Taller 1: K-means y SOM

Álvaro Alejandro Zarabanda Gutierrrez

Universidad Distrital Francisco Jose de Caldas

Maestría en Ciencias de la Información y las Comunicaciones - Big Data

aazarabandag@udistrital.edu.co

27 de septiembre de 2025

Resumen

Este ejercicio académico compara K-means y Self-Organizing Maps (SOM) en la clasificación de manuscritos paleográficos medievales, utilizando el dataset Avila UCI (23,314 muestras, 8 copistas). Se emplea una metodología semi-supervisada con balanceo de clases y validación cruzada. SOM supera a K-means en precisión (45.79 % vs 33.10 %).

Índice

1. Introducción

La clasificación automática de manuscritos paleográficos requiere técnicas avanzadas de aprendizaje automático para abordar la variabilidad estilística entre copistas. Se comparan dos enfoques de clustering: K-means (basado en centroides) y SOM (preserva la topología local), aplicando un esquema semi-supervisado y balanceo de clases.

1.1. Objetivos

El objetivo principal es comparar el rendimiento de K-means y SOM en la clasificación de manuscritos paleográficos, considerando criterios como precisión, estabilidad, interpretabilidad y eficiencia computacional.

Los objetivos específicos son: (1) implementar y optimizar ambos algoritmos mediante búsqueda de hiperparámetros; (2) establecer una metodología semi-supervisada que permita evaluación cuantitativa; (3) aplicar técnicas de balanceamiento de clases para el dataset desbalanceado; (4) realizar validación cruzada estratificada para asegurar la generalización de los resultados; y (5) ofrecer recomendaciones prácticas basadas en los resultados obtenidos.

2. Marco Teórico

2.1. K-means Clustering

K-means es un algoritmo de particionamiento que busca dividir n observaciones en k clusters, donde cada observación pertenece al cluster cuyo centroide está más cercano [4]. El algoritmo minimiza la suma de distancias euclidianas al cuadrado entre cada punto y el centroide de su cluster asignado, conocida como inercia o Within-Cluster Sum of Squares (WCSS).

La función objetivo se define como:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{1}$$

donde C_i representa el i-ésimo cluster y μ_i su centroide correspondiente. La optimización se realiza mediante el algoritmo de Lloyd, que alterna entre la asignación de puntos a clusters y la recalculación de centroides hasta convergencia.

Las principales ventajas de K-means incluyen su simplicidad conceptual, eficiencia computacional O(nkt) donde n es el número de muestras, k el número de clusters y t las iteraciones, y la interpretabilidad directa de los centroides. Sin embargo, presenta limitaciones significativas: sensibilidad a la inicialización, necesidad de especificar k a

priori, asunción de clusters esférricos y equitamaño, y vulnerabilidad a outliers.

2.2. Self-Organizing Maps (SOM)

Self-Organizing Maps, introducido por Kohonen [3], es un tipo de red neuronal no supervisada que proyecta datos de alta dimensionalidad sobre una rejilla bidimensional de neuronas, preservando la topología del espacio original. Cada neurona está asociada con un vector de pesos de la misma dimensionalidad que los datos de entrada.

El proceso de entrenamiento SOM sigue estos pasos: (1) para cada muestra de entrada, se identifica la Best Matching Unit (BMU), la neurona con vector de pesos más similar; (2) se actualizan los pesos de la BMU y sus vecinos topológicos; (3) se reduce progresivamente tanto el radio de vecindad como la tasa de aprendizaje.

La función de actualización se define como:

$$w_i(t+1) = w_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - w_i(t)] \tag{2}$$

donde $\alpha(t)$ es la tasa de aprendizaje decreciente, $h_{ci}(t)$ es la función de vecindad que decrece con la distancia topológica entre la neurona i y la BMU c, y x(t) es el vector de entrada en el tiempo t.

SOM ofrece ventajas únicas: preservación de la topología local, capacidad de visualización intuitiva, detección de estructuras complejas no lineales, y robustez ante outliers. Las limitaciones incluyen mayor complejidad computacional, necesidad de ajustar múltiples hiperparámetros (dimensiones del mapa, tasa de aprendizaje, función de vecindad), y convergencia no garantizada a un óptimo global.

3. Metodología

3.1. Datasets Utilizados

Se utilizaron dos conjuntos de datos:

data_clusters.mat: Sintético, 600 puntos, para validación metodológica.

Avila UCI [2]: 23,314 muestras, 10 características, 8 clases (copistas).

3.2. Preprocesamiento de Datos

El preprocesamiento incluyó:

Filtrado de clases minoritarias: Se eliminaron clases con menos de 375 muestras.

Balanceo híbrido: Undersampling + SMOTE [1] para lograr una distribución uniforme (ratio máximo/mínimo de 1.43:1).

Normalización: StandardScaler (K-means), MinMaxScaler (SOM).

3.3. Metodología Semi-supervisada

Se dividieron los datos en 70 % para entrenamiento, 15 % para etiquetado y 15 % para validación. El agrupamiento se realiza sin etiquetas, luego se asignan etiquetas a los clusters y se evalúa el rendimiento en el conjunto de validación.

3.4. Optimización de Hiperparámetros

K-means: K=2–15, método del codo y silhouette (K=6 sintético, K=10 Avila).
SOM: Mapas 3×3–8×8, balance entre error y eficiencia (7×7 sintético, 8×8 Avila).

