calibración del umbral de decisión mediante maximización de F1, lo que mejora de manera consistente el equilibrio precisión–recall frente al uso del umbral 0,5. En todos los experimentos, el F1 de validación aumentó tras la calibración y esta ganancia se trasladó a la prueba, evidenciando que el criterio de selección generaliza. La Tabla 1 sintetiza esta regularidad: las diferencias absolutas en F1 entre modelos son moderadas, pero la mejora relativa por ajuste de umbral es estable y operativamente relevante.

El análisis del mejor MLP regularizado aporta información adicional sobre sesgo y varianza. La Tabla 2 muestra que las métricas en entrenamiento y validación son cercanas, sin

brechas pronunciadas, lo cual sugiere un control adecuado del sobreajuste gracias a las capas de regularización (dropout), la penalización L2 (weight decay) y el uso de early stopping. En prueba, el uso del umbral óptimo determinado en validación eleva el F1 respecto al umbral 0,5 manteniendo el AUC prácticamente constante, lo que es coherente con la interpretación de AUC como métrica independiente del umbral.

Desde una perspectiva práctica, los resultados invitan a priorizar la simplicidad cuando los recursos de cómputo, la transparencia y la reproducibilidad son objetivos centrales. La regresión logística ofrece interpretabilidad directa de coeficientes y un desempeño competitivo. Cuando se busca exprimir ganancias marginales en F1 o maximizar la discriminación global, un MLP sencillo y regularizado es una alternativa razonable, especialmente si se gestiona correctamente la calibración del umbral y se mantiene una supervisión del desempeño fuera de muestra. En ambos casos, la adopción de umbrales sensibles al costo de errores de clasificación puede afinar aún más el desempeño operativo en contextos con desbalance de clases.

Cabe señalar limitaciones. Primero, los resultados dependen del esquema de particionado específico y de la definición del objetivo binario; variaciones en la partición o en el preprocesamiento podrían alterar marginalmente los valores numéricos, aunque es poco probable que cambien las conclusiones cualitativas. Segundo, la métrica F1 supone costos simétricos de falsos positivos y falsos negativos; si en la aplicación real estos costos difieren, convendría optimizar un criterio ponderado o utilizar métricas costo-sensibles. Tercero, el conjunto Adult incluye variables demográficas y laborales que, si se usan para decisión automatizada, exigen consideraciones éticas y de cumplimiento normativo; la evaluación de equidad por subgrupos y la calibración por segmentos debería contemplarse antes de cualquier despliegue.

Conclusiones

El estudio confirma tres hallazgos principales. Primero, la regresión logística es una línea base robusta: alcanza AUC elevado y, con calibración de umbral, F1 competitivo, lo que la convierte en una opción preferente cuando la interpretabilidad es crítica. Segundo, las redes neuronales multicapa, incluso con arquitecturas simples, aportan mejoras marginales y consistentes en F1 y, en ciertas configuraciones regularizadas, el mayor AUC; estas ganancias se materializan únicamente cuando se acompaña el entrenamiento con un proceso riguroso de validación y ajuste de umbral. Tercero, la etapa de validación es esencial no solo para comparar modelos, sino para establecer el punto de operación adecuado; trasladar a prueba el umbral aprendido en validación incrementa el F1 sin deteriorar la discriminación global.

En términos operativos, si la prioridad es la precisión global con transparencia, la recomendación es implementar la regresión logística y calibrar el umbral según validación, monitoreando su estabilidad. Si la prioridad es maximizar F1 o AUC con un costo marginal de complejidad, se sugiere un MLP pequeño y regularizado, también con umbral calibrado. Para trabajos futuros, se proponen tres líneas: incorporar calibración probabilística explícita de las salidas (Platt o isotónica) para mejorar la calidad de las probabilidades; explorar umbrales costo-sensibles o dependientes del contexto operativo; y evaluar métricas de equidad y estabilidad temporal para garantizar que el desempeño se mantenga en subgrupos y en presencia de cambios de distribución.

La evidencia sintetizada en la Tabla 1 y el detalle del mejor MLP en la Tabla 2 respaldan estas conclusiones: las diferencias absolutas entre modelos son moderadas, la calibración del umbral aporta el mayor retorno práctico y la regularización bien sintonizada permite controlar el riesgo de sobreajuste, obteniendo un desempeño que generaliza de manera consistente de la validación a la prueba.

