Luz y caos:

Análisis y clasificación automatizada de obras de arte

Laura Morales Guinart

IT ACADEMY

Resumen

Este proyecto explora la posibilidad de analizar obras de arte mediante metodologías avanzadas de análisis de datos.

El objetivo principal ha sido evaluar la viabilidad de transformar información cromática y estructural de imágenes digitales en descriptores numéricos capaces de identificar patrones estilísticos subyacentes.

Bajo este enfoque, se ha explorado la segmentación de obras en los grandes movimientos del canon artístico clásico de la Historia del Arte (Renacimiento, Barroco, Romanticismo y realismo, Impresionismo, Post-impresionismo y Vanguardias) mediante metodologías de machine learning no supervisado (PCA y K-means).

Complementariamente, se ha desarrollado un modelo predictivo de aprendizaje supervisado (*Random Forest*) y se ha evaluado su capacidad de generalización con una muestra externa de 24 cuadros de autores no presentes en el dataset. Los resultados indican una precisión del **51,25%,** determinando que los parámetros analizados permiten, de hecho, capturar la esencia de cada movimiento artístico (huella matemática medible) y clasificar, de manera limitada, obras de arte de manera automatizada.

1. Introducción

¿Es posible analizar obras de arte con análisis de datos? Tradicionalmente, la clasificación y atribución de obras se han basado en la observación experta y la comparación estilística cualitativa, métodos que, aunque fundamentales, poseen un componente intrínseco de subjetividad.

La historia del arte ha sido ya ampliamente estudiada desde hace siglos; el presente proyecto no ha pretendido revelar nueva información relativa a la evolución del arte, sino explorar una nueva metodología, una **metodología cuantitativa / computacional**, una aproximación desde el análisis de datos en la que se han usado también herramientas propias del campo de la ciencia de datos.

2. Metodología

2.1. Recopilación de datos

El dataset con el que se ha trabajado consta de una colección de obras de los 50 pintores más influyentes de la Historia del Arte. Concretamente, nos centramos en obras realizadas con técnica de pintura al óleo (descartando bocetos, grabados, frescos y otras técnicas que pudieran crear sesgos en el análisis por su naturaleza de soporte o técnica).

Por otro lado, las obras han sido organizadas en grandes estilos históricos, simplificando la complejidad del canon para facilitar esta primera aproximación técnica.

2.2. Librerías utilizadas

Para el procesamiento, análisis y visualización de los datos se emplearon las siguientes herramientas y librerías de Python:

- OpenCV y Scikit-image: para la extracción de las 9 características que he considerado como la base fundamental del análisis de las obras:

	id	artista	movement	paintings	sub_estilo	nombre_archivo	mean_r	mean_g	mean_b	saturation	brightness	entropy	contrast	homogeneity	edge_density
2361	6	Edouard Manet	Impresionismo	90	Impresionismo	Edouard_Manet_55.jpg	80.662500	42.896910	41.196452	144.270841	80.662976	6.737028	133.689444	0.188413	0.038806
1307	25	Caravaggio	Barroco	55	Barroco	Caravaggio_36.jpg	114.812657	79.977550	63.508026	110.342630	114.819894	7.026594	33.157695	0.280422	0.020484
2446	30	Edgar Degas	Impresionismo	702	Impresionismo	Edgar_Degas_14.jpg	135.045763	113.816989	65.451234	134.900198	135.471437	7.321020	26.175489	0.272027	0.011649
7506	5	Salvador Dali	Vanguardias	139	Surrealismo	Salvador_Dali_137.jpg	155.385364	168.864403	207.945710	65.947564	207.945713	6.795313	68.519355	0.326885	0.016151
7391	4	Rene Magritte	Vanguardias	194	Surrealismo	Rene_Magritte_33.jpg	97.380772	99.437604	125.140819	79.198865	132.217984	7.113835	105.743221	0.272327	0.078802

- Glob y Os: para gestionar carpetas e integrar imágenes.
- Pandas: como herramienta fundamental para la manipulación, limpieza y análisis de datos estructurados (tablas, CSV) mediante DataFrames.
- Matplotlib y Seaborn: generación de visualizaciones.
- Scikit-learn: para las herramientas de aprendizaje automático (machine learning) clásico, clasificación, clustering y reducción de dimensionalidad.