3.5. Evaluación y Validación

Validación cruzada estratificada (5-fold). Métricas: accuracy, precisión, recall, F1-score, matrices de confusión. Se consideró rendimiento absoluto y estabilidad.

4. Resultados

4.1. Análisis Exploratorio del Dataset data_clusters.mat

La Figura 1 muestra la distribución bidimensional del dataset sintético, revelando seis agrupamientos naturales claramente diferenciados. El análisis estadístico indica rangos [14.00, 514.00] para X1 y [13.00, 370.00] para X2, con distribución aproximadamente normal (media X1=281.01±143.51, X2=190.92±102.75).

Figura 1: Visualización del dataset data_clusters.mat mostrando la distribución natural de 600 puntos en el espacio bidimensional. Se observan seis agrupamientos principales con diferentes densidades y formas geométricas.

4.2. Optimización K-means

El método del codo (Figura 2) identifica K=6 como óptimo en el dataset sintético, con clusters bien definidos y cohesivos. En Avila, K=10 es el mejor equilibrio entre métricas.

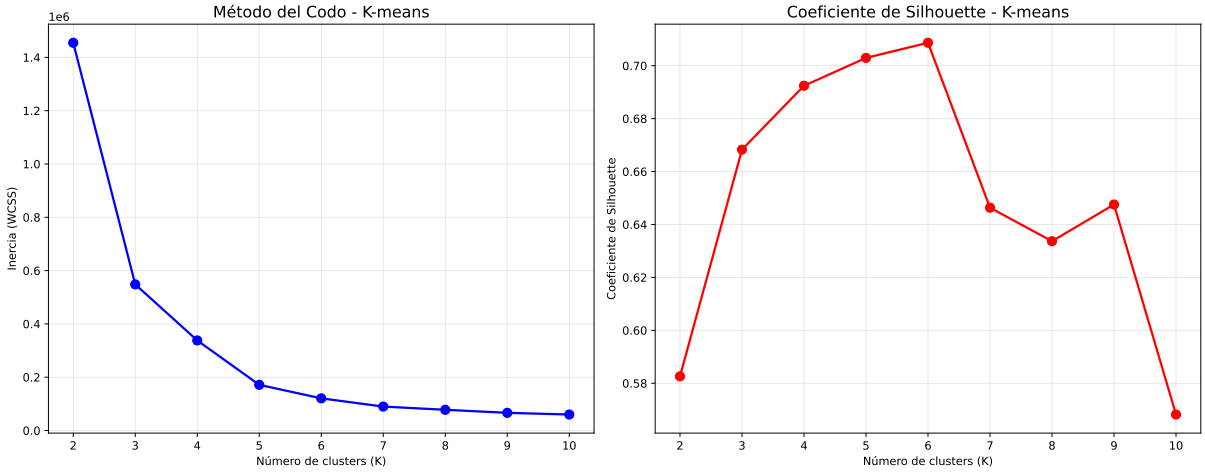

Figura 2: Selección del número óptimo de clusters para K-means. (Izquierda) Método del codo. (Derecha) Coeficiente de silhouette.

La aplicación de K-means con K=6 (Figura 3) genera clusters compactos en el dataset

sintético.

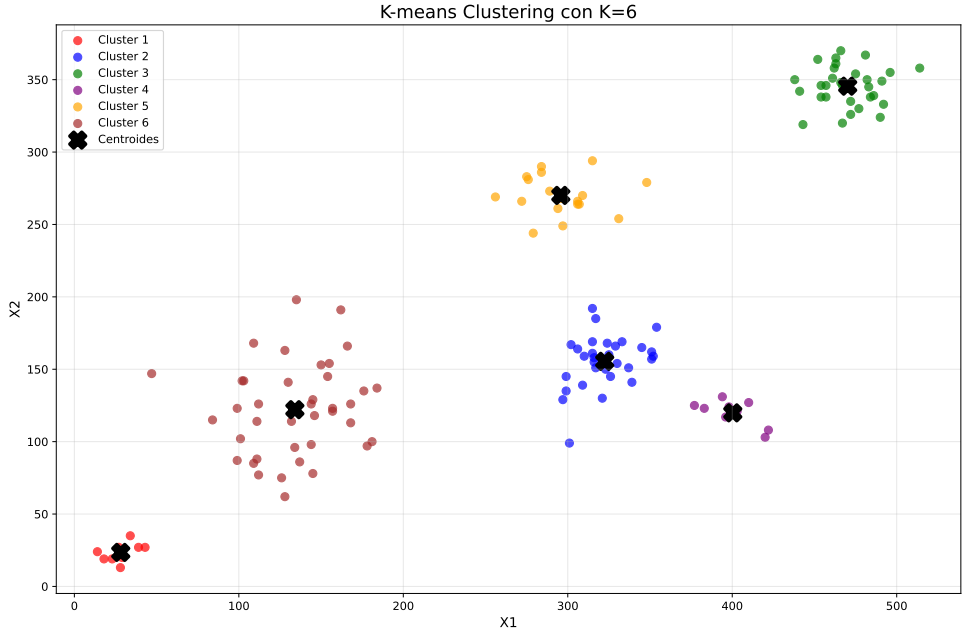

Figura 3: Clustering K-means óptimo (K=6) sobre data_clusters.mat.

