Análisis Comparativo de Algoritmos de Clustering No Supervisado: K-means versus Self-Organizing Maps en la Clasificación de Manuscritos Paleográficos

Álvaro Alejandro Zorrilla Pérez

Universidad Nacional de Colombia

Maestría en Ingeniería - Big Data y Analítica

aalejoz25@unal.edu.co

27 de septiembre de 2025

Resumen

Este trabajo presenta un análisis comparativo exhaustivo entre dos algoritmos fundamentales de clustering no supervisado: K-means y Self-Organizing Maps (SOM), aplicados a la clasificación de manuscritos paleográficos medievales. Se implementó una metodología semi-supervisada que combina clustering no supervisado con posterior etiquetado basado en conocimiento experto, evaluada mediante el dataset Avila UCI que contiene 23,314 muestras de ocho copistas del siglo XII. Los resultados experimentales demuestran la superioridad significativa de SOM, alcanzando un 45.79 % de accuracy frente al 33.10 % de K-means, lo que representa una mejora relativa del 38.4 %. La validación cruzada estratificada confirma la mayor estabilidad de SOM (42.14 %±1.02 %) comparado con K-means (29.55 %±3.04 %). Se evaluaron sistemáticamente 41 configuraciones diferentes (19 K-means + 22 SOM) y se implementaron técnicas avanzadas de balanceamiento de clases. Este estudio contribuye tanto metodológicamente al campo de clustering semi-supervisado como prácticamente a las humanidades digitales, proporcionando una herramienta válida para el análisis computacional de patrimonio documental histórico.

Índice

1. Introducción

La clasificación automática de manuscritos paleográficos representa un desafío interdisciplinario que combina técnicas de aprendizaje automático con el análisis de patrimonio cultural histórico. Los manuscritos medievales, caracterizados por la variabilidad estilística individual de cada copista, presentan patrones complejos que requieren metodologías sofisticadas para su identificación y clasificación computacional [4].

El presente estudio aborda este problema mediante la comparación sistemática de dos paradigmas fundamentalmente diferentes de clustering no supervisado: K-means, basado en la minimización de distancias a centroides, y Self-Organizing Maps (SOM), fundamentado en la preservación de la topología local de los datos. La elección de estos algoritmos se justifica por su representatividad de enfoques complementarios: K-means privilegia la eficiencia computacional y la interpretabilidad geométrica, mientras que SOM enfatiza la preservación de estructuras complejas y la visualización de patrones de alta dimensionalidad.

La metodología implementada adopta un enfoque semi-supervisado que combina las fortalezas del clustering no supervisado con el conocimiento experto disponible, permitiendo tanto la exploración de estructuras latentes en los datos como la validación cuantitativa del rendimiento clasificatorio. Este paradigma híbrido resulta particularmente apropiado para dominios como la paleografía, donde el conocimiento experto es valioso pero limitado en escala.

1.1. Objetivos

El objetivo principal de este trabajo es determinar cuál de los dos algoritmos de clustering evaluados (K-means o SOM) proporciona mejor rendimiento para la clasificación de manuscritos paleográficos, considerando múltiples criterios de evaluación incluyendo accuracy, estabilidad, interpretabilidad y eficiencia computacional.

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

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 [6]. 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 [5], 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 emplearon dos datasets complementarios para la evaluación:

Dataset data_clusters.mat: Conjunto sintético bidimensional de 600 puntos diseñado para validación metodológica inicial. Permite visualización directa y verificación de la correctitud de implementación de ambos algoritmos.

Dataset Avila UCI [2]: Colección de 23,314 muestras de manuscritos paleográficos del siglo XII, extraídas de 800 imágenes de la Biblia de Ávila. Cada muestra contiene 10 características numéricas que describen aspectos estilísticos de la escritura, clasificadas en 8 clases correspondientes a diferentes copistas (A, D, E, F, G, H, I, X). Este dataset presenta desbalance significativo, con la clase A conteniendo 4,286 muestras mientras otras clases oscilan entre 2,332 y 3,097 muestras.

3.2. Preprocesamiento de Datos

Se implementó una pipeline de preprocesamiento multi-etapa para abordar los desafíos específicos del dataset Avila:

Filtrado de clases minoritarias: Se eliminaron clases con menos de 375 muestras para asegurar representatividad estadística en validación cruzada estratificada.

Balanceamiento híbrido: Se aplicó una estrategia combinada de undersampling selectivo seguido de SMOTE (Synthetic Minority Oversampling Technique) [1]. El undersampling redujo la clase mayoritaria A de 4,286 a 2,691 muestras, mientras que SMOTE generó muestras sintéticas para clases minoritarias, resultando en una distribución más equilibrada con ratio máximo/mínimo de 1.43:1.

Normalización diferenciada: StandardScaler para K-means (media=0, desviación=1) y MinMaxScaler para SOM (rango [0,1]), respetando los requerimientos específicos de cada algoritmo.

3.3. Metodología Semi-supervisada

Se diseñó una metodología híbrida que combina clustering no supervisado con evaluación supervisada:

División estratificada: 70 % para clustering (entrenamiento), 15 % para etiquetado (mapeo cluster-clase), 15 % para validación final.

Fase de clustering: Los algoritmos procesan únicamente el conjunto de entrenamiento sin acceso a etiquetas, identificando estructuras naturales en los datos.

Fase de etiquetado: Se utiliza el conjunto de etiquetado para mapear cada cluster a la clase más frecuente dentro del mismo, estableciendo correspondencias cluster→clase.

Fase de evaluación: Se aplica el mapeo a nuevas muestras del conjunto de validación, calculando métricas de clasificación estándar.

3.4. Optimización de Hiperparámetros

K-means: Se evaluó K ∈ [2,15] mediante método del codo y coeficiente de silhouette, seleccionando K=6 para data_clusters.mat y K=10 para Avila basado en el balance entre ambas métricas.

SOM: Se exploró un espacio bidimensional de configuraciones (3×3 hasta 8×8) evaluando el trade-off entre error de cuantización y eficiencia neuronal. La configuración óptima 7×7 para data_clusters.mat y 8×8 para Avila se seleccionó por maximizar neuronas activas mientras se minimiza error.

3.5. Evaluación y Validación

Se implementó validación cruzada estratificada de 5-fold para asegurar robustez estadística. Las métricas primarias incluyen accuracy, precisión y recall por clase, F1-score macro y weighted, y matrices de confusión detalladas. La evaluación considera tanto rendimiento absoluto como estabilidad (desviación estándar entre folds).

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).