Anexos

Experimento	**thr_opt**	**VAL AUC**	**VAL F1**	**VAL Prec**	**VAL Rec**	**VAL Acc**	**TEST AUC**	**TEST F1**	**TEST Prec**	**TEST Rec**	**TEST Acc**
MLP [64] d=0.0 (best_epoch=8)	0.574	0.907	0.688	0.604	0.800	0.829	0.914	0.699	0.622	0.798	0.838
MLP [256,128] d=0.2 wd1e-3 (best_epoch=27)	0.607	0.910	0.692	0.618	0.787	0.835	0.914	0.697	0.619	0.797	0.836
MLP [128,64] d=0.2 (best_epoch=7)	0.604	0.908	0.693	0.619	0.787	0.835	0.912	0.694	0.610	0.804	0.833
MLP [256,128] d=0.5 (best_epoch=12)	0.581	0.908	0.690	0.603	0.805	0.829	0.911	0.691	0.590	0.835	0.824

LogReg (C=1.0, balanced)	0.636	0.901	0.680	0.639	0.725	0.838	0.906	0.685	0.637	0.742	0.839

Tabla 1. Resultados comparativos de los cinco experimentos (validación y prueba).

Tabla 2. Métricas del mejor MLP con regularización por partición y umbral.

Split	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC AUC**	**Umbral**
Train	0.836	0.617	0.848	0.714	0.924	0.560
Validación	0.826	0.596	0.815	0.689	0.909	0.560
Test @ 0.5	0.814	0.571	0.859	0.686	0.912	0.500
Test @ thr_opt	0.829	0.600	0.824	0.694	0.912	0.560

Nota: métricas redondeadas a 3 decimales para presentación.