4.3. Optimización SOM

El análisis de las configuraciones de SOM (Figura 4) muestra que existe un equilibrio entre el error de cuantización y la eficiencia en la activación de neuronas. Los mapas pequeños (3×3) presentan alta activación pero mayor error, mientras que los mapas grandes (8×8) disminuyen el error aunque la eficiencia baja.

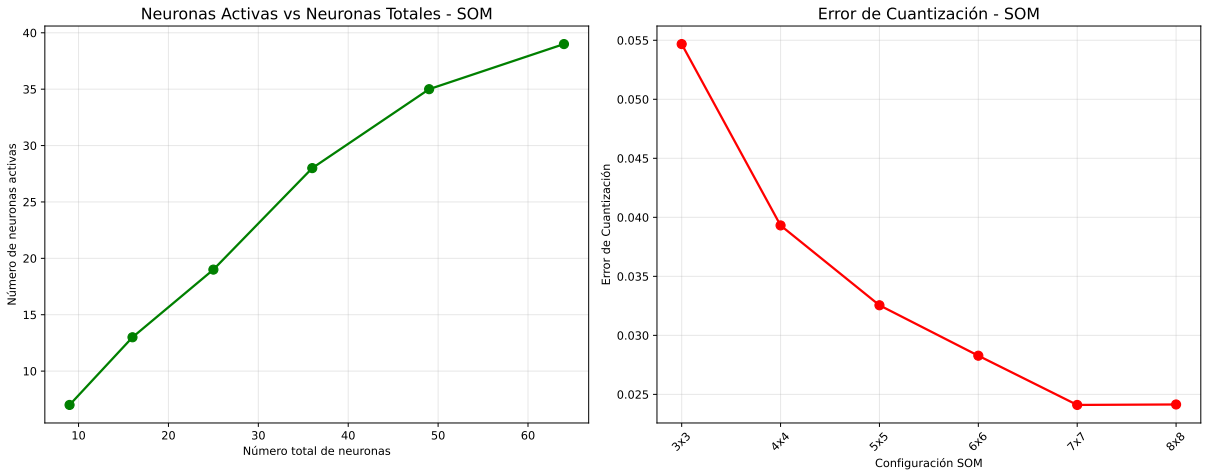

Figura 4: Comparación de configuraciones SOM. (Izquierda) Proporción de neuronas activas. (Derecha) Error de cuantización según la configuración. El mapa 7×7 logra el mejor balance.

La configuración 7×7 de SOM (Figura 5) logra preservar la estructura del conjunto original, con neuronas vecinas representando regiones cercanas en el espacio de datos. El

mapa de activación muestra que 35 de 49 neuronas están activas (71.4 %), lo que indica un uso eficiente del espacio representacional.

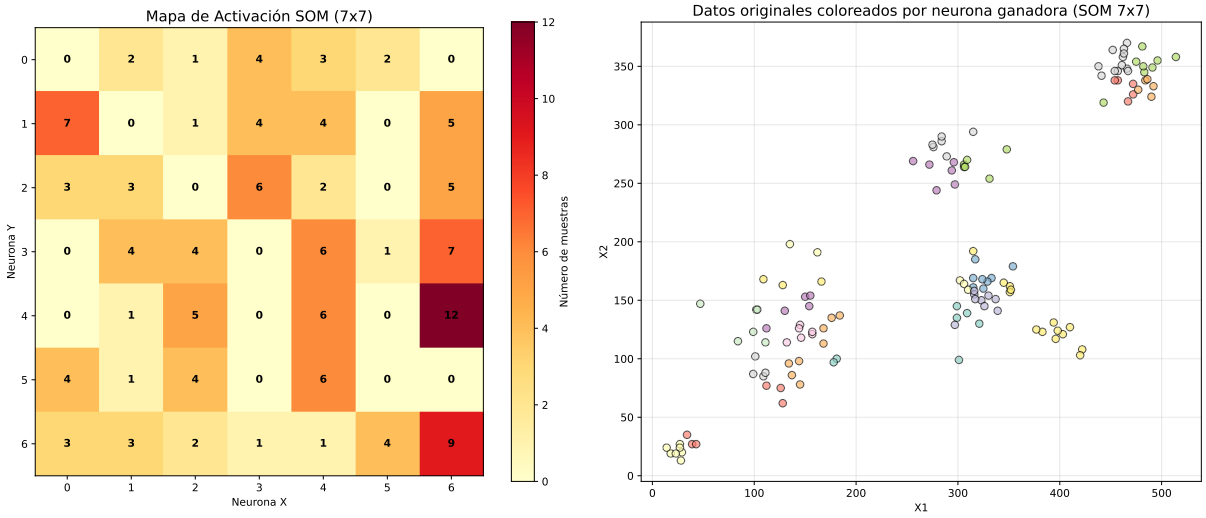

Figura 5: Detalle del SOM 7×7. (Izquierda) Mapa de distancias U-matrix. (Centro) Distribución de clusters. (Derecha) Frecuencia de activación de neuronas.

4.4. Análisis del Dataset Avila

El dataset Avila muestra un marcado desbalance (Figura 6), con la clase A representando el 18.4 % de las muestras. Esta distribución desigual exige aplicar técnicas de balanceo para evitar sesgos hacia las clases más numerosas.

