




Banknote Authentication Dataset

Alejandro Sánchez Monzón




Descripción del conjunto de datos

El conjunto nace de la recopilación de datos obtenidos de imágenes tomadas a ejemplares de billetes reales y falsos. Para ello, se utilizó una cámara industrial que es usualmente utilizada para la inspección de impresiones.

Las imágenes finales tienen un tamaño de 400x400 píxeles.

Para la extracción de las características de las imágenes se utilizó la transformada ondícula.





Referencias del conjunto de datos

El conjunto de datos es obra de Volker Lohweg, un ingeniero alemán y profesor de procesamiento de imágenes e información en la Universidad Técnica de Ostwestfalen-Lippe en Lemgo, Alemania. Se especializa en el campo de la automatización inteligente y ha sido director del Instituto de Tecnología de la Información Industrial de la TH OWL desde 2017.

El conjunto de datos ha sido obtenido de UC Irvine, una plataforma con cientos de conjuntos publicados, del cual he podido obtener información sobre el conjunto, y enlaces relativos a artículos de interés donde es mencionado.



```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 1372 entries, 0 to 1371
```

```
Data columns (total 5 columns):
```

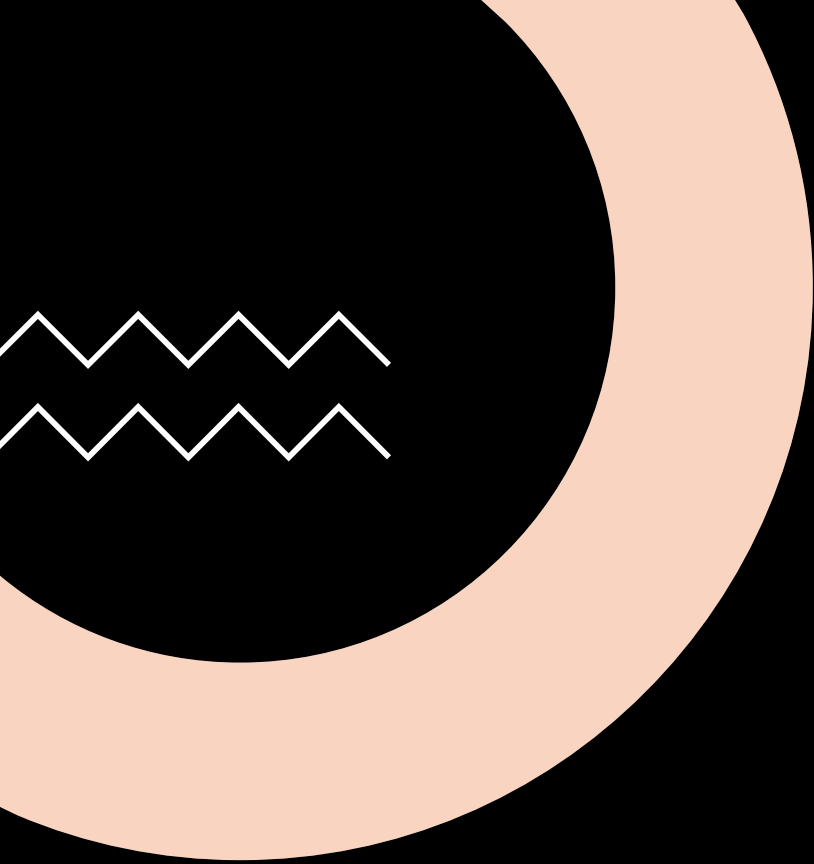
#	Column	Non-Null Count	Dtype
0	variance	1372 non-null	float64
1	skewness	1372 non-null	float64
2	curtosis	1372 non-null	float64
3	entropy	1372 non-null	float64
4	class	1372 non-null	int64

```
dtypes: float64(4), int64(1)
```

```
memory usage: 53.7 KB
```

Claves del conjunto de datos

El conjunto de datos se presenta como un reto perfecto para realizar un problema de clasificación, donde desarrollar un modelo capaz de identificar qué billetes son reales y cuáles falsos.



La Transformada Ondícula es una herramienta matemática que descompone una señal o un conjunto de datos en diferentes componentes de frecuencia, y luego estudia cada uno de estos componentes con una resolución que corresponde a su escala.

Transformada ondícula





Variables del conjunto de datos



Variance (varianza): La transformada ondícula convierte la imagen en una serie de funciones ondículas. La varianza proporciona una medida de cuánto varían los valores de las funciones en relación con su media.





Variables del conjunto de datos



Skewness (asimetría): La asimetría es una medida de la falta de simetría en la distribución de las funciones ondículas. La asimetría en este contexto se refiere a cómo esta distribución se desvía de esta simetría.





Variables del conjunto de datos



Curtois (curtosis): Es una medida de si los datos son pesados en las colas o ligeros en las colas en comparación con una distribución normal. Se refiere a cómo estas 'colas' se distribuyen en las funciones ondículas.





Variables del conjunto de datos



Entropy (entropía): La entropía es una medida estadística de la aleatoriedad que se puede utilizar para caracterizar la textura de la imagen de entrada. En otras palabras, es una medida de la cantidad de información o ‘desorden’ en la imagen.






Variables del conjunto de datos



Class (autenticidad): Esta es la variable objetivo que estamos tratando de predecir. Indica si un billete es auténtico o no.