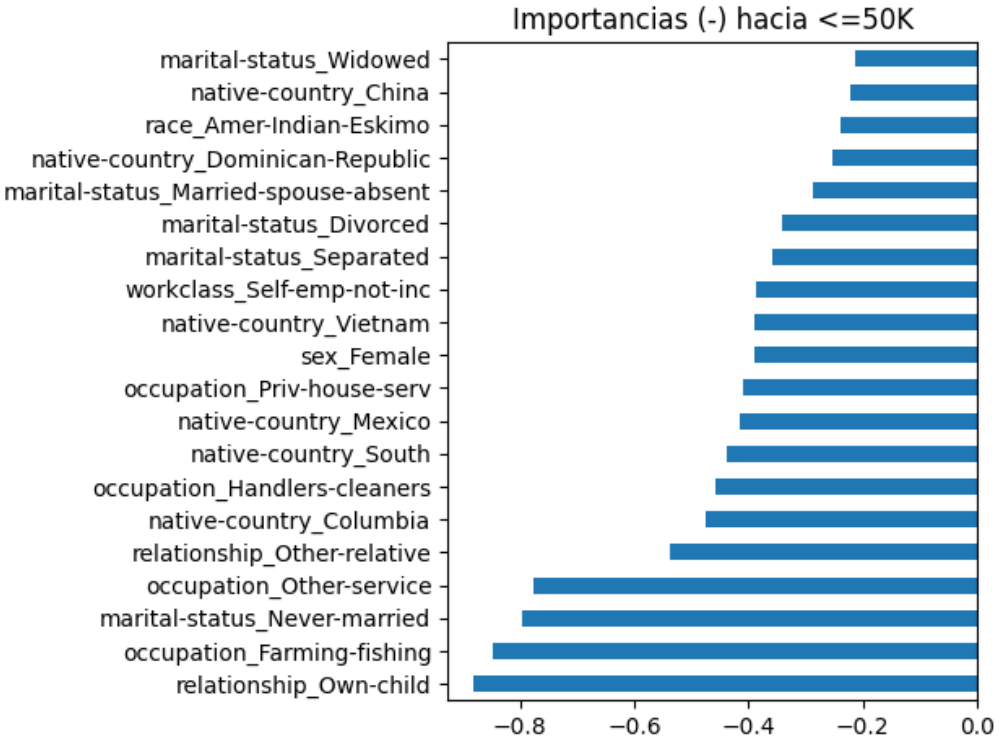

Figura 1. Importancia hacia <= 50k

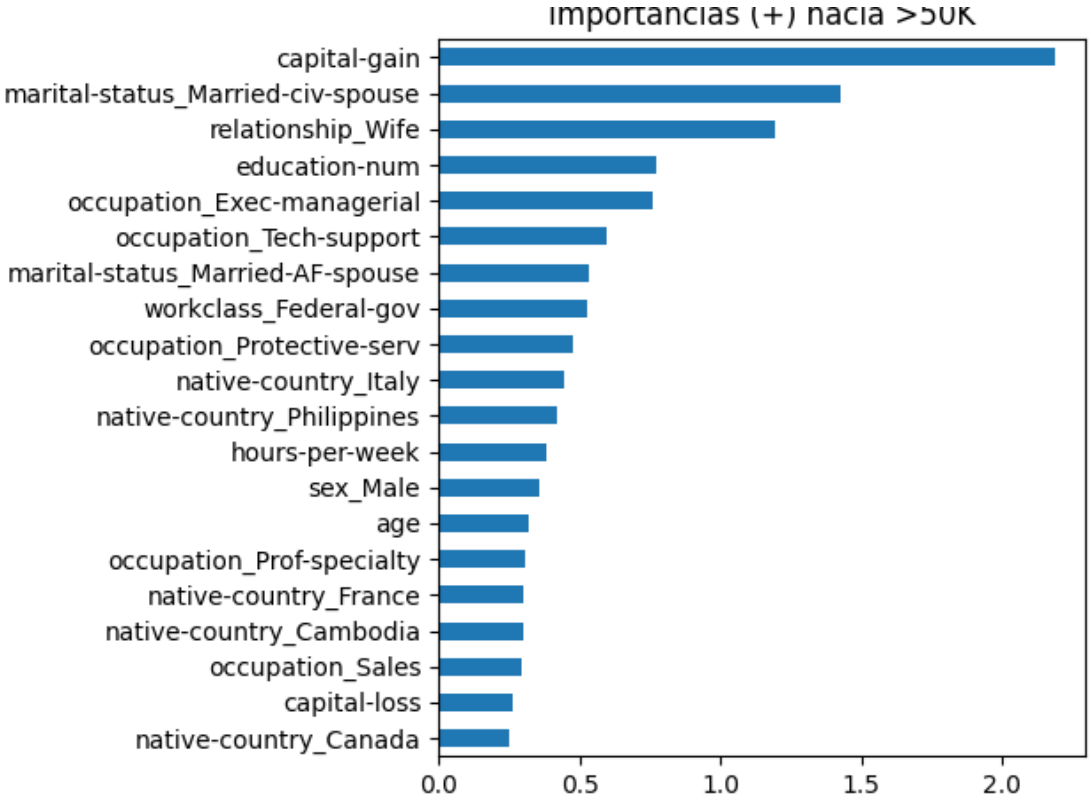

Figura 2. Importancia hacia > 50k


TEST - Precision-Recall
PR
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0

Figura 3.