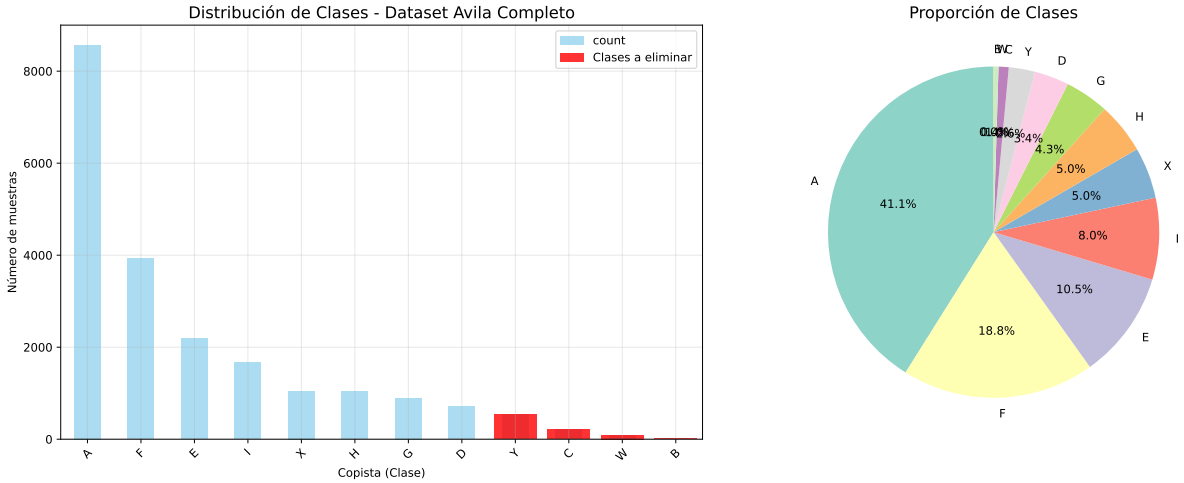

Figura 6: Distribución de clases en el dataset Avila UCI. (Arriba) Distribución original. (Abajo) Estadísticas por clase que muestran diferencias entre copistas.

Las técnicas de balanceo (Figura 7) modifican la distribución, reduciendo el ratio máximo/mínimo de 12.16:1 a 1.43:1. El undersampling mantiene la información relevante y SMOTE genera muestras sintéticas que respetan la estructura original.