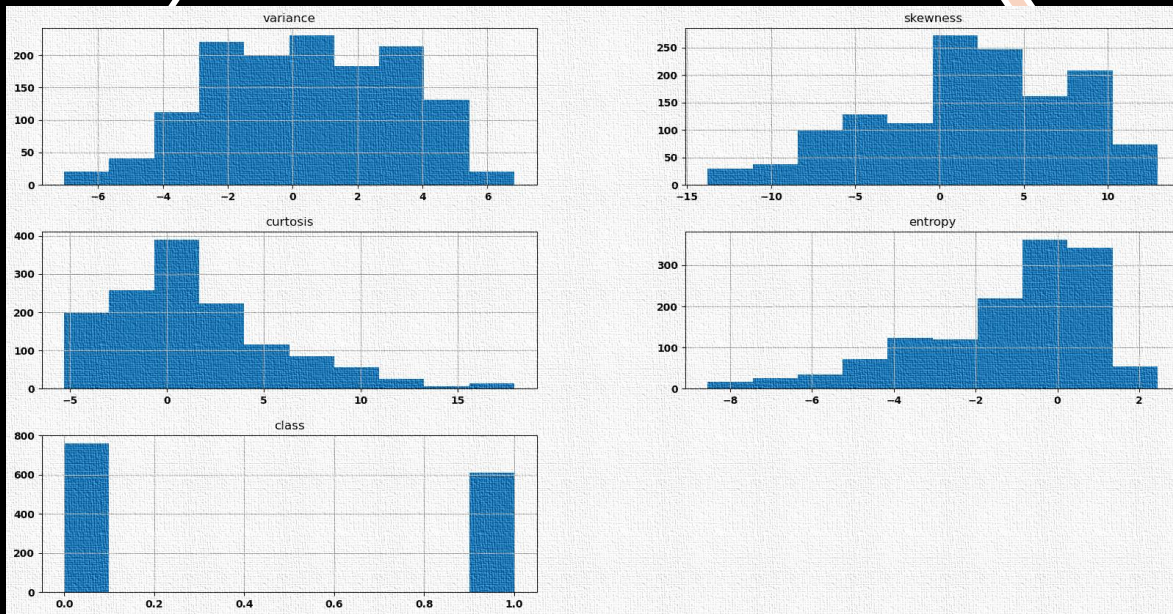




Solución del problema

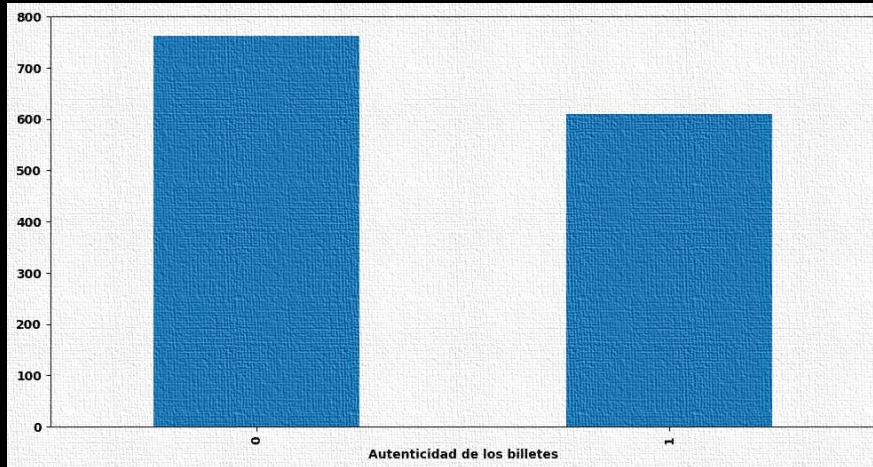
Detectar si un billete es verdadero o falso.

1. Análisis general del conjunto de datos



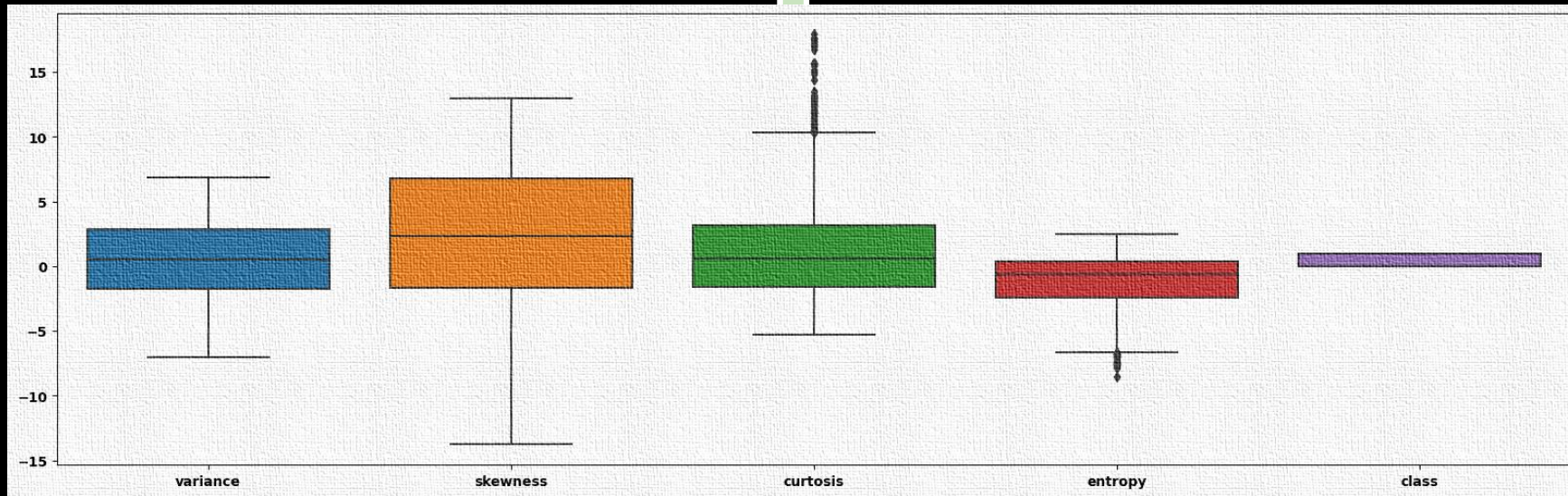
	variance	skewness	kurtosis	entropy	class
0	3.62160	8.66610	-2.8073	-0.44699	0
1	4.54590	8.16740	-2.4586	-1.46210	0
2	3.86600	-2.63830	1.9242	0.10645	0
3	3.45660	9.52280	-4.0112	-3.59440	0
4	0.32924	-4.45520	4.5718	-0.98880	0
...
1367	0.40614	1.34920	-1.4501	-0.55949	1
1368	-1.38870	-4.87730	6.4774	0.34179	1
1369	-3.75030	-13.45860	17.5932	-2.77710	1
1370	-3.56370	-8.38270	12.3930	-1.28230	1
1371	-2.54190	-0.65804	2.6842	1.19520	1
1372 rows × 5 columns					

