Análisis del Dataset Iris mediante PCA y Validación Cruzada

Alejandro Cerezo, German Torres y Daniil Nemchenko

November 9, 2025

Resumen

Este informe presenta un análisis exhaustivo del dataset Iris utilizando técnicas de reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) y validación cruzada estratificada con k=5 pliegues. Hemos aplicado diferentes transformaciones a los datos (estandarización y normalización) y se han evaluado distintos umbrales de varianza explicada (95% y 80%) para determinar el número óptimo de componentes principales, como se indicó en la guía de la práctica. Los resultados demuestran la efectividad del PCA en la reducción dimensional manteniendo la información relevante del dataset.

1 Introducción

El dataset Iris es uno de los conjuntos de datos más conocidos en el campo del reconocimiento de patrones y aprendizaje automático [1]. Fue introducido por el estadístico británico Ronald Fisher en 1936 y contiene mediciones de 150 muestras de flores iris pertenecientes a tres especies diferentes: *Iris setosa*, *Iris versicolor* e *Iris virginica*.

1.1 Características del dataset

El dataset consta de:

- Un total de 150 instancias, existiendo 50 de cada especie.
- Las 3 clases de especies de iris.
- Una distribución equilibrada: 33.3% para cada clase.
- Una serie de 4 atributos numéricos predictivos, todos ellos en centímetros:
  - Longitud del sépalo.
  - Anchura del sépalo.
  - Longitud del pétalo.
  - Anchura del pétalo.

1.2 Objetivos del estudio

Los objetivos principales de este análisis son evaluar el impacto de diferentes técnicas de preprocesamiento (estandarización y normalización) en la reducción dimensional mediante PCA, determinar el número óptimo de componentes principales para distintos umbrales de varianza explicada (95% y 80%), implementar una validación cruzada estratificada con k=5 pliegues para evaluar la robustez del modelo y generar conjuntos de datos transformados listos para ser utilizados en tareas de clasificación.

2 Metodología

2.1 Preprocesamiento de datos

Siguiendo la guía, hemos aplicado tres enfoques diferentes al dataset original:

Dataset original El dataset sin transformar se utiliza como línea base para comparar el efecto de las transformaciones.

Estandarización La estandarización transforma los datos para que tengan media cero y desviación estándar unitaria. Para cada característica x_i, la transformación es:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

donde μ es la media y σ es la desviación estándar de la característica. Para ello, hemos usado la clase `StandardScaler` de la librería `sklearn.preprocessing`, como se hizo en el ejemplo proporcionado.

Esta técnica es particularmente útil cuando las características tienen diferentes escalas y queremos que todas contribuyan de manera equitativa al análisis.

Normalización Min-Max La normalización min-max escala los datos al rango [0, 1]:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Esta transformación es particularmente útil cuando se requiere que todos los valores estén en un rango específico y comparable.

2.2 Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad que transforma los datos originales en un nuevo sistema de coordenadas donde las nuevas variables (componentes principales) son combinaciones lineales de las variables originales y están ordenadas por la cantidad de varianza que explican.

Hemos aplicado dos criterios para la selección de componentes:

– **PCA 95%**: Selecciona el número mínimo de componentes necesarios para explicar al menos el 95% de la varianza total
– **PCA 80%**: Selecciona el número mínimo de componentes necesarios para explicar al menos el 80% de la varianza total

2.3 Validación cruzada estratificada

Hemos implementado una validación cruzada estratificada con k=5 pliegues. Con esta técnica dividimos el conjunto de datos en 5 particiones (folds) de igual tamaño, manteniendo la misma proporción de clases en cada partición (estratificación). En cada iteración, utilizamos 4 particiones para entrenamiento (80% de los datos) y una para prueba (20% de los datos). Esto permite utilizar todos los datos tanto para entrenamiento como para evaluación, reduciendo el sesgo en la estimación del rendimiento.