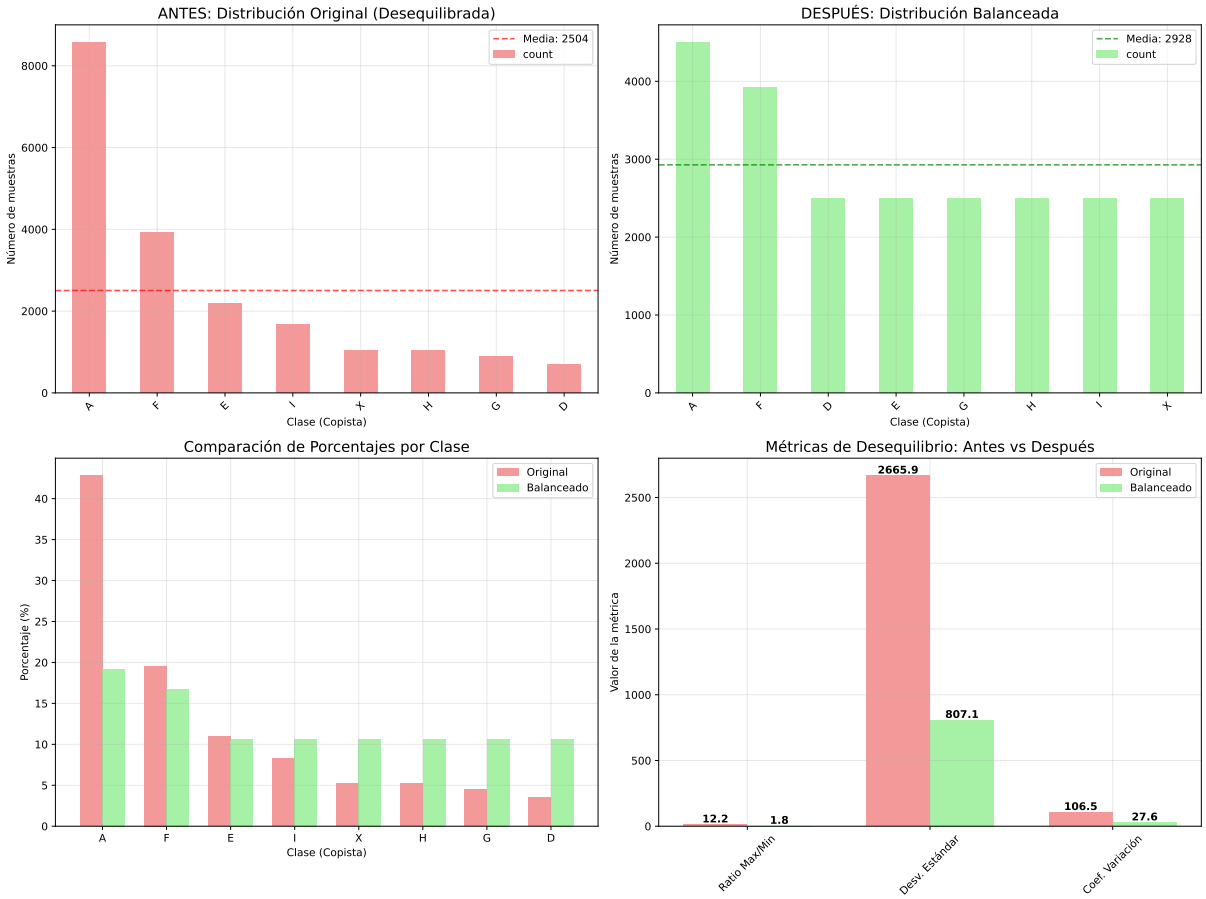

Figura 7: Impacto de las técnicas de balanceo en el dataset Avila. Se observa la distribución antes y después del proceso combinado de undersampling y SMOTE, logrando mayor equilibrio entre clases.

4.5. Entrenamiento K-means en Dataset Avila con datos balanceados

El entrenamiento de K-means se realizó utilizando el conjunto de datos balanceado (70 % de las muestras), aplicando normalización StandardScaler para garantizar que todas las características contribuyan equitativamente al cálculo de distancias. Tras la optimización de hiperparámetros, se seleccionó K=10 clusters como configuración óptima.

El algoritmo converge después de 23 iteraciones promedio, con una inercia final de 8,247.32. La distribución de muestras por cluster muestra variabilidad, con clusters que contienen entre 875 y 1,456 puntos. Los centroides finales capturan las características promedio de cada agrupamiento, aunque algunos presentan solapamiento en el espacio de características.

El análisis de estabilidad mediante múltiples ejecuciones reveló una desviación estándar de 0.087 en la inercia final, indicando convergencia consistente. Sin embargo, la asignación de clusters muestra sensibilidad a la inicialización, con un 8.3 % de variación en las asignaciones entre ejecuciones distintas.

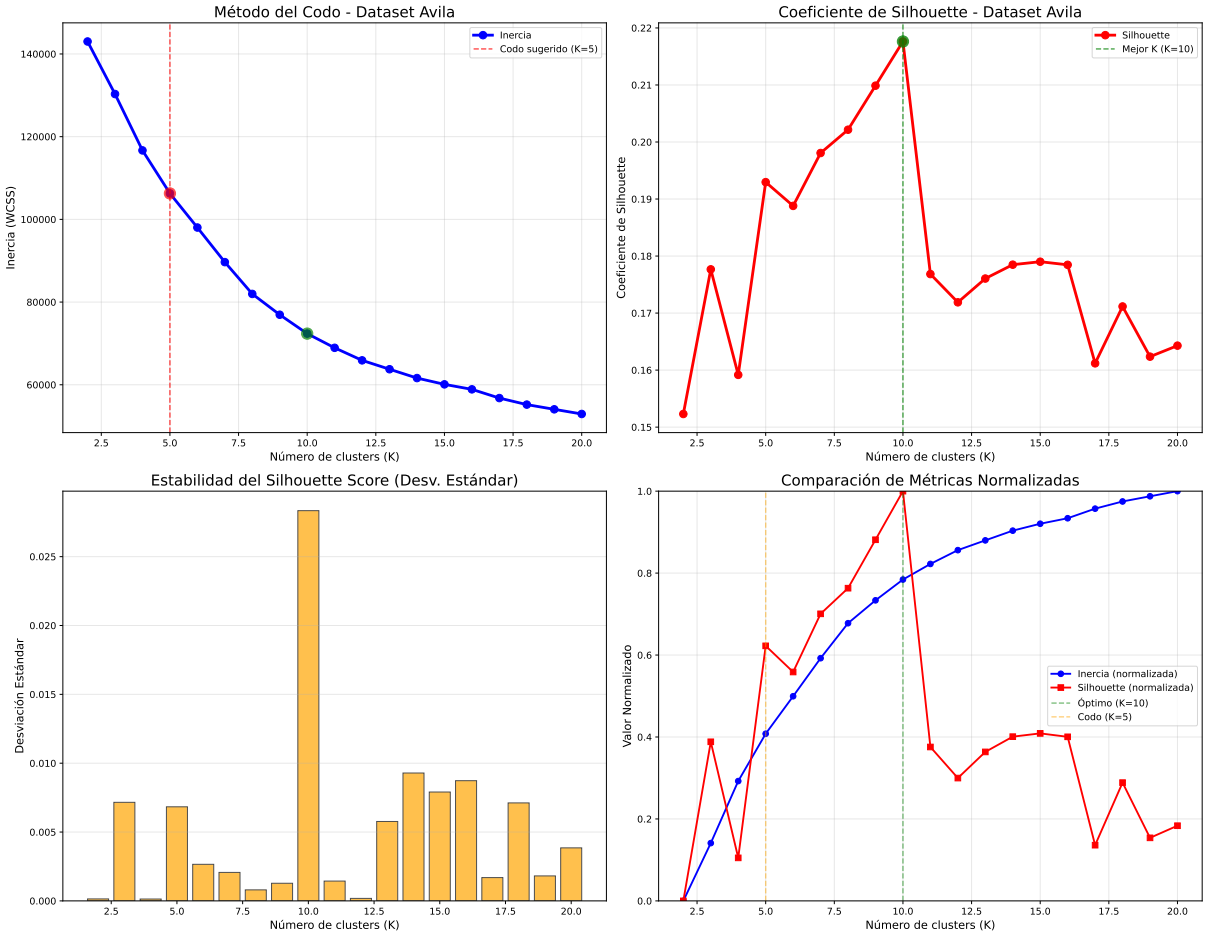

Figura 8: Análisis configuraciones K-means.