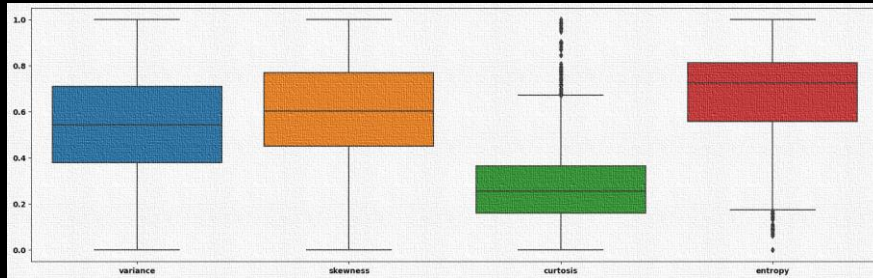
1. Análisis general del conjunto de datos



2. Ajustes generales

¿Es necesario normalizar los datos?





2. Ajustes generales

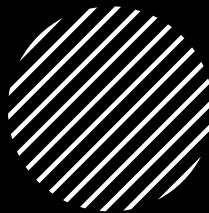
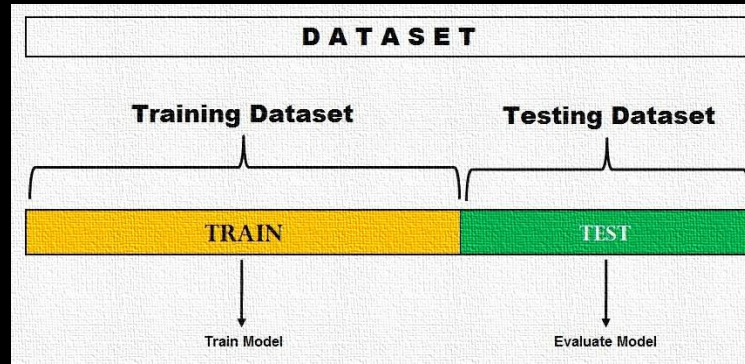
Utilizamos **MinMaxScaler** para llegar a la escala de **0-1**.

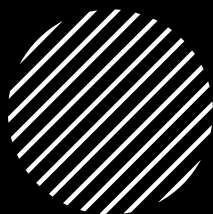
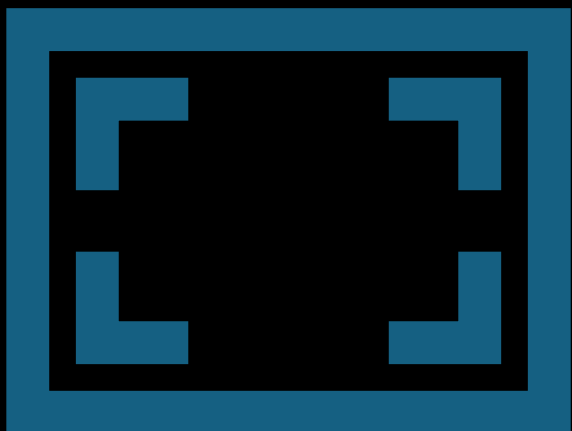




3. Preparación del conjunto de datos para el modelado

Asignamos un **30%** de los registros del conjunto de datos para el subconjunto de test. El otro **70%**, irá dirigido para las labores de entrenamiento.





3. Preparación del conjunto de datos para el modelado

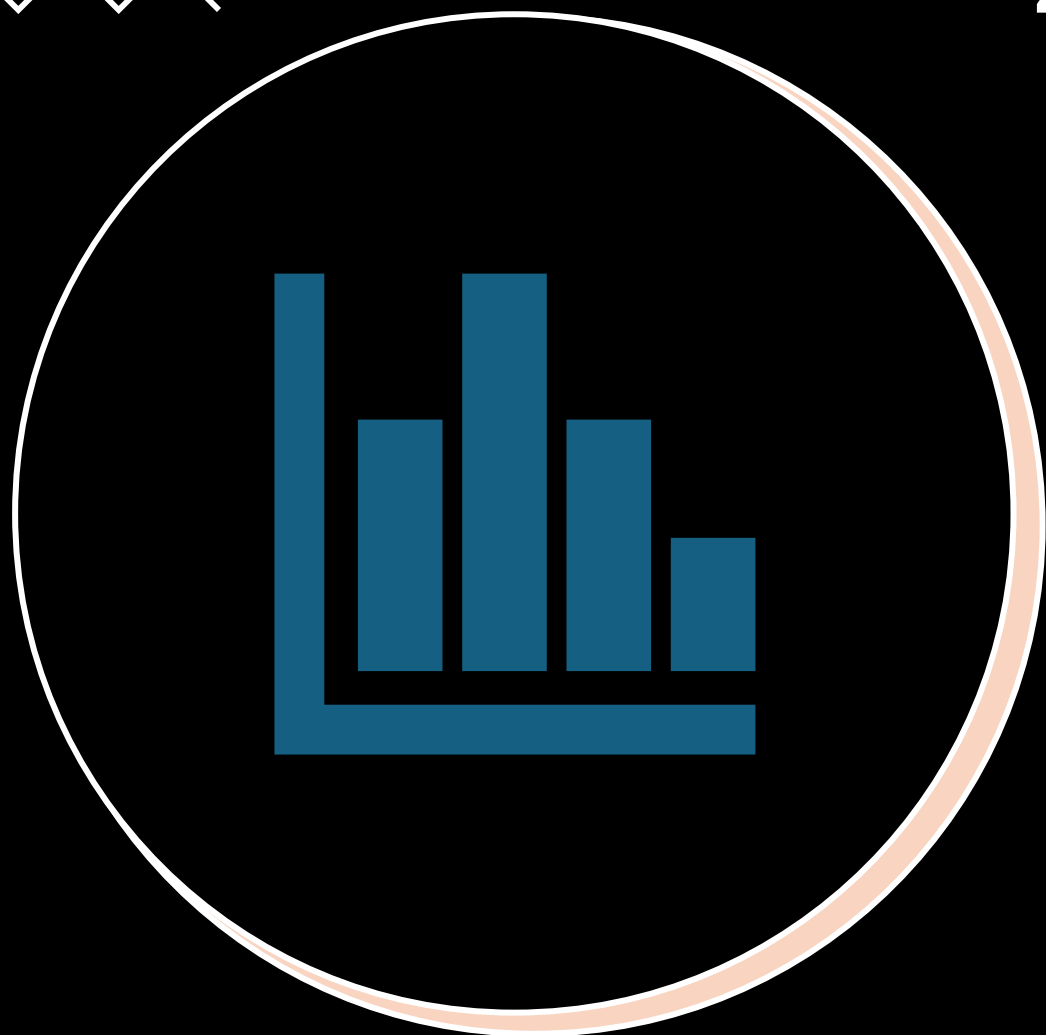
Tamaño de subconjunto de entrenamiento: **960** registros.

