Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Módulo 2:

Análisis y Reporte sobre el desempeño del modelo

Profesor:

Jorge Adolfo Ramírez Uresti

Alumno:

Pablo Spínola López

A01753922

11 de septiembre de 2024

Introducción

A continuación, se presenta un análisis del desempeño de un modelo de machine learning, respaldado con métricas, gráficas y comparaciones entre estas, con la finalidad de que, posteriormente, se pueda realizar un diagnóstico de aquellas medidas que definen y afectan la precisión y generalización del modelo, como lo pueden llegar a ser el sesgo y la varianza. Todo esto para que, finalmente, se pueda realizar un ajuste informado a dicho modelo, y se puedan ajustar ciertos parámetros o utilizar ciertas técnicas de regularización en los datos para que el modelo aprenda de forma más efectiva y alcance un mejor desempeño tras haber realizado un riguroso análisis.

Selección del modelo

Cabe mencionar que, para el correcto desarrollo de este análisis y mejora de un modelo de aprendizaje máquina, se utilizaron dos algoritmos de entrenamiento con datasets similares y la misma lógica de clasificación. Esto con la finalidad de diferenciar a aquél que tuviese un mejor desempeño y poder realizar mejoras para alcanzar un mejor nivel de desempeño. En lo personal, los dos algoritmos empleados para entrenar los modelos fueron *Regresión Logística*, el cual fue implementado de forma manual, y *Random Forest*, implementado con ayuda de la librería de Python *SciKit Learn*.

El modelo de entrenamiento que se analizará y se afinará en el documento presente es aquél implementado manualmente, es decir, el que implementó **Regresión Logística**. Esta decisión se basa en que, tras ver los resultados de desempeño y apreciar que con el Random Forest se alcanzó una muchísimo mejor f1-score y una precisión muy superior (con casi el 100% en ambas), no cabe mucha mejora en este modelo, sin embargo, al tomar en cuenta las métricas de la regresión logística y tener en cuenta que su implementación fue manual, existe una tasa de mejora superior y un análisis más visible en torno a los valores arrojados, tanto en gráficas como las métricas impresas en consola. Por esto mismo es que se realizará el análisis con base en este algoritmo: *Regresión Logística*.

Dataset

Elección del Dataset utilizado

> *X: {90, 89, 16, 87, 86, 42, 81, 83, 32, 40, 94, 72, 88, 92, 100, 24, 30, 73, 66, 58, 35, 22, 15, 69, 70, 74, 86, 62, 94, 77, 10, 88, 82, 79, 61, 15, 96, 71, 13, 85, 24, 49, 76, 84, 99, 2, 80, 67, 97, 99}*
>
> *Y: {1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1}*

Se optó por utilizar este dataset para el entrenamiento, teniendo a la variable independiente dentro conjunto de valores numéricos enteros dentro de un rango que va del 0 al 100, y la

variable dependiente un valor de entre 0 y 1; un valor binario. La elección del dataset viene de la mano con el algoritmo en cuestión, ya que, al ser un valor binario el que se ha de predecir, se observa claramente que es un problema clasificatorio, otorgando a cierto tipo de datos 0 o 1 dependiendo del criterio de discriminación de los datos, mismo criterio del que el modelo debe ser capaz de identificar, y al ser una regresión logística, se espera poder clasificar de manera precisa. En este caso en concreto, el dataset representa el conjunto de calificaciones que obtuvieron 50 alumnos en un examen, y el caso de haber aprobado o no dicho examen, dependiendo si esta calificación es igual o mayor que 70 (caso aprobatorio 1), o si es menor (caso reprobatorio 0).

Por eso es por lo que se ha seleccionado este dataset, ya que queremos clasificar a futuros alumnos que tomen el examen de forma correcta en caso de que hayan aprobado o reprobado con respecto al criterio impuesto. Así que el dataset de entrenamiento contiene valores aleatorios en un rango alrededor del 70 (0 - 100) en comparación con los valores que se esperan predecir en el dataset de evaluación que contienen valores dentro del mismo rango, pero distintos a los utilizados en el conjunto previo, con la finalidad de que el modelo entrenado pueda generalizar bien, teniendo predicciones precisas con datos no vistos durante la fase de entrenamiento y aprenda correctamente la forma de la clasificación. Por esto es que, para la evaluación final del modelo, se utilizó un nuevo dataset de 30 valores distintos dentro de un mismo rango de valores, comprobando la generalización y evaluando así la precisión dentro de este nuevo dataset:

X evaluación: {47, 82, 14, 29, 56, 91, 67, 23, 34, 89, 6, 52, 77, 16, 40, 3, 68, 95, 24, 70, 86, 19, 44, 59, 84, 28, 11, 97, 81, 33}

Y evaluación: {0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0}

División del dataset

Como es de esperarse, para entrenar este modelo, el dataset de entrenamiento mostrado previamente fue dividido en 3 partes; entrenamiento, validación y prueba. El subconjunto de entrenamiento fue el más extenso, representando un **60%** de la información, y utilizado para actualizar los pesos y entrenar correctamente el modelo. El subconjunto de validación, al igual que el de prueba, contiene el **20%** de los datos, y fue utilizado para calcular la precisión del modelo a lo largo del entrenamiento, de modo que este continuase hasta alcanzar una convergencia aceptable. El conjunto de prueba fue utilizado para comprobar la precisión y desempeño del modelo tras finalizar el entrenamiento, ya que, al contener datos distintos a los utilizados para entrenar y validar el entrenamiento, muestra métricas más claras de las capacidades del modelo.

Cabe mencionar que, al ser un ambos datasets ya aleatorizados, no fue necesario "randomizar" los valores para realizar la división en los subconjuntos ya mencionada, por lo que simplemente se especificó la proporción que deberían tener los conjuntos del dataset de entrenamiento.

En la figura 1.1, se puede observar una clara distribución de los datos del entrenamiento en los diferentes subconjuntos, habiendo valores de los tres subconjuntos en ambas clasificaciones, y distribuidos aleatoriamente, por lo que se espera que se generalice correctamente y haya una gran aproximación de las estimaciones a los valores reales y un mejor desempeño general del modelo al momento de probarlo con el nuevo dataset.

