

A man in a white lab coat is seen from behind, sitting in a chair and interacting with a large, futuristic medical data interface. His hands are raised, touching the interface. The interface displays various medical data, including a human figure with internal organs, a grid of brain scans, and various charts and graphs. The overall aesthetic is high-tech and futuristic, with blue and white tones. The text "EDA" is overlaid in the center of the image.

EDA

LITERATURA DEL TARGET

BI-RADS 1 (Negativo):

- Clase de anomalidad: `NORM`
- Severidad: Sin valor, ya que son normales.

BI-RADS 2 (Benigno):

- Clase de anomalidad: `CALC`, `CIRC`, `MISC`, `ARCH`, `ASYM`
- Severidad: `B` (Benigno)

BI-RADS 3 (Probablemente Benigno):

- En MIAS, esta categoría no está explícitamente representada, pero casos de `CALC` y `CIRC` con `B` podrían requerir seguimiento en la práctica clínica.

BI-RADS 4 (Sospechoso):

- Clase de anomalidad: `CALC`, `CIRC`, `MISC`, `SPIC`, `ARCH`, `ASYM`
- Severidad: `M` (Maligno), ya que requieren mayor evaluación, como biopsia.

BI-RADS 5 (Altamente sospechoso de malignidad):

- Clase de anomalidad: `SPIC`, `ARCH`
- Severidad: `M` (Maligno)

BI-RADS 6 (Confirmado maligno):

- Casos en los que se ha confirmado la malignidad mediante estudios adicionales (no disponible directamente en MIAS, pero los casos `M` ↓ proximan a esta categoría en términos de sospecha).

CATEGORIZAR CARACTERÍSTICAS

```
current_entry = {  
    "img_id": None,  
    "tipo_tejido": None,  
    "clase_anomalia": None,  
    "severidad": None,  
    "x_coord": None,  
    "y_coord": None,  
    "radio": None,  
    "BI_RADS": None,  
    "dimension": None,  
    "lado": None,  
    "clase_amomalia_2": None,  
    "severidad_2": None,  
    "x2": None,  
    "y2": None,  
    "radio_2": None,  
    "clase_amomalia_3": None,  
    "severidad_3": None,  
    "x3": None,  
    "y3": None,  
    "radio_3": None,  
}
```

PROCESAMIENTO DE DATOS

Limpieza de datos

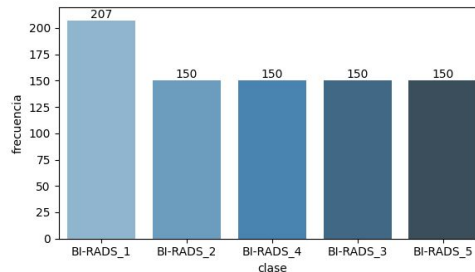
```
Imágenes cargadas iniciales: (117, 224, 224, 3)
Etiquetas cargadas iniciales: (117,)
img_id      0
tipo_tejido  0
clase_anomalia  0
severidad    0
x_coord      8
y_coord      8
radio        3
BI_RADS      0
dimension    0
lado         0
dtype: int64
img_id      0
tipo_tejido  0
clase_anomalia  0
severidad    0
x_coord      0
y_coord      0
radio        0
BI_RADS      0
dimension    0
lado         0
dtype: int64
```