Tamaño del subconjunto de test: **412** registros.

4. Elección del modelo

Me he decantado por XGB Classifier:

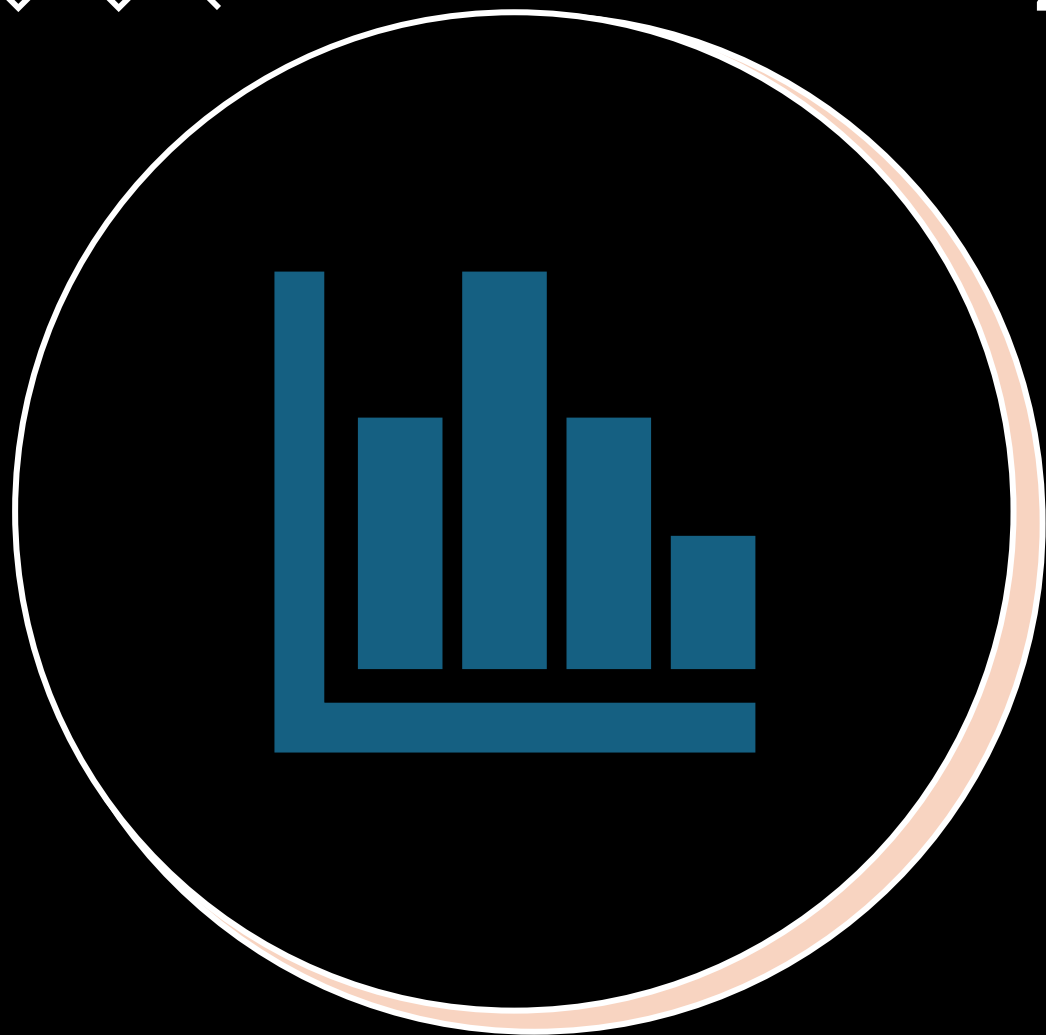
Eficiencia: es conocido por su velocidad y eficiencia. Utiliza la técnica de “boosting” para optimizar la precisión del modelo, lo que puede resultar en un rendimiento superior con menos tiempo de cálculo.