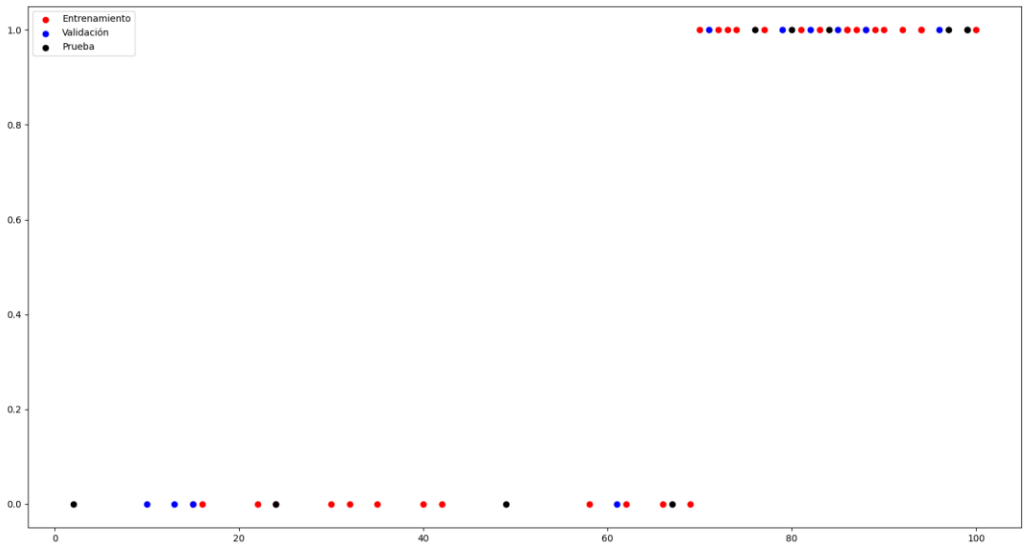

Figura 1.1

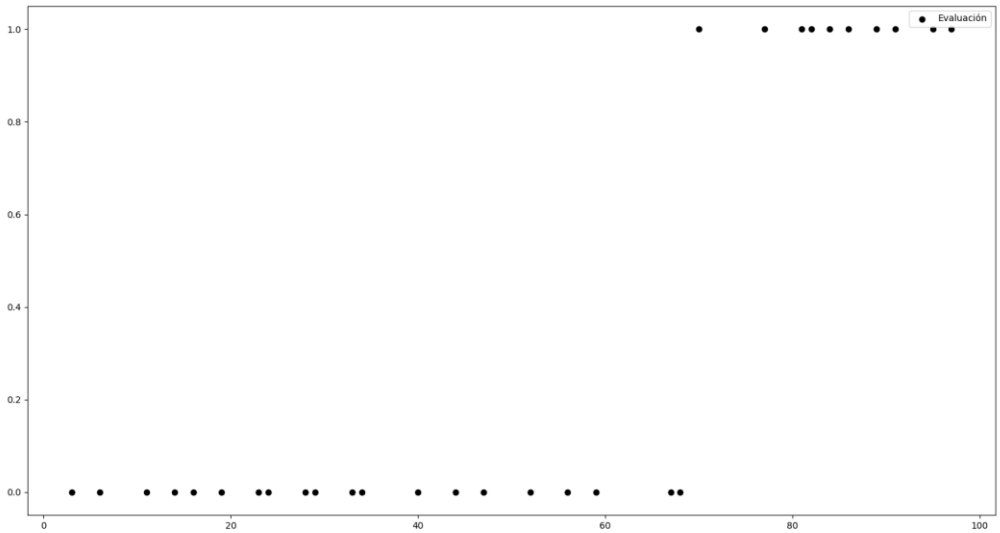

Figura 1.2

En la figura 1.2, se muestra igual una distribución a lo largo del rango especificado de valores y distinta a los valores mostrados en la figura 1.1, teniendo números válidos para evaluar precisión y comprobar generalización.

A continuación, se presentan los subconjuntos del dataset de entrenamiento:

X entrenamiento: [90, 89, 16, 87, 86, 42, 81, 83, 32, 40, 94, 72, 88, 92, 100, 24, 30, 73, 66, 58, 35, 22, 15, 69, 70, 74, 86, 62, 94, 77]
Y entrenamiento: [1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1]

X validación: [10, 88, 82, 79, 61, 15, 96, 71, 13, 85]
Y validación: [0, 1, 1, 1, 0, 0, 1, 1, 0, 1]

X test: [24, 49, 76, 84, 99, 2, 80, 67, 97, 99]
Y test: [0, 0, 1, 1, 1, 0, 1, 0, 1, 1]

Diagnóstico del modelo y análisis de métricas y gráficos

Para realizar un análisis correcto del desempeño del modelo para así poder diagnosticarlo y dar una solución sólida para mejorar su precisión y puntaje F1, se presenta una serie de métricas obtenidas desde la consola, al igual que gráficas representativas, desde las pérdidas hasta la curva de aprendizaje final. Gracias a estos valores y gráficas, nos es posible comprender el **bias** y **varianza** que presenta el modelo, diagnosticando así la presencia de un overfit, underfit o, en el mejor de los casos, un fit normal. En torno a esta declaración es que se podrá ajustar el modelo en aquellas métricas ya mencionadas que afecten el desempeño y exactitud de nuestro modelo.

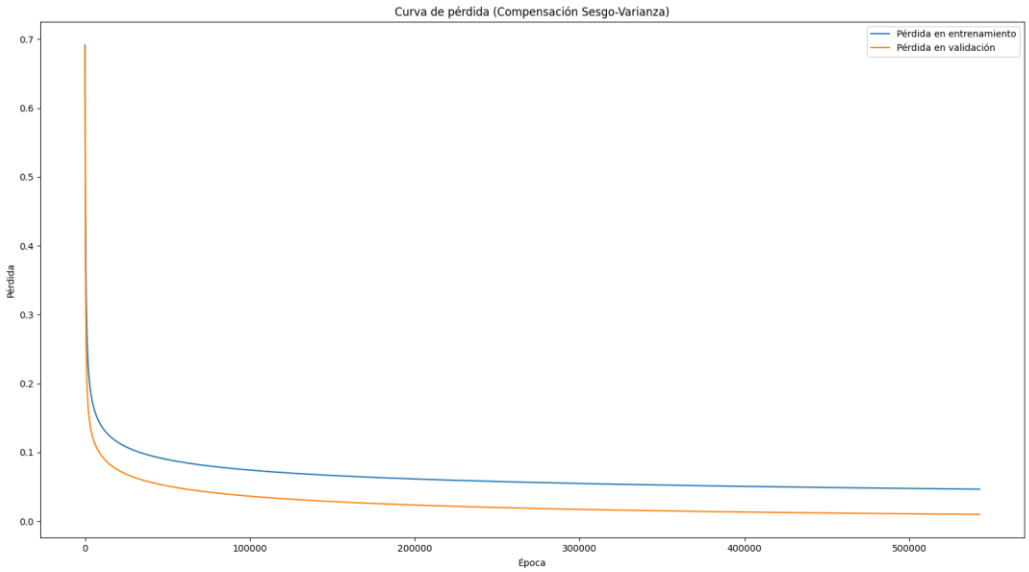

Figura 2.1

La gráfica que se muestra en la figura 2.1representa la curva de pérdida a lo largo del entrenamiento. Se puede ver claramente cómo es que ambas pérdidas, tanto la de entrenamiento como la de validación disminuyen suavemente desde un inicio, con cierto paralelismo y con una posible tendencia a una convergencia. El error bajo en la curva de error en la validación muestra una buena generalización del modelo, es decir, se acomoda bien para los datos no vistos durante el entrenamiento y es capaz de realizar buenas predicciones, un buen indicativo inicial de poco bias y varianza. Gracias a la forma de ambas curvas y que su descenso es suave y curvo, sin subir posteriormente o tener valores altos, empezamos a tener indicios de que el modelo está bien, sin que el underfitting u overfitting sean una preocupación de momento, sin embargo, puede mejorar aún más en caso de que se observe una convergencia en ambas curvas, indicando una notable mejora.

