PROYECTO FINAL MACHINE LEARNING II

MELISSA MAYÉN QUIROZ

RELEVANCIA BIOLÓGICA

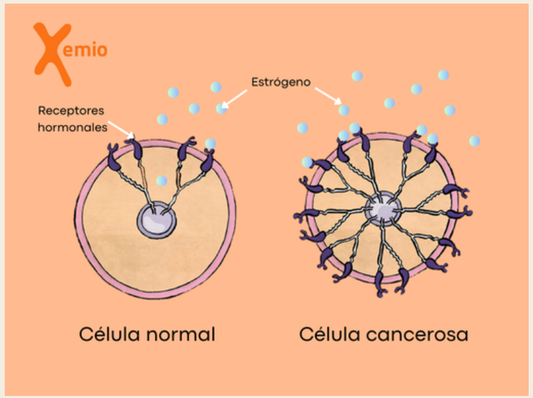
RECEPTORES HORMONALES COMO PREDICTORES DE LA RESPUESTA A LA QUIMIOTERAPIA

TRATAMIENTOS DIFERENTES SEGÚN EL ESTÁTUS DE RECEPTOR DE ESTRÓGENO

FACTOR PREDICTIVO DE RESPUESTA A LA HORMONOTERAPIA

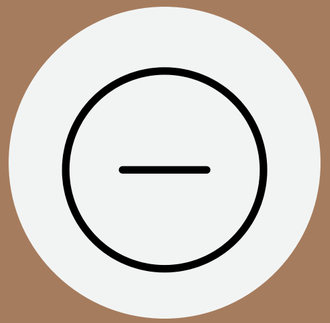
RECEPTORES HORMONALES EN CÁNCER DE MAMA

- **El cáncer de mama con receptor de estrógenos positivos depende del estrógeno para su crecimiento celular.**
- **El estrógeno se une al receptor de estrógeno que a continuación se une con otro receptor de estrógeno, formando un dímero que interactúa con el DNA, promoviendo el crecimiento celular.**



Factor predictivo de respuesta a la hormonoterapia (resistencia a terapia hormonal)

El cáncer de mama (ER+) estrógeno responde bien a la terapia hormonal



El cáncer de mama (ER-) puede requerir enfoques de tratamiento diferentes

SET DE DATOS

Los datos usados corresponden a una cohorte de 271 tejidos de cáncer de mama, 204 receptores de estrógeno positivos (ER+), y 67 receptores de estrógeno negativos (ER−).

Los datos fueron recopilados del biobanco del Departamento de Patología del Hospital Charité, Berlín, Alemania.

En cada muestra se evaluaron 162 metabolitos.

La descarga y e importación de los datos se realizó desde archivos .csv y .txt divididos de la siguiente manera:

Datos de entrenamiento (80%) - 216 instancias
Datos de prueba (20%) - 55 instancias

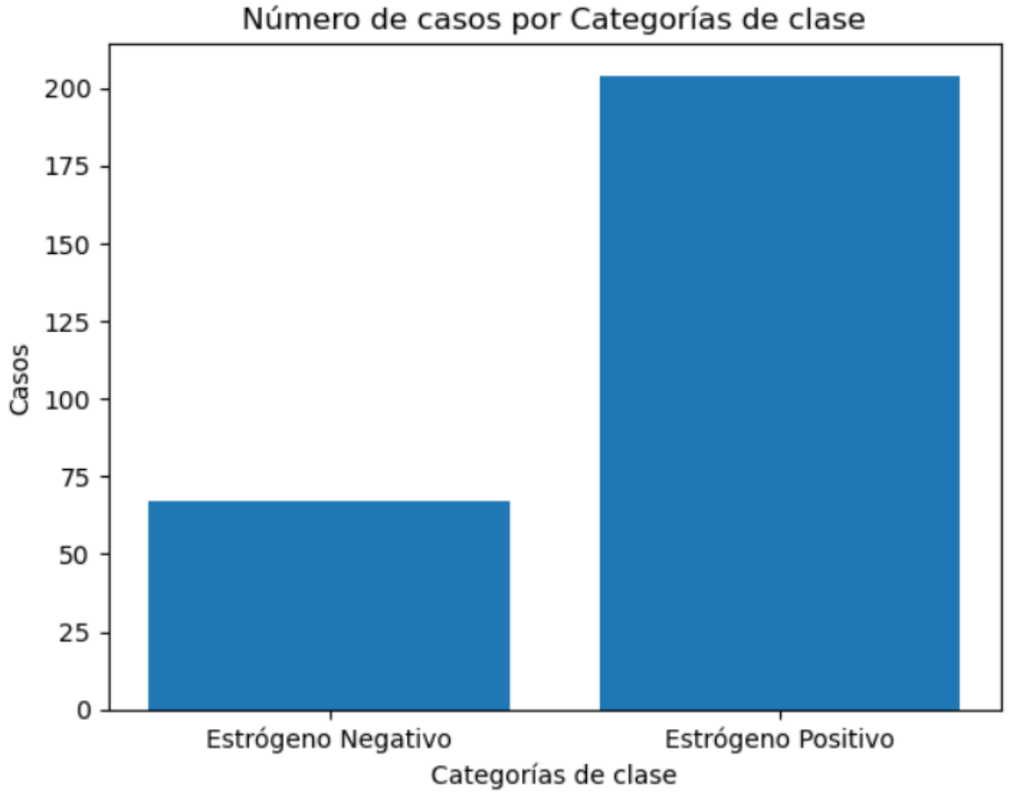
- 2 archivos .csv
  - Datos de entrenamiento - 162 características
  - Datos de evaluación - 162 características
- 2 archivos .txt
  - Etiquetas correspondientes a los datos de entrenamiento (+/-)
  - Etiquetas correspondientes a los datos de evaluación (+/-)

Ejemplos por etiqueta (entrenamiento)

0	55
1	161

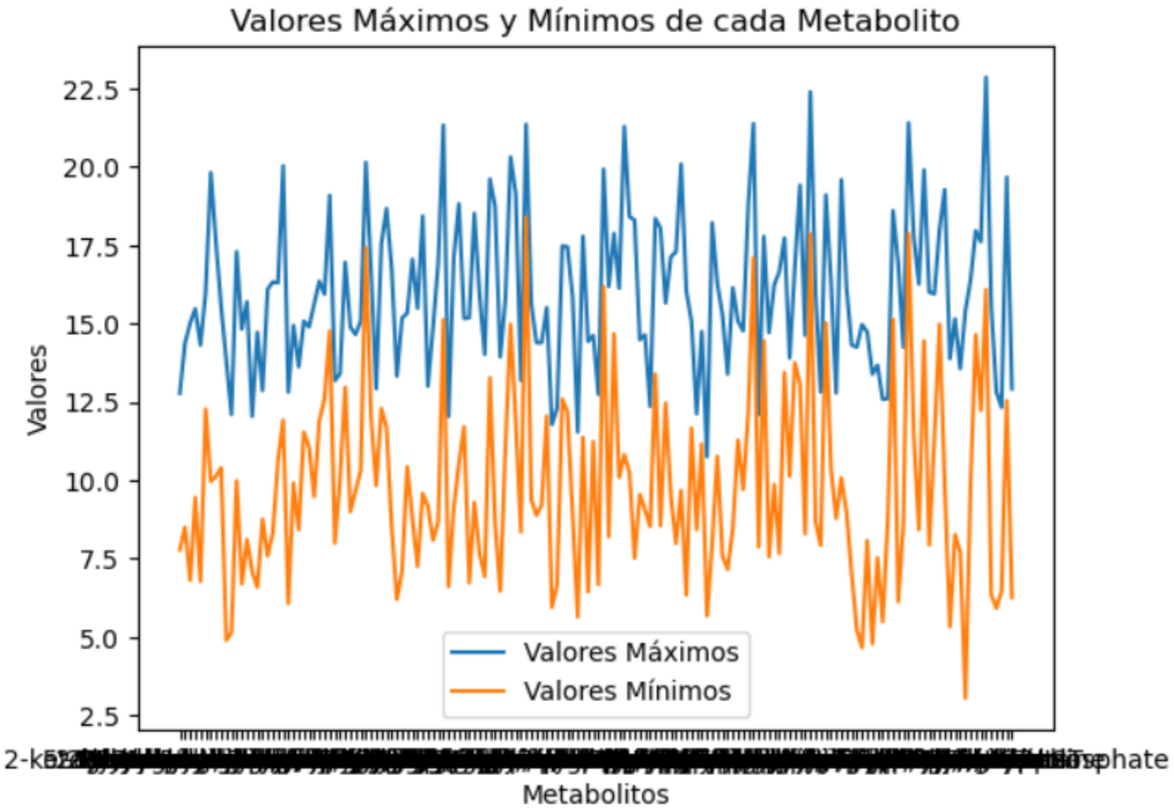
Ejemplos por etiqueta (evaluación)

0	12
1	43



SET DE DATOS

RANGO DE VALORES DE TODOS LOS METABOLITOS EN TODAS LAS MUESTRAS



"1. BASELINE FEEDFORWARD NEURAL NETWORK"

DIVISIÓN DE DATOS

ENTRENAMIENTO (80%) - PRUEBA (20%)

ARQUITECTURA DE LA RED

Red neuronal feedforward (FFNN) con tres capas ocultas y una capa de salida Softmax.