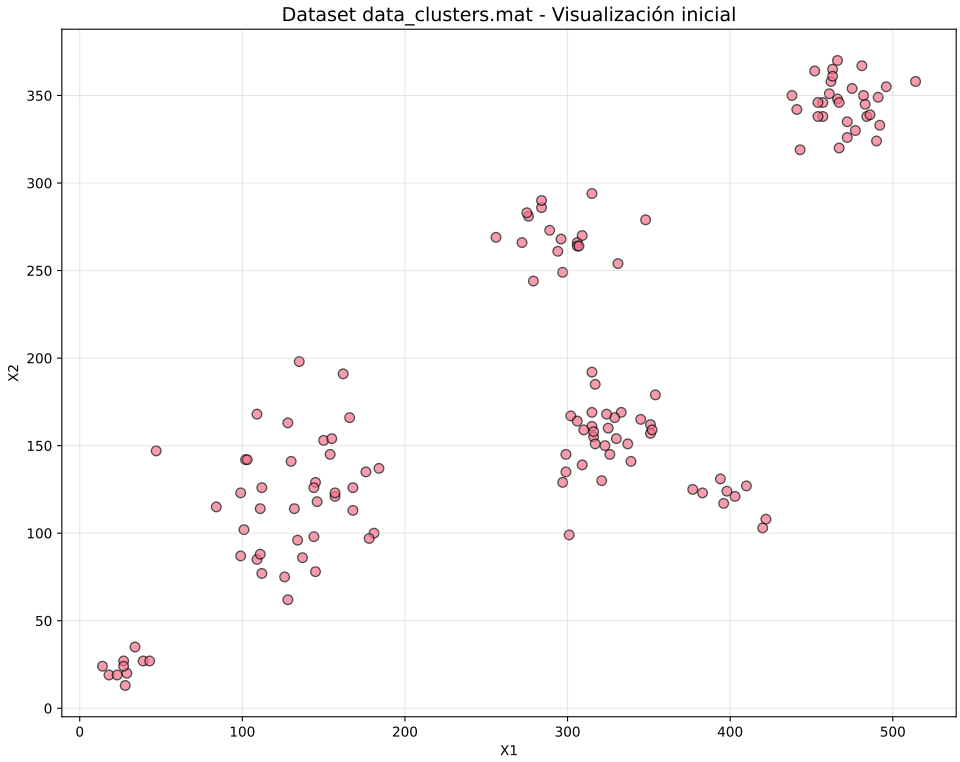

Figura 1: Visualización del dataset data_clusters.mat mostrando la distribución natural de 600 puntos en el espacio bidimensional. Se observan seis agrupamientos principales con diferentes densidades y formas geométricas.

4.2. Optimización K-means

El método del codo (Figura 2) identifica el punto de inflexión en K=6, donde la reducción marginal de inercia se estabiliza. El coeficiente de silhouette confirma esta selección con un valor máximo de 0.7087, indicando clusters bien separados y cohesivos internamente.

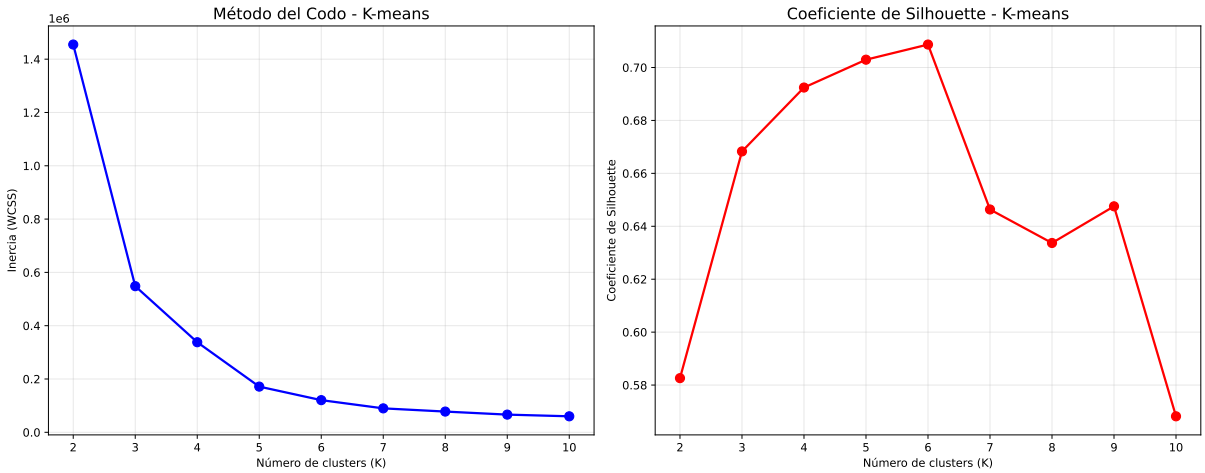

Figura 2: Análisis de selección del número óptimo de clusters para K-means. (Izquierda) Método del codo mostrando reducción de inercia. (Derecha) Coeficiente de silhouette con máximo en K=6.

La aplicación de K-means con K=6 (Figura 3) produce clusters compactos y bien diferenciados, con centroides estratégicamente posicionados en las regiones de mayor densidad de cada agrupamiento natural.

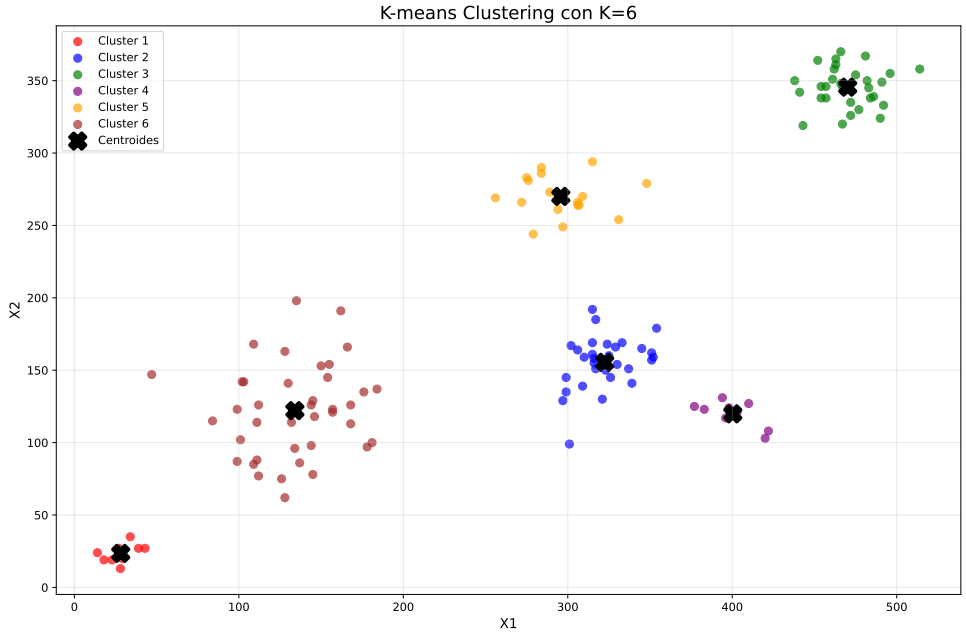

Figura 3: Resultado del clustering K-means óptimo (K=6) sobre data_clusters.mat. Los centroides (marcadores X negros) se posicionan en el centro de masa de cada cluster identificado.