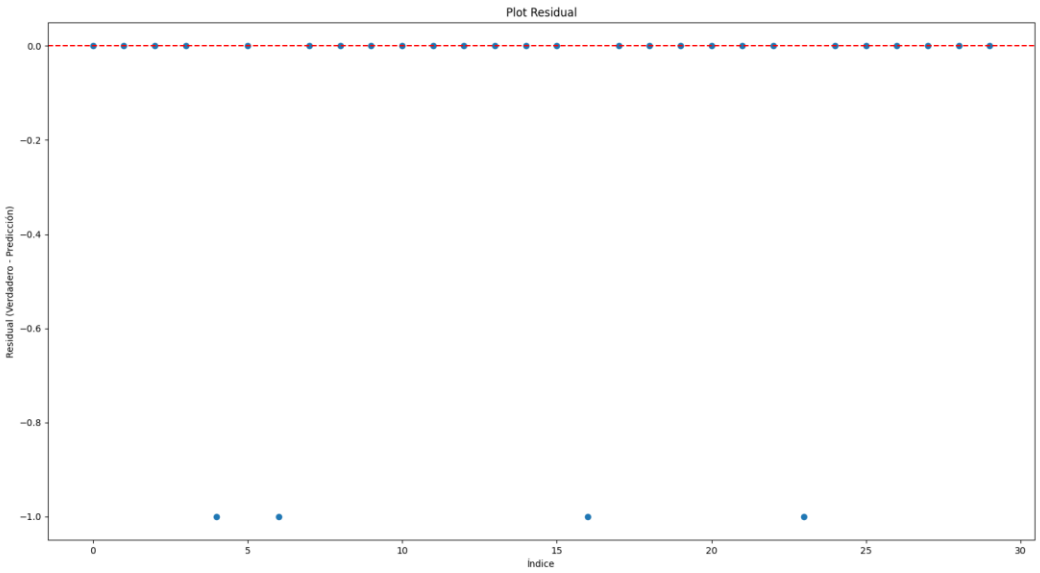

Figura 3.1

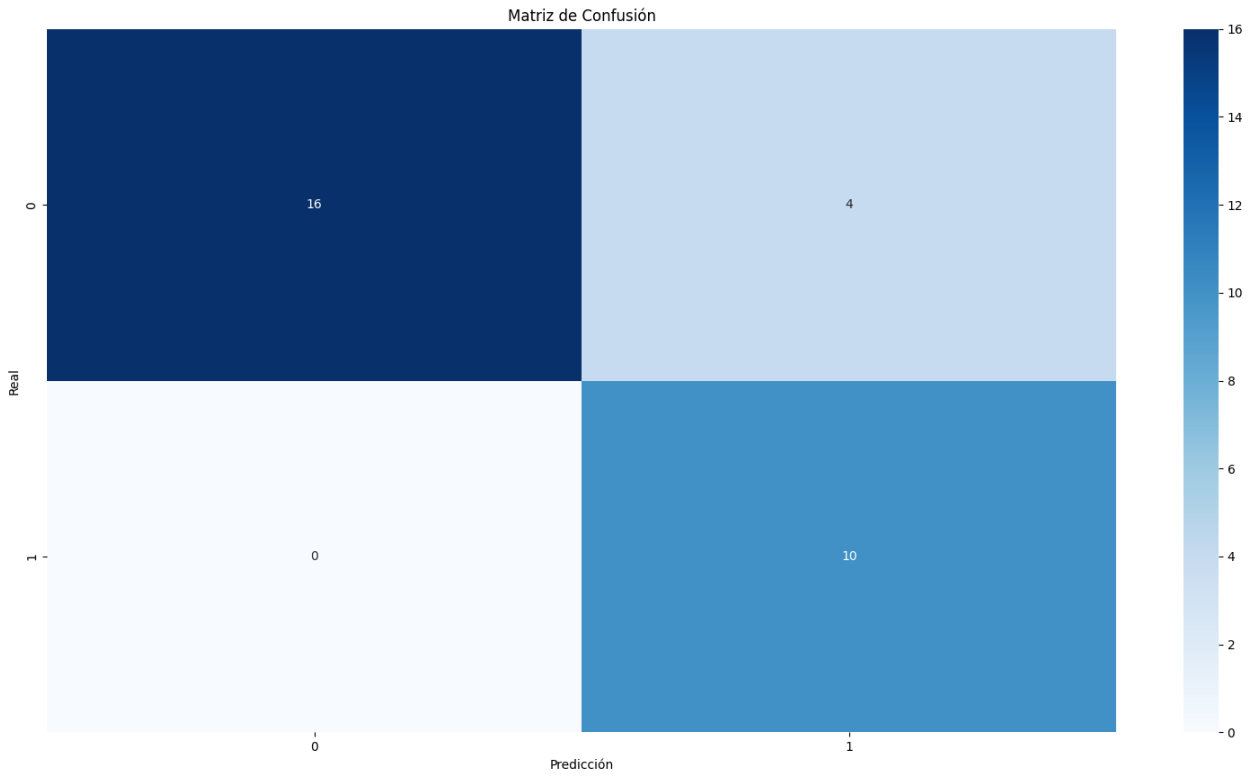

Figura 4.1

En estas dos figuras (3.1 y 4.1), las cuales son una gráfica de valores residuales o erróneos y una matriz de confusión, respectivamente, se ve una gran densidad en las áreas de valores correctos; verdaderos positivos y verdaderos negativos, con cuatro valores erróneos que representan una porción considerable tomando en cuenta el tamaño del dataset empleado, y los cuales fueron catalogados como positivos siendo en realidad negativos, es decir, falsos positivos. Esto nos daría una precisión de poco más de 2 tercios, por lo que es importante afinar esta situación, ya que buscamos una mayor precisión al predecir la clase positiva.