El uso de estas herramientas permitió transformar imágenes en un conjunto de datos estructurados.

2.3. Reducción de Dimensionalidad (PCA) y clustering (K-means)

Dado que las variables presentaban alta correlación, y para simplificar las 9 variables originales en dos dimensiones manejables, se aplicó **PCA**. El análisis reveló que la varianza de los datos se explica principalmente en dos ejes:

1. **Eje de Luminosidad/Color:** Define el impacto luz/oscuridad.
2. **Eje de Estructura y Forma:** Mide el grado de complejidad vs. homogeneidad.

Esta transformación eliminó el ruido y permitió visualizar las "fronteras" técnicas entre estilos.

También aplicamos **K-means** para identificar si los estilos formaban grupos naturales y validar si los estilos se separan de forma lógica sin etiquetas previas.

Antes de realizar la reducción de dimensionalidad y el clustering, se procedió a un *downsampling* de las clases mayoritarias. Al igualar el número de muestras por movimiento artístico, se garantizó que los componentes principales y los centroides de los clusters no se vieran sesgados por el volumen de datos de un solo estilo.

2.4. Interpretación de Componentes y Análisis Exploratorio (EDA)

Se analizaron las distribuciones y la evolución basándonos en los ejes propuestos por la PCA. Este proceso permitió confirmar empíricamente patrones históricos, como la baja luminosidad característica del Barroco frente a la claridad del Impresionismo o bien cómo han evolucionado a nivel de complejidad los diferentes períodos.

2.5. Random forest

Para la clasificación automatizada, se implementó un algoritmo de **Random Forest** entrenado mediante una partición de **train/test split** sobre el dataset original. Con el fin de mitigar el sesgo hacia las clases mayoritarias, se aplicó la técnica **SMOTE** (Synthetic Minority Over-sampling Technique) en la fase de entrenamiento. Este sobremuestreo sintético fue crítico para que el modelo aprendiera los rasgos distintivos de movimientos con baja representatividad, como el **Realismo** y el **Romanticismo**, que anteriormente pasaban desapercibidos.

Como resultado de este proceso, el modelo alcanzó una **precisión (accuracy) del 51,25% en el set de test**. Este porcentaje representa una mejora sustancial

respecto al azar y supera el rendimiento de la clase mayoritaria, confirmando que el algoritmo ha extraído reglas de clasificación válidas a partir de las 9 variables técnicas iniciales.

Para verificar la robustez de este aprendizaje, se realizó una **validación externa a posteriori** mediante una muestra de **24 cuadros inéditos** (4 por estilo) de autores ausentes en el dataset original. Esta prueba adicional permitió observar la capacidad de generalización del modelo, confirmando que el predictor es capaz de reconocer la "huella matemática" de un movimiento artístico incluso en obras totalmente desconocidas para el sistema.

3. Resultados

Reducción Dimensional (PCA)

El análisis dimensional mediante PCA permitió proyectar el dataset en un plano bidimensional donde el **Componente 1 (Luminosidad)** y el **Componente 2 (Complejidad)** explicaron la mayor parte de la varianza.