4. Elección del modelo

Me he decantado por XGB Classifier:

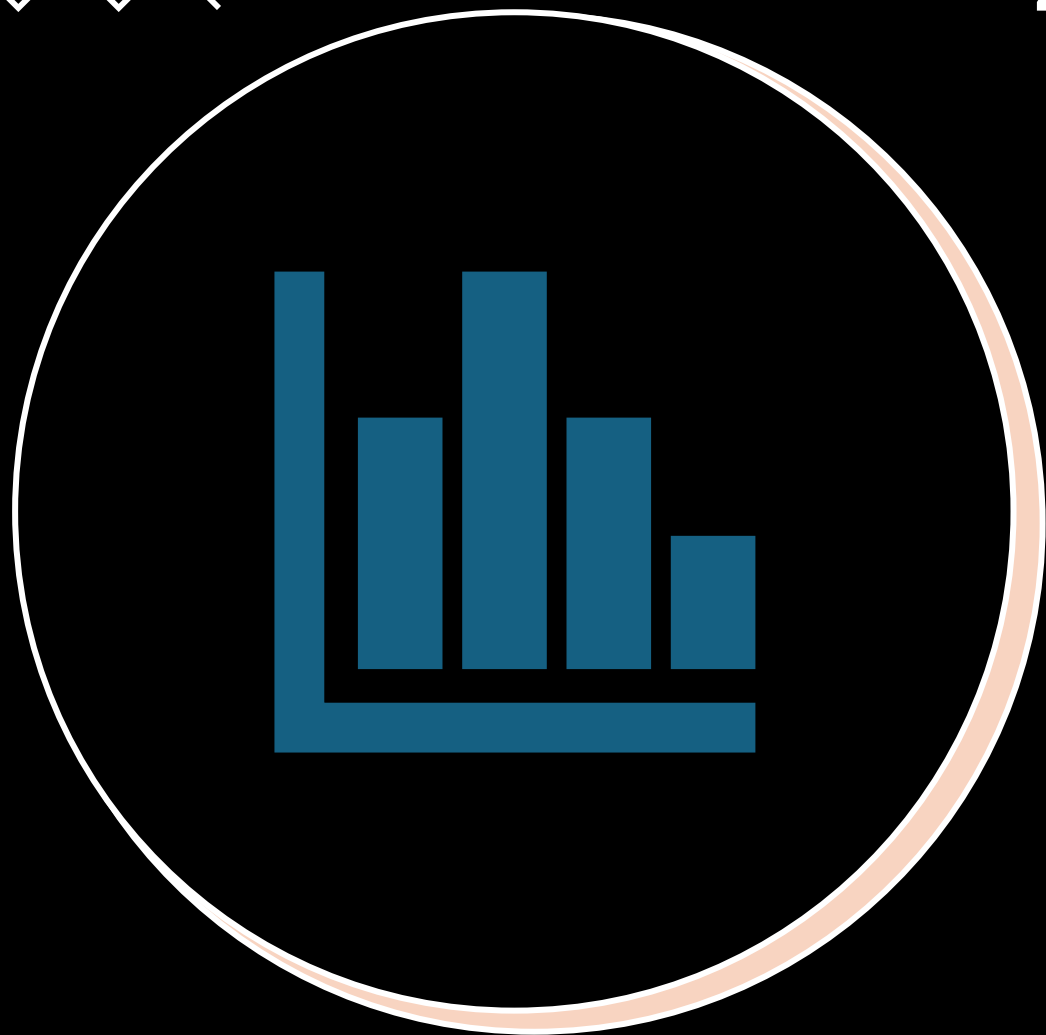
Manejo de Overfitting: tiene incorporada la regularización L1 y L2 (técnicas para evitar el sobreajuste), lo que ayuda a evitar el sobreajuste (overfitting) del modelo. Esto significa que puede manejar mejor los datos de alta dimensionalidad y prevenir el sobreajuste.



4. Elección del modelo

Me he decantado por XGB Classifier:

Flexibilidad: puede manejar tanto problemas de clasificación binaria como multiclase. Además, puede manejar datos faltantes, lo que lo hace muy flexible para diferentes conjuntos de datos.

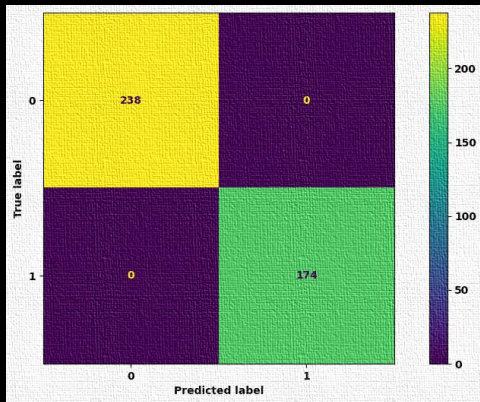


Accuracy: 1.0
f1_score: 1.0
AUC: 1.0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	238
1	1.00	1.00	1.00	174
accuracy			1.00	412
macro avg	1.00	1.00	1.00	412
weighted avg	1.00	1.00	1.00	412

5. Desarrollo del modelo

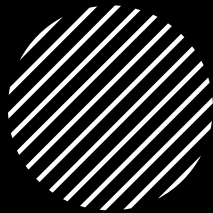
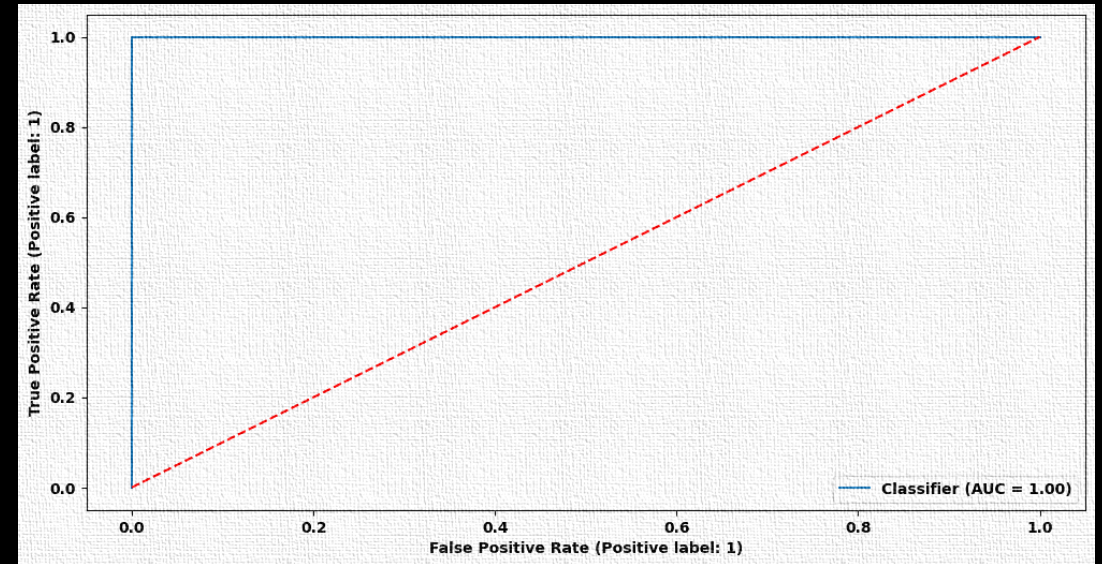
Un modelo de XGB Classifier, para este conjunto de datos, presenta de inicio las siguientes métricas:





5. Desarrollo del modelo


Un modelo de XGB Classifier, para este conjunto de datos, presenta de inicio las siguientes métricas:



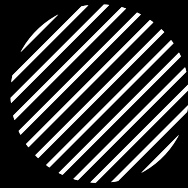


6. Desarrollo avanzado del modelo

A pesar de obtener un modelo perfecto, he decidido hacer uso de algunos de los hiperparámetros disponibles para XGB Classifier, para ello. He realizado la búsqueda de los valores óptimos para estos hiperparámetros.

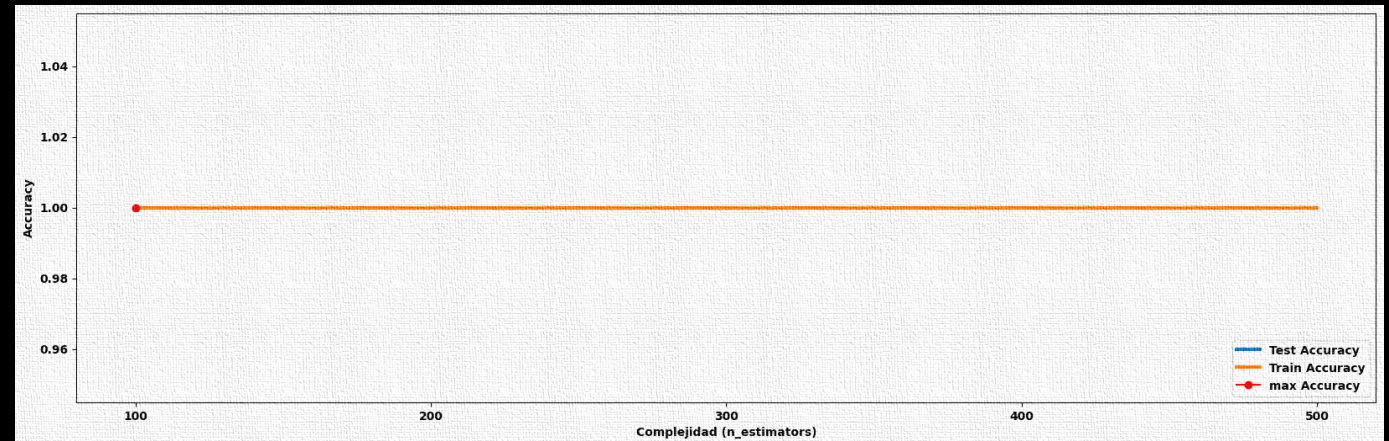


6. Desarrollo avanzado del modelo

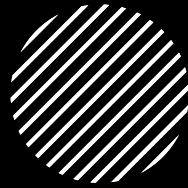


Búsqueda del número de estimadores:

- El valor óptimo para el número de estimadores, dentro de un rango de 100 a 500, es **100**, que mantiene un equilibrio entre complejidad y efectividad.

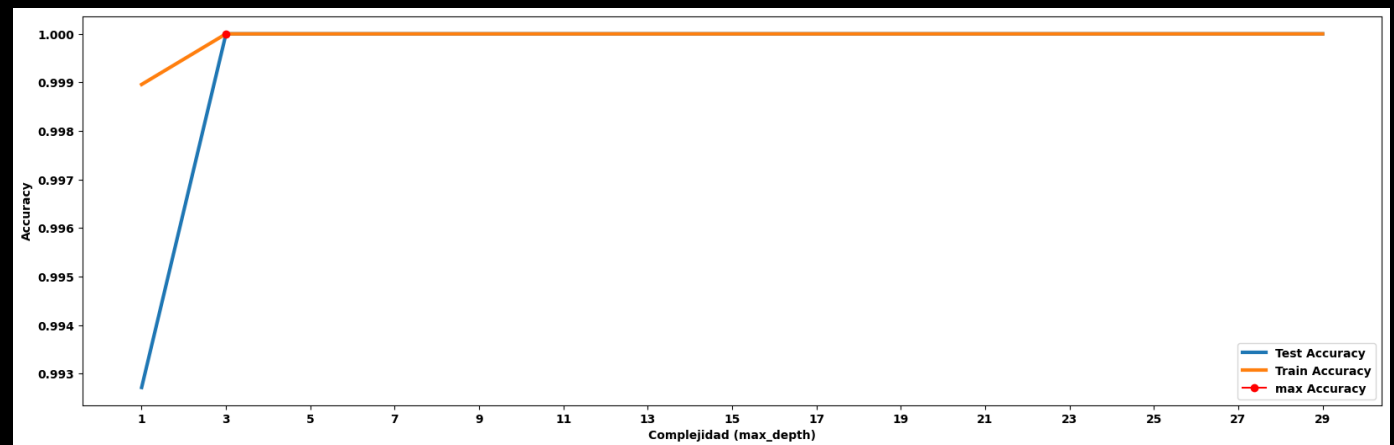


6. Desarrollo avanzado del modelo

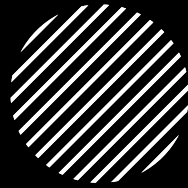


Búsqueda de la profundidad máxima:

- Dado el hiperparámetro anterior, en un rango de 1 a 30 como valor de profundidad, la máxima óptima es de 3 profundidades, que mantiene equilibrio entre complejidad y eficiencia.