En las siguientes 2 figuras se muestran 2 tipos de curvas estadísticas que relacionan la precisión con el recall de diferentes formas. La primera (Figura 5.1) representa una curva PR (Precisión-Recall), la cual se enfoca en el desempeño con respecto a la clase positiva. El hecho de que inicie con un recall de 1 y se mantenga a lo largo del crecimiento de la precisión, nos dice que se catalogó correctamente la información positiva desde un inicio, posible señal de overfitting, sin embargo, al aumentar la precisión y llegar a poco más de 0.7 (más de dos tercios como se mencionó con anterioridad al analizar la matriz de confusión), indicando que poco a poco fueron habiendo menos falsos negativos. Esto puede significar que el modelo es algo débil para la clase positiva y un dataset poco balanceado, por lo que es importante notar estos datos para la próxima mejora de este modelo. La segunda figura (6.1) representa una curva ROC, enfocada en relacionar la tasa de verdaderos positivos

contra falsos positivos, y al proyectar un área de casi 0.8, nos indica que el modelo es bueno clasificando, misma característica que se observa en la curva de PR, viendo un posible espacio de mejora.

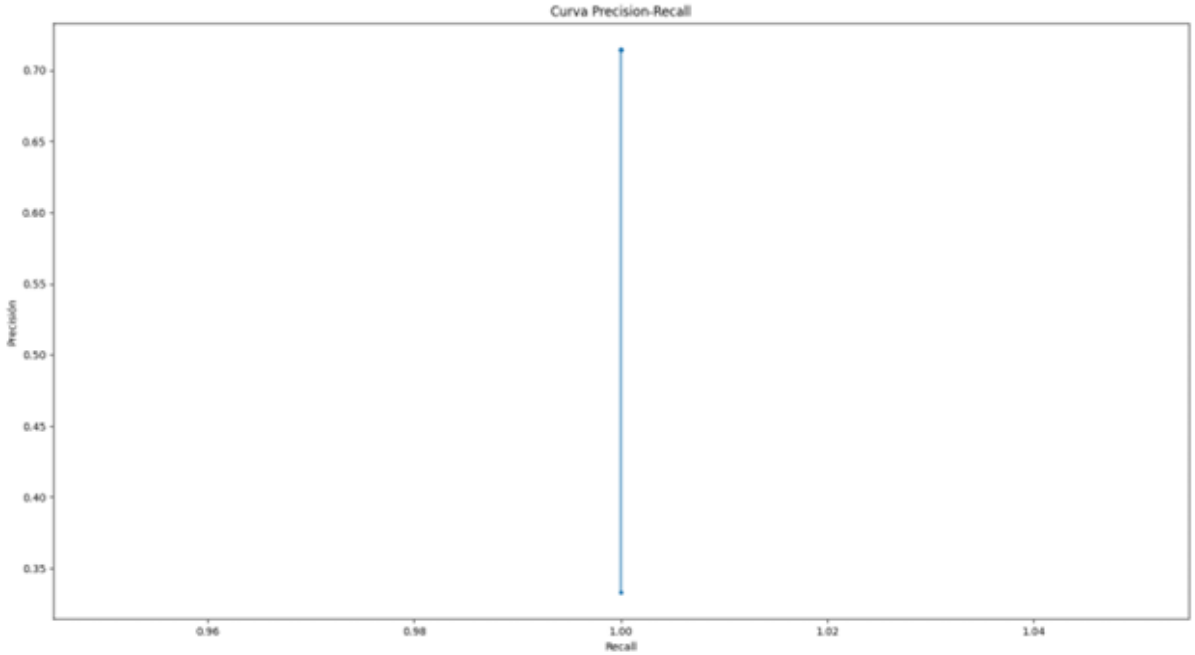

Figura 5.1

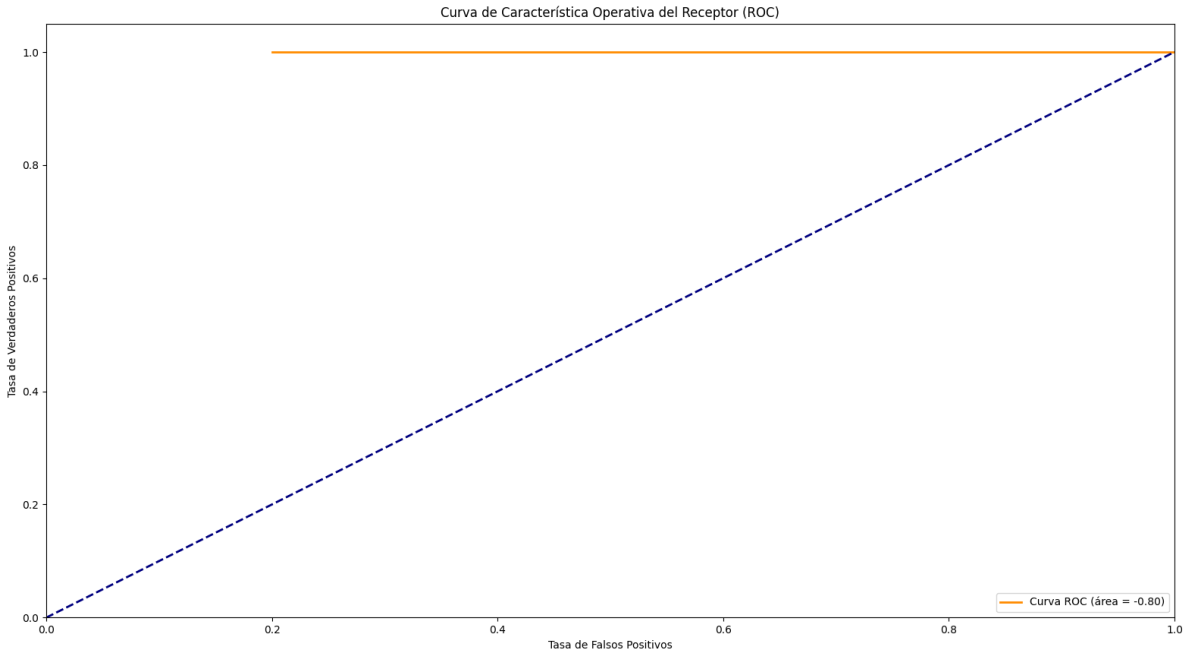

Figura 6.1

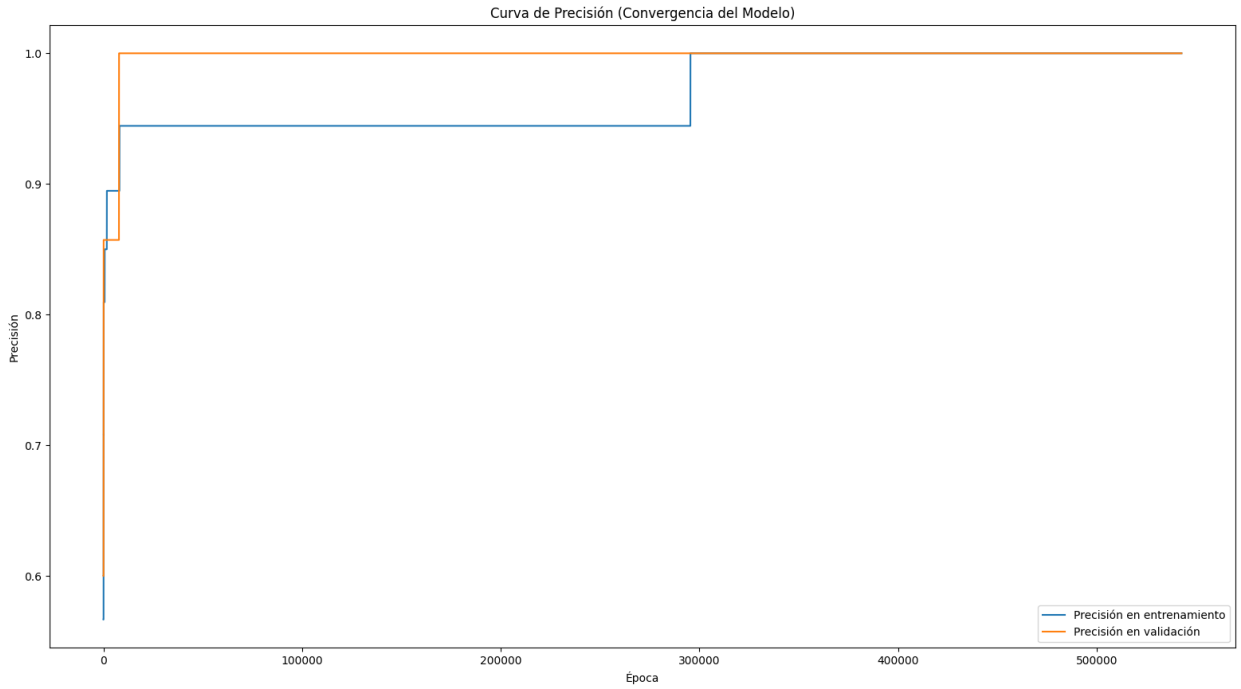

Figura 7.1

```
Entrenamiento finalizado en la época 542768
 Pesos finales ajustados del modelo:
        w0 = -3.061258469489301
        w1 = 71.63624922301815

 Pérdida con data de validación: 0.0100
 Pérdida con data de prueba: 0.0082
 Precisión: 0.7142857142857143
 Recall: 1.0
 Puntuación F1: 0.8333333333333333
```

Figura 8.1