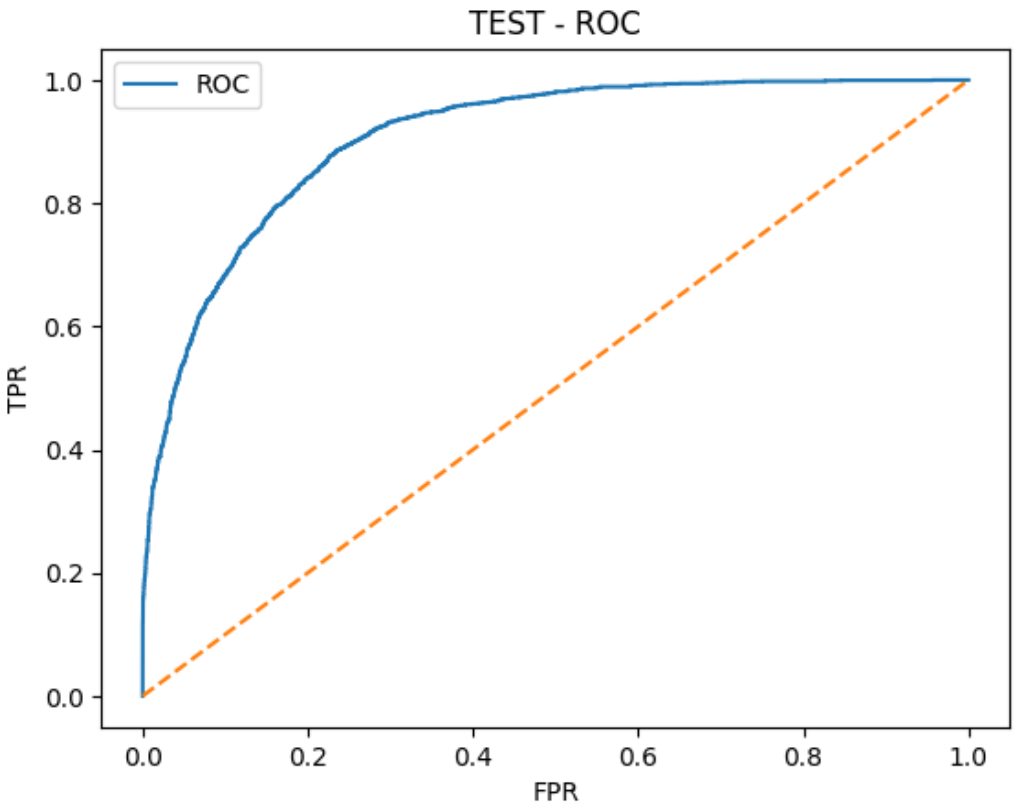

Figura 4.

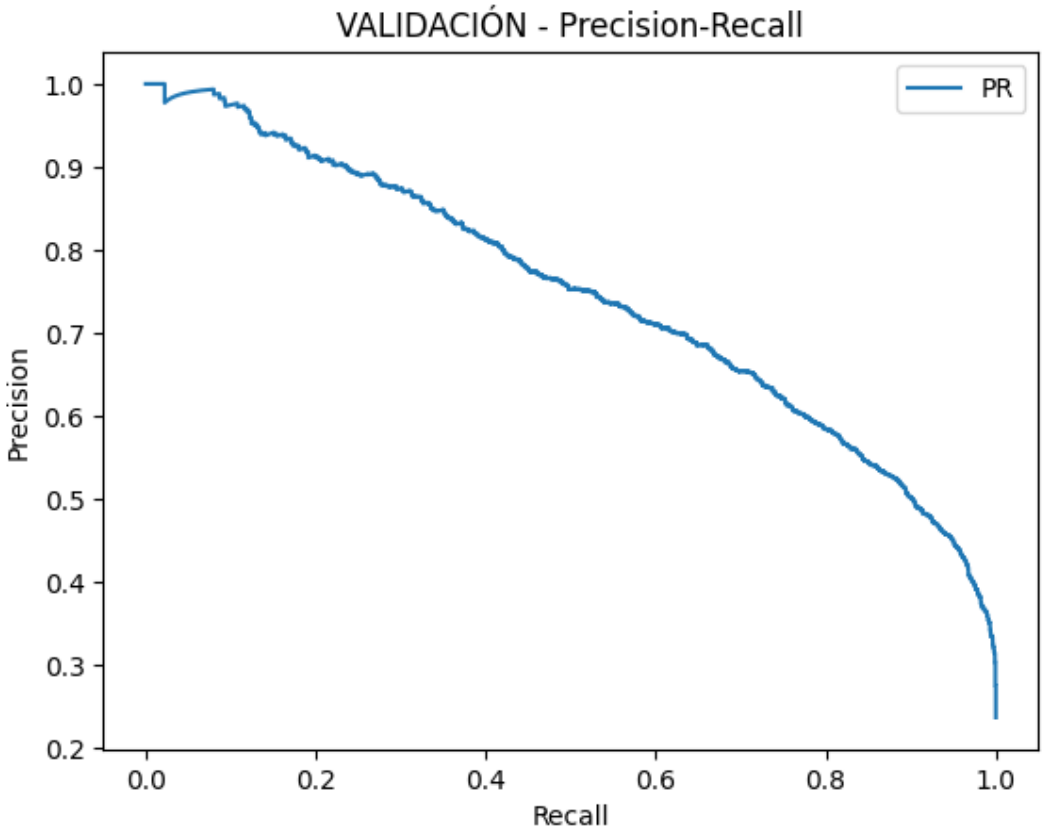
VALIDACIÓN - Precision-Recall
PR
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0

Figura 5.