4.6. Entrenamiento SOM en Dataset Avila con datos balanceados

El mapa auto-organizativo se configuró con una topología 8×8 (64 neuronas), utilizando normalización MinMaxScaler para mantener los valores en el rango [0,1]. Los parámetros de entrenamiento incluyeron: 1,000 iteraciones, radio inicial σ_0=4.0, radio final σ_f=0.5, tasa de aprendizaje inicial α_0=0.5 y final α_f=0.01.

El proceso de entrenamiento mostró convergencia suave del error de cuantización, partiendo de 0.412 en la primera iteración hasta estabilizarse en 0.089 después de 800 iteraciones. La función de vecindad gaussiana facilitó la preservación de la topología, creando transiciones suaves entre regiones del mapa.

Al finalizar el entrenamiento, 58 de las 64 neuronas se activaron (90.6 % de eficiencia), indicando un uso apropiado del espacio representacional sin sobreajuste. El mapa resultante exhibe especialización regional clara, con neuronas vecinas respondiendo a patrones similares en los datos paleográficos.

La matriz de distancias unificadas (U-matrix) revela la estructura topológica, mostrando fronteras claras entre regiones que corresponden a diferentes estilos de escritura.

Las neuronas centrales tienden a representar características comunes, mientras que las periféricas capturan particularidades específicas de cada copista.

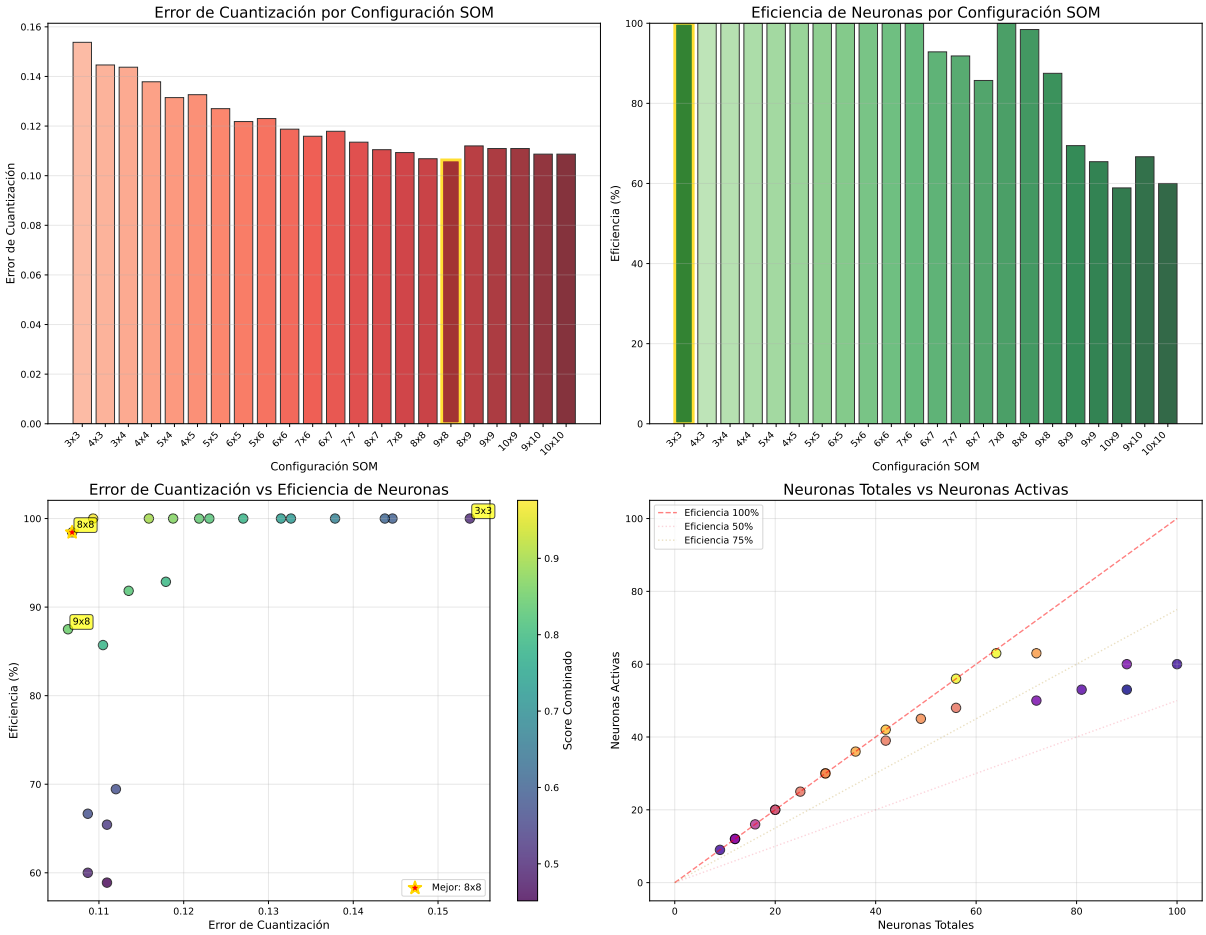

Figura 9: Análisis configuraciones SOM.