Finalmente, vemos la curva de aprendizaje mostrada en la figura 7.1, con una convergencia alcanzada alrededor de la época 550000. Aunque no se ve un aumento tan suave en la precisión, se puede observar una convergencia entre la precisión del entrenamiento y de la validación. Esto quiere decir que ha generalizado correctamente, sin embargo, se corre el riesgo de un overfit con la información de entrenamiento, aunque poco, ya que al ver las métricas arrojadas en la figura 8.1, finalmente observamos que la puntuación F1 es del 83%, indicando un buen modelo para predicciones, con una media armónica entre precisión y recall considerable, y una pérdida baja en conjunto de validación y de validación. Sin embargo, es verdad que la puntuación F1 salió relativamente alta debido a que el recall es muy alto y la precisión regular, por lo que hay que buscar mejorar la precisión para que uno no eleve al otro.

Una vez vistas las diagnosticado el modelo e identificadas sus fallas y áreas de mejora, es posible tener una posible modificación, para reducir así la presencia de varianza

y bias, reduciendo de igual forma el overfitting y underfitting. Entre las posibles mejoras se encuentra:

- Normalizar la información: Con esto, se espera tener una convergencia más pronta al convertir los valores en porcentajes relativos, además de mejorar el gradiente al normalizar los datos. Esto añade exactitud y reduce el bias en general, aportando al desempeño general y optimización del modelo.
- Aumentar y balancear los datos utilizados en el conjunto de evaluación: Esto nos ayuda a conocer qué tan bien es que se ha generalizado el modelo, por lo que ayuda a reducir tanto underfit como overfit, disminuyendo bias. Además, ayuda a que la matriz de confusión esté balanceada, siendo una buena métrica para evaluar el desempeño del modelo y calcular precisión.
- Realizar un grid search: Para encontrar los mejores parámetros que más se acerquen a una predicción precisa del modelo y un fit normal, puede convenir un grid search, en el que se barran diferentes valores de tasa de aprendizaje para encontrar la que de una mejor convergencia.