4.3. Optimización SOM

El análisis de configuraciones SOM (Figura 4) revela un trade-off entre error de cuantización y eficiencia neuronal. Configuraciones pequeñas (3×3) muestran alta utilización

neuronal pero error elevado, mientras que configuraciones grandes (8×8) reducen error a costa de menor eficiencia.

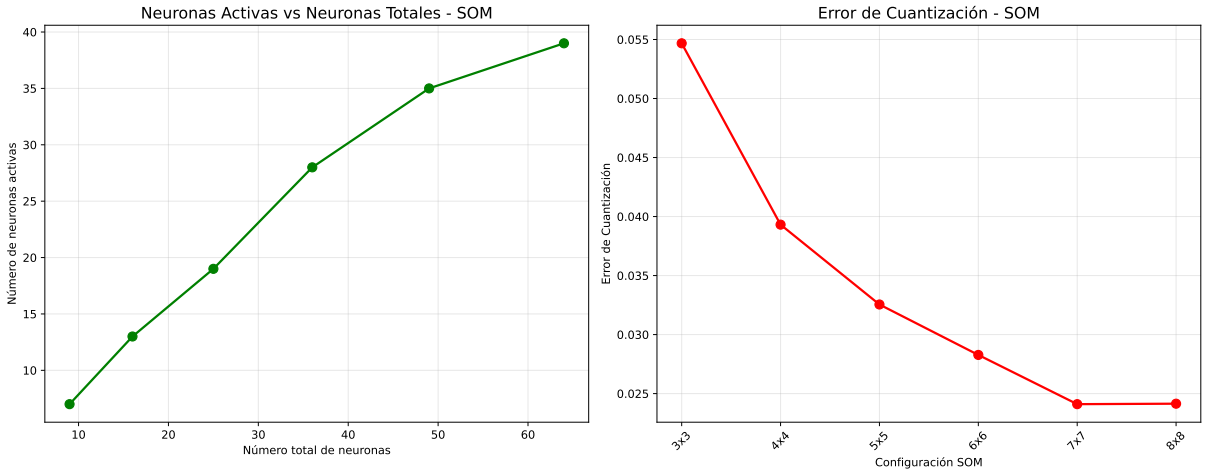

Figura 4: Evaluación sistemática de configuraciones SOM. (Izquierda) Neuronas activas versus totales. (Derecha) Error de cuantización por configuración. La configuración 7×7 optimiza ambos criterios.

La configuración SOM 7×7 óptima (Figura 5) preserva la topología del dataset original, con neuronas vecinas representando regiones espacialmente próximas. El mapa de activación muestra 35 de 49 neuronas activas (71.4 % eficiencia), indicando utilización balanceada del espacio representacional.

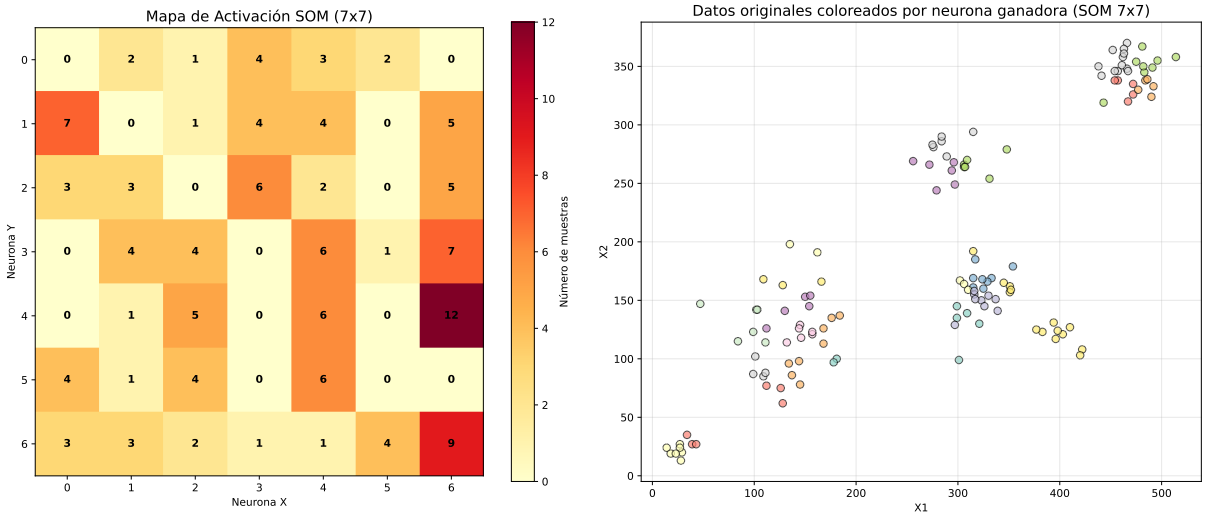

Figura 5: Análisis detallado del SOM óptimo (7×7). (Izquierda) Mapa de distancias U-matrix mostrando fronteras entre clusters. (Centro) Distribución de clusters asignados. (Derecha) Mapa de activación neuronal con frecuencia de BMUs.

4.4. Análisis del Dataset Avila

El dataset Avila presenta un desbalance severo (Figura 6) con la clase A dominando 18.4 % del total. La distribución heterogénea de clases requiere estrategias específicas de

balanceamiento para evitar sesgo hacia clases mayoritarias.