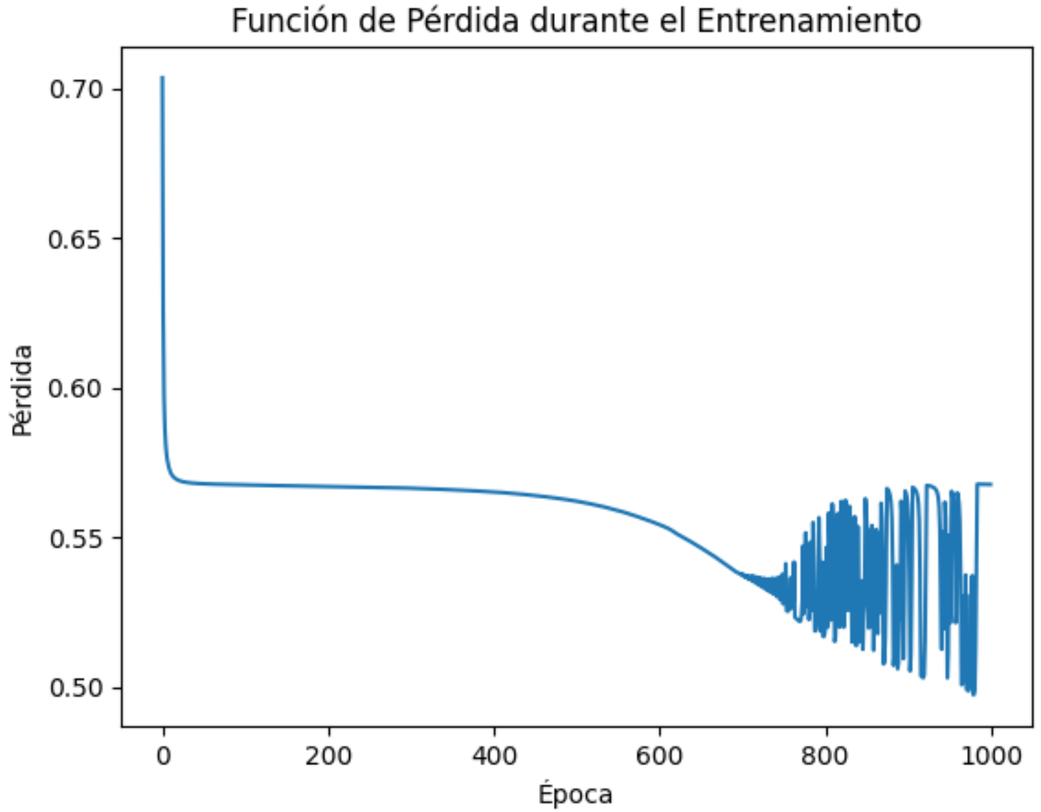
Capa de Entrada: Consta de162 nodos. Cada nodo corresponde a una característica.

Primera Capa Oculta: Realiza una transformación lineal de las 162 características de entrada a 30 nodos ocultos, seguido por la aplicación de la función de activación ReLU.

Segunda Capa Oculta: Realiza otra transformación lineal de los 30 nodos ocultos anteriores a otro conjunto de 30 nodos ocultos, seguido por la aplicación de la función de activación ReLU.

Capa de Salida: Realiza una transformación lineal de los 30 nodos ocultos anteriores a 2 nodos de salida, que corresponden a las clases (1 o 0) . Se aplica la función de activación Softmax que proporciona una distribución de probabilidad sobre las clases.

Se utiliza un método forward para definir el paso hacia adelante de la red.



El valor mínimo de pérdida fue 0.4973188638687134 en la época 979
El valor máximo de pérdida fue 0.7035389542579651 en la época 0

Pérdida final: 0.5676087141036987

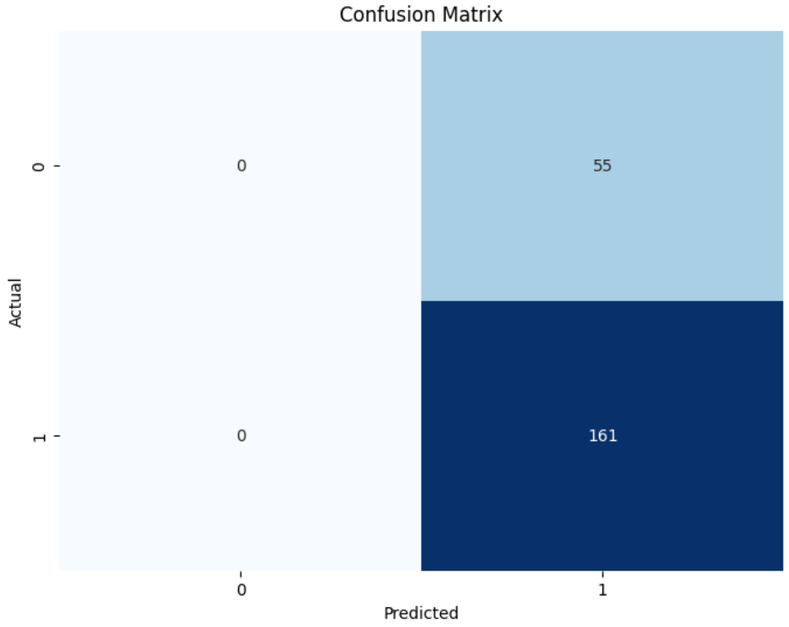
HIPERPARÁMETROS

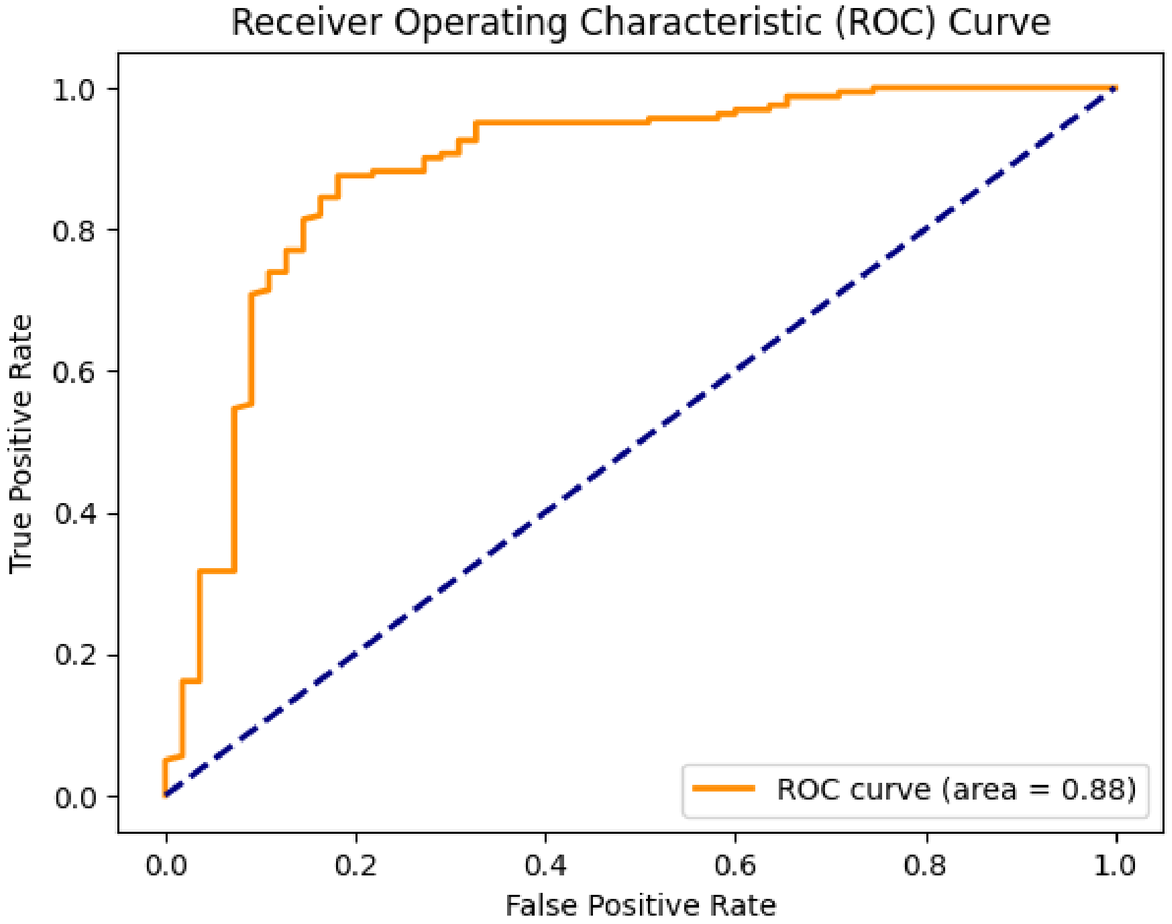
- Número de capas y unidades: 162 nodos de entrada, 3 capas ocultas, 2 nodos de salida.
- Función de activación en las capas ocultas: ReLU
- Función de activación en la capa de salida: Softmax
- Tamaño de lote: 1
- Épocas de entrenamiento: 1000
- Optimizador: Descenso de Gradiente Estocástico (SGD)
- Función de pérdida: Entropía cruzada
- "*Learnning rate*" : 0.01