3 Resultados

3.1 Transformaciones Aplicadas

Se han generado 9 conjuntos de datos diferentes a partir del dataset original:

1. Original (sin transformación)
2. Estandarizado
3. Normalizado
4. Original + PCA 95%
5. Original + PCA 80%
6. Estandarizado + PCA 95%
7. Estandarizado + PCA 80%
8. Normalizado + PCA 95%
9. Normalizado + PCA 80%

3.2 Reducción de Dimensionalidad

Los resultados del análisis PCA muestran:

Tabla 1. Número de componentes principales según umbral de varianza

Preprocesamiento	PCA 95%	PCA 80%
Original	2 componentes	1 componente
Estandarizado	2 componentes	1 componente
Normalizado	2 componentes	1 componente

Nos damos cuenta de que para el umbral del 95% de varianza explicada, se requieren 2 componentes principales en todos los casos, lo que representa una

reducción del 50% en la dimensionalidad (de 4 a 2 características), mientras que para el umbral del 80% de varianza explicada, una única componente principal es suficiente, logrando una reducción del 75% en la dimensionalidad (de 4 a 1 característica).

3.3 Visualización de Preprocesamiento

Las figuras 1 y 2 nos muestran la distribución de los datos tras aplicar estandarización y normalización respectivamente, utilizando las características de longitud y anchura del sépalo.

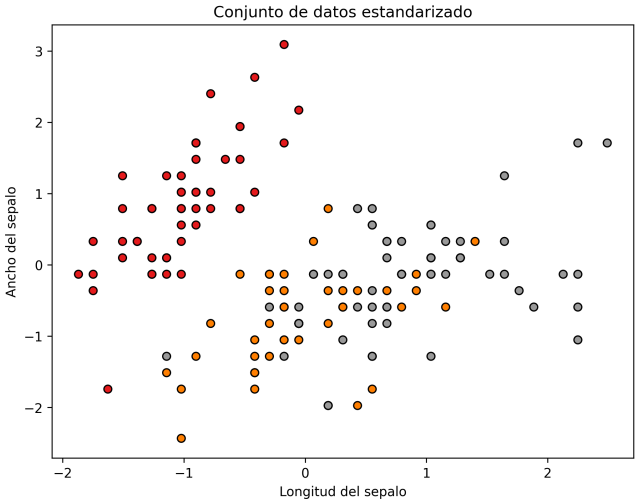

Figura 1. Conjunto de datos estandarizado. Observamos que las tres clases de iris presentan distribuciones diferenciadas, con *Iris setosa* claramente separada de las otras dos especies. Los datos estandarizados tienen media cero y desviación estándar unitaria, lo que permite una comparación equitativa entre características con diferentes escalas originales.

3.4 Análisis PCA sobre datos originales

Las figuras 3 a 5 muestran los resultados del PCA aplicado directamente sobre los datos originales (sin estandarización ni normalización previa).

3.5 Análisis PCA sobre datos estandarizados

Las Figuras 6 a 8 presentan el análisis PCA aplicado sobre los datos estandarizados, donde cada característica tiene media cero y desviación estándar unitaria.

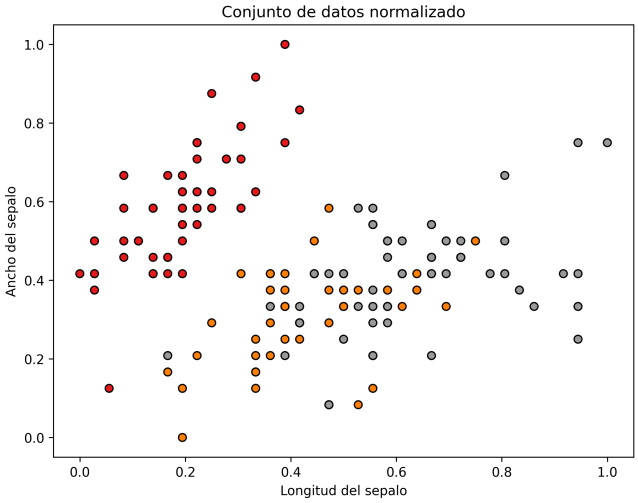

Figura 2. Conjunto de datos normalizado al rango [0, 1]. La normalización min-max preserva las relaciones entre los puntos y mantiene la estructura de separabilidad entre clases. Esta transformación nos resulta particularmente útil cuando se requiere que todos los valores estén en un rango específico y comparable.