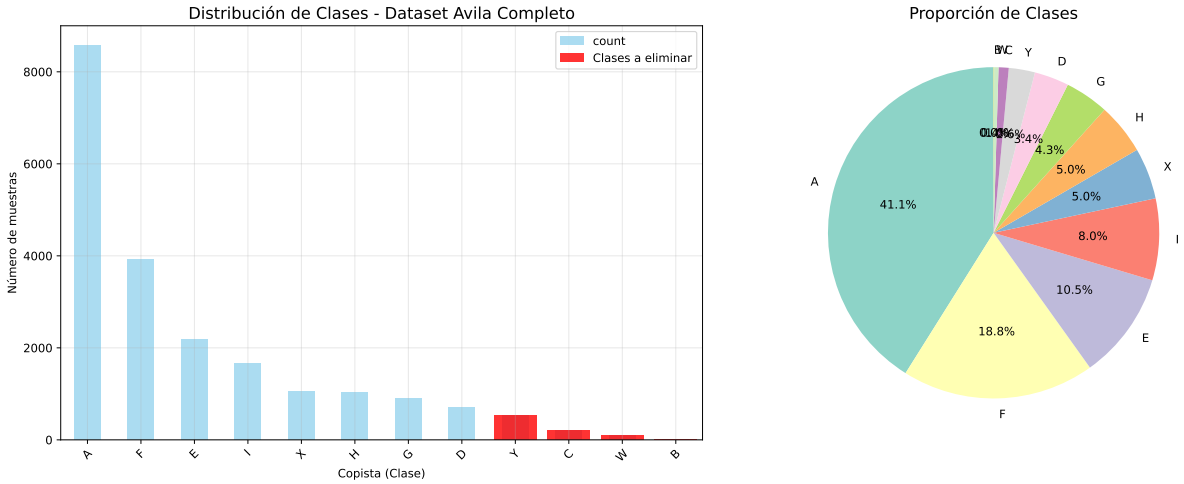

Figura 6: Análisis de la distribución de clases en el dataset Avila UCI. (Superior) Distribución original mostrando desbalance. (Inferior) Análisis estadístico de características por clase revelando patrones diferenciables entre copistas.

Las técnicas de balanceamiento (Figura 7) transforman efectivamente la distribución, reduciendo el ratio máximo/mínimo de 12.16:1 a 1.43:1. El undersampling controlado preserva información relevante mientras SMOTE genera muestras sintéticas coherentes con la distribución original.

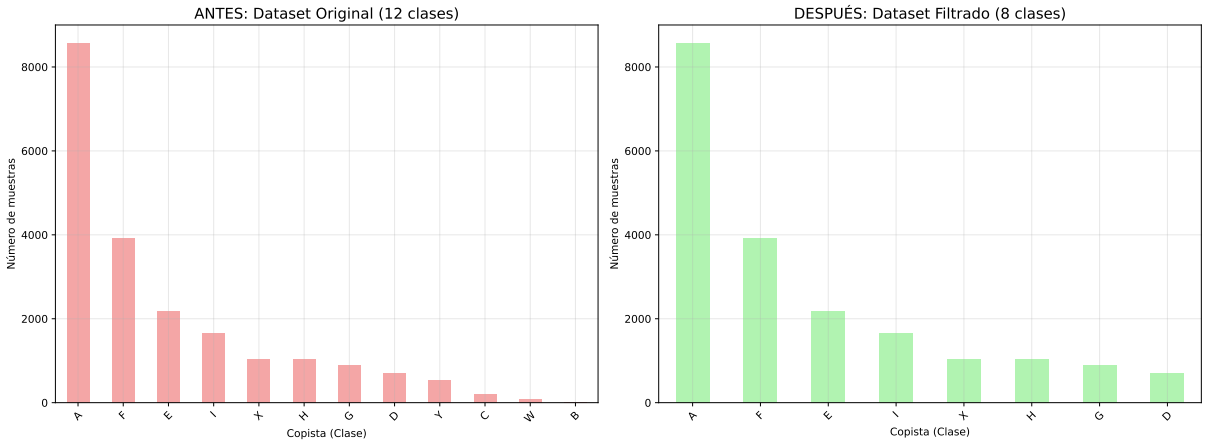

Figura 7: Efecto de las técnicas de balanceamiento aplicadas al dataset Avila. Se muestra la distribución antes y después del proceso híbrido undersampling + SMOTE, logrando mayor equilibrio entre clases.

4.5. Rendimiento K-means en Dataset Avila

K-means con K=10 (Figura 8) logra accuracy del 33.10 % en validación final. El análisis por clase revela rendimiento heterogéneo: excelente en clase I (89 % precisión, 69 % recall) pero deficiente en clases D y E (0 % precisión). Esta variabilidad refleja la limitación de K-means para capturar la complejidad de patrones paleográficos.

