

Práctica 2

Análisis de Componentes Principales y Validación Cruzada

Miguel Alba Ruiz¹, Gonzalo Pantín de Miguel², and Carlos García Guzmán³

¹albaruiz.miguel@uma.es

²gonzalopantindm@uma.es

³carlosgarciag552@uma.es

Aprendizaje Automático. Universidad de Málaga.

Abstract. En este informe se realizan distintas técnicas de preprocesamiento de datos sobre el *dataset Iris* para la clasificación de flores según su especie. Se incluye un breve análisis exploratorio de datos, junto a una explicación de las distintas técnicas de preprocesamiento aplicadas y los resultados obtenidos. Entre estas técnicas destacamos la **normalización** y **estandarización** de los datos y la reducción de la dimensionalidad mediante el **análisis de componentes principales** o PCA. Finalmente, haciendo uso de todos los preprocesamientos de datos, se somete a evaluación un algoritmo supervisado de clasificación mediante **validación cruzada de k iteraciones** o *k-fold cross validation*.

1 Introducción

El dataset *Iris* es uno de los más conocidos en la literatura sobre reconocimiento de patrones, introducido originalmente por el estadístico Ronald A. Fisher en un artículo de 1936. Este sigue siendo muy utilizado hoy en día porque representa un problema de clasificación supervisada ideal para la validación de algoritmos.

Contiene un conjunto de flores *iridáceas* compuesto por 150 muestras, divididas equitativamente en tres clases: **setosa**, **virginica** y **versicolor**. Cada muestra tiene cuatro características numéricas: **sepal length**, **sepal width**, **petal length** y **petal width**.

El objetivo de este proyecto es utilizar el *dataset* para estudiar los beneficios de las técnicas de preprocesamiento de datos mediante la normalización y estandarización, reducción de la dimensionalidad mediante el análisis de componentes principales y la evaluación de algoritmos supervisados mediante la validación cruzada de k iteraciones.

2 Análisis exploratorio de datos

Antes del preprocesamiento se ha realizado un análisis exploratorio de los datos a fin de encontrar posibles *outliers*, entender mejor las variables y definir sus tipos. De esta manera se podrá comprender mejor qué técnicas de preprocesamiento son necesarias y así realizar un mejor análisis de los datos. En la tabla 1 se pueden observar las tres primeras muestras del conjunto.

A continuación, en la figura 1 mostramos la distribución del *dataset* para cada par de variables. Es interesante pues se pueden ver qué características coinciden entre clases y hacer una estimación de cómo de discriminantes son las susodichas.

Table 1. Dataframe de las 3 primeras muestras con 4 features

Samples	Sepal length	Sepal width	Petal length	Petal width
x_1	5.1	3.5	1.4	0.2
x_2	4.9	3.0	1.4	0.2
x_3	4.7	3.2	1.3	0.2