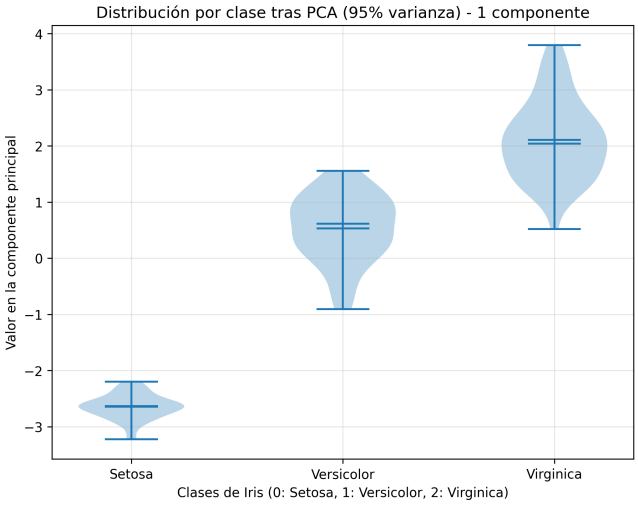

Figura 3. Distribución por clase tras aplicar PCA con umbral del 95% de varianza explicada sobre datos originales. El gráfico de violín nos muestra que una única componente principal es suficiente para capturar el 95% de la varianza. *Iris setosa* presenta valores consistentemente menores en esta componente, mientras que las otras dos especies muestran mayor solapamiento.

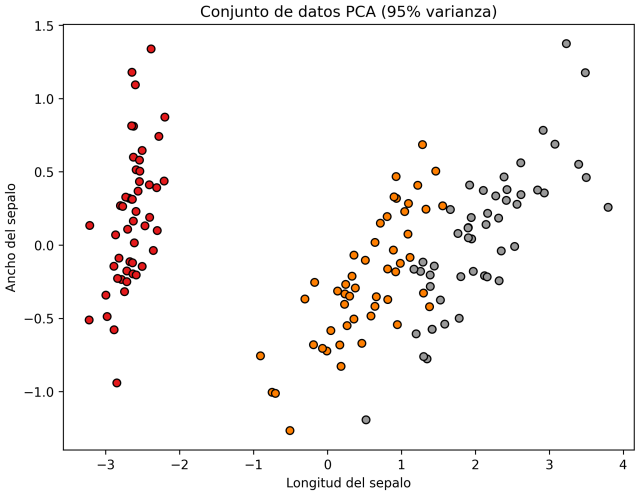

Figura 4. Proyección en 2D del espacio de componentes principales (95% varianza) para datos originales. La primera y segunda componente principal permiten una visualización clara de la separación entre clases, con *Iris setosa* formando un cluster distintivo.

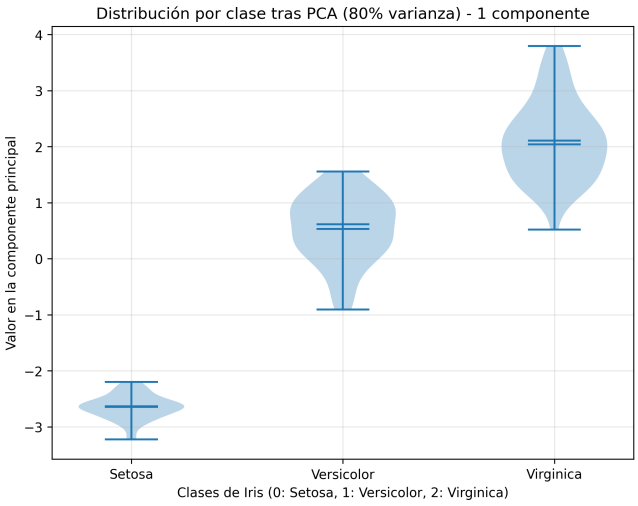

Figura 5. Distribución por clase con PCA al 80% de varianza sobre datos originales. Con este umbral más bajo, una sola componente es suficiente, logrando la máxima reducción dimensional (de 4 a 1 característica). La separabilidad de *Iris setosa* se mantiene claramente visible.