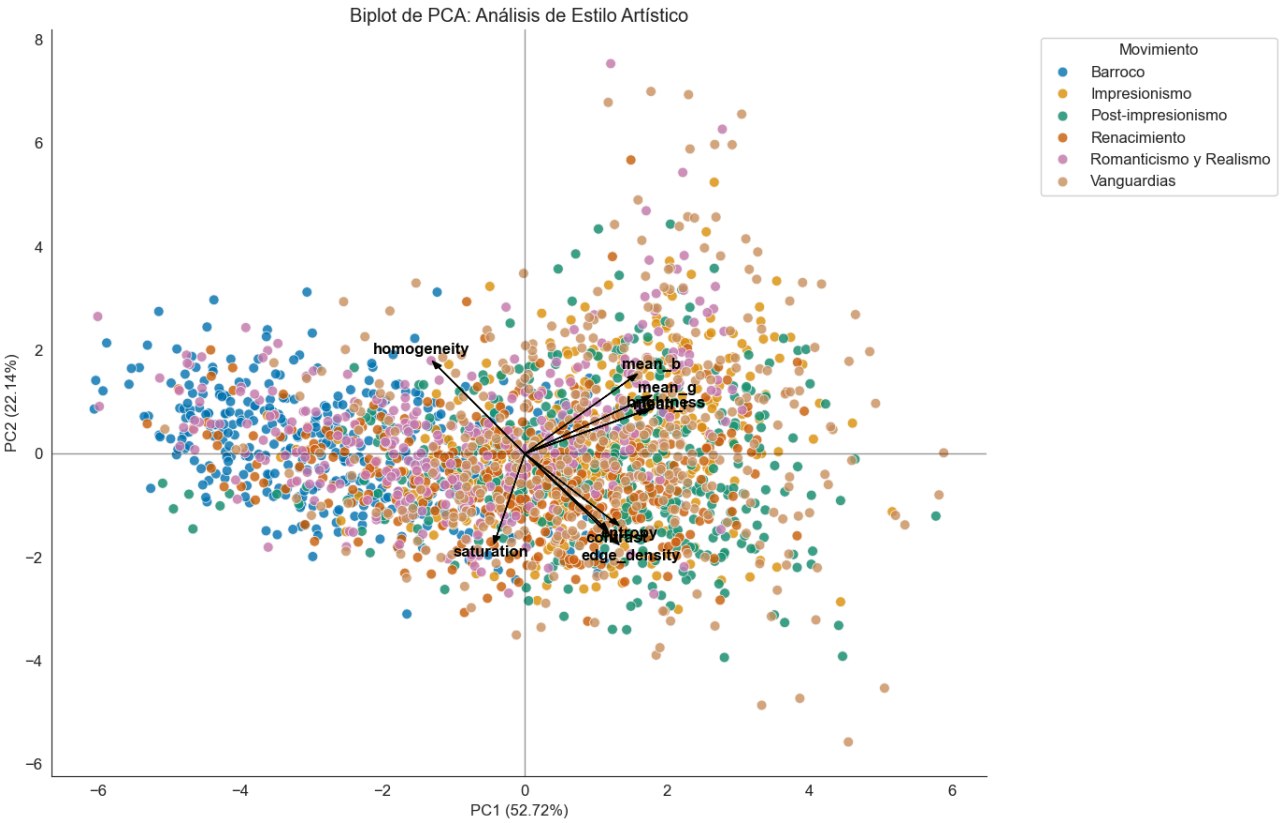

Figura 1.

Con el fin de facilitar la interpretación visual y mitigar el solapamiento entre categorías, se procedió a la **segmentación del Biplot por movimiento artístico**. Esta descomposición permitió analizar de forma aislada la dispersión de cada estilo respecto a sus **centroides (medias vectoriales)**, evidenciando las diferencias en la distribución de las variables sobre el espacio latente del PCA (Figura 2).