Resultados finales de mejoras

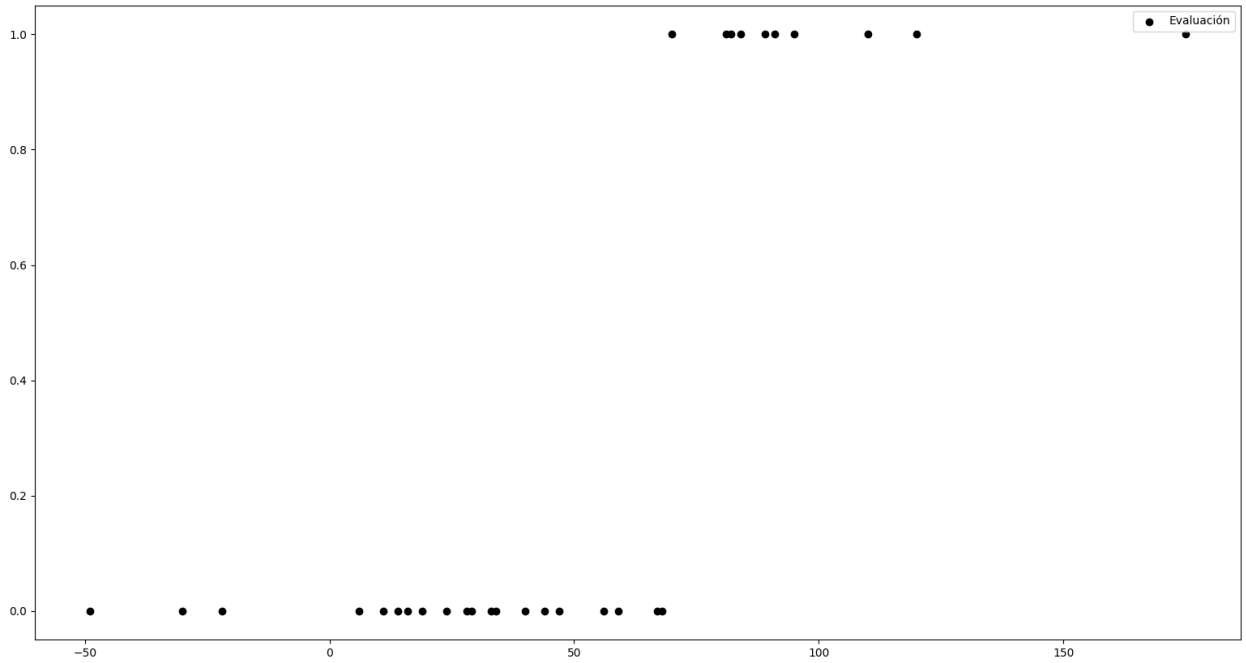

Figura 1.3

Aquí se observa que los datos están más distribuidos en el conjunto de evaluación, entre la clase negativa y positiva, con un mayor rango de valores que van de -50 a 180. Esto con la esperanza de una mayor generalización y mejor relación entre Recall y Precisión (puntuación F1). Además de que al evaluar el modelo con datos muy distintos a los vistos

en el dataset de entrenamiento, puede esperarse una mucho menor cantidad de bias y varianza en caso de tener una alta precisión.

En la gráfica de pérdidas mostrada a continuación, se observa un error aún menor, y una curva igual de suave que casi converge, señalando un bajo riesgo de overfit.

