Taller 3b – Análisis de Componentes Principales (PCA)

Estudiante: Juan Felipe Gómez Carmona
Maestría en Analítica Aplicada
Universidad de La Sabana
Profesor: Hugo Franco, Ph.D.

Octubre 2025

Índice

1. Introducción **2**

2. Metodología **2**
2.1. Dataset y Preprocesamiento 2
2.2. Procedimiento 2

3. Resultados **3**
3.1. Iris Dataset 3
3.2. Distribución y Varianza – Wine Dataset 4
3.3. Visualización PCA 2D – Wine Dataset 5
3.4. Desempeño de Modelos 5

4. Discusión **5**

5. Conclusiones **6**

6. Referencias **6**

1. Introducción

Este taller tiene como propósito aplicar la técnica de **Análisis de Componentes Principales (PCA)** dentro de un flujo de trabajo de aprendizaje supervisado. A través de los datasets *Iris* y *Wine Quality*, se busca analizar el efecto de la reducción de dimensionalidad sobre el rendimiento de modelos *k-NN*, tanto en problemas multiclase como binarios.

El objetivo central es comprender cómo el PCA transforma los datos originales en nuevas variables (componentes principales) que concentran la mayor cantidad de varianza posible, reduciendo la complejidad sin sacrificar significativamente la capacidad predictiva.

2. Metodología

2.1. Dataset y Preprocesamiento

Iris Dataset: contiene 150 muestras de tres especies de flores (*Setosa, Versicolor, Virginica*) con cuatro características morfológicas. Se utiliza para demostrar la reducción de dimensionalidad y la visualización 2D mediante PCA.

Wine Quality Dataset: incluye 1599 muestras de vino tinto, con 11 variables fisicoquímicas y una etiqueta de calidad numérica. Esta variable se transformó en tres categorías: *Poor, Fair y Good*, y posteriormente en una clasificación binaria (*Good* ¿6).

Los datos fueron divididos en entrenamiento (80 %) y prueba (20 %), aplicando escalamiento con `StandardScaler` antes del PCA.

2.2. Procedimiento

El flujo de trabajo seguido fue:

1. Entrenar un modelo base (*k-NN*) sin PCA.
2. Aplicar PCA y repetir el entrenamiento con 2 componentes.
3. Evaluar el desempeño mediante *accuracy, recall, precision.*
4. Visualizar la varianza explicada y las proyecciones 2D.
5. Extender el análisis al caso binario (*Good ¿6*).

3. Resultados

3.1. Iris Dataset

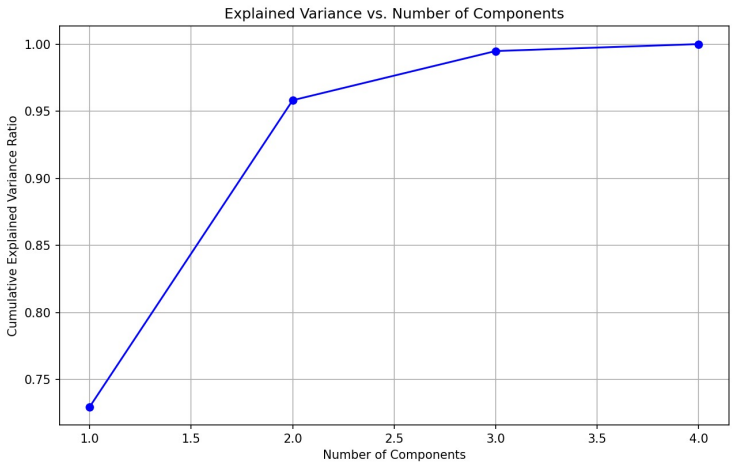

Figura 1: Varianza explicada acumulada según el número de componentes (Iris Dataset).

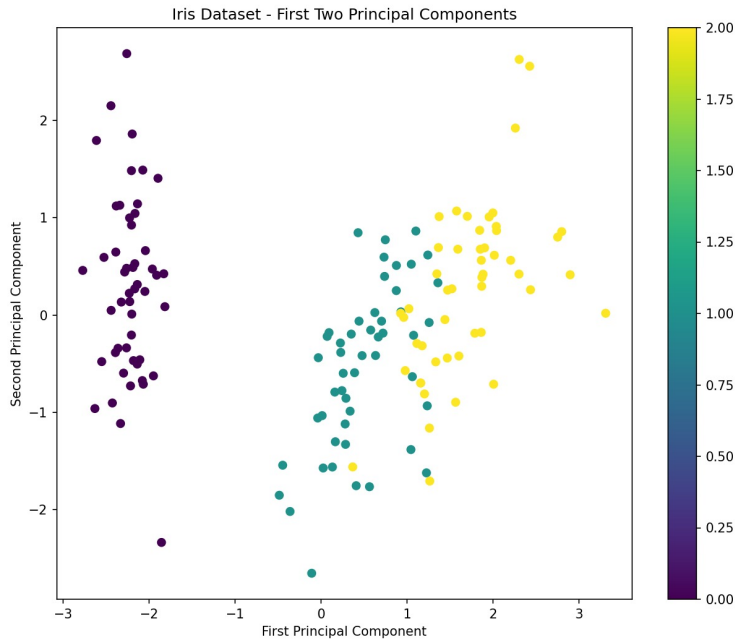

Figura 2: Proyección 2D del Iris Dataset en los dos primeros componentes principales.