CONJUNTO DE ENTRENAMIENTO

ACCURACY: 0.7454

AUC: 0.8830

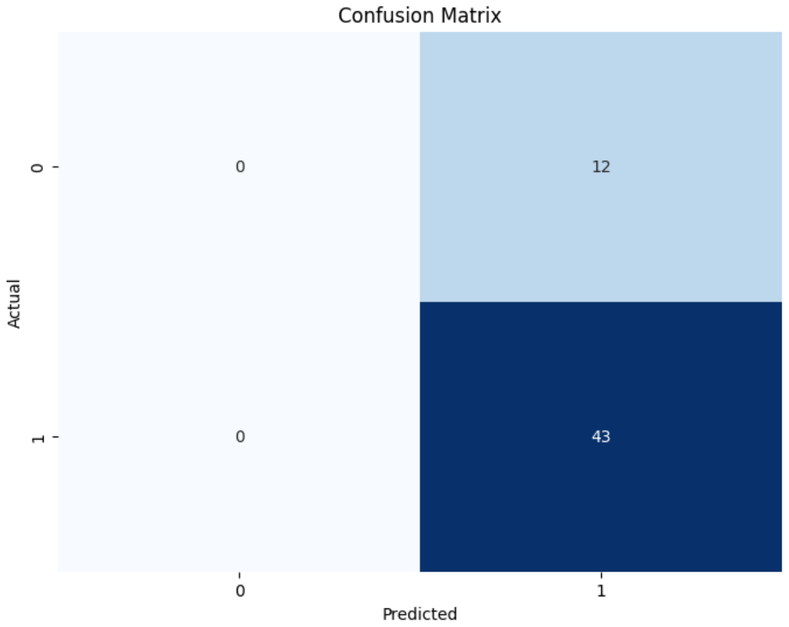


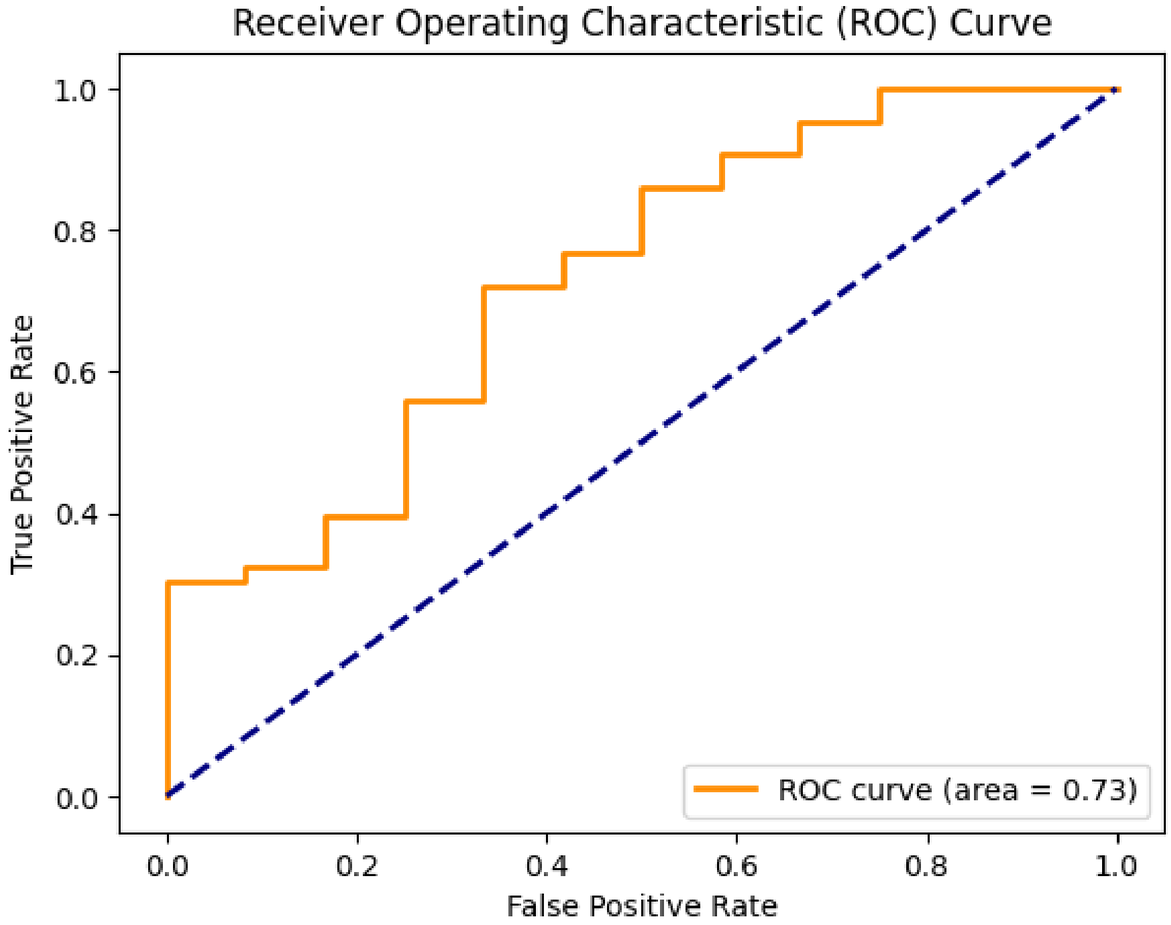


CONJUNTO DE PRUEBA

ACCURACY: 0.7818

AUC: 0.7326





"2. BINARY FEEDFORWARD NEURAL NETWORK"

VALIDACIÓN + "*EARLY STOPING* "

DIVISIÓN DE DATOS

ENTRENAMIENTO (60%) - VALIDACIÓN(20%) - PRUEBA (20%)

ARQUITECTURA DE LA RED

Red neuronal feedforward (FFNN) con tres capas ocultas y una capa de salida Sigmoide.

La primera capa así como las capas ocultas se mantuvieron sin ningun cambió.

Se modificó la capa de Salida:
Se usó un solo nodo de salida para clasificación binaria. Posteriormente se aplica la función de activación Sigmoide que proporciona un número con rango entre 0-1 que se interpreta como la probabilidad de que la entrada pertenezca a la clase positiva

BUCLE DE ENTRENAMIENTO

Validación

Se dividió el set de datos para realizar un paso de validación en el entrenamiento usando "*train_test_split*" perteneciente al módulo "*sklearn.model_selection*" (semilla 42).

Early stoping

Se modificó el bucle de entrenamiento para recibir un parámetro "paciencia" (en este caso de 15). Al completarse cada época, se compara la pérdida obtenida con la mejor hasta el momento, y en caso de no mejorar en número de casos indicados por el parámetro paciencia, se detiene el entrenamiento y se indica la pérdida final así como la época en la que se obtuvo.