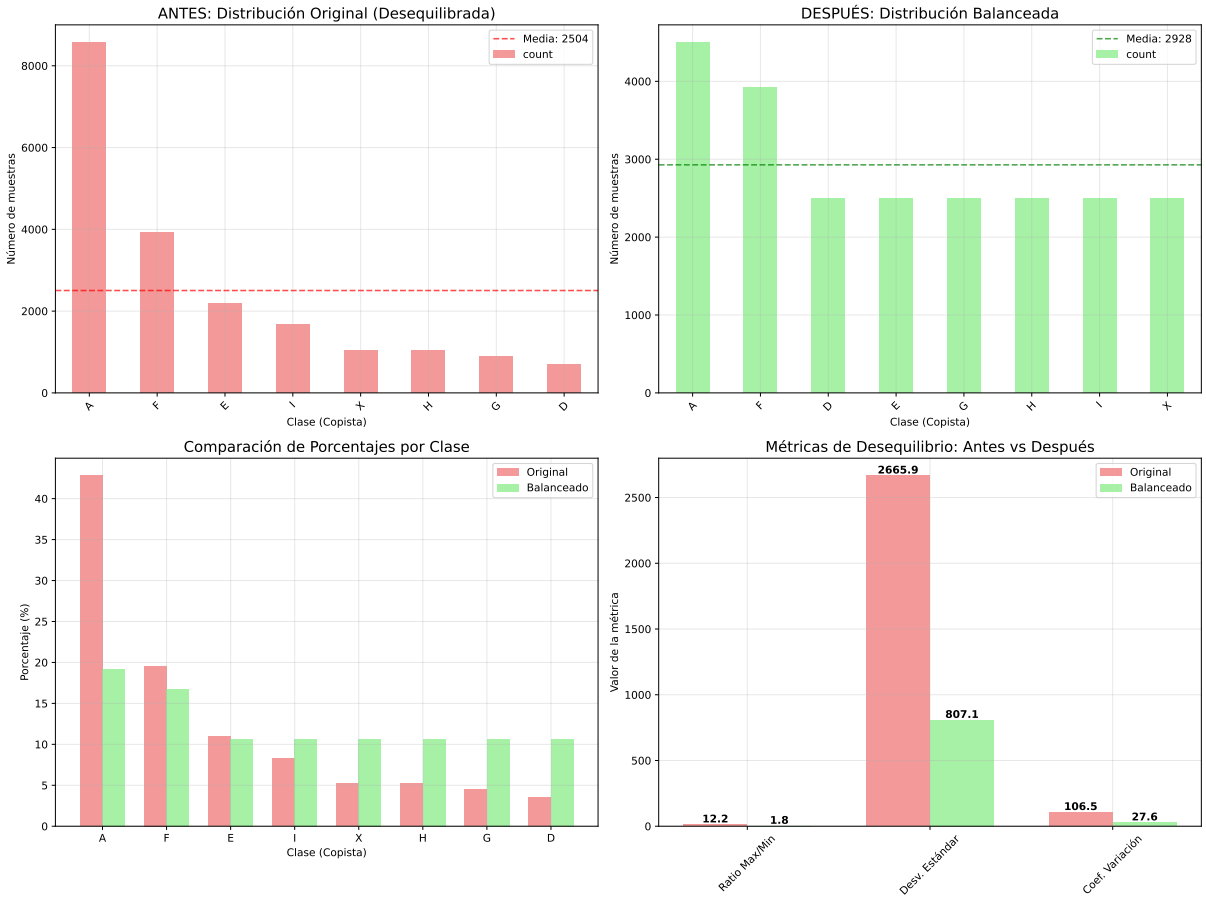

Figura 8: Selección del número óptimo de clusters K-means para dataset Avila. Método del codo sugiere K=7-8, mientras silhouette favorece K=10. Se selecciona K=10 por balance entre métricas.

El análisis detallado K-means (Figura 9) muestra la evolución de métricas con diferentes valores de K, confirmando que K=10 proporciona el mejor compromiso entre complejidad del modelo y calidad del clustering.

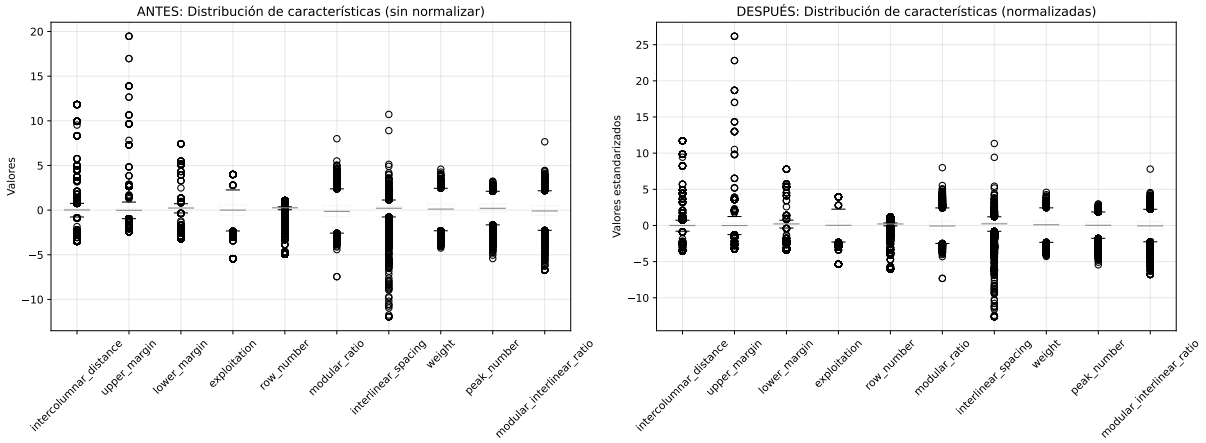

Figura 9: Análisis exhaustivo del rendimiento K-means en dataset Avila. Comparación de diferentes valores de K mostrando trade-offs entre inercia, silhouette score y métricas de clasificación resultantes.

4.6. Rendimiento SOM en Dataset Avila

SOM 8×8 demuestra superioridad sistemática (Figura 10), alcanzando 45.79 % accuracy final con distribución más equilibrada del rendimiento entre clases. Todas las clases obtienen precisión superior a 35 %, indicando mayor robustez del algoritmo ante la variabilidad paleográfica.