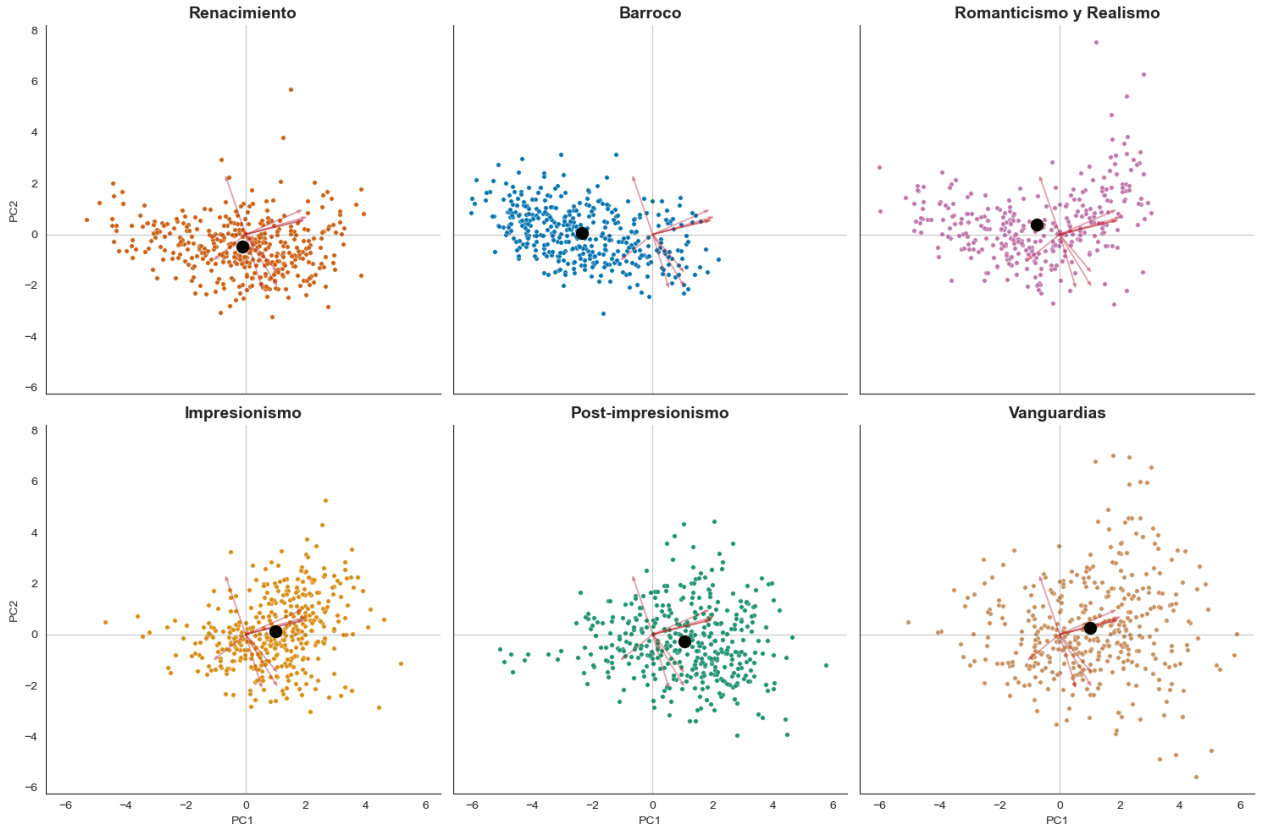

Figura 2.

Agrupación Natural de Estilos (K-Means)

cluster	0	1	2	3
movement				
Renacimiento	49.000000	7.500000	28.500000	15.000000
Barroco	34.750000	2.000000	4.500000	58.750000
Romanticismo y Realismo	46.708464	23.197492	5.329154	24.764890
Impresionismo	38.500000	34.750000	25.000000	1.750000
Post-impresionismo	31.750000	24.000000	39.250000	5.000000
Vanguardias	32.000000	33.000000	28.000000	7.000000

Figura 3. Clústeres normalizados por filas (porcentajes)

Evolución ejes principales de análisis

Eje 1 (Luminosidad):