HIPERPARÁMETROS

- Número de capas y unidades: 162 nodos de entrada, 3 capas ocultas, 1 nodo de salida.
- Función de activación en las capas ocultas: ReLU
- Función de activación en la capa de salida: Sigmoide
- Tamaño de lote: 1
- Épocas de entrenamiento (máx.): 10,000
  - Implementación de "*Early Stoping*"
  - Paciencia: 40
- Optimizador: Descenso de Gradiente Estocástico (SGD)
- Función de pérdida: Entropía cruzada
- "Learnning rate" : 0.0085

"Early Stoping"

Detención temprana en la época 1323
No se ha observado mejora en 40 épocas consecutivas

Entrenamiento

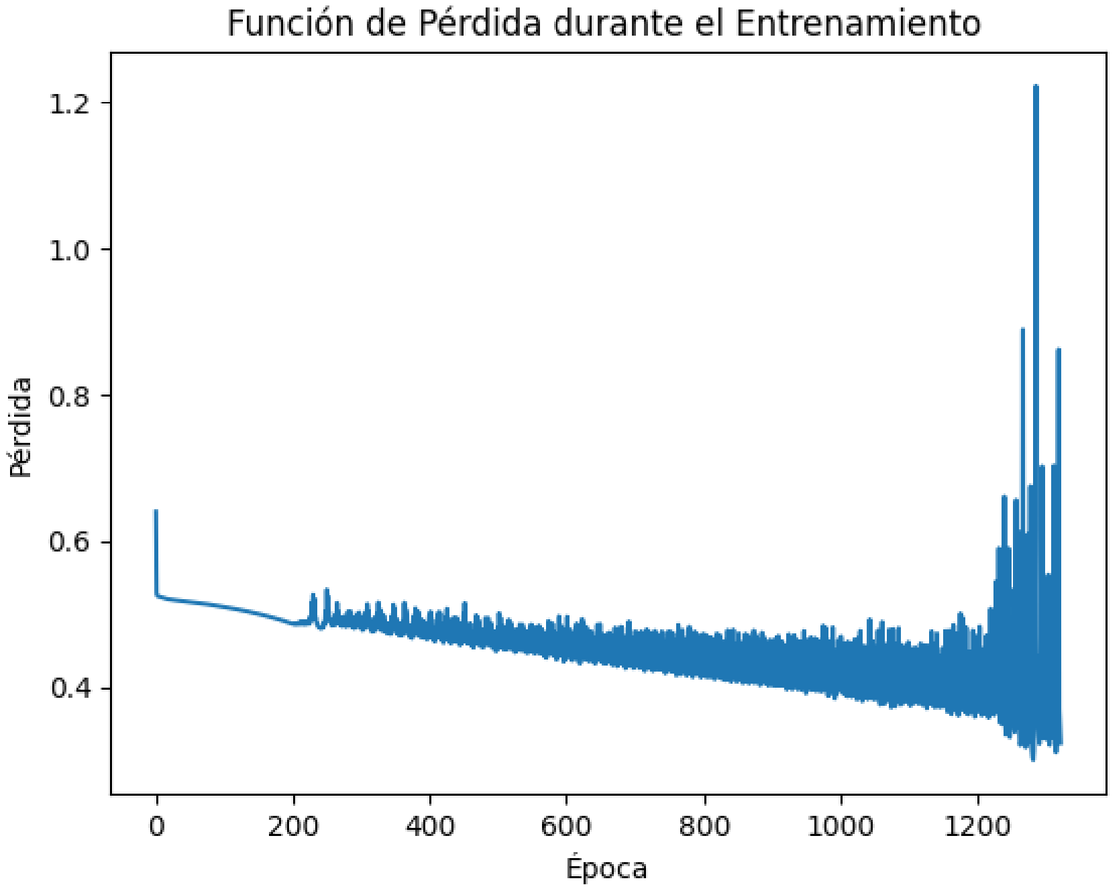
El valor mínimo de pérdida durante el entrenamiento fue 0.3000054359436035 en la época 1282

El valor máximo de pérdida durante el entrenamiento fue 1.2226577997207642 en la época 1286

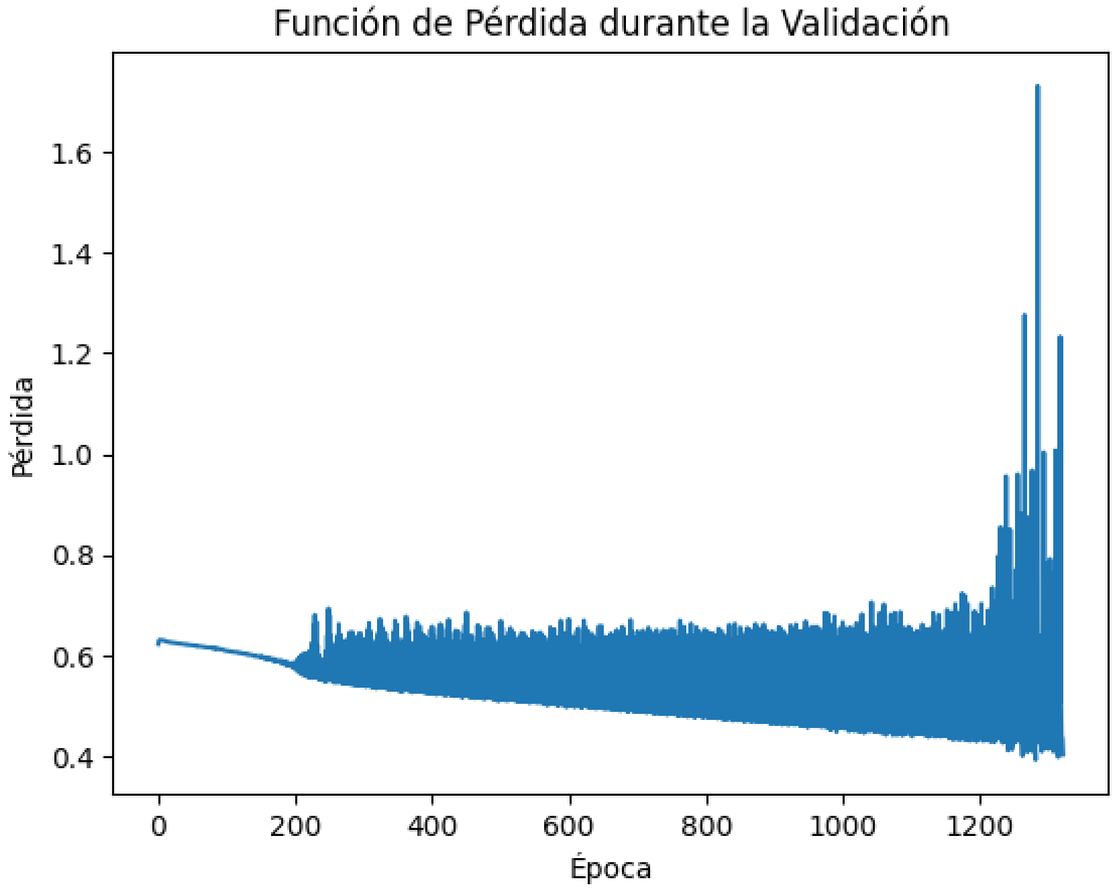
Validación

El valor mínimo de pérdida durante la validación fue 0.3928053677082062 en la época 1282

El valor máximo de pérdida durante la validación fue 1.7320414781570435 en la época 1285