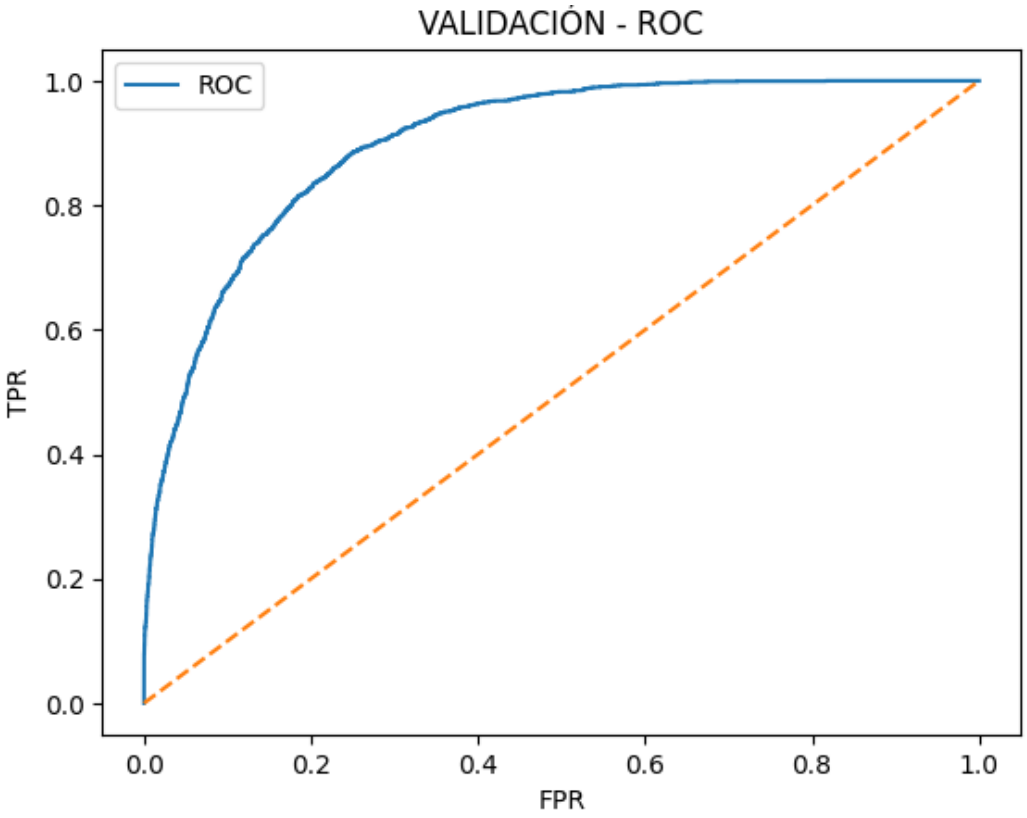

Figura 6.

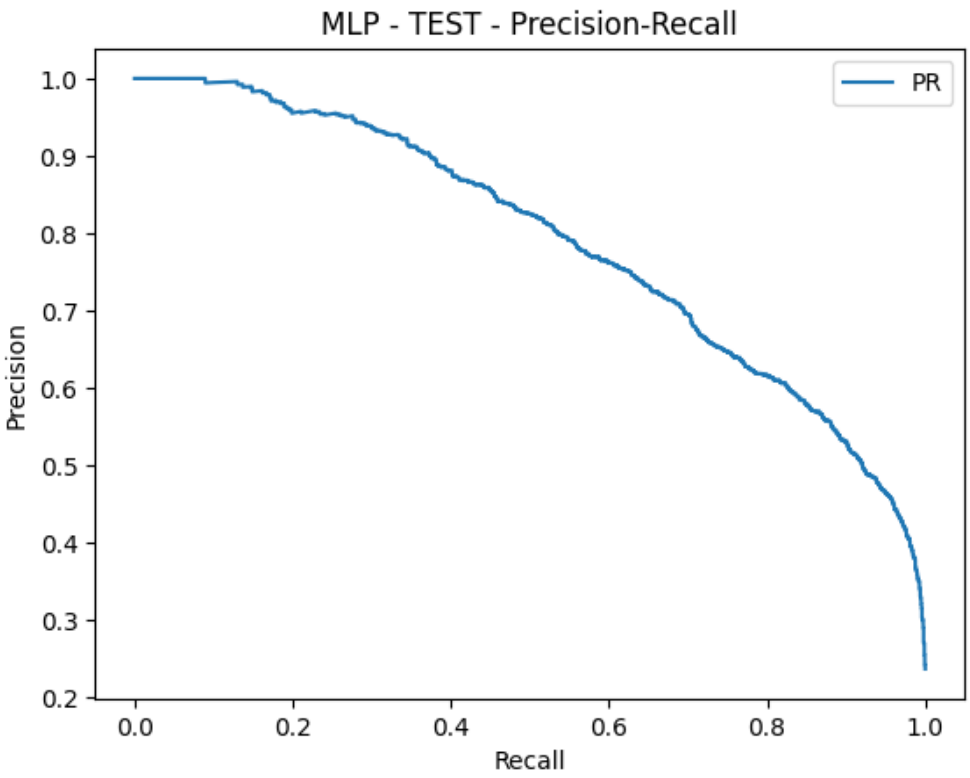
MLP - TEST - Precision-Recall
PR
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0