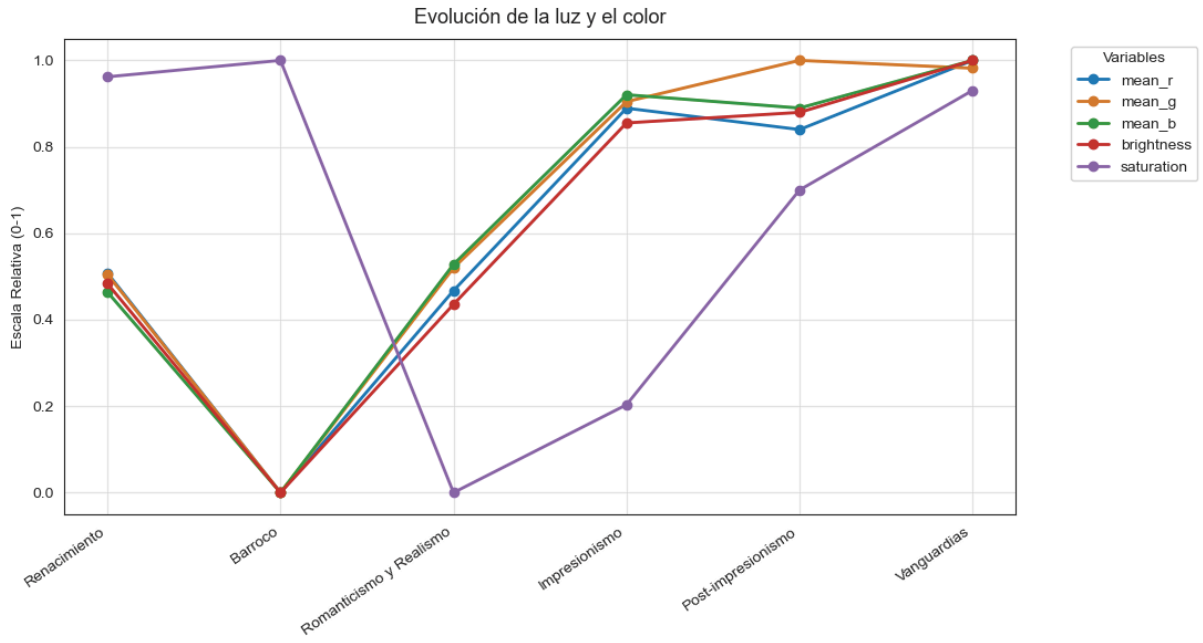

Figura 4. Evolución de la luminosidad por período.

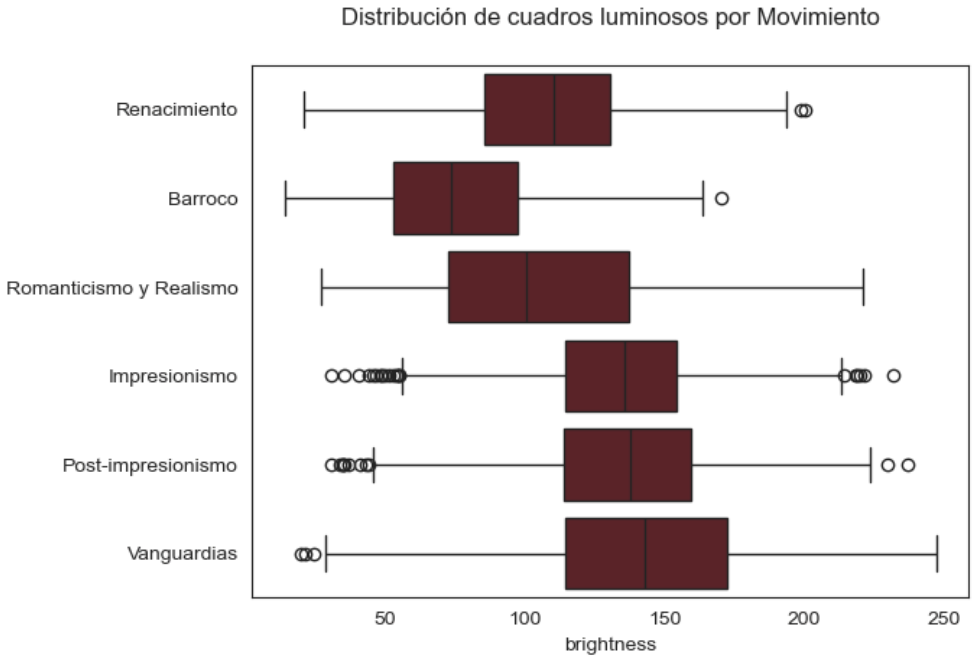

Figura 5. Distribución de la luminosidad por período.

Eje 2 (Complejidad y Estructura):