Pérdida final: 0.3231065273284912

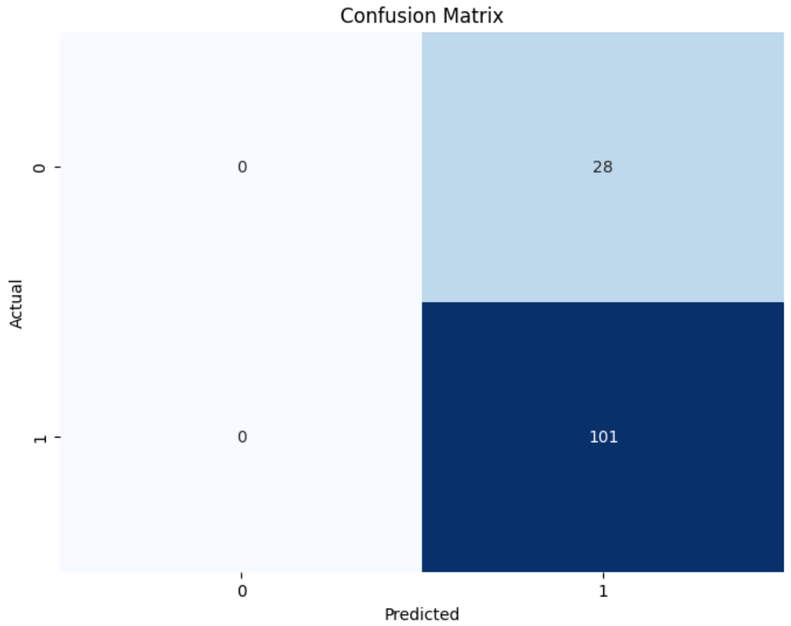


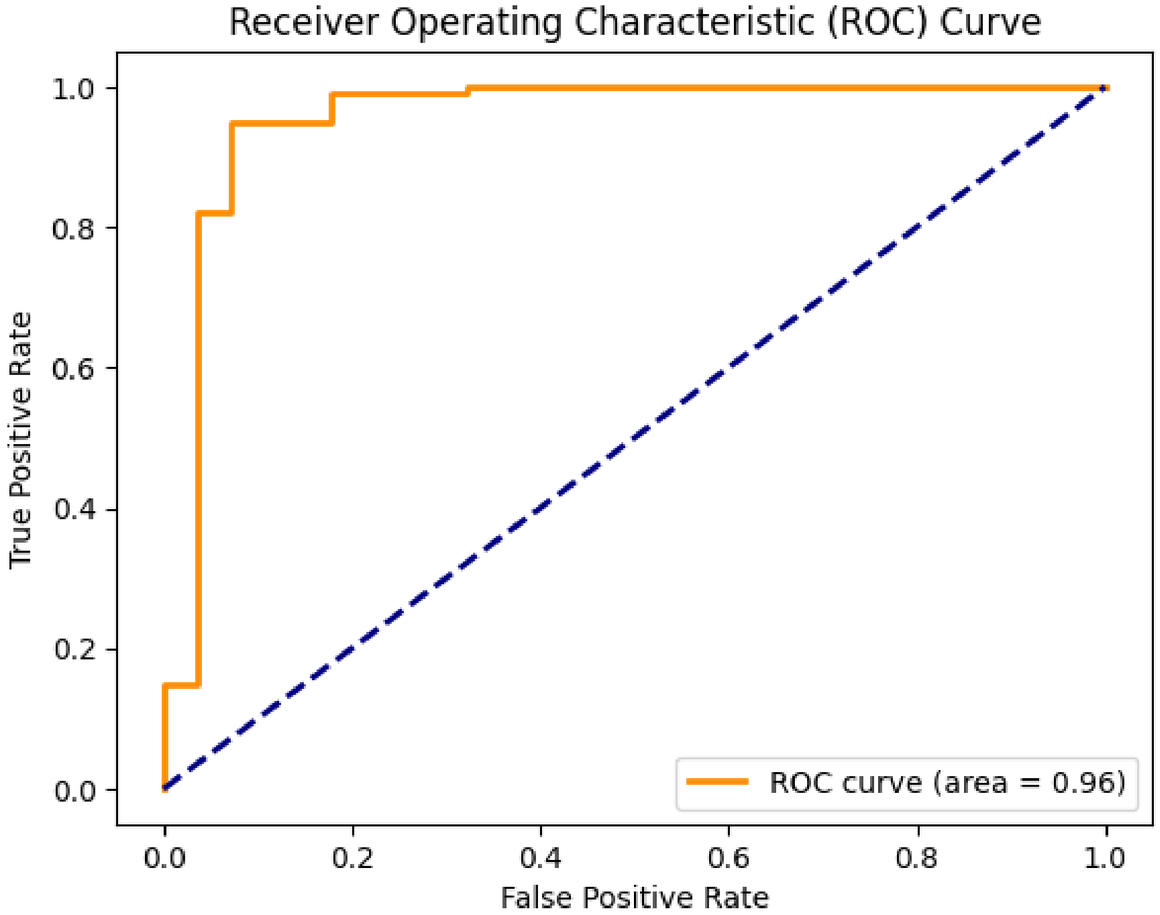
Pérdida final: 0.40331706404685974

CONJUNTO DE ENTRENAMIENTO

ACCURACY: 0.7829

AUC: 0.9565

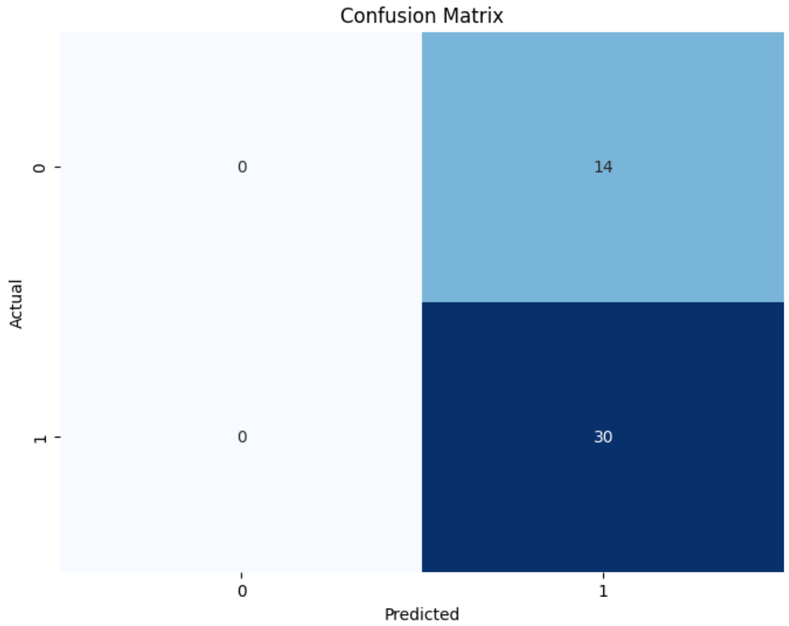


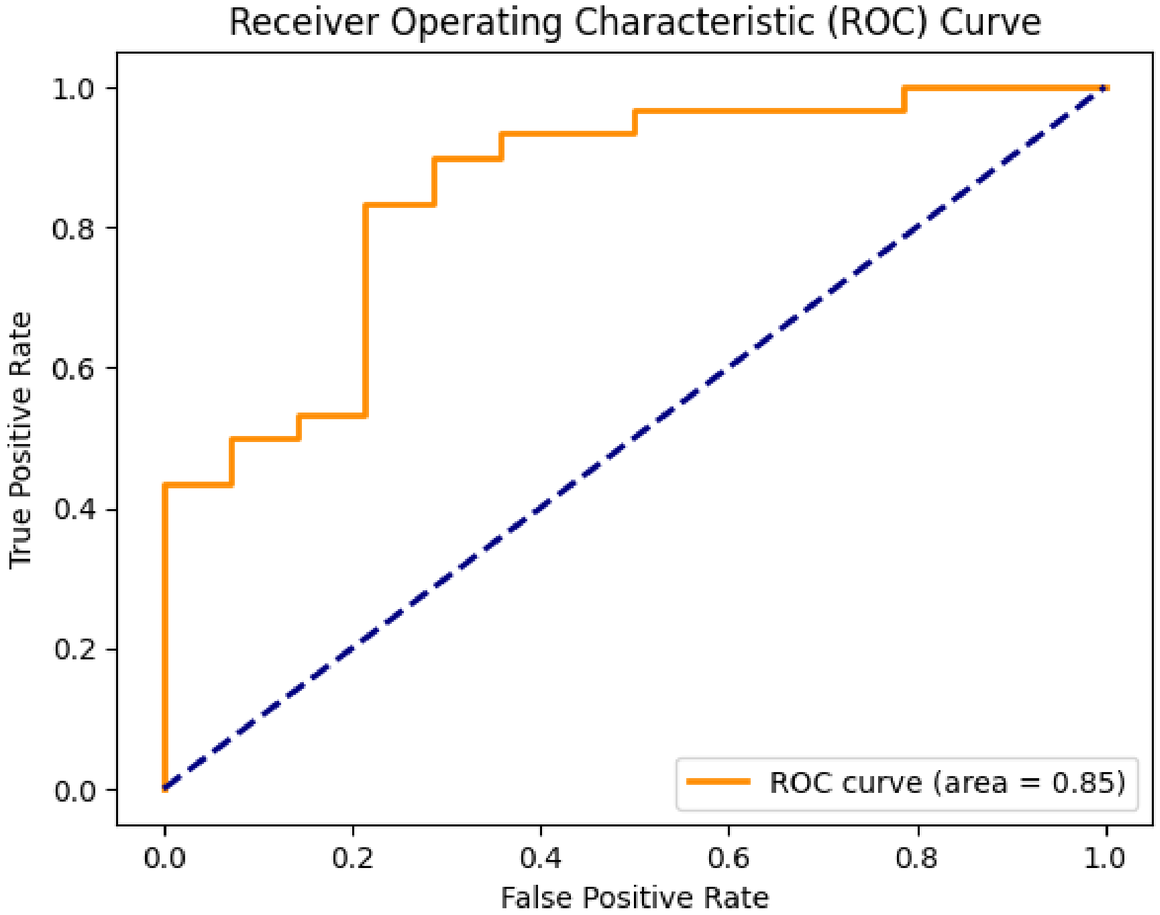


CONJUNTO DE PRUEBA

ACCURACY: 0.6818

AUC: 0.8524





"3. BINARY FEEDFORWARD NEURAL NETWORK"