Figura 7.

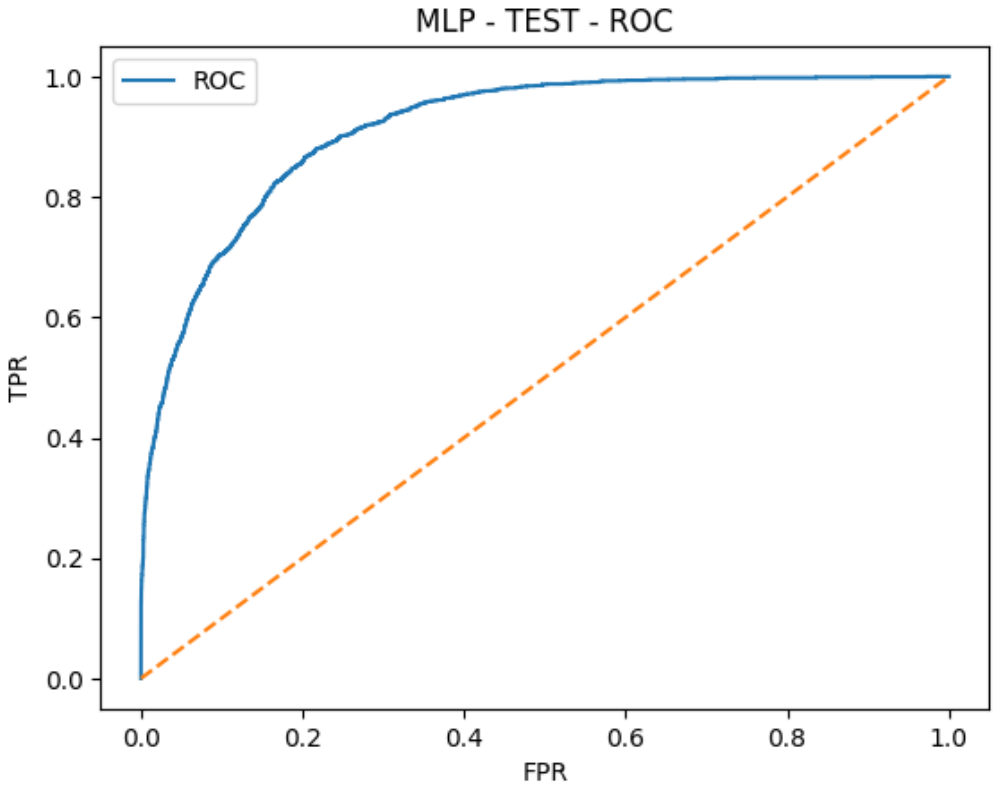

Figura 8.