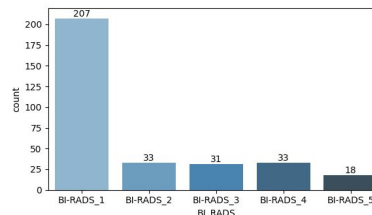
```
img_id tipo_tejido clase_anomalia severidad x_coord y_coord radio BI_RADS dimension lado
0 mdb3111l F NORM MIAS NaN NaN 12.0 BI-RADS_1 1 izquierda
1 mdb2771m G NORM MIAS NaN NaN 11.0 BI-RADS_1 m izquierda
2 mdb078r1 F NORM MIAS NaN NaN 5.0 BI-RADS_1 1 derecho
3 mdb044rs G NORM MIAS NaN NaN 4.0 BI-RADS_1 s derecho
4 mdb112r1 D NORM MIAS NaN NaN 6.0 BI-RADS_1 1 derecho
Imágenes cargadas iniciales: (117, 224, 224, 3)
Etiquetas cargadas iniciales: (117,)
img_id tipo_tejido clase_anomalia severidad x_coord y_coord radio BI_RADS dimension lado
0 mdb3111l F NORM MIAS 0.0 0.0 12.0 BI-RADS_1 1 izquierda
1 mdb2771m G NORM MIAS 0.0 0.0 11.0 BI-RADS_1 m izquierda
2 mdb078r1 F NORM MIAS 0.0 0.0 5.0 BI-RADS_1 1 derecho
3 mdb044rs G NORM MIAS 0.0 0.0 4.0 BI-RADS_1 s derecho
4 mdb112r1 D NORM MIAS 0.0 0.0 6.0 BI-RADS_1 1 derecho
```

Balanceo de datos

Distribución Balanceada de BI-RADS

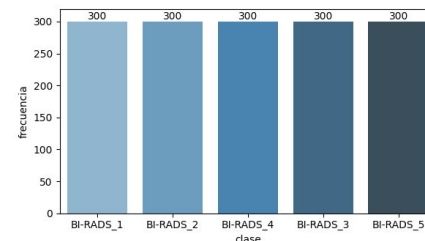


Distribución Inicial de BI-RADS



Aumento de datos

Distribución Aumentada de BI-RADS





Se resalta la importancia del aumento de datos sin usar ni brillo, zoom y corte. Ya que estos 3 factores en dataset médicos afecta el rendimiento del modelo

EXTRACCIÓN DE CARACTERÍSTICAS

The background image is a composite of futuristic medical and technological elements. On the right, a young male doctor in a white lab coat and blue tie smiles while holding a tablet. On the left, a white humanoid robot with glowing blue eyes and joints stands. The background is a dark blue grid with various medical icons: a heart rate line, a caduceus, a heart in a hexagon, a gear, and a brain. A large, glowing blue circular pattern is in the center. In the bottom right, there's a small inset showing a waveform graph with labels 75, 80, 85, 90, 95, 100.

NOMBRE DE LAS CAPAS Y NÚMERO DE CARACTERÍSTICAS EXTRAÍDAS

MODELOS	LAYERS	NÚMERO CARACTERÍSTICAS	UBICACIÓN
DenseNet121	avg_pool	1024	-2
EfficientNetV2B0	avg_pool	1280	-3
Inception_v4	flatten	1536	-4
MobileNetV3Large	flatten	1000	-2
resnext50_32x4d	flatten	2048	-2
VGG16	flatten	25088	-4
tumor_cerebral	flatten	200704/51889	-5

 1º lugar
 2º lugar

NÚMERO DE UMBRALES

0.01

0.02

0.03

0.05

0.06

0.1

MODELOS	AC	F1-score	AC	F1-score	AC	F1-score	AC	F1-score	AC	F1-score	AC	F1-score
DenseNet121	93%	93%	94%	93%	91%	91%	85%	84%	63%	63%	20%	7%
EfficientNetV2B0	47%	47%	48%	48%	49%	48%	45%	45%	41%	41%	41%	41%
Inception_v4	84%	84%	84%	84%	79%	79%	43%	43%	20%	7%	20%	7%
MobileNetV3Large	52%	52%	51%	51%	52%	53%	41%	40%	26%	25%	20%	7%
resnext50_32x4d	90%	90%	89%	89%	85%	85%	60%	59%	32%	30%	20%	7%
VGG16	79%	78%	77%	77%	78%	77%	73%	72%	36%	36%	20%	7%
Tumor_Cerebral												