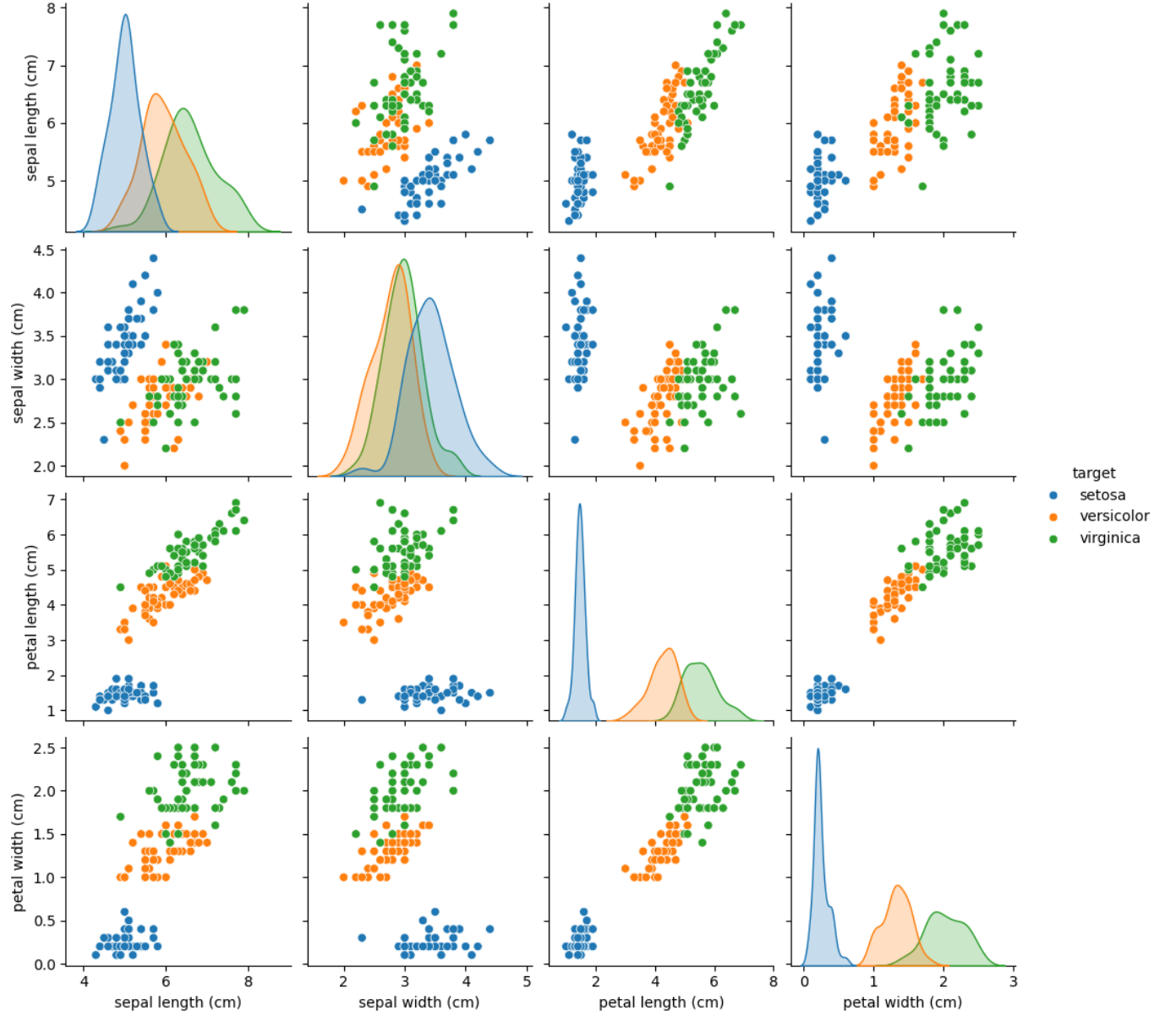


Fig. 1. Distribución de las distintas características y sus clases

3 Preprocesamiento de datos

Al conjunto de datos original, se han aplicado los métodos de estandarización y normalización de los datos, para que la escala de las variables sea homogénea. Se muestra a continuación la distribución de las variables longitud y ancho de sépalo, estandarizadas (figura 2) y normalizadas (figura 3).

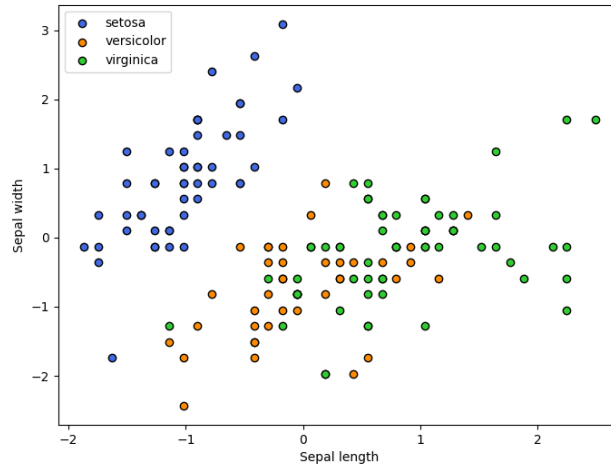


Fig. 2. Distribución por longitud y ancho de sépalo estandarizado.

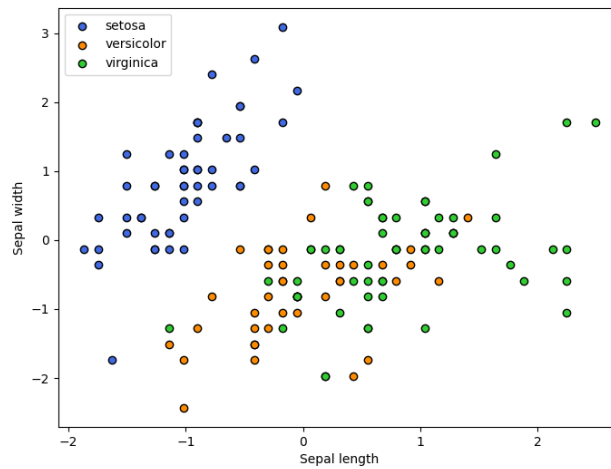


Fig. 3. Distribución por longitud y ancho de sépalo normalizado.