El modelo PCA alcanzó una precisión similar al modelo base, mostrando que la información principal del Iris se retiene en las dos primeras dimensiones (˜95 % de la varianza). Esto evidencia que PCA es altamente eficiente cuando las variables originales están fuertemente correlacionadas, como ocurre en este conjunto.

3.2. Distribución y Varianza – Wine Dataset

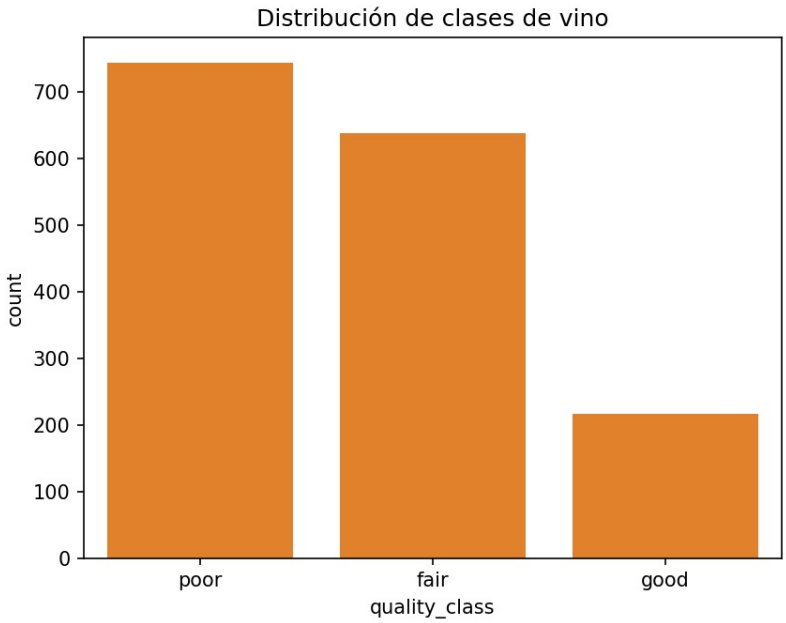

Figura 3: Distribución de clases en el dataset de vino tinto.

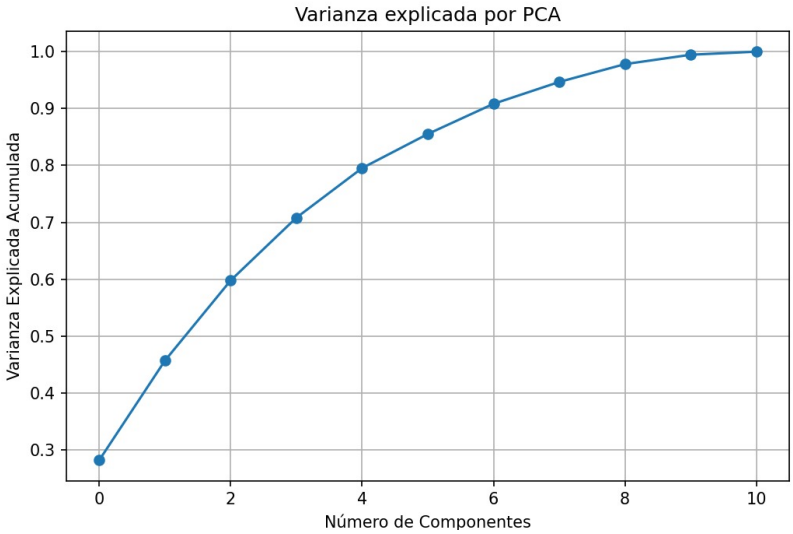

Figura 4: Varianza explicada acumulada por número de componentes (Wine Dataset).

Las dos primeras componentes retienen aproximadamente el 46 % de la varianza total, lo que implica una pérdida de información relevante para la clasificación. A diferencia del Iris, este conjunto presenta alta multicolinealidad y mayor complejidad, por lo que PCA no logra capturar de manera compacta toda la estructura informativa.