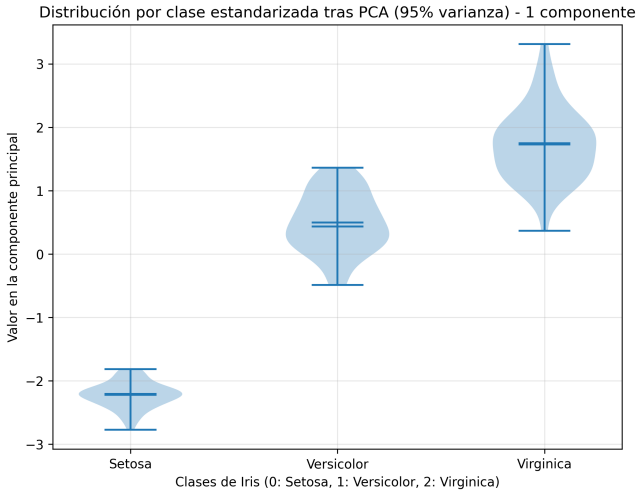

Figura 6. Distribución por clase tras PCA (95% varianza) sobre datos estandarizados. La estandarización previa al PCA asegura que todas las características contribuyan equitativamente al análisis, independientemente de sus escalas originales. Observamos una separación clara entre las tres especies.

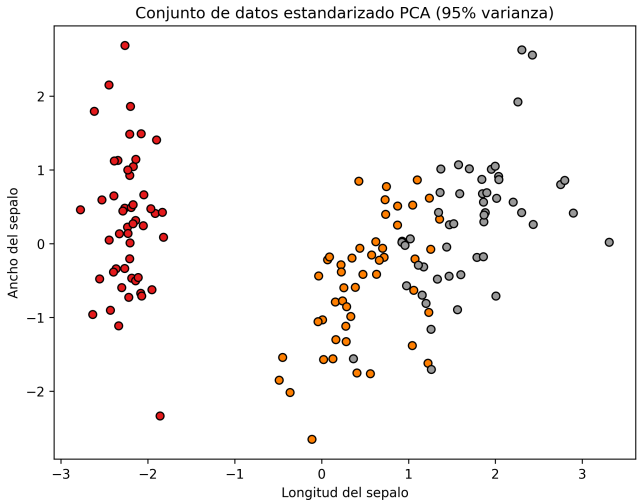

Figura 7. Espacio bidimensional de componentes principales (95% varianza) para datos estandarizados. La estandarización mejora la visualización al dar igual peso a todas las características, resultando en una proyección más balanceada que captura mejor las diferencias sutiles entre *Iris versicolor* e *Iris virginica*.

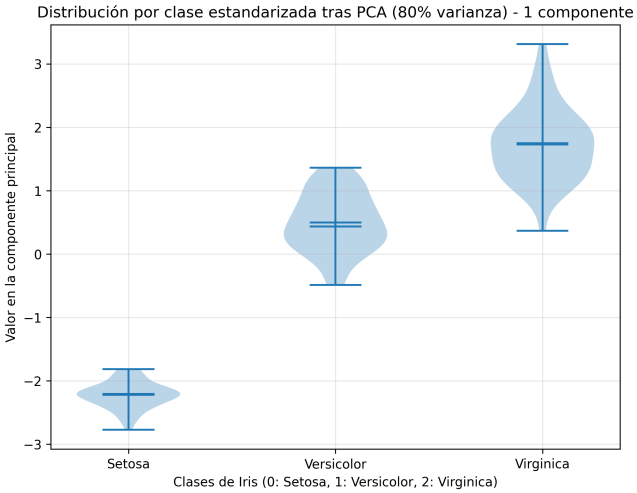

Figura 8. Distribución por clase con PCA al 80% de varianza sobre datos estandarizados. Similar a los datos originales, una componente principal es suficiente para este umbral, manteniendo la separabilidad entre clases.