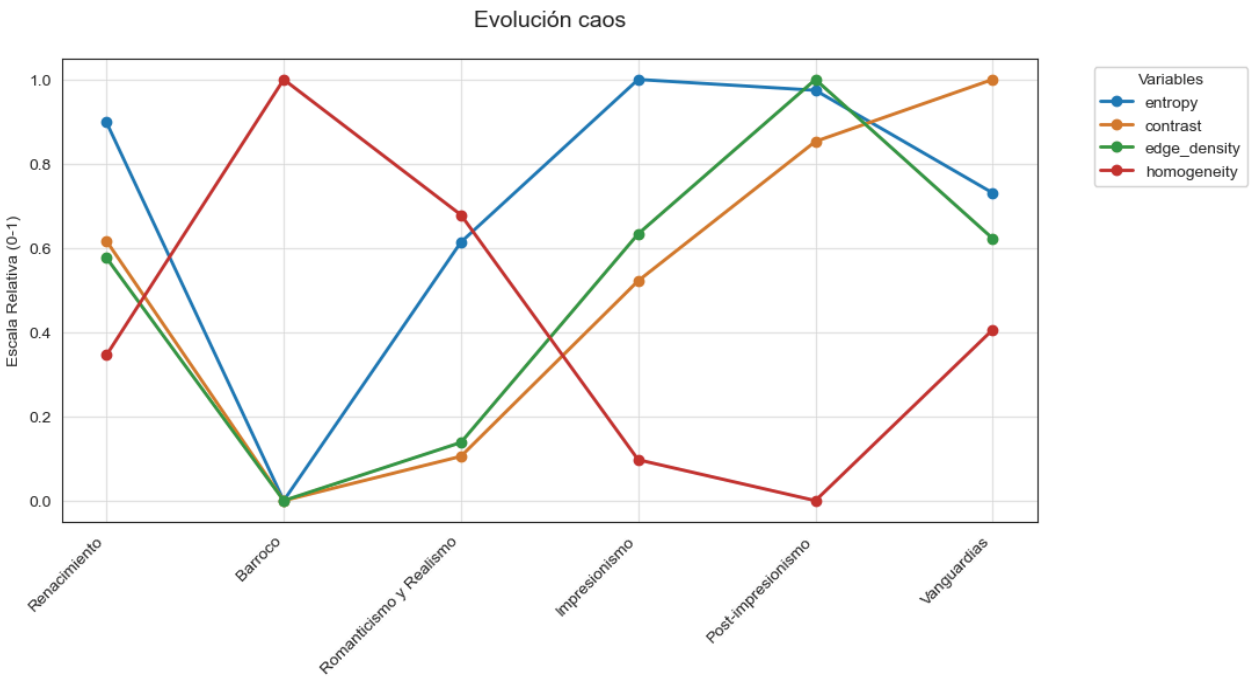

Figura 6. Evolución de los parámetros de textura, estructura y complejidad por período.

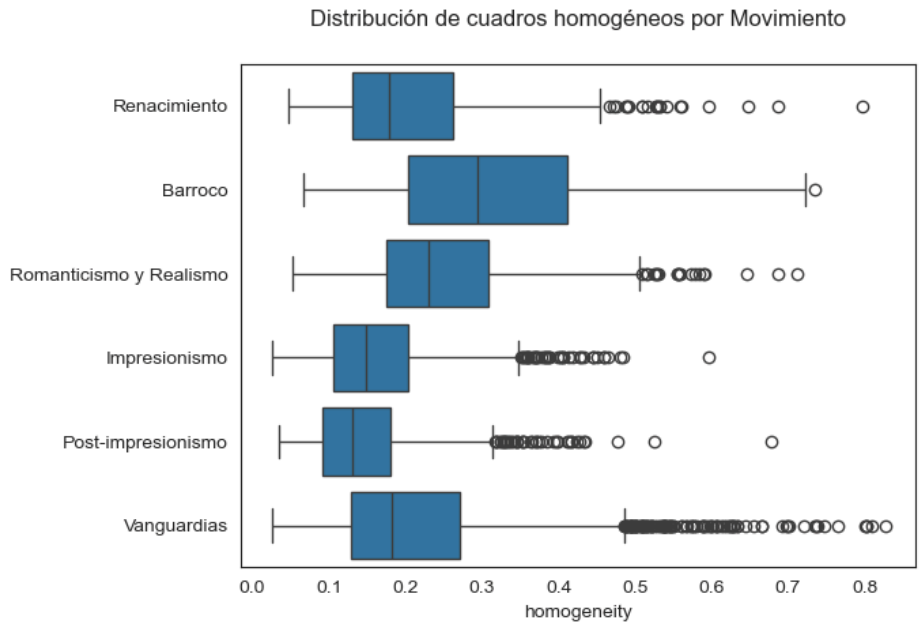

Figura 7. Distribución de la homogeneidad por período.

Rendimiento del Clasificador y SMOTE (Random Forest)

El modelo **Random Forest** alcanzó una **precisión (accuracy) del 51,25%** en el conjunto de test.