3.3. Visualización PCA 2D – Wine Dataset

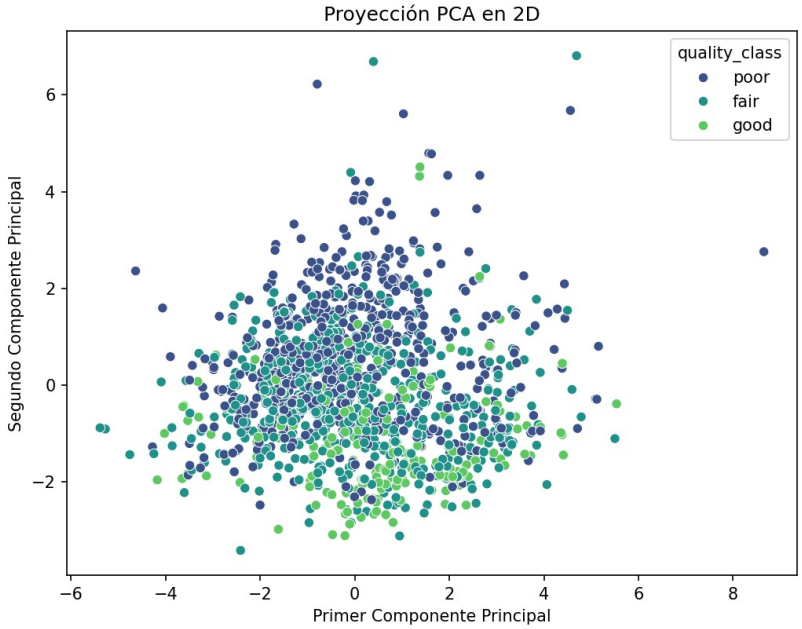

Figura 5: Proyección PCA en 2D del Wine Dataset.

Las clases *poor*, *fair* y *good* presentan una alta superposición, reflejando la dificultad del modelo para separar clases de calidad similares en un espacio reducido de dos dimensiones. Esto refuerza la idea de que PCA, aunque útil para exploración visual, no siempre mejora el poder discriminativo.

3.4. Desempeño de Modelos

Modelo	**Exactitud (Accuracy)**	**Tipo de Clasificación**
Base (sin PCA)	0.55	Multiclase
Con PCA (2 comps.)	0.52	Multiclase
Base Binario	0.86	Binaria
PCA Binario (2 comps.)	0.86	Binaria

Cuadro 1: Comparación general de desempeño entre modelos con y sin PCA.

En la clasificación binaria, los modelos mejoran considerablemente su exactitud (86 %), confirmando que la simplificación del problema reduce el error de clasificación.

4. Discusión

Los resultados evidencian que el PCA:

- Facilita la **visualización** y simplifica el modelo, aunque con pérdida de información relevante en datasets complejos.
- No mejora el rendimiento predictivo en el problema multiclase, ya que las relaciones no lineales entre variables no se capturan adecuadamente en componentes lineales.
- En la versión binaria, conserva una precisión alta, manteniendo eficiencia computacional sin comprometer el rendimiento.

Además, al comparar los resultados entre los dos conjuntos de datos, se pueden notar diferencias claras:

- En el caso del **Iris Dataset**, los datos están bien organizados y las clases se distinguen con facilidad. Esto hace que el PCA funcione muy bien, ya que con solo dos componentes logra conservar casi toda la información importante del conjunto.
- En cambio, el **Wine Quality Dataset** es mucho más complejo. Las clases no están tan separadas y las variables se relacionan de forma más enredada. Al reducir los datos a solo dos componentes, se pierde parte de la información necesaria para distinguir bien los vinos de diferente calidad.
- Esto muestra que el PCA no siempre es la mejor opción. Su utilidad depende mucho del tipo de datos que se esté analizando y del objetivo del trabajo: si se busca entender mejor los datos y visualizarlos, el PCA ayuda bastante; pero si lo que se necesita es la máxima precisión posible, puede no ser suficiente por sí solo.

5. Conclusiones

- El PCA es útil para exploración y reducción de dimensionalidad, aunque no siempre mejora la exactitud de los modelos predictivos.
- En el caso multiclase, la pérdida de información fue significativa (solo 46 % de varianza retenida), afectando la capacidad del modelo.
- La clasificación binaria mejora notablemente la precisión general (86 %), evidenciando la importancia de simplificar el problema cuando es posible.
- Se recomienda usar PCA como herramienta de apoyo para visualización, análisis de correlaciones y eficiencia, más que como optimizador de modelos.

6. Referencias

- UCI Machine Learning Repository. (2025). *Wine Quality Dataset*. Disponible en: `https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/`
- Scikit-learn Documentation. (2025). *Decomposition: PCA*. Disponible en: `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`