3-2

1-3

2-1

CONJUNTO DE DATOS	TIPO CONCATENACIÓN	ANTES DE LA SELECCIÓN DE CARACTERÍSTICAS	DESPUÉS DE LA SELECCIÓN DE CARACTERÍSTICAS	# UMBRAL	ACCURACY	F1-SCORE	MODELO
MIAS	Total	31976		0.01	89%	89%	RidgeClas...CV
MIAS	Seleccionado	31976	11507	0.01, 0.02 Y 0.03	90%	89%	LGBMClas... RidgeClas...CV
MIAS	Seleccionado c/u umbral	31976	2047	0.03	90%	89%	NuSVC LGBMClas...
MIAS	dense+incep+resnet+vgg		1414	0.03	92%	91%	NuSVC LGBMClas...
MIAS	dense+incep+resnet	2324	2043	0.01	95%	95%	NuSVC, SVC
MIAS							

Resumen:

- Se seleccionaron 7 modelos y se extrajeron las características de la capa de aplanamiento de cada uno. A continuación, se aplicó la métrica de Información Mutua a las características de cada modelo. Luego, se evaluaron una serie de umbrales positivos [0.01,0.02,0.03,0.05,0.06,0.1] para determinar cuál proporcionaba los mejores resultados al ser probado con las imágenes ya procesadas. Para evaluar el desempeño, se utilizó LazyPredict, que permitió observar cada modelo junto con su métrica de evaluación aproximada. Finalmente, se guardaron las características extraídas de cada modelo.
- Se probaron varias estrategias para optimizar las métricas. La estrategia más efectiva fue seleccionar únicamente la información mutua correspondiente a los umbrales que ofrecieron los mejores resultados.
- Se aplicó un filtro y se seleccionaron solo 3 de los 7 modelos, aquellos que ofrecieron resultados superiores al 80% en f1-score.

=== Resultados para umbral: 0.01 ===						
	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken	
Model						
NuSVC	0.95	0.95	None	0.95	4.93	
SVC	0.94	0.94	None	0.94	4.23	
CalibratedClassifierCV	0.91	0.91	None	0.91	35.03	
KNeighborsClassifier	0.90	0.90	None	0.89	0.54	
PassiveAggressiveClassifier	0.90	0.90	None	0.90	0.92	
=== Resultados para umbral: 0.02 ===						
	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken	
Model						
SVC	0.95	0.95	None	0.95	4.17	
NuSVC	0.94	0.94	None	0.94	6.04	
KNeighborsClassifier	0.91	0.91	None	0.90	0.23	
CalibratedClassifierCV	0.90	0.90	None	0.90	12.95	
PassiveAggressiveClassifier	0.89	0.89	None	0.89	1.44	
=== Resultados para umbral: 0.03 ===						
	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken	
Model						
NuSVC	0.94	0.94	None	0.93	1.11	
SVC	0.93	0.93	None	0.93	1.28	
KNeighborsClassifier	0.90	0.90	None	0.89	0.12	
LGBMClassifier	0.89	0.89	None	0.89	4.75	
ExtraTreesClassifier	0.89	0.89	None	0.88	1.05	
=== Resultados para umbral: 0.05 ===						
	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken	
Model						
KNeighborsClassifier	0.85	0.85	None	0.84	0.07	
ExtraTreesClassifier	0.85	0.85	None	0.85	0.42	



ENSEMBLE LEARNING

BOOSTING

SELECTION

STACKING

FINAL

CONSTRUCCIÓN DEL MODELO

BOOSTING

BOOSTING & STACKING

BAGGING BOOSTING

PREDICTOR MODEL

STASTING

STACKING

DIVISIÓN DE DATOS

DATOS TOTALES

Valor: BI-RADS_1, Conteo: 500
Valor: BI-RADS_2, Conteo: 500
Valor: BI-RADS_3, Conteo: 500
Valor: BI-RADS_4, Conteo: 500
Valor: BI-RADS_5, Conteo: 500