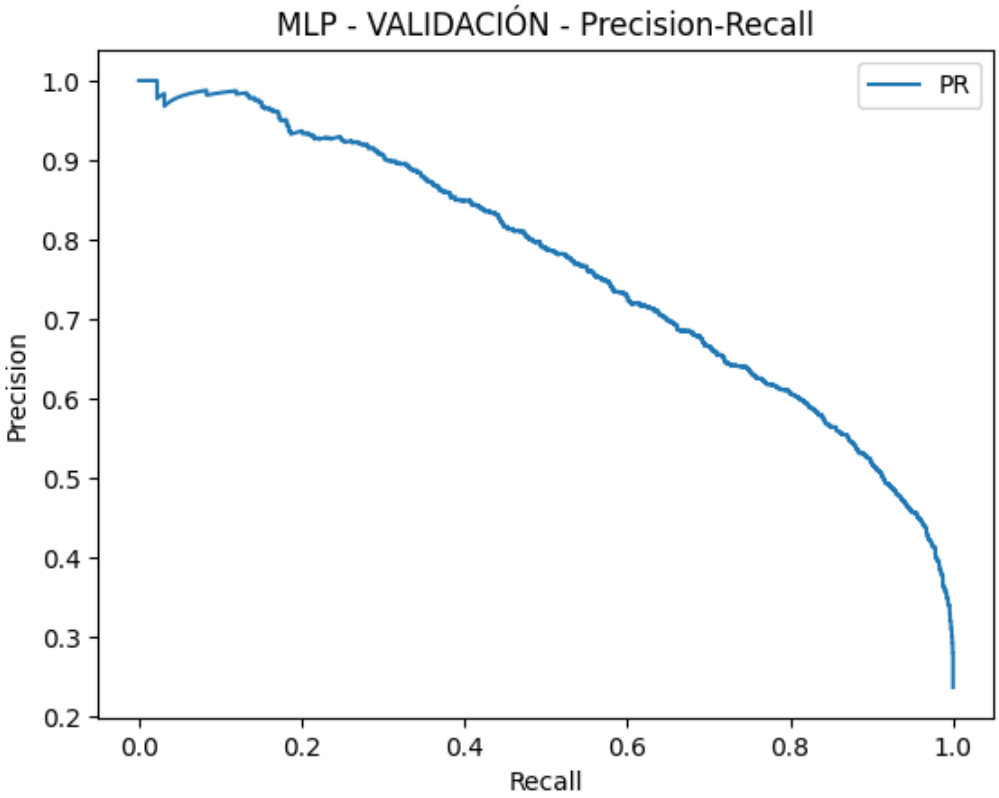
MLP - VALIDACIÓN - Precision-Recall
PR
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0

Figura 9.

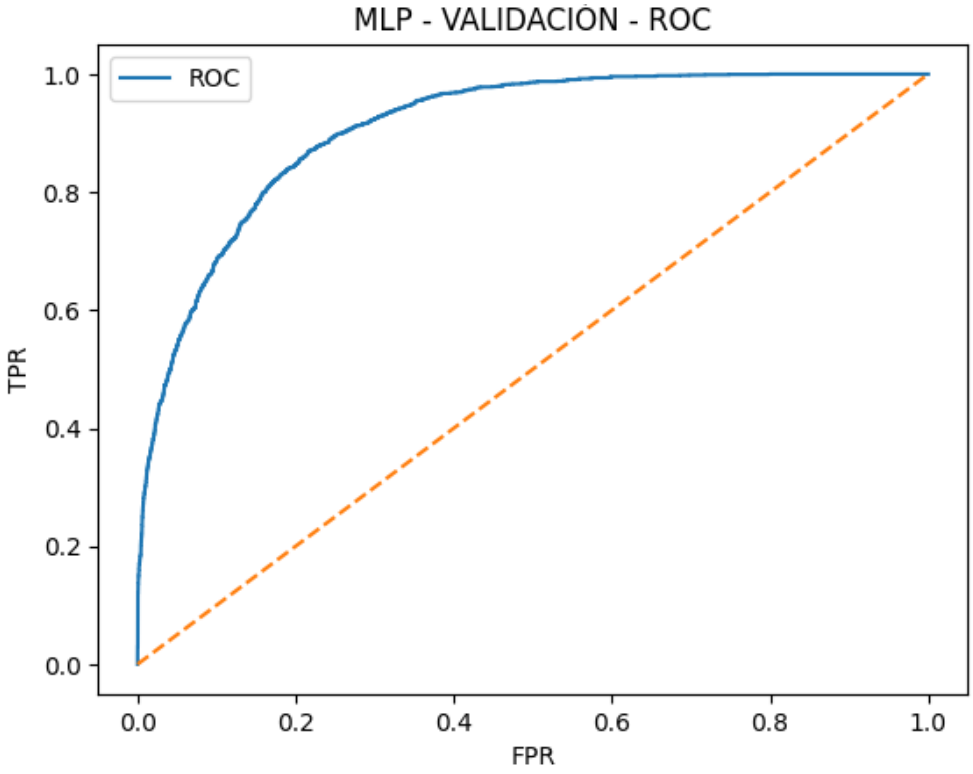

Figura 10.