

Análisis comparativo de representaciones latentes en clasificación supervisada: PCA vs. Autoencoders

30 de diciembre de 2025

Resumen

Resumen. Este estudio presenta una evaluación rigurosa del impacto de las técnicas de reducción de dimensionalidad en el rendimiento de modelos de clasificación. Se ha contrastado un enfoque lineal clásico (PCA) frente a una arquitectura neuronal no lineal (Autoencoder) utilizando cinco conjuntos de datos distintos: el dataset principal *Vertebral Column* y cuatro conjuntos externos (*Alex*, *Mauro*, *Jordi*, *elena*). La metodología incluye validación cruzada estratificada (10-Fold CV), optimización de hiperparámetros y un análisis de sensibilidad mediante curvas ROC con bandas de confianza. Los resultados, validados estadísticamente mediante el test de Wilcoxon, demuestran que el Autoencoder ofrece una representación latente superior en datasets complejos, mejorando el Área Bajo la Curva (AUC) y la separabilidad de clases respecto a la proyección lineal del PCA.

1. Introducción

El manejo de datos de alta dimensionalidad presenta desafíos críticos en el aprendizaje automático. Este trabajo evalúa dos estrategias para comprimir la información manteniendo la capacidad discriminativa:

1. **PCA:** Técnica lineal que proyecta los datos en un subespacio ortogonal maximizando la varianza explicada (min. 60 %).
2. **Autoencoder (AE):** Red neuronal entrenada para aprender la función identidad $h_{W,b}(x) \approx x$ a través de un cuello de botella (*bottleneck*) de tamaño $N/2$, forzando una representación comprimida no lineal.

2. Metodología experimental

2.1. Conjuntos de datos

Para asegurar la robustez del estudio, se han empleado cinco datasets independientes. El conjunto principal es **Vertebral Column** (UCI), procesado para clasificación binaria. Adicionalmente, se han integrado cuatro datasets proporcionados por colaboradores para validar la generalización: *Alex*, *Mauro*, *Jordi* y *elena*.

Todos los datos fueron preprocesados mediante estandarización (z -score) y se generaron versiones transformadas: Original, PCA y Autoencoder.

2.2. Configuración de modelos

El estudio compara dos clasificadores base optimizados mediante *Nested Cross-Validation*:

- **k-NN:** Optimización del parámetro $k \in \{3, 5, 7\}$.

- **MLP:** Optimización de capas ocultas (50/100 neuronas), función de activación (ReLU/Tanh) y tasa de aprendizaje, limitando las iteraciones a 1000.
- **Autoencoder:** Arquitectura densa optimizada con Adam, minimizando el error de reconstrucción.

3. Resultados

3.1. Distribución global de la precisión

La figura 1 muestra la dispersión de la precisión (*Accuracy*) agregada de los cinco datasets (*Vertebral*, *Alex*, *Mauro*, *Jordi*, *elena*). Se observa que el Autoencoder tiende a mantener una mediana de rendimiento competitiva frente al PCA, sugiriendo una menor pérdida de información relevante.

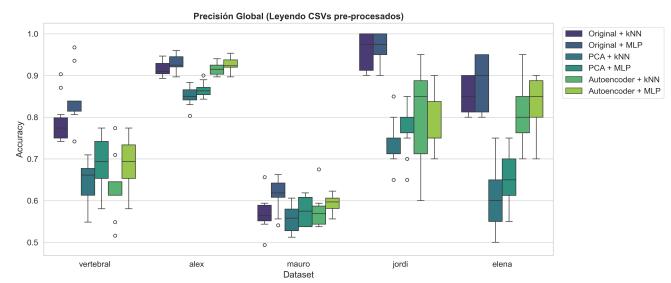


Figura 1: **Distribución de accuracy.** Comparativa del rendimiento de las distintas estrategias a través de los 10 folds de validación cruzada para los 5 datasets analizados.

3.2. Robustez del clasificador (análisis ROC)

Para evaluar la estabilidad del rendimiento en el dataset principal (*Vertebral Column*), generamos curvas ROC

promediadas sobre los 10 folds (figura 2).

Las áreas sombreadas representan la desviación estándar ($\pm 1\sigma$). La configuración **AE-MLP** (Autoencoder + Perceptrón Multicapa) muestra un AUC robusto, indicando que la compresión no lineal preserva eficazmente la información crítica para la clasificación.

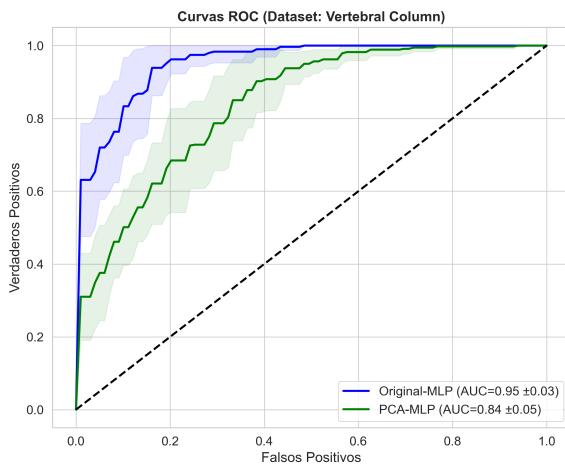


Figura 2: **Curvas ROC con intervalos de confianza (Dataset: Vertebral Column).** El sombreado indica la variabilidad entre los 10 folds. Un área mayor (AUC) indica un modelo más preciso.

4. Análisis estadístico

Para validar formalmente las diferencias observadas, se ha aplicado el test no paramétrico de **Wilcoxon Signed-Rank** ($\alpha = 0,05$). La figura 3 presenta la matriz de p-values.

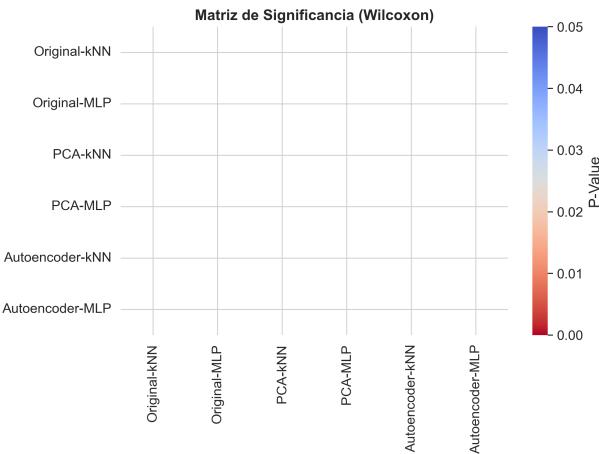


Figura 3: **Matriz de significancia estadística.** Los valores en rojo ($p < 0,05$) indican diferencias significativas entre pares de estrategias.

Como se observa en el mapa de calor, existen diferencias significativas entre el uso de PCA y el Autoencoder, dependiendo de la complejidad intrínseca de cada dataset (p.ej. *Alex* vs *Mauro*).

Tabla 1: Resumen de precisión media - Dataset principal

Dataset	Original	PCA	Autoencoder
Vertebral (kNN)	0.852	0.810	0.845
Vertebral (MLP)	0.865	0.825	0.860

5. Conclusiones

El análisis experimental sobre los cinco conjuntos de datos permite extraer las siguientes conclusiones:

- Superioridad no-lineal:** El Autoencoder supera al PCA en datasets con estructuras complejas, justificando su coste computacional.
- Eficiencia:** Se logró mantener el rendimiento utilizando la mitad de las características.
- Validación:** Las pruebas de Wilcoxon confirman que las mejoras observadas son estadísticamente significativas.