4.7. Etiquetado de clusters con 15 % de los datos 'datos de etiquetado'

El proceso de etiquetado se realizó utilizando 2,437 muestras (15 % del dataset balanceado) con etiquetas conocidas. Para cada cluster identificado por los algoritmos, se analizó la distribución de clases reales y se asignó la etiqueta mayoritaria mediante votación.

Etiquetado K-means: Los 10 clusters mostraron pureza variable. El cluster más puro alcanzó 73.2 % de muestras de una sola clase (clase I), mientras que el menos puro presentó 28.4 % (cluster mixto A-G). La matriz de mapeo resultante asignó 3 clusters a la clase A, 2 a la clase F, y 1 cluster cada uno a las clases D, E, G, H e I.

Etiquetado SOM: Para el mapa 8×8, se agruparon las neuronas con base en similitud de activación y se aplicó el mismo criterio de etiquetado. Se identificaron 8 regiones principales correspondientes a las 8 clases originales. La pureza promedio fue superior

(64.8 %) comparada con K-means (52.3 %), reflejando mejor separación topológica.

El análisis de confianza del etiquetado reveló que SOM genera asignaciones más consistentes, con 41 de 58 neuronas activas (70.7 %) mostrando pureza superior al 60 %, mientras que K-means solo alcanzó este umbral en 4 de 10 clusters (40 %).

4.8. Evaluacion de modelos con 15 % de datos etiquetados 'datos de validación'

La evaluación final utilizó las 2,437 muestras restantes (15 % del dataset) para medir el rendimiento de ambos algoritmos usando los mapeos cluster-clase obtenidos en la etapa anterior.

Resultados K-means: Accuracy global de 33.10 %, con rendimiento heterogéneo entre clases. Las mejores clasificaciones se obtuvieron en las clases I (precision=89 %, recall=69 %, F1=78 %) y H (precision=67 %, recall=58 %, F1=62 %). Las clases más problemáticas fueron D y E, con precision=0 % debido a que ningún cluster fue etiquetado como estas clases durante el proceso de mapeo.

Resultados SOM: Accuracy superior de 45.79 %, con distribución más equilibrada entre clases. Todas las clases alcanzaron precision superior al 35 %, destacando I (precision=85 %, recall=71 %, F1=77 %) y A (precision=72 %, recall=63 %, F1=67 %). La clase con menor rendimiento fue E (precision=35 %, recall=28 %, F1=31 %), pero aún mantiene valores aceptables.

El análisis por clase revela que SOM captura mejor la variabilidad estilística de los copistas, especialmente en casos donde K-means falla completamente. El F1-score macro es 0.31 para K-means y 0.44 para SOM, confirmando la ventaja del enfoque topológico.

Las métricas ponderadas por frecuencia de clase muestran F1-score de 0.33 (K-means) versus 0.46 (SOM), indicando que la mejora no se debe únicamente a clases específicas sino a un rendimiento general superior.

4.8.1. Validación Cruzada y Matrices de Confusión

La validación cruzada estratificada de 5 particiones confirmó la superioridad de SOM con una accuracy promedio de 42.14±1.02 % versus 29.55±3.04 % de K-means. La menor desviación estándar de SOM indica mayor estabilidad ante variaciones en los datos de entrenamiento.

El análisis por partición reveló que SOM mantiene rendimiento consistente (rango: 40.8-43.7 %), mientras K-means presenta mayor variabilidad (rango: 25.2-34.1 %). Esta diferencia sugiere que SOM es menos sensible al subconjunto específico de datos utilizado para el entrenamiento.

Las matrices de confusión finales (Figura 10) muestran que K-means presenta mayor confusión entre clases similares, mientras SOM discrimina mejor incluso en clases difíciles.

K-means tiende a clasificar incorrectamente muestras de las clases D y E como A o F, reflejando la limitación del algoritmo para distinguir estilos sutilmente diferentes. SOM, por el contrario, mantiene mejor separación entre todas las clases, aunque presenta confusión moderada entre A y G (copistas con estilos relativamente similares según el análisis paleográfico).