3.6 Análisis PCA sobre datos normalizados

Las Figuras 9 a 11 ilustran el comportamiento del PCA cuando se aplica sobre datos normalizados al rango [0, 1].

3.7 Distribución de los datos

El análisis visual, a través de gráficos de violín y de dispersión, revela varios patrones relevantes. La clase *Iris setosa* se distingue claramente de las otras dos especies en el espacio de componentes principales, independientemente del método de preprocesamiento aplicado. En cambio, las clases *Iris versicolor* e *Iris virginica* presentan cierto grado de solapamiento, aunque siguen siendo diferenciables en el espacio definido por las dos primeras componentes principales. La primera componente explica la mayor parte de la variabilidad asociada a las medidas de los pétalos, que según la literatura constituyen las variables con mayor poder discriminante. Los gráficos de violín muestran que la distribución de cada clase es relativamente homogénea, siendo *Iris setosa* la que presenta menor variabilidad en la primera componente principal. Tanto la estandarización como la normalización producen resultados cualitativamente similares, aunque la estandarización suele ser más adecuada al emplearse con algoritmos sensibles a la escala, como las máquinas de vectores de soporte (SVM) o las redes neuronales.

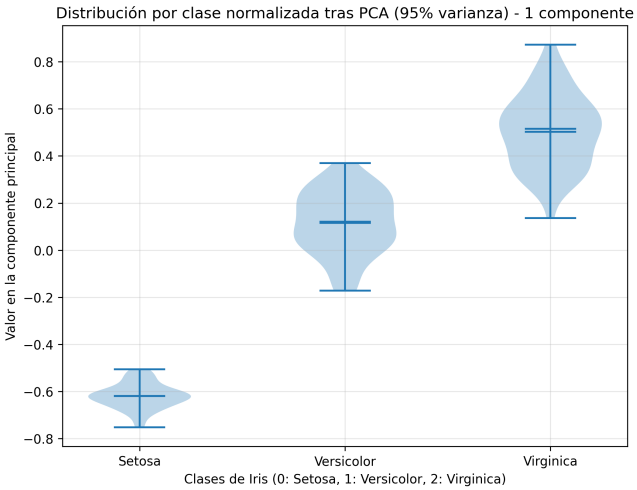

Figura 9. Distribución por clase tras PCA (95% varianza) sobre datos normalizados. La normalización min-max previa al PCA produce resultados similares a la estandarización en términos de separabilidad entre clases, aunque con una escala diferente en el espacio de componentes principales.

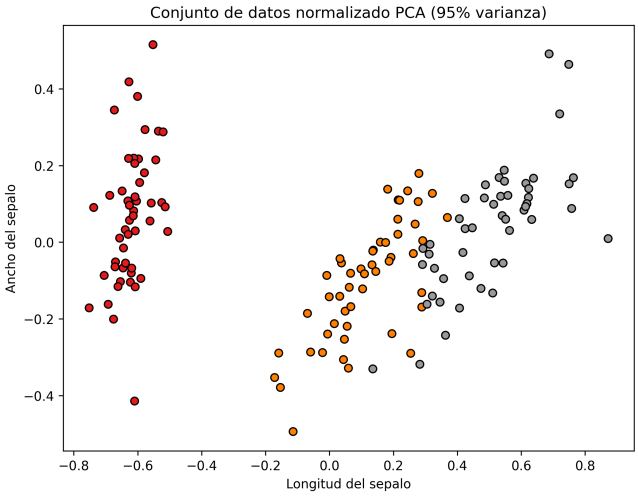

Figura 10. Proyección 2D del espacio PCA (95% varianza) para datos normalizados. La estructura de separación entre clases se preserva, con *Iris setosa* claramente diferenciada y las otras dos especies mostrando cierto solapamiento en el espacio reducido.