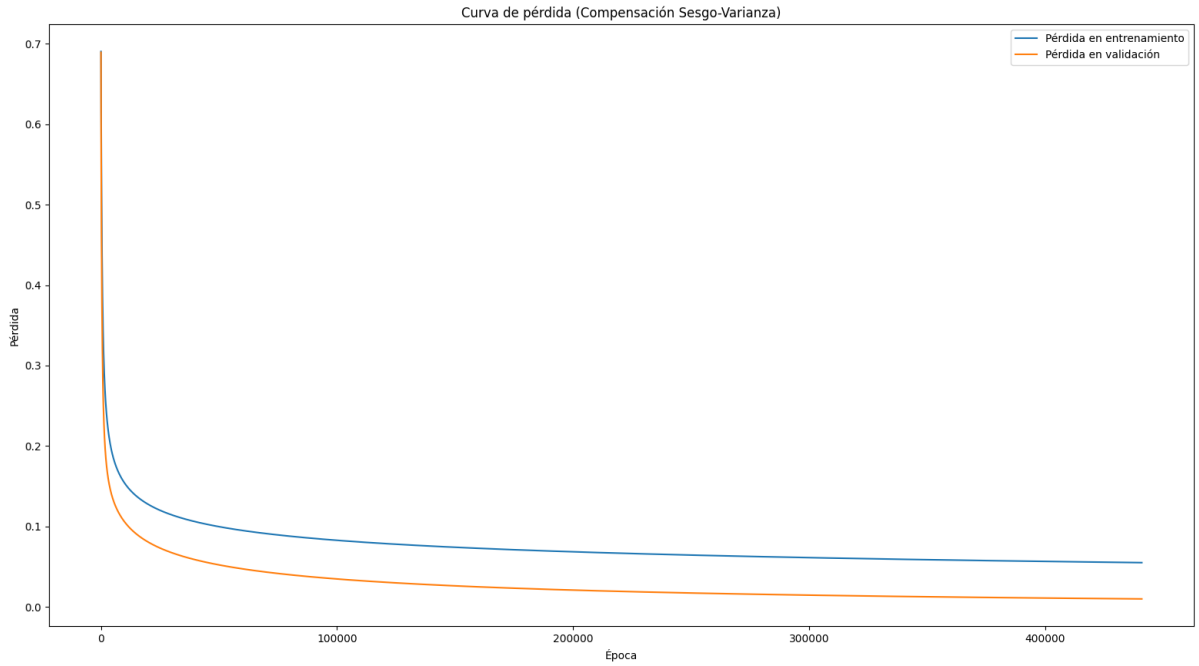

Figura 2.2

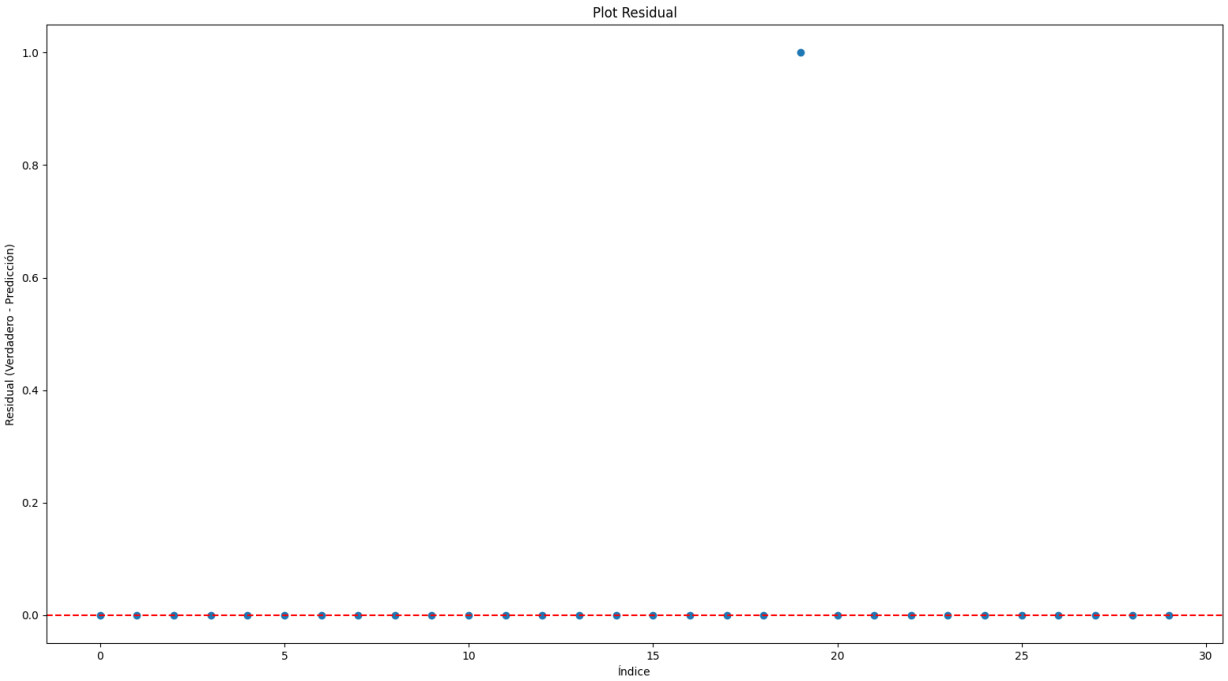

Figura 3.2

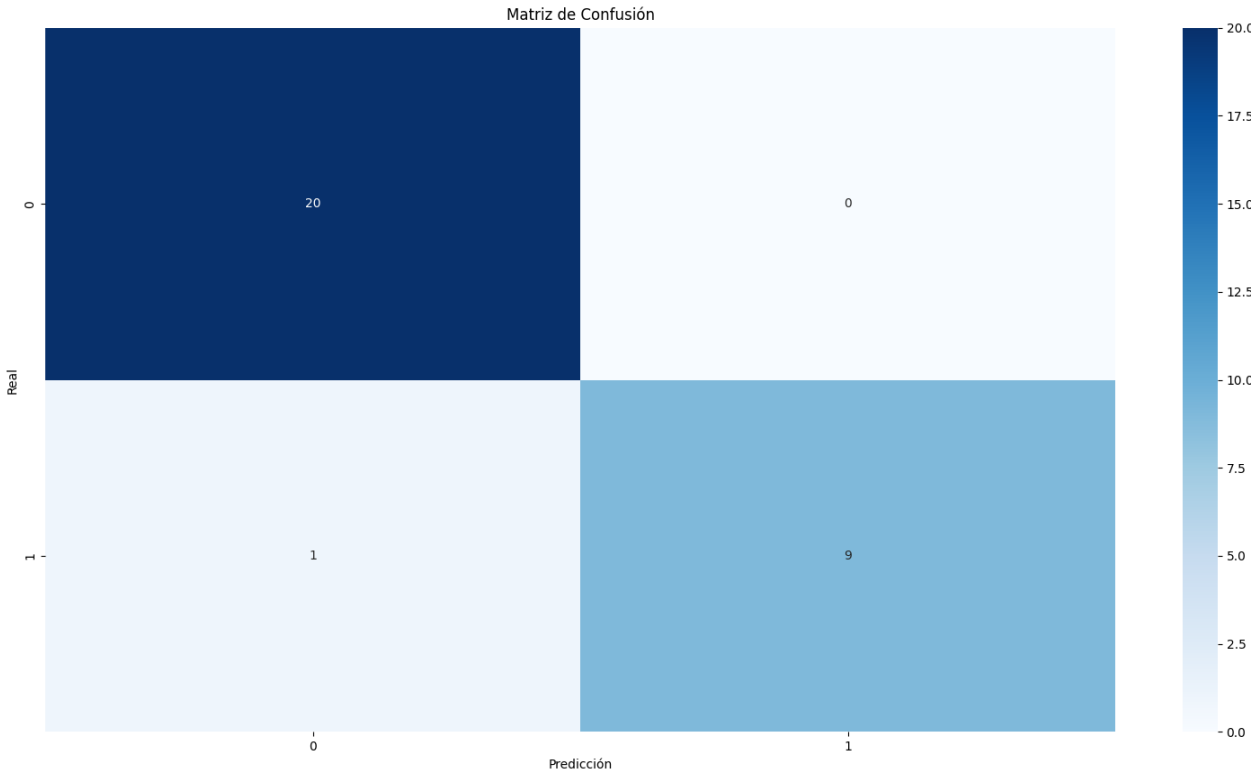

Figura 4.2

En los valores predichos, observamos una clara diferencia, al ser capaz de predecir correctamente muchos más valores, incluso al haber incluido valores fuera del rango. Esto refuerza la idea de generalización y la ausencia de un posible overfit o un exceso de varianza, mostrando que no se ha memorizado los valores del conjunto de entrenamiento inicial, con un rango y cantidad de valores reducidos en comparación al conjunto de evaluación.

En la curva de PR se puede apreciar un claro cambio, llegando a una precisión del 100% junto con un recall del 90%. Podemos asumir que es un mejor clasificador, gracias a lo mostrado en esta curva y al área debajo de la curva ROC, la cual es de 0.95. No clasifica de forma aleatoria e indica correctamente tanto positivos como negativos. Al tener un balance más claro entre estas dos métricas y al haberlas elevado, podemos esperar que la puntuación F1 sea apremiante y muestre una exactitud del modelo mayor.