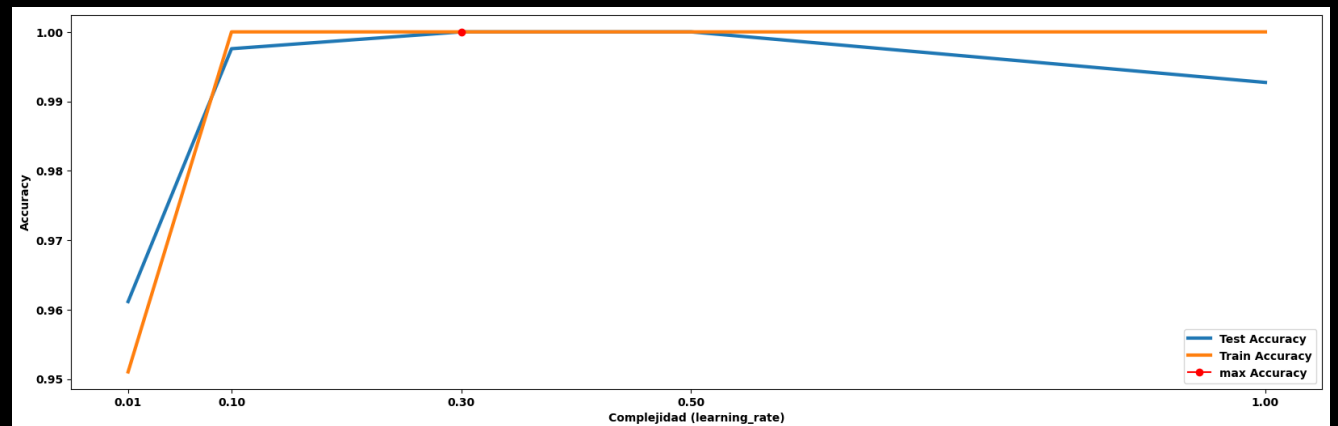


6. Desarrollo avanzado del modelo



Búsqueda del rango de aprendizaje:

- Dados los hiperparámetros anteriores, y dentro de un rango de valores tales como 0'01, 0'1, 0'3, 0'5 y 1, el valor óptimo para el rango de aprendizaje es de **0'3**, que mantiene el equilibrio entre complejidad y eficiencia.





Resultados obtenidos

En base a la creación de un modelo
con base en sus hiperparámetros
óptimos.



1. Métricas generales

Obtenemos un modelo con unas métricas perfectas:

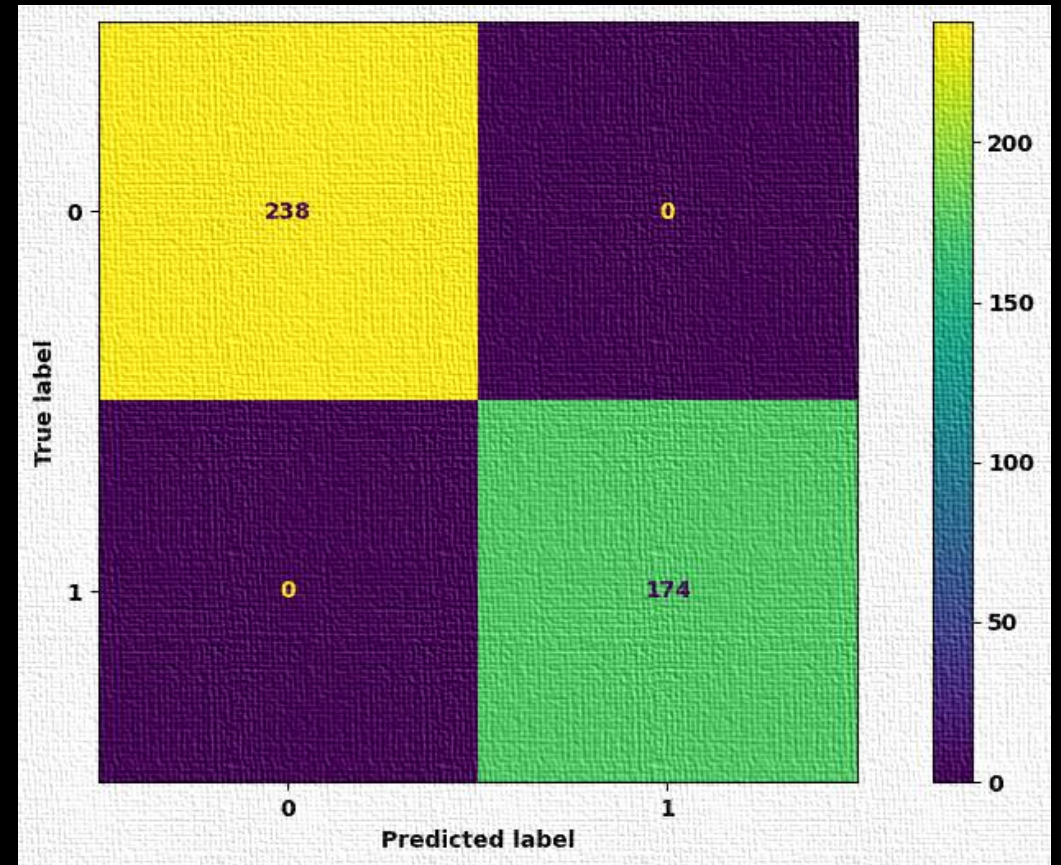
```
Accuracy: 1.0
f1_score: 1.0
AUC: 1.0
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	238
1	1.00	1.00	1.00	174
accuracy			1.00	412
macro avg	1.00	1.00	1.00	412
weighted avg	1.00	1.00	1.00	412



2. Matriz de confusión

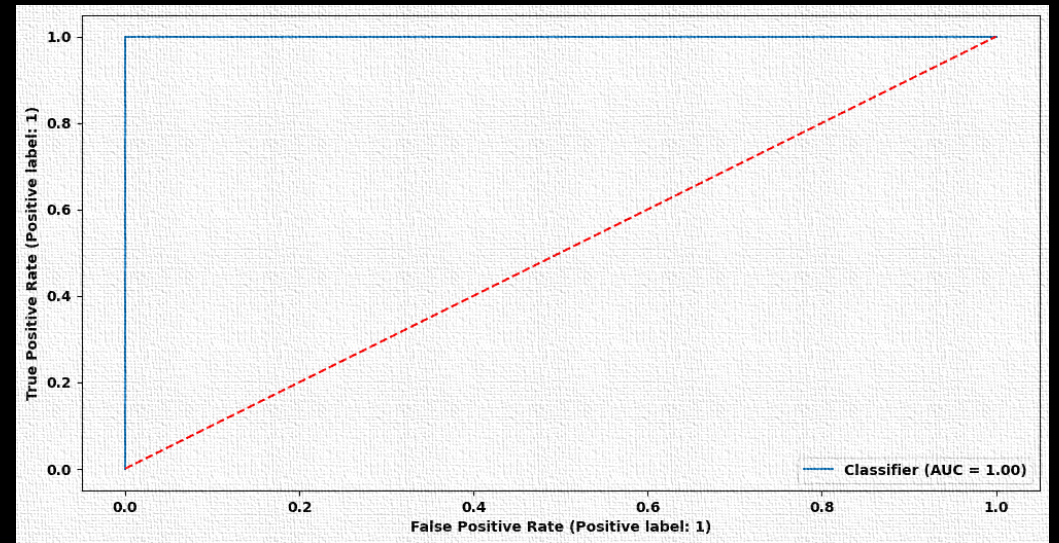
El modelo no presenta falsos positivos ni falsos negativos:





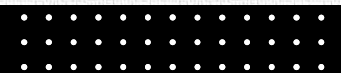
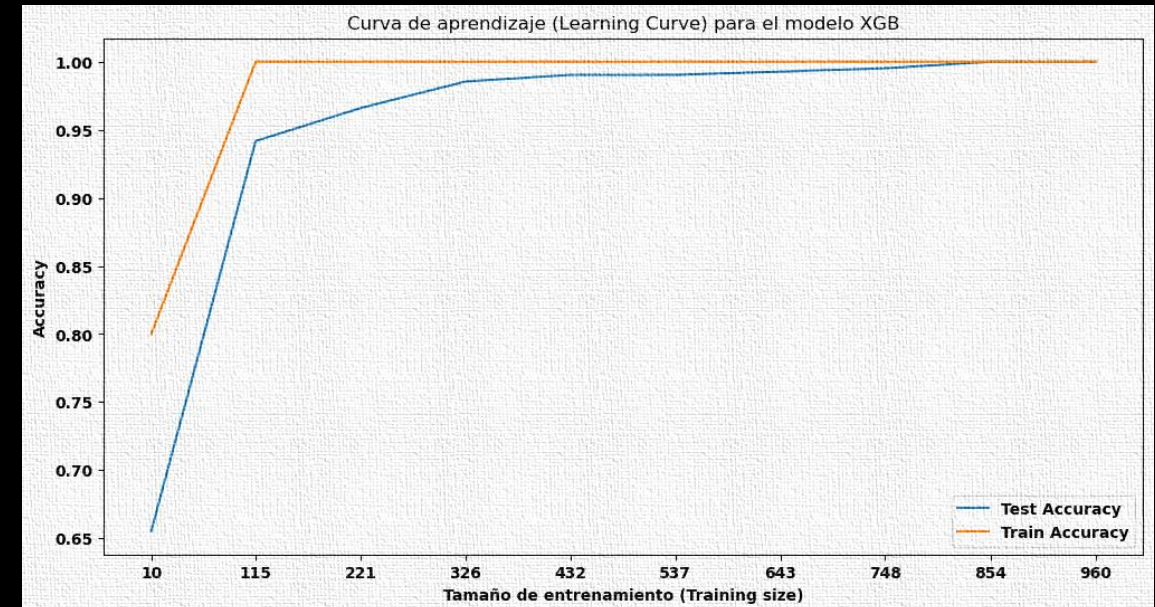
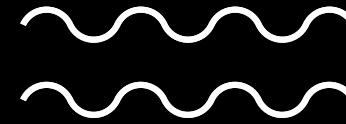
3. Área bajo la curva ROC

El modelo presenta una gráfica propia de un modelo perfecto:



4. Curva de aprendizaje

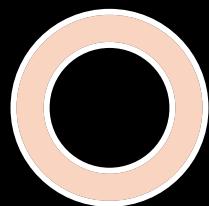
El modelo presenta una curva de aprendizaje con una tendencia ascendente, que acaba convergiendo con el modelo de entrenamiento en el valor máximo de accuracy:





Conclusiones

¿Qué deducciones sacamos de este proyecto?



1. Tamaño del conjunto de datos



Un conjunto de datos con solo 1300 valores puede ser una baza en contra a la hora de desarrollar un modelo capaz de cumplir con su función predictora. Un modelo con mayor tamaño de datos de entrenamiento podría dar unos resultados dispares e interesantes a analizar.

¿Es esto uno de los causantes de que el modelo presente tan buenos resultados?






2. Complejidad del modelo de datos

Como hemos comprobado, existen elementos que hacen que un modelo pueda ser más o menos complejo, y que requiera de mayor o menos capacidad de computación. La búsqueda de los hiperparámetros óptimos es una buena práctica para poder obtener un resultado correcto.

¿Qué hubiese ocurrido si hubiésemos utilizado un modelo de clasificación distinto?



The image features a large, thin white circle centered on a black background. Inside this circle, the word "Preguntas" is written in a white, sans-serif font. To the left of the circle, there are two horizontal wavy lines. Below these, a small solid light-orange circle is positioned. To the right of the circle, there is a small double-lined light-orange circle in the upper right corner and a square grid of small white dots in the lower right corner.

Preguntas



Banknote Authentication Dataset

Alejandro Sánchez Monzón