Se puede observar, que ambos conjuntos obtienen una distribución proporcional a la original, pero con distintas escalas de datos, lo cual era el objetivo que se perseguía con la estandarización y normalización.

A los tres conjuntos de datos que se han obtenido hasta ahora, se ha realizado una reducción de dimensionalidad mediante el cálculo de componentes principales (PCA).

Para cada conjunto, se han aplicado dos valores mínimos de varianza explicada que debían presentar los conjuntos resultantes, uno con 0.8, y otro con 0.95. Dicha varianza explicada, indica la proporción de información que se conserva con respecto al conjunto de datos original; antes de aplicar la reducción de dimensiones.

4 Resultados obtenidos

Se muestran los resultados obtenidos tras realizar el cálculo de componentes principales a los 3 conjuntos de datos, y para los dos valores de varianza mencionados anteriormente (figura 4)

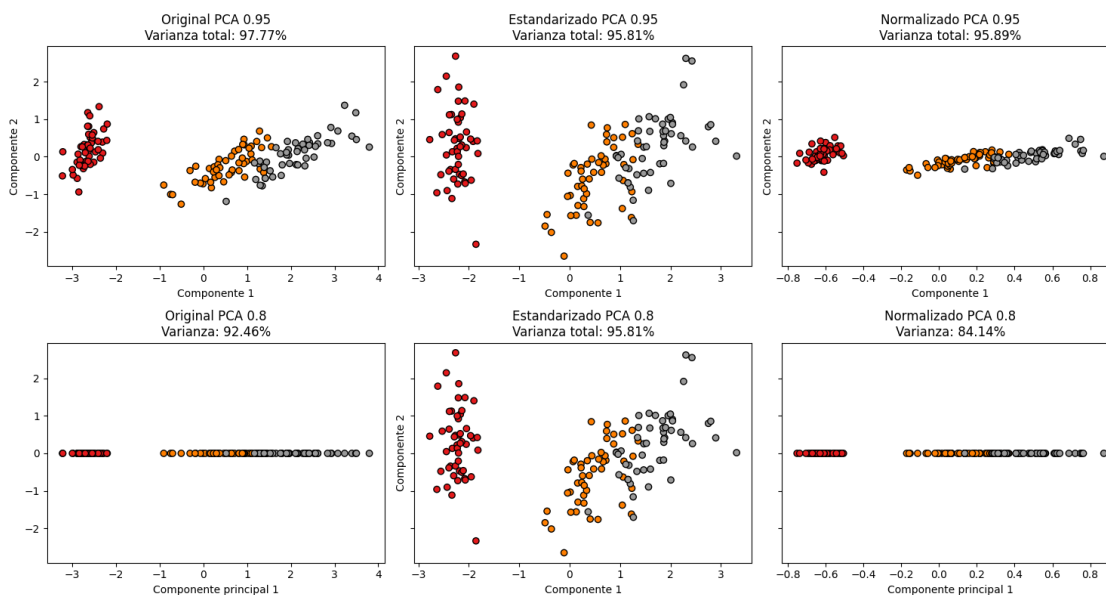


Fig. 4. Comparación de distribuciones con PCA de los distintos conjuntos de datos

Comparando los resultados obtenidos para 0.8 y 0.95 de varianza explicada, se puede observar que para los datos originales y normalizados, tan solo es necesario una componente principal para representar más del 80% de la varianza, y se necesitan dos para el 95%. Por tanto, si se observan las gráficas, la componente verdaderamente discriminante es la primera. Se puede concluir entonces que tan solo es necesaria una componente principal para poder establecer una separación de las categorías tan precisa como la que se obtiene con los 4 atributos originales, pudiendo añadir una componente más para lograr un ligero aumento de la separabilidad.

En cuanto a los datos estandarizados, para un 80% de varianza explicada, son necesarias las dos primeras componentes principales. Por tanto, en el conjunto estandarizado, la primera componente principal proporciona menos información que en los otros dos.

Por tanto, en este *dataset*, el PCA para la normalización presenta mejor desempeño que la estandarización, ya que con una sola componente se obtiene más información.

5 Validación cruzada

La validación cruzada es una técnica muy utilizada para evaluar los resultados de un determinado método de aprendizaje supervisado y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en calcular la media aritmética de las medidas de evaluación sobre las diferentes particiones del conjunto de datos.

Para este proyecto se ha simulado una validación cruzada de $k=5$ iteraciones en cada uno de los conjuntos de datos: original, normalizado, estandarizado, originalPCA80, etc. generando 10 ficheros csv (5 para entrenamiento y 5 para *testing*) representando las 5 particiones train-test que una validación cruzada utilizaría para evaluar un modelo.