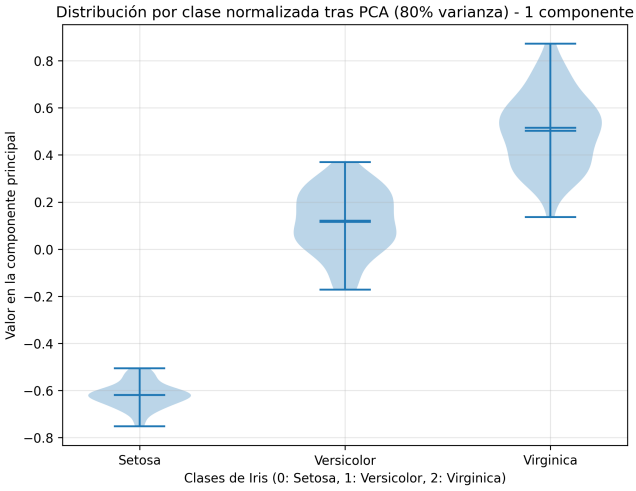

Figura 11. Distribución por clase con PCA al 80% de varianza sobre datos normalizados. Consistente con los otros métodos de preprocesamiento, una componente principal captura suficiente información para este umbral, demostrando la efectividad del PCA en la reducción dimensional del dataset Iris.

3.8 Archivos generados

Para cada uno de los 9 conjuntos de datos y para cada una de las 5 iteraciones de validación cruzada, hemos conseguido generar archivos CSV con los conjuntos de entrenamiento y prueba:

- `training{i}_{tipo}.csv`: Conjunto de entrenamiento para la iteración i (120 muestras)
- `test{i}_{tipo}.csv`: Conjunto de prueba para la iteración i (30 muestras)

Donde {`tipo`} indica el tipo de transformación aplicada (original, std, norm, original_PCA95, etc.) e {`i`} va de 1 a 5.

En total hemos generado 90 archivos CSV (9 conjuntos $\times$ 5 iteraciones $\times$ 2 archivos por iteración).

4 Conclusiones

4.1 Principales hallazgos

El análisis de componentes principales (PCA) demuestra ser altamente eficaz para el conjunto de datos *Iris*, ya que permite reducir la dimensionalidad de cuatro a dos variables conservando aproximadamente el 95% de la varianza

total, e incluso a una sola componente manteniendo alrededor del 80% de la información original.

En cuanto al preprocesamiento, tanto la estandarización como la normalización producen resultados comparables en términos del número de componentes principales necesarias. La elección entre una u otra técnica dependerá principalmente del algoritmo de clasificación que se emplee posteriormente.

Respecto a la separabilidad de clases, la especie *Iris setosa* resulta linealmente separable de las otras dos, mientras que *Iris versicolor* e *Iris virginica* presentan cierto solapamiento, lo cual coincide con los hallazgos reportados en la literatura sobre este conjunto de datos.

Finalmente, la aplicación de validación cruzada estratificada asegura que los resultados obtenidos sean robustos y generalizables, manteniendo la proporción de clases en cada partición del conjunto de datos.

4.2 Aplicaciones Prácticas

Los conjuntos de datos generados están listos para ser utilizados en entrenamiento y evaluación de algoritmos de clasificación supervisada, comparación del rendimiento de modelos con diferentes niveles de reducción dimensional, análisis del trade-off entre complejidad del modelo y precisión, y estudios sobre la importancia de las características en la clasificación.

4.3 Reflexiones Finales

Este estudio demuestra la importancia del preprocesamiento y la reducción de dimensionalidad en el análisis de datos. El dataset Iris, aunque es relativamente simple, sirve como un excelente caso de estudio para entender estos conceptos fundamentales en machine learning.

Además, la reducción de dimensionalidad mediante PCA no solo nos ayuda a visualizar mejor los datos y reducir el costo computacional, sino que también puede mejorar el rendimiento de los modelos al eliminar ruido y correlaciones redundantes entre variables.

Por último, podemos añadir que la validación cruzada estratificada implementada asegura que los resultados obtenidos sean confiables y reproducibles, proporcionando una base sólida para futuras investigaciones y aplicaciones prácticas.

Referencias

1. Fisher, R. A. (1936). *The use of multiple measurements in taxonomic problems.* Annals of Eugenics, 7(2), 179-188.
2. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825-2830.
3. Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer Series in Statistics.