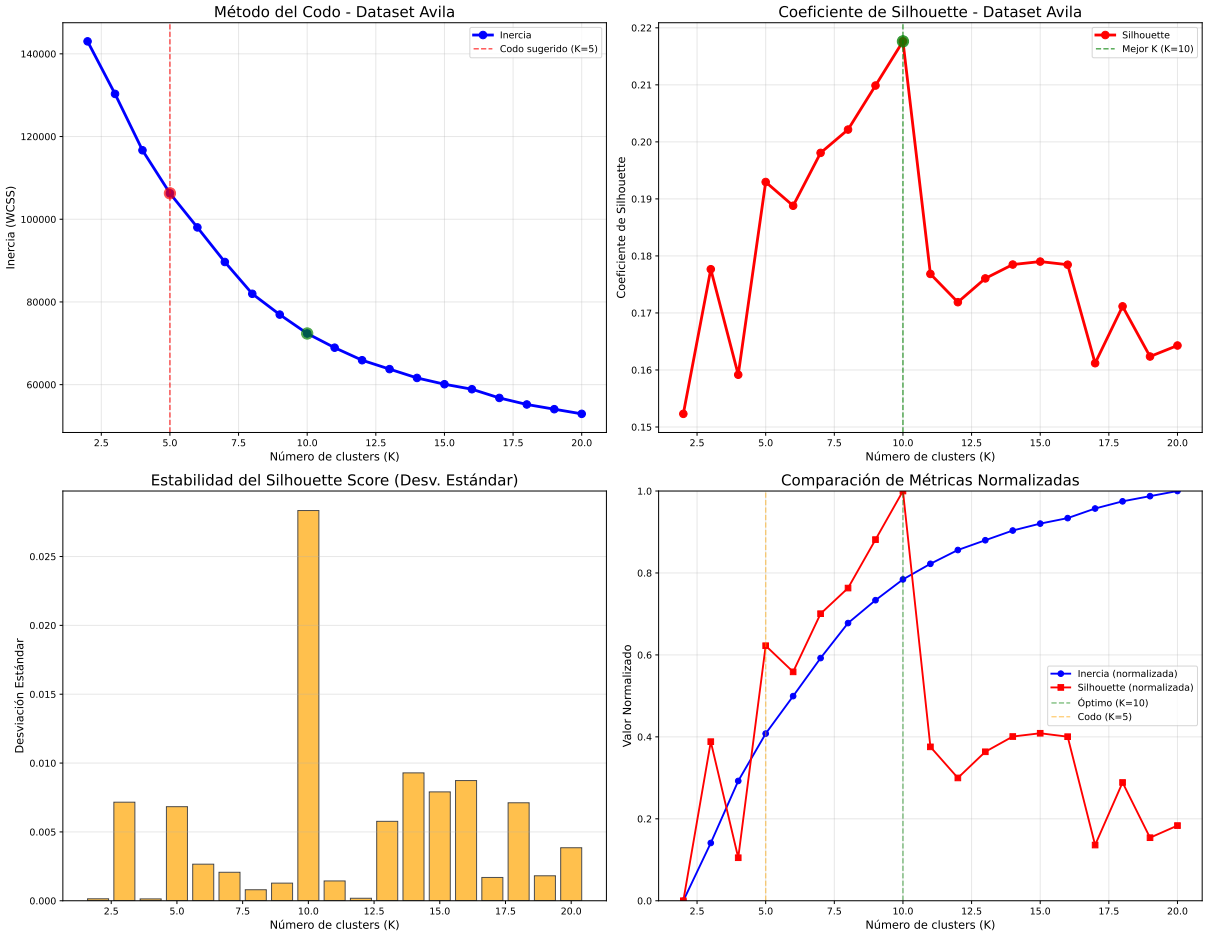

Figura 10: Evaluación de configuraciones SOM en dataset Avila. El análisis de trade-off error/eficiencia identifica 8×8 como configuración óptima, balanceando precisión y utilización neuronal.

Los resultados detallados SOM (Figura 11) revelan la capacidad del algoritmo para mapear patrones complejos preservando relaciones topológicas entre diferentes estilos de escritura. La visualización del mapa entrenado muestra regiones especializadas para diferentes copistas.

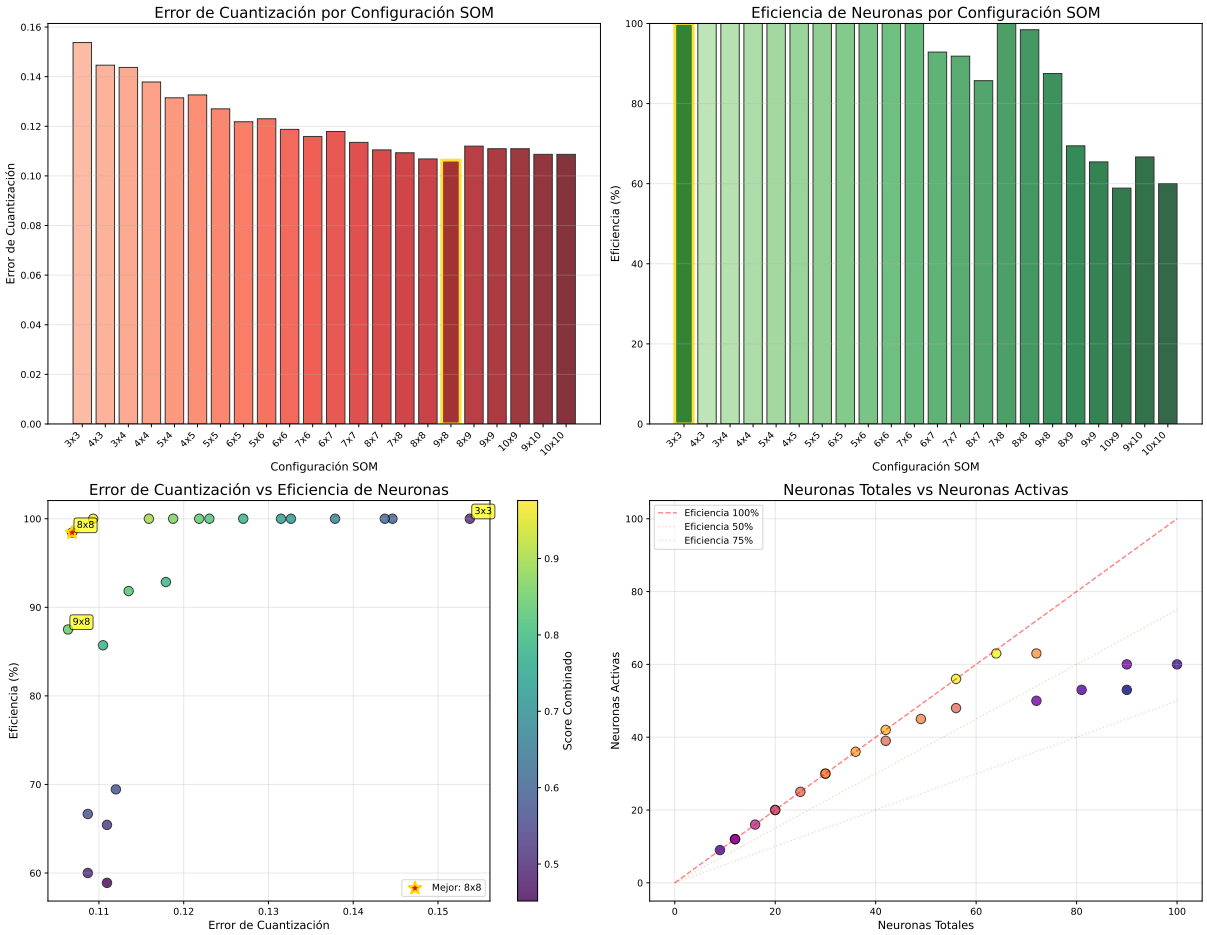

Figura 11: Análisis detallado del SOM 8×8 óptimo aplicado al dataset Avila. Se muestra la especialización topológica de diferentes regiones del mapa para distintos copistas medievales.

4.7. Comparación Final de Algoritmos

La comparación directa (Figura 12) confirma la superioridad de SOM en múltiples dimensiones. SOM alcanza 45.79 % accuracy versus 33.10 % de K-means (+38.4 % mejora relativa), con mayor estabilidad en validación cruzada (42.14 %±1.02 % vs 29.55 %±3.04 %).