VALIDACIÓN + "*EARLY STOPING* " + DIVISIÓN POR LOTES

DIVISIÓN DE DATOS

ENTRENAMIENTO (60%) - VALIDACIÓN(20%) - PRUEBA (20%)

ARQUITECTURA DE LA RED

Red neuronal feedforward (FFNN) con tres capas ocultas y una capa de salida Sigmoide.

La arquitectura general de la red fue la misma que la anterior, usada para clasificación binaria.
Cuenta con una capa de entrada de 162 nodos, 3 capas ocultas que usan la función de activación ReLU,
y una capa de salida con 1 nodo y función de activación Sigmoide.

BUCLE DE ENTRENAMIENTO

Validación

Early stoping

División por lotes

Se modificó el bucle de entrenamiento para recibir una lista correspondiente a los distintos tamaños de lote que se quieren probar: 16, 32, 48, 64, 80, 96, 112, 128, 144 y 160.

HIPERPARÁMETROS

- Número de capas y unidades: 162 nodos de entrada, 3 capas ocultas, 1 nodo de salida.
- Función de activación en las capas ocultas: ReLU
- Función de activación en la capa de salida: Sigmoide
- Tamaños de lote: 16, 32, 48, 64, 80, 96, 112, 128, 144 y 160
- Épocas de entrenamiento (máx.): 10,000
    - Implementación de "*Early Stoping*"
    - Paciencia: 48
- Optimizador: Descenso de Gradiente Estocástico (SGD)
- Función de pérdida: Entropía cruzada
- "Learnning rate" : 0.001

"EARLY STOPPING"

- Tamaño de lote 16: Detención temprana en la época 469.
- Tamaño de lote 32: Detención temprana en la época 69.
- Tamaño de lote 48: Detención temprana en la época 1195.
- Tamaño de lote 64: Detención temprana en la época 55.
- Tamaño de lote 80: Detención temprana en la época 563.
- Tamaño de lote 96: Detención temprana en la época 49.
- Tamaño de lote 112: Detención temprana en la época 173.
- Tamaño de lote 128: Detención temprana en la época 69.
- Tamaño de lote 144: Detención temprana en la época 425.
- Tamaño de lote 160: Detención temprana en la época 116.

VALORES MÁXIMOS Y MÍNIMOS DE PÉRDIDA

Tamaño de lote: 16

Valor mínimo de pérdida durante el entrenamiento: 0.04109667241573334 en la época 461

Valor máximo de pérdida durante el entrenamiento: 1.7775651216506958 en la época 291

Valor mínimo de pérdida durante la validación: 0.46110886335372925 en la época 420

Valor máximo de pérdida durante la validación: 0.6729153792063395 en la época 448

Tamaño de lote: 32

Valor mínimo de pérdida durante el entrenamiento: 0.1011974886059761 en la época 32

Valor máximo de pérdida durante el entrenamiento: 2.5962424278259277 en la época 33

Valor mínimo de pérdida durante la validación: 0.42351652681827545 en la época 20

Valor máximo de pérdida durante la validación: 0.6877630949020386 en la época 2

Tamaño de lote: 48

Valor mínimo de pérdida durante el entrenamiento: 0.06122822314500809 en la época 1119

Valor máximo de pérdida durante el entrenamiento: 0.7696738243103027 en la época 988

Valor mínimo de pérdida durante la validación: 0.33685848116874695 en la época 1146

Valor máximo de pérdida durante la validación: 1.5072886943817139 en la época 991

Tamaño de lote: 64

Valor mínimo de pérdida durante el entrenamiento: 0.0028422893956303596 en la época 15

Valor máximo de pérdida durante el entrenamiento: 2.140990734100342 en la época 39

Valor mínimo de pérdida durante la validación: 0.33917924761772156 en la época 6

Valor máximo de pérdida durante la validación: 2.499406099319458 en la época 14

Tamaño de lote: 80

Valor mínimo de pérdida durante el entrenamiento: 0.055716075003147125 en la época 347

Valor máximo de pérdida durante el entrenamiento: 1.51190984249115 en la época 510

Valor mínimo de pérdida durante la validación: 0.33027973771095276 en la época 514

Valor máximo de pérdida durante la validación: 0.8916879892349243 en la época 509

Tamaño de lote: 96

Valor mínimo de pérdida durante el entrenamiento: 0.06625760346651077 en la época 21

Valor máximo de pérdida durante el entrenamiento: 1.989759922027588 en la época 13

Valor mínimo de pérdida durante la validación: 0.3297366797924042 en la época 0

Valor máximo de pérdida durante la validación: 1.1065781116485596 en la época 12

Tamaño de lote: 112

Valor mínimo de pérdida durante el entrenamiento: 0.032109618186950684 en la época 137

Valor máximo de pérdida durante el entrenamiento: 2.2436764240264893 en la época 171

Valor mínimo de pérdida durante la validación: 0.31631872057914734 en la época 124

Valor máximo de pérdida durante la validación: 1.4096449613571167 en la época 111

Tamaño de lote: 128

Valor mínimo de pérdida durante el entrenamiento: 8.476139919366688 e-05 en la época 30

Valor máximo de pérdida durante el entrenamiento: 4.657848358154297 en la época 40

Valor mínimo de pérdida durante la validación: 0.3244923949241638 en la época 20

Valor máximo de pérdida durante la validación: 3.882502317428589 en la época 39

Tamaño de lote: 144

Valor mínimo de pérdida durante el entrenamiento: 0.12111659348011017 en la época 424

Valor máximo de pérdida durante el entrenamiento: 0.28714290261268616 en la época 0

Valor mínimo de pérdida durante la validación: 0.3328634202480316 en la época 376