====NUEVA NOMENCLATURA DE BI-RADS-

[0 1 2 3 4]
BI-RADS_1 = 0
BI-RADS_2 = 1
BI-RADS_3 = 2
BI-RADS_4 = 3
BI-RADS_5 = 4

DATOS ENTRENAMIENTO

Valor: 0, Conteo: 350
Valor: 1, Conteo: 350
Valor: 2, Conteo: 350
Valor: 3, Conteo: 350
Valor: 4, Conteo: 350

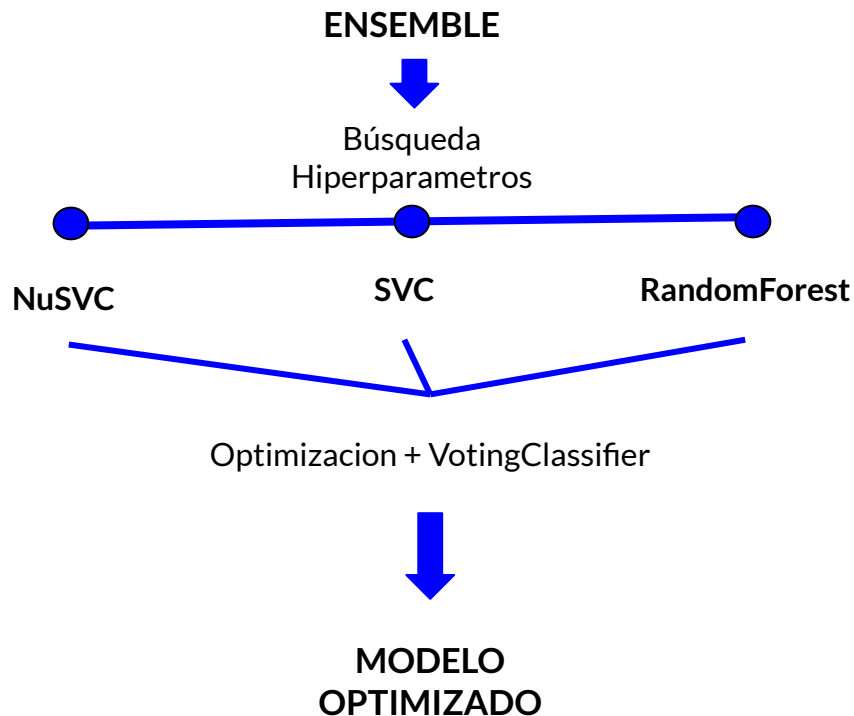
DATOS TEMPO

Valor: 0, Conteo: 150
Valor: 1, Conteo: 150
Valor: 2, Conteo: 150
Valor: 3, Conteo: 150
Valor: 4, Conteo: 150

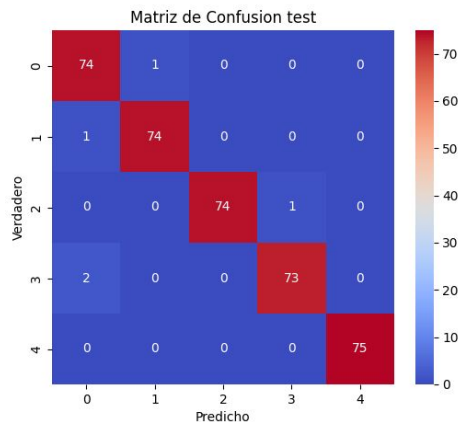
DATOS VALIDACION Y TEST

Valor: 0, Conteo: 75
Valor: 1, Conteo: 75
Valor: 2, Conteo: 75
Valor: 3, Conteo: 75
Valor: 4, Conteo: 75

CONSTRUCCIÓN MODELO



MÉTRICAS DE EVALUACIÓN

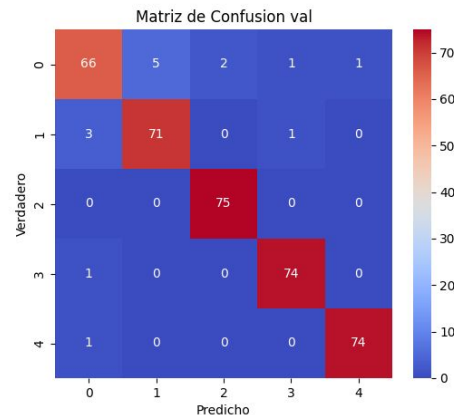


METRICAS DE EVALUACION MODELO:

Reporte de clasificacion test

	precision	recall	f1-score	support
0	0.96	0.99	0.97	75
1	0.99	0.99	0.99	75
2	1.00	0.99	0.99	75
3	0.99	0.97	0.98	75
4	1.00	1.00	1.00	75
accuracy			0.99	375
macro avg	0.99	0.99	0.99	375
weighted avg	0.99	0.99	0.99	375

TEST

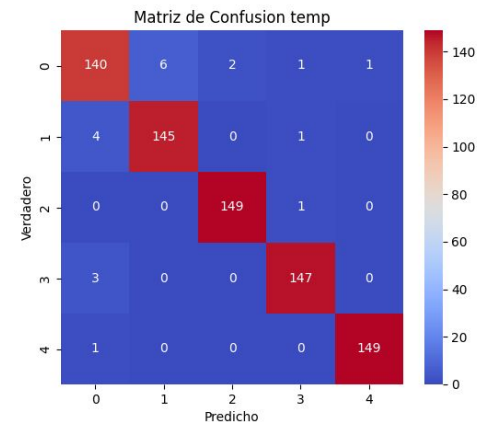


METRICAS DE EVALUACION MODELO:

Reporte de clasificacion val

	precision	recall	f1-score	support
0	0.93	0.88	0.90	75
1	0.93	0.95	0.94	75
2	0.97	1.00	0.99	75
3	0.97	0.99	0.98	75
4	0.99	0.99	0.99	75
accuracy			0.96	375
macro avg	0.96	0.96	0.96	375
weighted avg	0.96	0.96	0.96	375

VAL



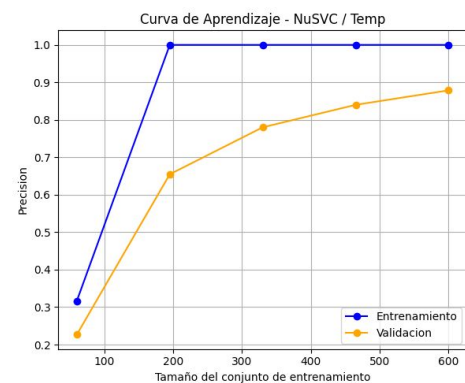
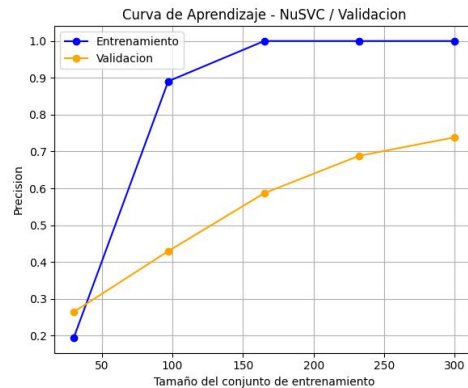
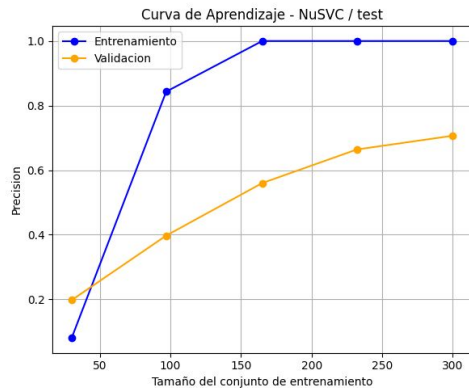
METRICAS DE EVALUACION MODELO:

Reporte de clasificacion temp

	precision	recall	f1-score	support
0	0.95	0.93	0.94	150
1	0.96	0.97	0.96	150
2	0.99	0.99	0.99	150
3	0.98	0.98	0.98	150
4	0.99	0.99	0.99	150
accuracy			0.97	750
macro avg	0.97	0.97	0.97	750
weighted avg	0.97	0.97	0.97	750

TEMP

EVALUACIÓN DEL OVERFITTING



Tamaño conjunto de datos de entrenamiento: [30 97 165 232 300]

CURVA DE APRENDIZAJE DE ENTRENAMIENTO:

[0.4	0.	0.	0.]
[1.	0.86597938	0.78350515	0.78350515	0.78350515]
[1.	1.	1.	1.	1.]
[1.	1.	1.	1.	1.]
[1.	1.	1.	1.	1.]
Promedios:	[0.08	0.84329897	1.	1.]