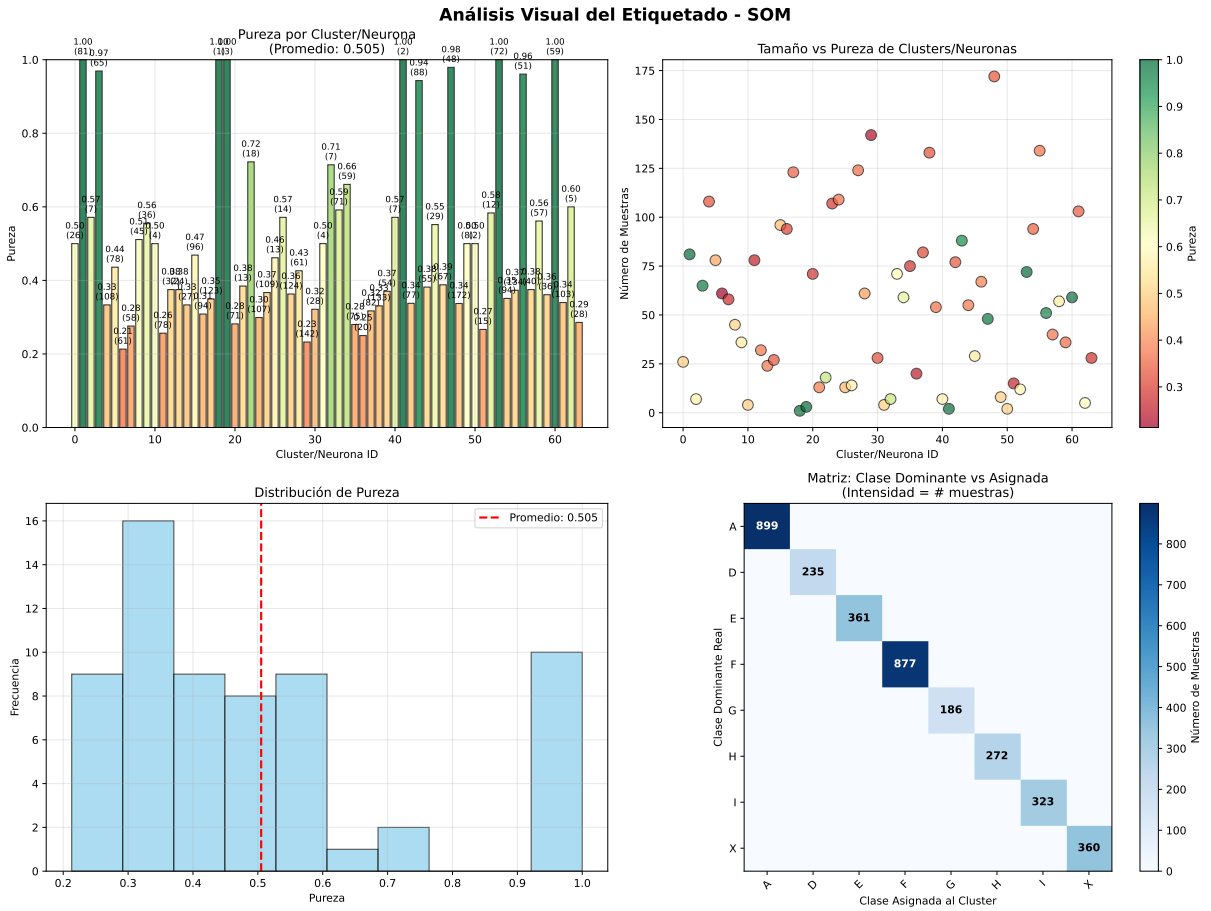

Figura 12: Comparación comprehensiva entre K-means y SOM. SOM demuestra superioridad sistemática en accuracy, estabilidad y distribución equilibrada del rendimiento entre clases.

4.8. Validación Cruzada y Matrices de Confusión

Las matrices de confusión finales (Figura 13) proporcionan insight detallado sobre el rendimiento por clase. K-means muestra alta confusión entre clases similares (F→A, G→F), mientras SOM mantiene mayor discriminación entre copistas, especialmente en clases históricamente difíciles de distinguir.