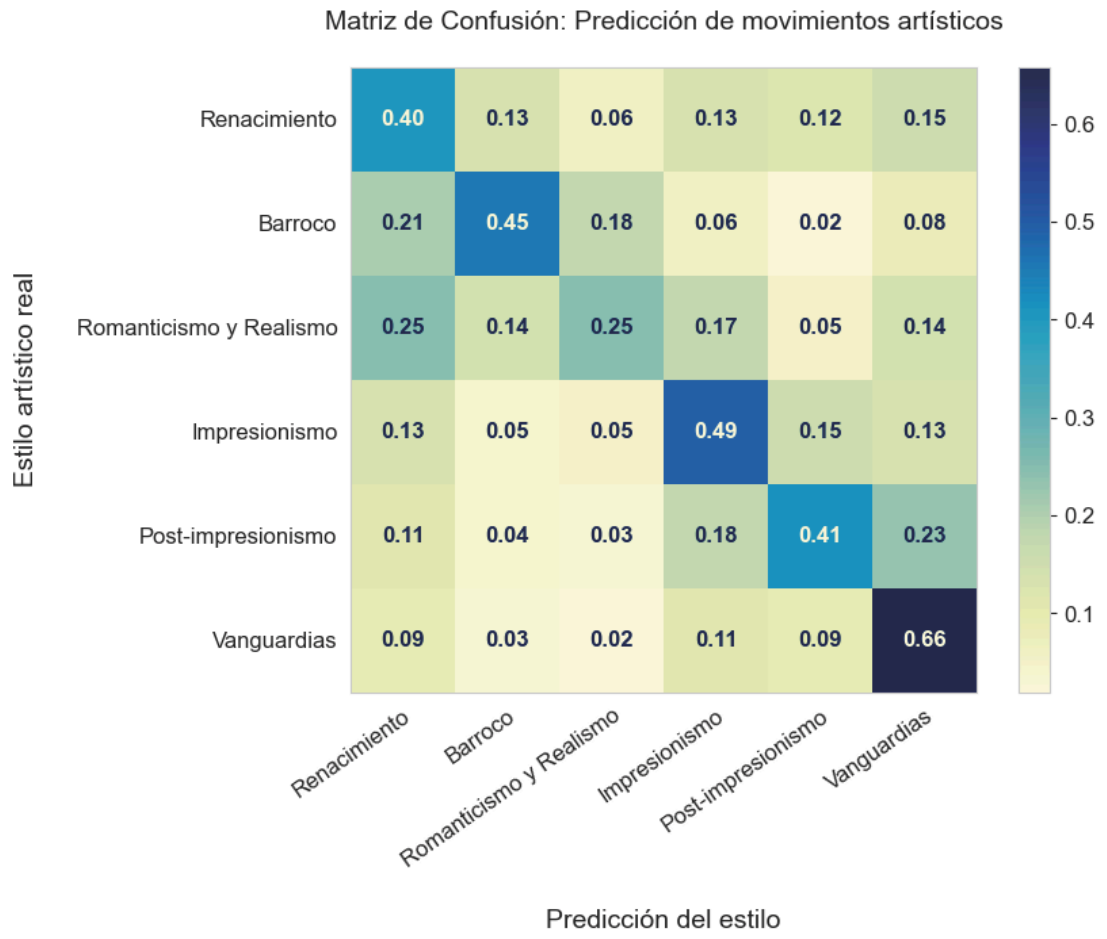

Figura 8. Matriz de confusión.

La implementación de **SMOTE** fue el factor determinante para lograr este equilibrio. Antes del sobremuestreo, el modelo tendía a sesgarse hacia las Vanguardias, ignorando las categorías minoritarias.

```
Después de SMOTE: movement
Renacimiento                2058
Barroco                     2058
Romanticismo y Realismo     2058
Impresionismo               2058
Post-impresionismo          2058
Vanguardias                 2058
```

Validación de Generalización (Test de 24 cuadros)

La prueba de fuego consistió en la validación con **24 cuadros externos** (4 por estilo) de autores no incluidos previamente.

Figura 9. Resultados del test para el Barroco.

4. Discusión

La interpretación de los componentes del PCA sugiere que la historia del arte se puede mapear en un espacio de "Complejidad vs. Luz". La separación clara de estilos como el Impresionismo en el eje de la estructura valida la hipótesis de que la pincelada fragmentada tiene una firma estadística única.