Valor máximo de pérdida durante la validación: 0.38947057723999023 en la época 0

Tamaño de lote: 160

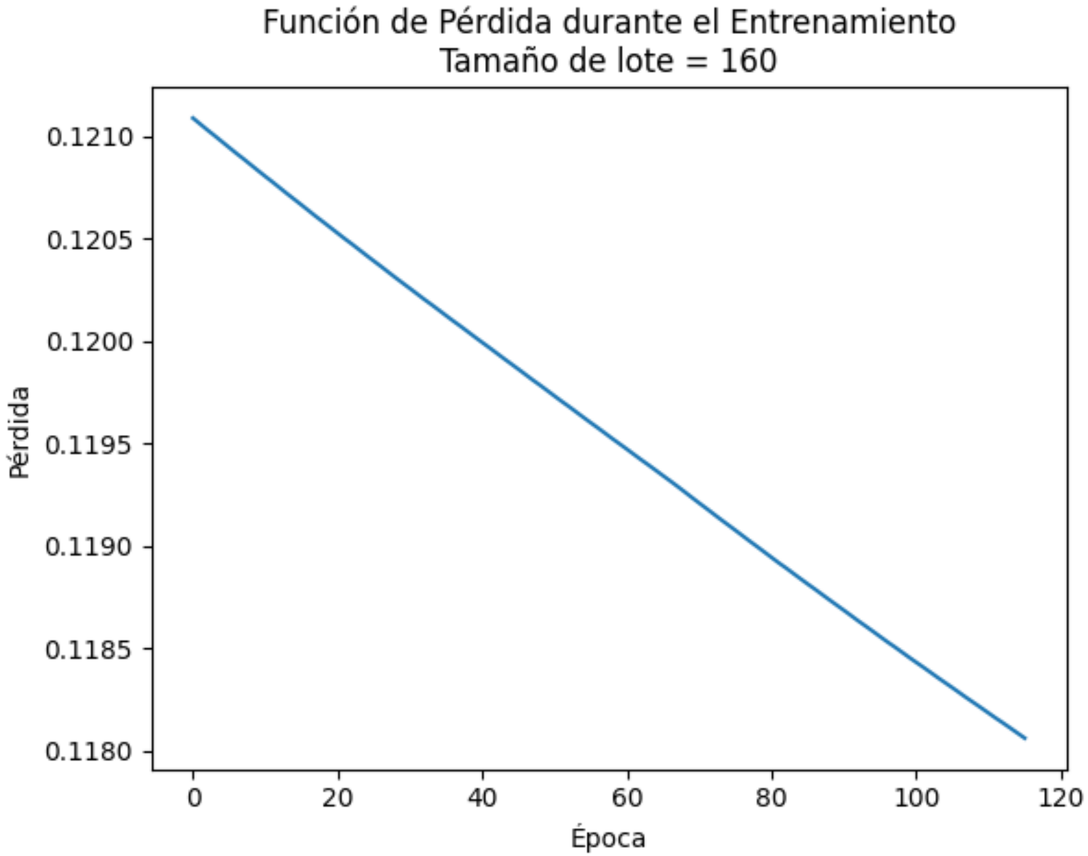
Valor mínimo de pérdida durante el entrenamiento: 0.11806139349937439 en la época 115

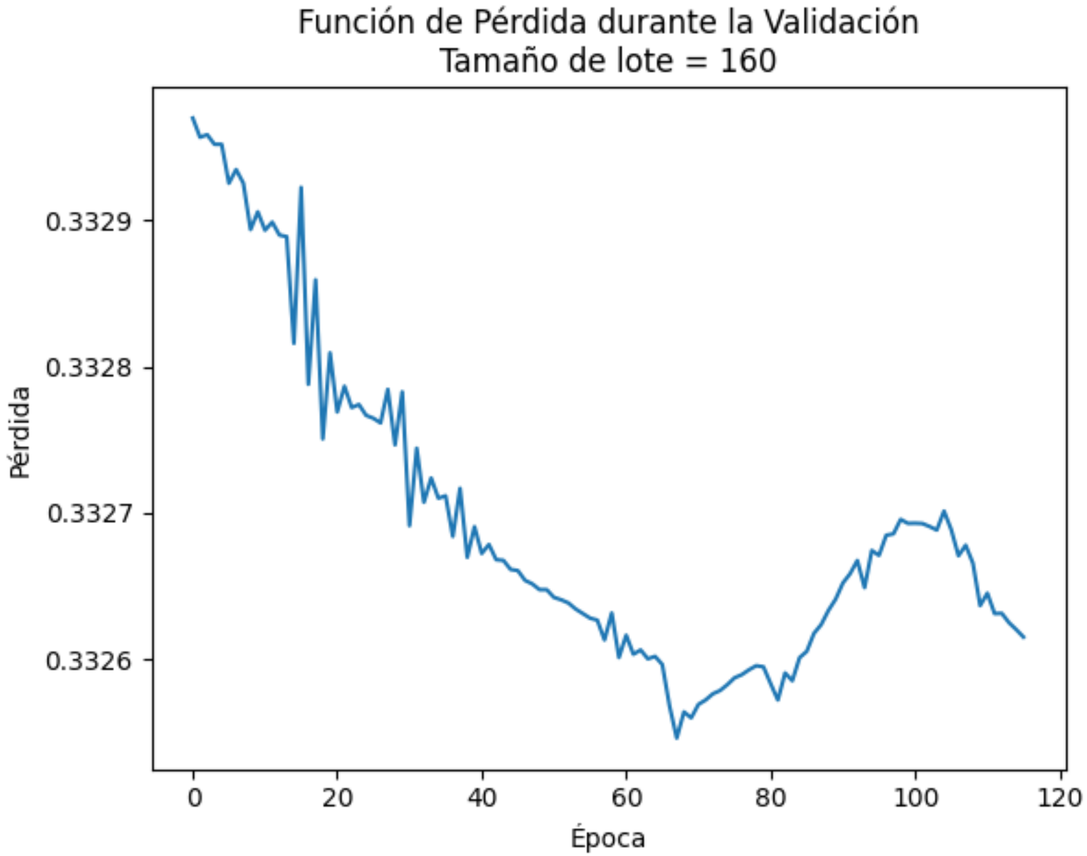
Valor máximo de pérdida durante el entrenamiento: 0.12108813226222992 en la época 0