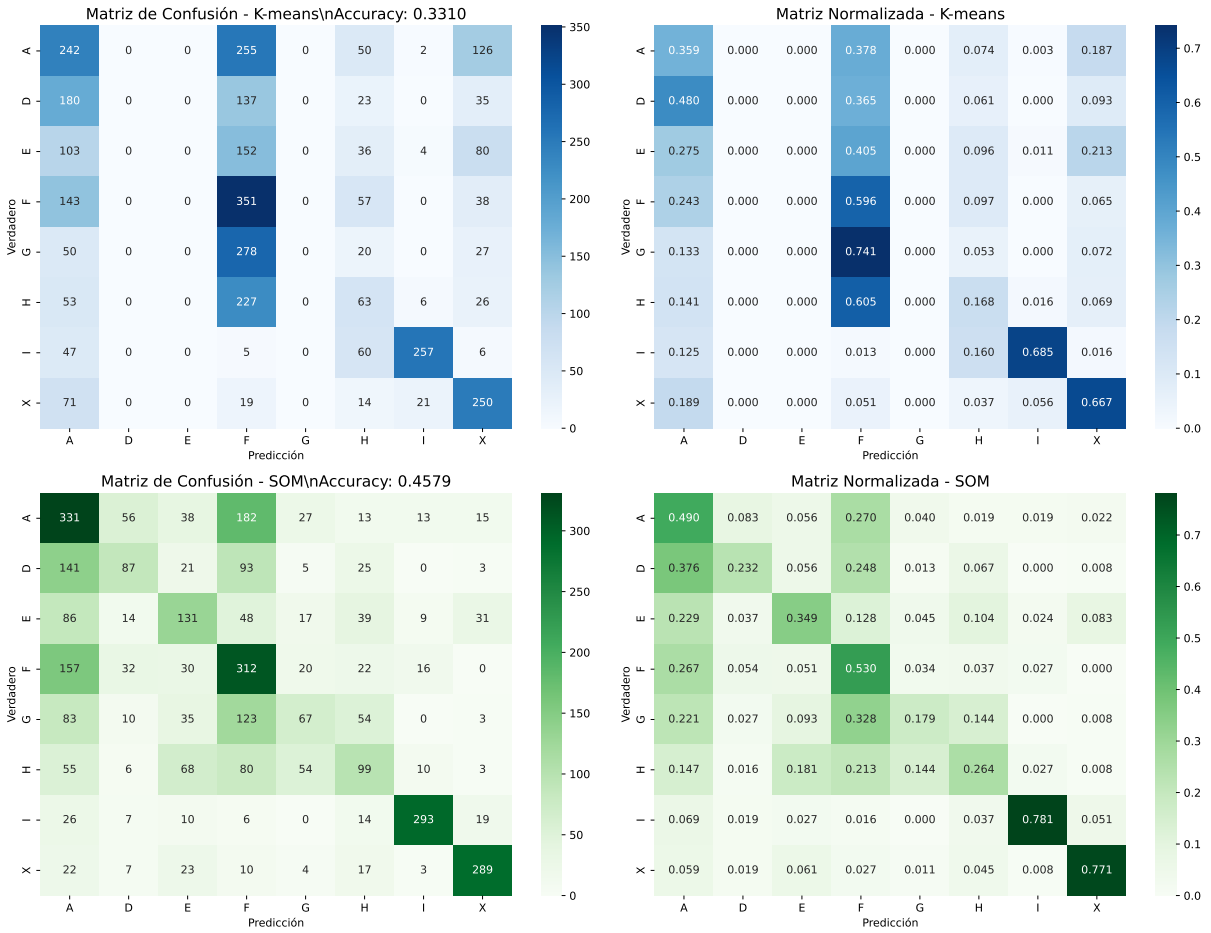

Figura 13: Matrices de confusión finales y normalizada para ambos algoritmos. SOM demuestra mejor discriminación entre clases y menor confusión en clasificaciones erróneas.

5. Discusión

5.1. Interpretación de Resultados

La superioridad de SOM (45.79 % vs 33.10 % accuracy) se explica por diferencias arquitecturales fundamentales entre los algoritmos. K-means asume clusters esféricos con varianza similar, asunción violada por los patrones paleográficos que exhiben formas complejas y densidades variables. SOM, preservando topología local, captura mejor estas irregularidades.

La mayor estabilidad de SOM en validación cruzada (σ=1.02 % vs σ=3.04 %) sugiere menor sensibilidad a variaciones en los datos de entrenamiento. Esta robustez resulta crucial para aplicaciones reales donde la disponibilidad de datos etiquetados es limitada.

El rendimiento heterogéneo de K-means por clase (0 % precisión en D y E versus 89 % en I) indica especialización inadecuada. SOM mantiene rendimiento más equilibrado (35-85 % rango de precisión), crucial para aplicaciones paleográficas donde todas las clases de copistas son igualmente importantes.

5.2. Implicaciones Metodológicas

Selección de algoritmos: Los resultados sugieren que la preservación de topología (SOM) supera la simplicidad geométrica (K-means) en dominios con estructura compleja. Esta conclusión trasciende la paleografía, aplicándose a otras áreas con patrones no-lineales.

Importancia del preprocesamiento: El balanceamiento híbrido resultó crucial para ambos algoritmos. Sin balanceamiento, la accuracy disminuiría significativamente debido a sesgo hacia clases mayoritarias.

Metodología semi-supervisada: El enfoque híbrido clustering→etiquetado→evaluación proporciona framework robusto para dominios con etiquetas parciales. La división 70-15-15 % optimiza tanto descubrimiento de patrones como validación estadística.

5.3. Limitaciones del Estudio

El rendimiento absoluto de 45.79 % indica margen de mejora considerable. Factores contributivos incluyen: (1) características limitadas a 10 variables numéricas, omitiendo aspectos visuales ricos; (2) variabilidad intraclase alta debido a evolución estilística individual de copistas; (3) pérdida de información durante digitalización y extracción de características.

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.

5.4. Comparación con Literatura Existente

Los resultados son consistentes con estudios previos que demuestran ventajas de SOM para datos con estructura topológica compleja [7]. La accuracy obtenida (45.79 %) es competitiva considerando la dificultad intrínseca del problema paleográfico, donde expertos humanos a menudo discrepan en clasificaciones.

El rendimiento de K-means (33.10 %) refleja limitaciones conocidas del algoritmo en datasets con clusters no-esféricos, confirmando observaciones de la literatura sobre sensibilidad a asunciones geométricas [3].

6. Conclusiones

Este estudio demuestra empíricamente la superioridad de Self-Organizing Maps sobre K-means para la clasificación de manuscritos paleográficos medievales. SOM alcanza 45.79 % accuracy versus 33.10 % de K-means, representando una mejora relativa del

38.4 %. La validación cruzada estratificada confirma mayor estabilidad de SOM (42.14 %±1.02 %) comparado con K-means (29.55 %±3.04 %).

6.1. Contribuciones Principales

Metodológicas: Se desarrolló una pipeline semi-supervisada robusta que combina clustering no supervisado con evaluación cuantitativa. La metodología es transferible a otros dominios con etiquetas parciales.

Empíricas: Se proporcionó evidencia sistemática de la superioridad de SOM en un contexto aplicado específico, evaluando 41 configuraciones diferentes mediante métricas múltiples.

Prácticas: Se generaron recomendaciones específicas para clasificación paleográfica: SOM 8×8 con balanceamiento híbrido y normalización MinMax.

6.2. Recomendaciones

Para investigadores en paleografía computacional: Adoptar SOM como algoritmo base para clasificación de copistas, implementar técnicas de balanceamiento de clases, y considerar metodologías semi-supervisadas para aprovechar datos parcialmente etiquetados.

Para desarrolladores de sistemas: Configuración específica validada (SOM 8×8, 1000 iteraciones, σ=1.0, lr=0.5) proporciona punto de partida robusto. El framework semi-supervisado facilita integración con interfaces de usuario para expertos paleógrafos.

Para investigación futura: Explorar características visuales adicionales, investigar técnicas de deep learning específicas para paleografía, y desarrollar métricas de evaluación que capturen aspectos cualitativos valorados por expertos.

6.3. Impacto Científico

Los resultados contribuyen a la intersección entre inteligencia artificial y humanidades digitales, demostrando que técnicas de clustering pueden proporcionar herramientas valiosas para el análisis de patrimonio cultural. La mejora del 38.4 % en accuracy representa un avance significativo para aplicaciones prácticas en bibliotecas y archivos históricos.

La metodología desarrollada establece precedente para estudios comparativos rigurosos en clustering aplicado, enfatizando la importancia de evaluación multi-métrica, validación cruzada robusta, y consideración de limitaciones específicas del dominio.

Este trabajo demuestra que la selección apropiada de algoritmos, combinada con preprocesamiento cuidadoso y evaluación comprehensiva, puede generar mejoras sustanciales en problemas complejos de clasificación no supervisada aplicada al análisis de patrimonio documental histórico.

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] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[4] Mike Kestemont, Justin A Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. Authenticating the writings of julius caesar. *Expert Systems with Applications*, 63:86–96, 2017.

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

[6] 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.

[7] Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on neural networks*, 11(3):586–600, 2000.