CURVA DE APRENDIZAJE DE VALIDACION:

[0.2	0.2	0.2	0.18666667	0.2]
[0.38666667	0.44	0.4	0.46666667	0.29333333]	
[0.54666667	0.53333333	0.49333333	0.70666667	0.52]
[0.65333333	0.65333333	0.53333333	0.77333333	0.70666667]	
[0.72	0.72	0.57333333	0.81333333	0.70666667]	
Promedios:	[0.19733333	0.39733333	0.56	0.664	0.70666667]

TEST

Tamaño conjunto de datos de entrenamiento: [30 97 165 232 300]

CURVA DE APRENDIZAJE DE ENTRENAMIENTO:

[0.96666667	0.	0.	0.]
[0.94845361	1.	0.83505155	0.83505155	0.83505155]
[1.	1.	1.	1.	1.]
[1.	1.	1.	1.	1.]
[1.	1.	1.	1.	1.]
Promedios:	[0.19333333	0.89072165	1.	1.]

CURVA DE APRENDIZAJE DE VALIDACION:

[0.34666667	0.32	0.17333333	0.2	0.28]
[0.46666667	0.46666667	0.37333333	0.37333333	0.46666667]	
[0.58666667	0.57333333	0.54666667	0.62666667	0.6]
[0.65333333	0.72	0.64	0.73333333	0.69333333]	
[0.70666667	0.82666667	0.66666667	0.77333333	0.72]
Promedios:	[0.264	0.42933333	0.58666667	0.688	0.73866667]

VAL

Tamaño conjunto de datos de entrenamiento: [60 195 330 465 600]

CURVA DE APRENDIZAJE DE ENTRENAMIENTO:

[0.91666667	0.16666667	0.16666667	0.16666667	0.16666667]
[1.	1.	1.	1.	1.]
[1.	1.	1.	1.	1.]
[1.	1.	1.	1.	1.]
[1.	1.	1.	1.	1.]
Promedios:	[0.31666667	1.	1.	1.]

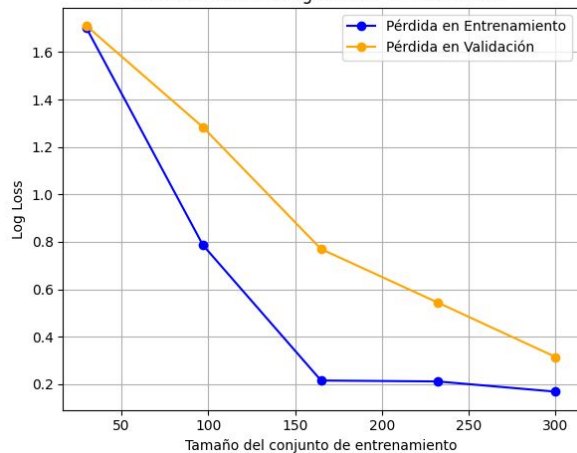
CURVA DE APRENDIZAJE DE VALIDACION:

[0.27333333	0.20666667	0.20666667	0.21333333	0.23333333]	
[0.62	0.58	0.67333333	0.68	0.72]	
[0.78	0.72	0.78666667	0.78	0.83333333]	
[0.82	0.78	0.86	0.83333333	0.90666667]	
[0.86666667	0.83333333	0.88	0.89333333	0.92]	
Promedios:	[0.22666667	0.65466667	0.78	0.84	0.87866667]