Valor mínimo de pérdida durante la validación: 0.3325458765029907 en la época 67

Valor máximo de pérdida durante la validación: 0.33296999335289 en la época 0

MEJOR TAMAÑO DE LOTE: 160





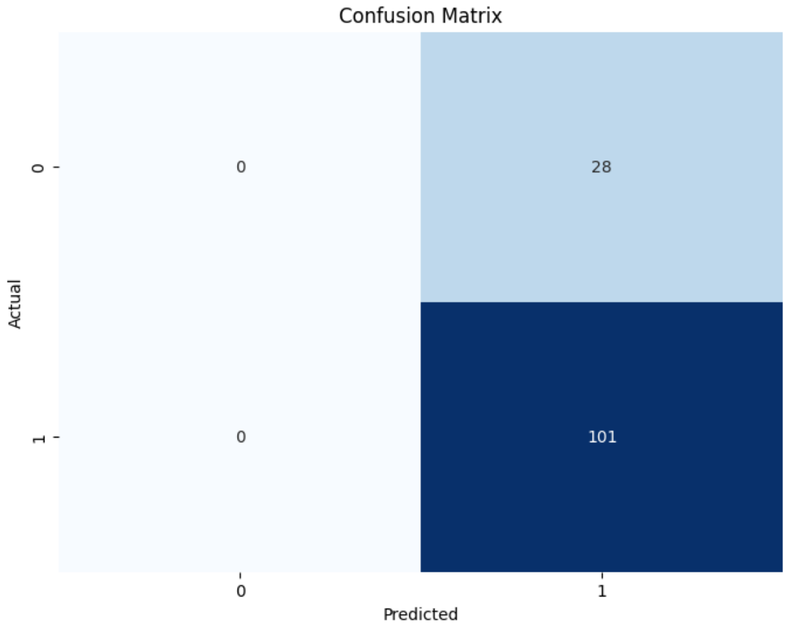
Pérdida final: 0.11806139349937439

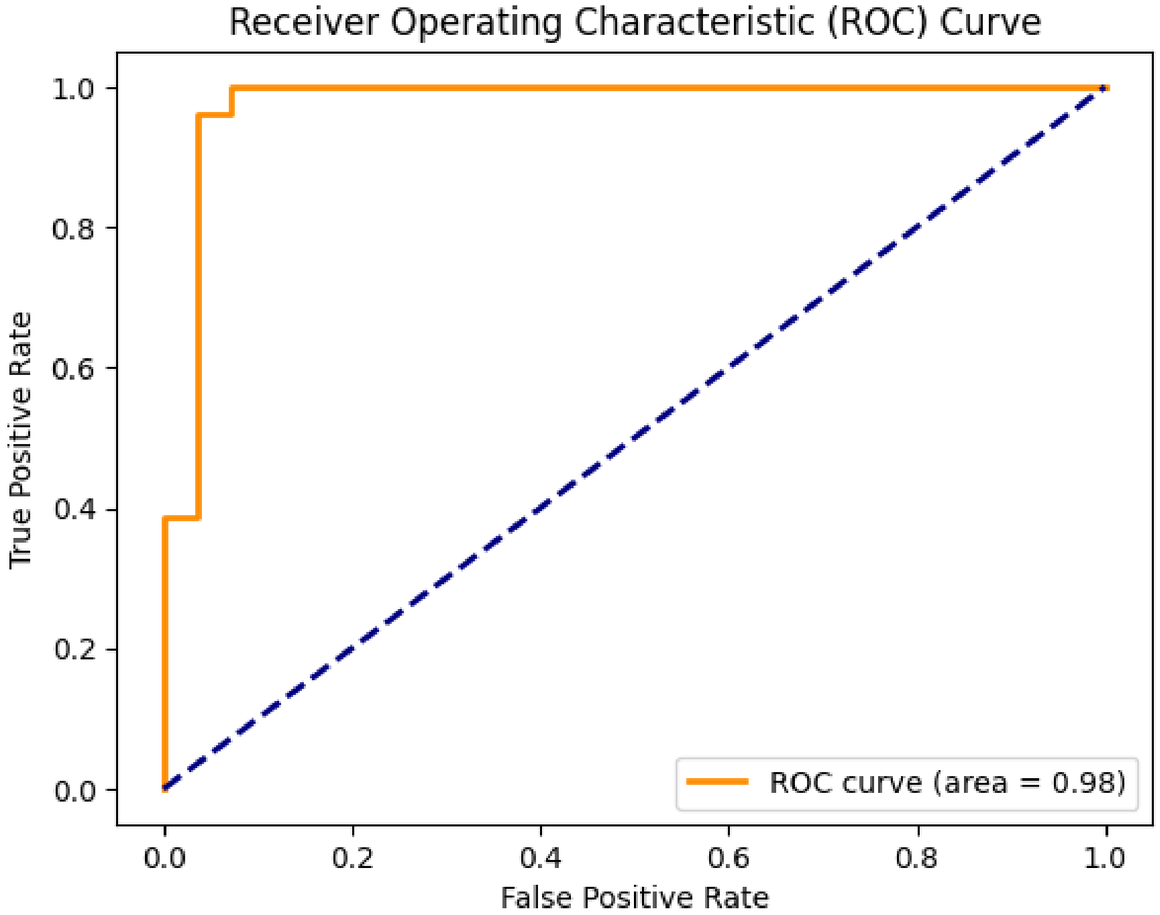
Pérdida final: 0.3326150178909302

CONJUNTO DE ENTRENAMIENTO

ACCURACY: 0.7829

AUC: 0.9767

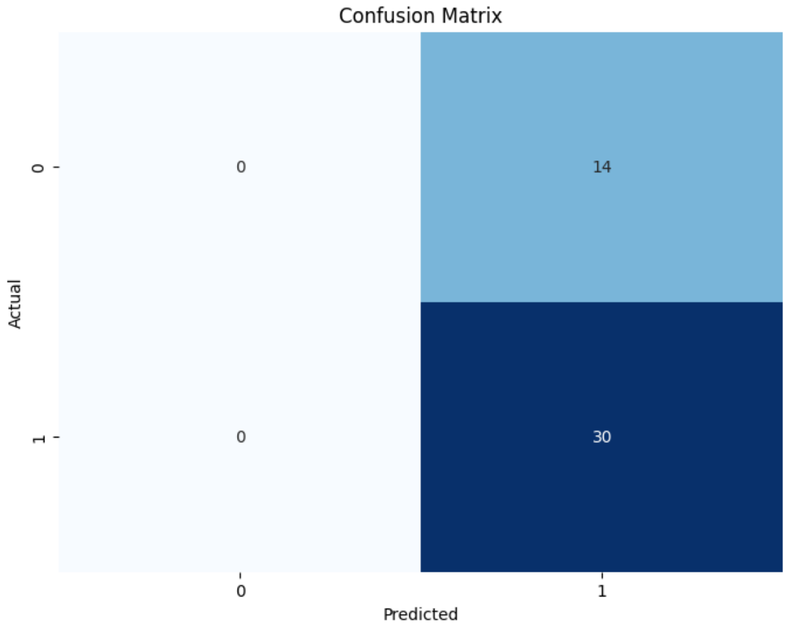


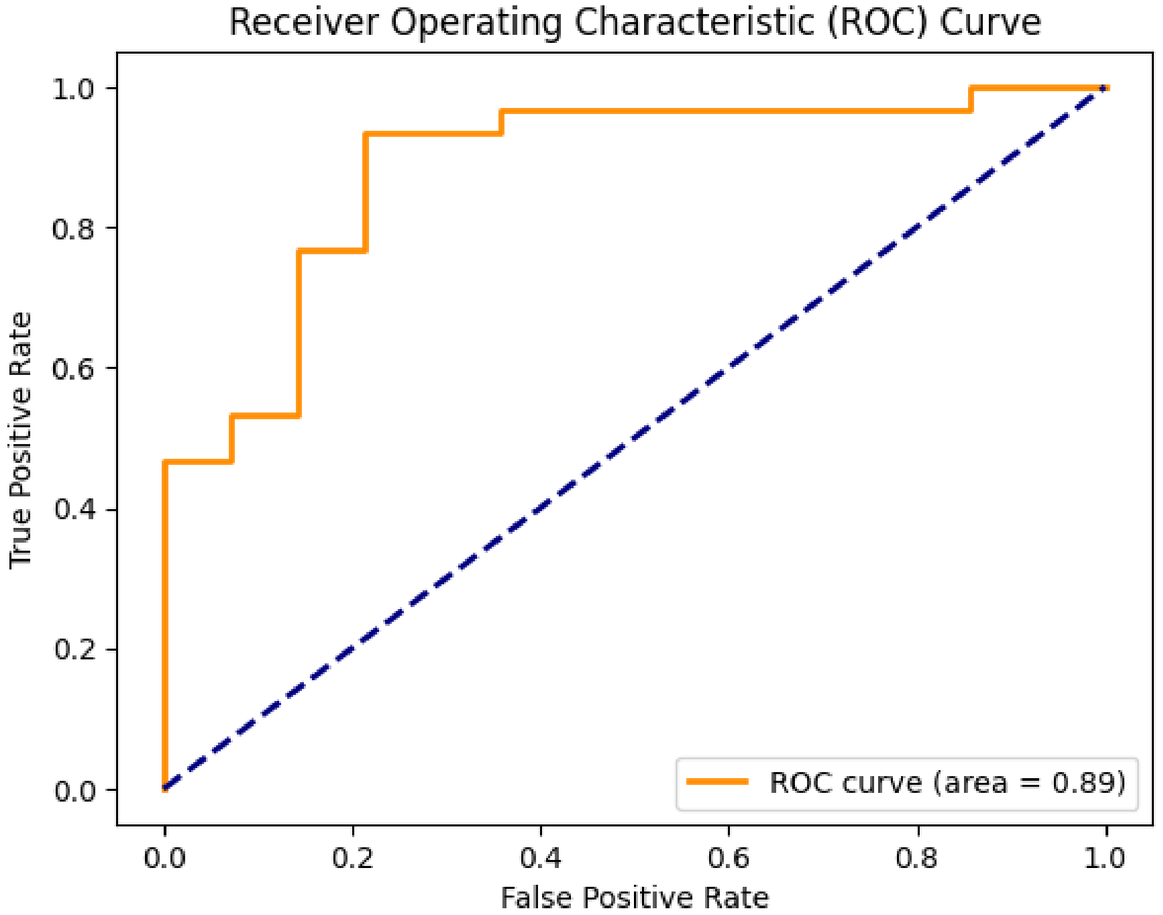


CONJUNTO DE PRUEBA

ACCURACY: 0.6818

AUC: 0.8857





CONCLUSIONES

Desbalance de Clases:

El conjunto de datos presenta un desbalance significativo entre las clases (hacia los casos positivos), lo cual es un desafío común en problemas de clasificación. En el caso de aprendizaje de máquinas tradicional, este problema puede abordarse con “*undersampling*”, “*oversampling*”, o el ajuste de hiperparámetros (Ej. 'class_weight':['balanced']) lo cual requiere de la manipulación de los datos, cosa que se pretende evitar con *deep learning.*

Comportamiento del Modelo Actual:

El modelo actual tiende a predecir todas las instancias como pertenecientes a la clase mayoritaria, lo cual es una limitación significativa.

Expectativas con Deep Learning:

La expectativa al utilizar técnicas de Deep Learning es que el modelo sea capaz de manejar de manera más efectiva el desbalance de clases, aprendiendo patrones complejos automáticamente.

Repositorio de GitHub

https://github.com/Melii99/ProyectoFinal_MLII

GRACIAS

Referencias:

Alakwaa, F. M., Chaudhary, K., & Garmire, L. X. (2018). Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. Journal of proteome research, 17(1), 337-347.

Breastcancer.org. (s.f.). Informe patológico: Estado de los receptores hormonales. Breastcancer.org. https://www.breastcancer.org/es/informe-patologico/estado-receptores-hormonas

Fernández Parra, J., & Bernet Vegué, E. (2002). Receptores hormonales en cáncer de mama. Rev. senol. patol. mamar.(Ed. impr.), 115-122.