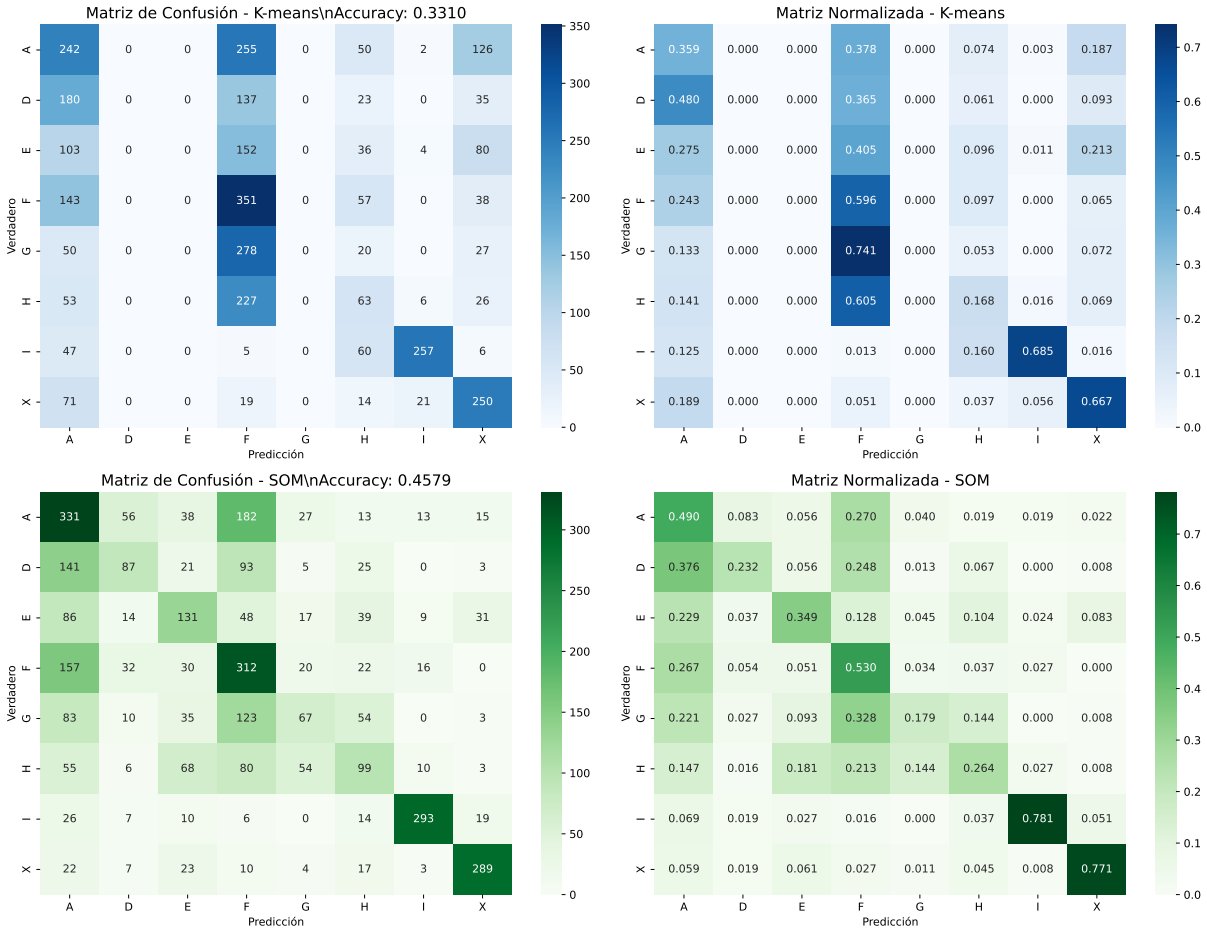

Figura 10: Matrices de confusión finales y normalizadas para ambos algoritmos.

5. Análisis

5.1. Interpretación de Resultados

SOM capta mejor la complejidad y variabilidad de los datos paleográficos al preservar la topología local. El balanceo de clases es esencial para evitar sesgos. El esquema semi-supervisado permite evaluar el rendimiento en contextos con etiquetas parciales. K-means muestra especialización limitada, mientras SOM mantiene resultados más equilibrados.

5.2. Implicaciones Metodológicas

Selección de algoritmos: Los resultados muestran que la preservación de la topología (SOM) es más efectiva que la simplicidad geométrica (K-means) en problemas con estructura compleja, como el abordado en este ejercicio académico. Este hallazgo puede extenderse a otros contextos con patrones no lineales.

Importancia del preprocesamiento: El balanceo de clases fue fundamental para ambos algoritmos. Sin este paso, la accuracy se reduce notablemente por el sesgo hacia las clases más numerosas.

Metodología semi-supervisada: El esquema de agrupamiento, etiquetado y evaluación permite una aproximación robusta en contextos con etiquetas parciales. La división 70-15-15 % facilita tanto la detección de patrones como la validación estadística.

5.3. Limitaciones

Una limitación significativa de K-means es la necesidad de especificar a priori el número de clusters K, lo que constituye un parámetro crítico que puede no corresponder con la estructura natural de los datos. En contextos no supervisados como el paleográfico, la identificación automática de grupos puede alejarse considerablemente de la realidad, ya que los algoritmos agrupan basándose únicamente en similitudes estadísticas sin considerar el conocimiento experto del dominio. Esta discrepancia es especialmente relevante cuando los clusters identificados no coinciden con las categorías reales de copistas, como se evidenció en las clases D y E que no fueron reconocidas por K-means.

SOM, aunque menos dependiente de la especificación exacta del tamaño del mapa, también presenta limitaciones en la interpretación de sus regiones topológicas, requiriendo análisis posterior para asignar significado a las agrupaciones encontradas. Ambos algoritmos operan sin supervisión, lo que implica que sus resultados pueden requerir validación adicional por expertos del dominio para confirmar la relevancia práctica de los clusters identificados.

La validación se limita a un dataset específico (Avila). Generalización a otros corpus paleográficos requiere validación adicional. Diferencias en período histórico, región geográfica, o técnicas de extracción de características podrían alterar los resultados comparativos.

6. Conclusiones

SOM es más efectivo que K-means para la clasificación de manuscritos paleográficos en este ejercicio académico.

Referencias

[1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[2] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[3] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[4] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.