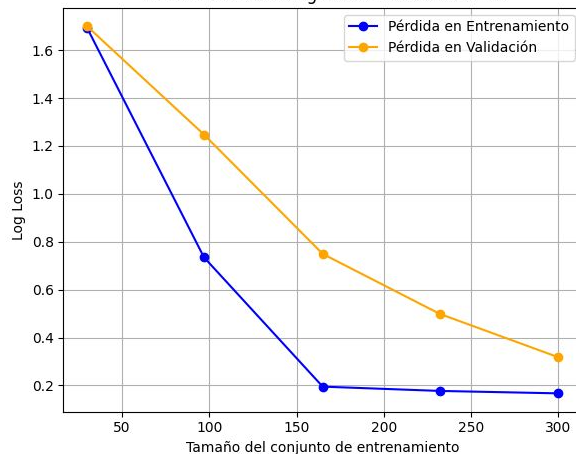
TEMP

EVALUACIÓN DEL ERROR LOGARÍTMICO

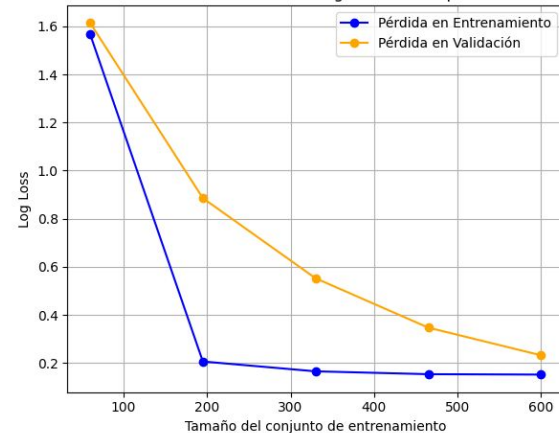
Curva de Pérdida Logarítmica Validacion test



Curva de Pérdida Logarítmica Validacion Val



Curva de Pérdida Logarítmica Temp



Tamaño conjunto de datos de entrenamiento: [30 97 165 232 300] ←

Tamaño conjunto de datos de entrenamiento: [30 97 165 232 300] ←

Tamaño conjunto de datos de entrenamiento: [60 195 330 465 600] ←

CURVA DE APRENDIZAJE DE ENTRENAMIENTO:

[1.6937394712719174, 0.7349629130745837, 0.19491028173295122, 0.17690438558960855, 0.16679297479556549]

CURVA DE APRENDIZAJE DE ENTRENAMIENTO:

[1.7007321622119422, 0.7863270811202374, 0.21553165595675444, 0.21166070763650027, 0.16851265389399936]

CURVA DE APRENDIZAJE DE ENTRENAMIENTO:

[1.5604059325240838, 0.20590552412022836, 0.16495350178573123, 0.1530601777574224, 0.15140471874265052]

CURVA DE APRENDIZAJE DE VALIDACION:

[1.699750536432232, 1.2474683794206964, 0.7492084940550225, 0.4988076855286228, 0.3174925930997478]

CURVA DE APRENDIZAJE DE VALIDACION:

[1.710038231131813, 1.2834641740022643, 0.7697479440166103, 0.5446496183544562, 0.3146820730867339]

CURVA DE APRENDIZAJE DE VALIDACION:

[1.6150094167990909, 0.8857390051585084, 0.5521531927860149, 0.34677078507226117, 0.23159366477404533]

TEST

VAL

TEMP

CONCLUSIONES

Para evaluar el sobreajuste en cada curva de aprendizaje, se utilizó tres fragmentos de datos. Se observó que, a medida que aumenta la cantidad de datos utilizados, los porcentajes de validación, prueba (test) y temp mejoran, acercándose cada vez más al rendimiento en entrenamiento.

Por último, al comparar los resultados finales, noto que el dataset con mayor cantidad de datos es el más cercano a los resultados del entrenamiento. Esto indica que mi modelo generaliza mejor cuando se utiliza un mayor número de datos.

FLUJO DEL MODELO