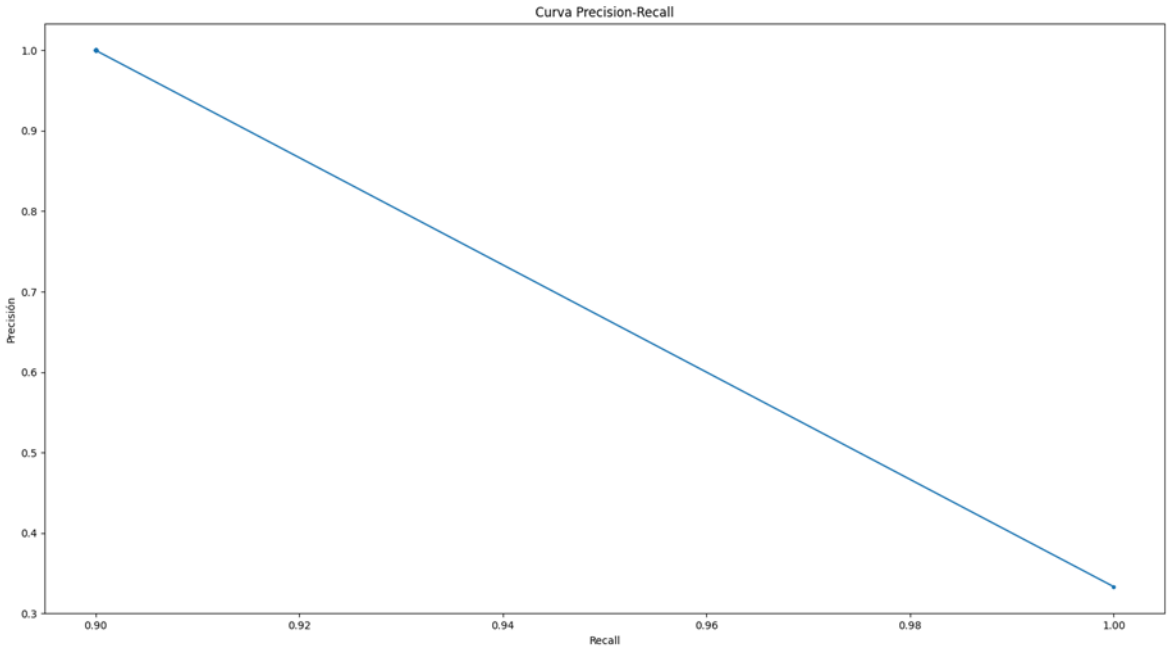

Figura 5.2

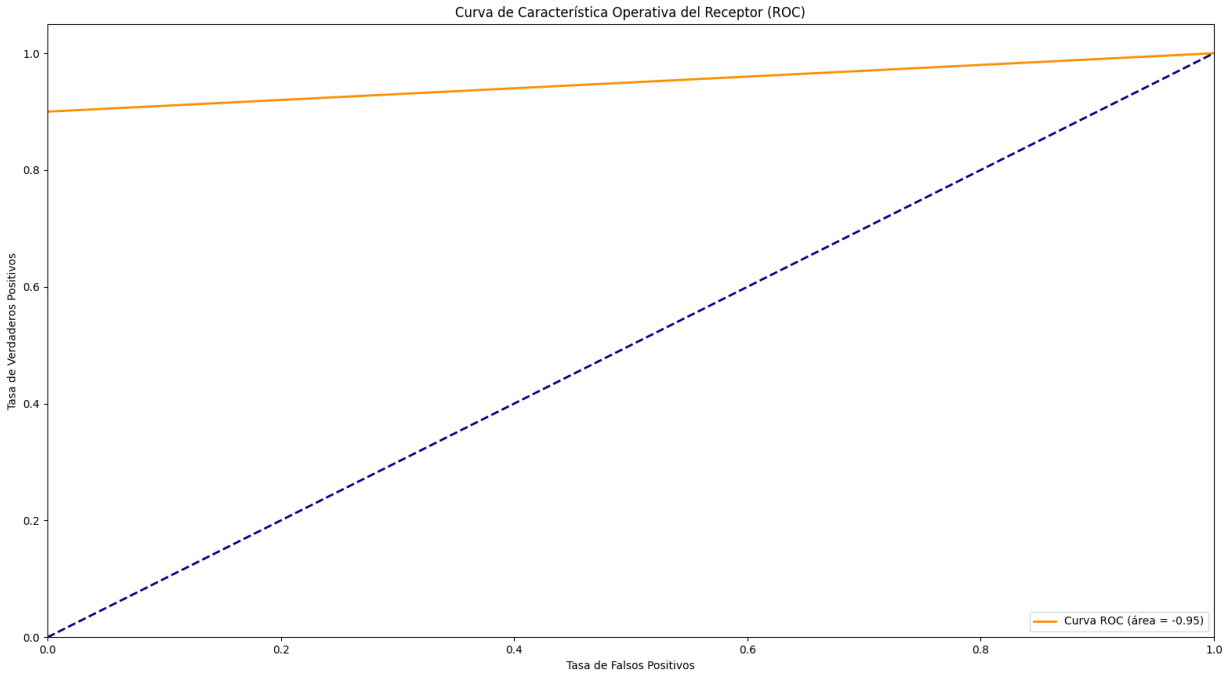

Figura 6.2

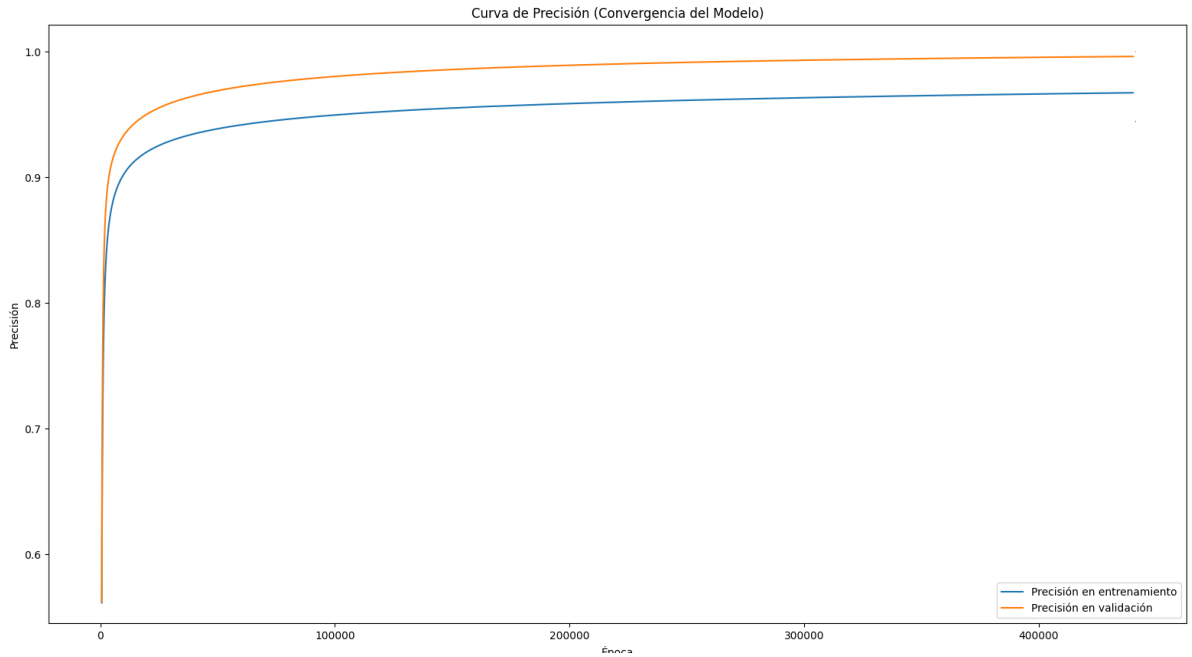

Figura 7.2

```
Entrenamiento finalizado en la época 440902
 Pesos finales ajustados del modelo:
        w0 = -39.538302970182635
        w1 = 57.315193940277595

 Pérdida con data de validación: 0.0100
 Pérdida con data de prueba: 0.0400
 Precisión: 1.0
 Recall: 0.9
 Puntuación F1: 0.9473684210526316
```

Figura 8.2

Finalmente, se puede observar que las métricas han mejorado en sobremanera y que la curva de aprendizaje es muy precisa. Esto nos muestra que el modelo fue diagnosticado correctamente y que se ha mejorado el fit, alcanzando una mejor normalización y disminuyendo el bias y la varianza. Se observa que los pesos finales difieren considerablemente de los vistos anteriormente, mostrando un ajuste en las predicciones del modelo y el error en los pesos del anterior. El F1 score es mucho mayor, aumentando en más del 10%, calificando a este modelo de clasificación como un modelo bien entrenado y mostrando una buena tasa de predicción, tanto para valores positivos como negativos. Además, gracias a las gráficas y su evolución, se puede decir con seguridad de que cualquier tipo de bias o varianza disminuyó significativamente, aumentando de forma casi proporcional la capacidad de generalización y reducción de underfitting y overfitting.

El modelo mejoró gracias a las técnicas utilizadas, seleccionadas a partir del análisis y consideración de las métricas y gráficas que aportó el desempeño del modelo.