La visualización de estos componentes (Figura 2) revela que los estilos se distribuyen de forma coherente con la teoría del arte: el **Barroco** ocupa el cuadrante de menor luminosidad (confirmando el uso técnico del claroscuro), mientras que el **Impresionismo** y las **Vanguardias** se desplazan hacia niveles superiores de claridad y mayor fragmentación estructural. Este resultado confirma que es posible mapear la historia del arte sobre un plano matemático objetivo.

La aplicación del algoritmo K-Means (Figura 3) reveló que el Barroco forma un clúster natural bastante claro, seguido del Renacimiento y del Realismo y Romanticismo, mientras que los movimientos más modernos mostraron más segmentación.

Este resultado es clave, ya que demuestra que A diferencia del Renacimiento o el Barroco, que tenían reglas más estrictas, el arte moderno experimenta con todo el espectro de color y textura. El K-means detecta esta "dispersión" como una falta de un clúster único dominante.

Respecto a la evolución de los componentes principales, los resultados muestran una evolución clara en el eje lumínico (Figura 4) donde el **Barroco** se sitúa en los valores mínimos de intensidad lumínica, contrastando con las **Vanguardias**, que dominan los rangos más altos de claridad, mientras que en el eje complejidad (Figura 5) permite trazar se observa una progresión desde la **homogeneidad y orden** de los períodos clásicos (Barroco) hacia una mayor **entropía y fragmentación** en el Impresionismo y en las Vanguardias, reflejando estadísticamente la ruptura de la pincelada tradicional y la diversidad de estilos.

En relación a los resultados predictivos del Random Forest, la precisión (Accuracy) del 51, 25% representa una **mejora incremental de aproximadamente el 40%** respecto al *baseline* de la clase mayoritaria (las Vanguardias, que representaban el 35% del dataset). Superar significativamente este umbral demuestra que el modelo no se limita a predecir la clase más frecuente, sino que ha desarrollado una capacidad real de discriminación entre los distintos movimientos.

La matriz de confusión (Figura 8) revela que el modelo posee una alta capacidad discriminativa para las **Vanguardias (66%)**, el **Impresionismo (49%)** y el **Barroco (45%)**, confirmando que estos movimientos poseen rasgos técnicos lo suficientemente distintivos para ser capturados algorítmicamente.

La capacidad del modelo para clasificar obras de autores desconocidos (Figura 9) que observamos en el Test, con solo 9 variables demuestra que el algoritmo ha "entendido" las reglas técnicas generales de cada movimiento, superando el simple aprendizaje de memoria.

5. Conclusiones

La pregunta que dio origen a este trabajo era tan simple como ambiciosa: ¿puede una pintura —materia, gesto, intuición— analizarse desde una metodología propia del análisis de datos? La respuesta es afirmativa. Siempre que entendamos que toda cuantificación es una forma de mirada: una reducción consciente que no agota la obra, pero sí revela estructuras invisibles a simple vista.

Con solo nueve variables estadísticas —medias de color, textura, entropía y estructura— ha sido posible capturar una huella matemática de cada movimiento artístico. Los resultados muestran que los estilos no son únicamente categorías históricas, sino configuraciones formales medibles. La proyección en el espacio PCA sugiere incluso una cartografía de la historia del arte organizada en torno a un eje interpretable como “Complejidad vs. Luz”. La separación nítida de corrientes como el Impresionismo valida que la pincelada fragmentada posee una firma estadística propia, mientras que la proximidad entre Realismo y Romanticismo confirma su base técnica compartida.

Desde el plano metodológico del “clasificador de cuadros”, el modelo no se limita a memorizar obras: generaliza. La correcta clasificación de artistas no presentes en el dataset demuestra que el algoritmo ha aprendido reglas formales, no nombres. La mejora del 40% respecto al baseline de la clase mayoritaria refuerza que incluso una representación extremadamente comprimida puede contener información discriminativa significativa.

Este estudio no reduce el arte a números; demuestra que el arte también tiene estructura. Que la intuición deja rastro medible. Y que, observada desde esta doble perspectiva —estética y matemática— la historia pictórica aparece no como una sucesión caótica, sino como una trayectoria coherente hacia mayores niveles de luz y diversidad formal.

6. Referencias bibliográficas

Rubayo, S. (2021). *Te gusta el arte aunque no lo sepas*. Paidós.