5.1 Aplicación real en un clasificador

Finalmente, se ha evaluado un clasificador *Naive Bayes* mediante validación cruzada para todos los conjuntos de datos obtenidos anteriormente. Debido a que el *dataset* está perfectamente balanceado, se ha computado únicamente el *accuracy* medio.

Nótese que el cálculo del *accuracy* en un problema de clasificación de 3 o más clases es similar, pero no idéntico, a uno de clasificación binaria. Suponiendo que en una de las iteraciones se genera la siguiente matriz de confusión:

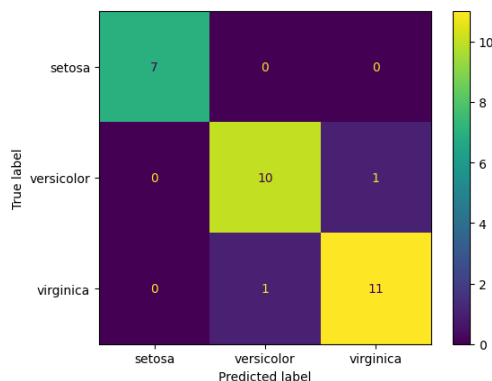


Fig. 5. Matriz de confusión real obtenida en una de las iteraciones sobre el conjunto original

Entonces, el cálculo vendría dado por:

$$accuracy = \frac{7 + 10 + 11}{7 + 10 + 11 + 1 + 1} = \frac{28}{30} = 0.933$$

En la figura 6 se muestran los *accuracy* medios obtenidos en cada uno de los conjuntos de datos. Se puede observar que los modelos generados a partir de todas las características del *dataset* tienen un mejor desempeño a la hora de clasificar muestras. Aunque cabe destacar que en el conjunto de datos normalizado con PCA del 80% se consigue igualar el resultado con un *accuracy* el 96%.

Conjunto	original	originalPCA80	originalPCA95
<i>Accuracy</i>	0.960	0.940	0.893

Conjunto	normalized	normPCA80	normPCA95
<i>Accuracy</i>	0.960	0.960	0.913

Conjunto	estandarized	estandPCA80	estandPCA95
<i>Accuracy</i>	0.960	0.893	0.893

Fig. 6. Resultados de la validación cruzada

6 Conclusiones

En lo referente al cálculo de las PCA con distintos datos (originales, estandarizados y normalizados) tanto para el 80% y 95% de la varianza explicada, se puede concluir que la componente verdaderamente discriminante es la *primera componente* al aportar el 90% de dicha varianza explicada (algo más o algo menos, depende del preprocesado de los datos utilizado). La segunda componente aporta algo de separabilidad entre las dos clases de la derecha de las gráficas (Figura 4), ya que se superponen parcialmente. Por ende, en base a estos datos y en lo referente a este conjunto, la mejor opción parece ser utilizar los datos **normalizados u originales** pues con los estandarizados se pierde información.

No obstante, al entrenar el clasificador de *Naive-Bayes* con los distintos datos podemos apreciar que para este conjunto *normalizar, estandarizar y calcular las PCA* no marca la diferencia. Este valor ha sido obtenido gracias a la utilidad del método de validación cruzada *K-Fold*, el cual permite mostrar el desempeño de un algoritmo de aprendizaje independientemente del conjunto de datos con el que se entrene. El máximo **accuracy** obtenido es 0.960, el cual se obtiene con los **datos originales** (y en otras transformaciones también).

En conclusión, por útil que es el preprocesado de datos, también se puede apreciar que para este conjunto no es completamente necesario. A pesar de esto, siempre conviene aplicarlas con tal de disponer de todas las herramientas necesarias para un problema en específico.

7 Declaración de autoría

Nosotros, los abajo firmantes, Carlos García Guzmán, Gonzalo Pantín de Miguel y Miguel Alba Ruiz declaramos que el presente proyecto, es resultado de nuestro trabajo original.

Certificamos que el código fuente, los análisis y los resultados presentados en este documento han sido desarrollados íntegramente por los autores.

En Málaga, a 01 de noviembre de 2025.

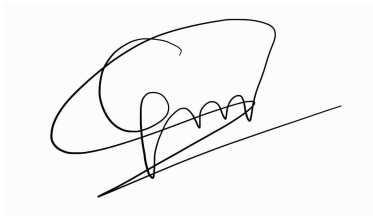
Handwritten signature of Carlos in black ink.Handwritten signature of Gonzalo in black ink.Handwritten signature of Miguel in blue ink.

Fig. 7. Logo de la Universidad de Málaga