Instituto Politécnico Nacional

Escuela Superior de Cómputo

Ingeniería en Inteligencia Artificial

PCA y SOM

Machine Learning

Integrantes:

- Hernández Jiménez Erick Yael
- Patiño Vázquez Samuel
- Robert Garayzar Arthur

11 de septiembre de 2024

Semestre 2025-1

Índice

1. Introducción

En el presente reporte se aplican dos técnicas de reducción y agrupación de datos: el Análisis de Componentes Principales (PCA) y los Mapas Autoorganizados (SOM, por sus siglas en inglés), empleando dos conjuntos de datos clásicos: Wine y Glass Identification. Estas técnicas se utilizan para obtener una mejor comprensión de las características subyacentes en los datos, permitiendo simplificar la representación de los mismos y visualizar cómo se distribuyen y agrupan en espacios de menor dimensionalidad.

El PCA es una técnica que busca proyectar los datos originales en un nuevo espacio de menor dimensión, capturando la mayor parte de la varianza de los datos. Por su parte, los SOM son redes neuronales no supervisadas que organizan los datos en un mapa bidimensional, agrupándolos de acuerdo con similitudes intrínsecas.

En este reporte mostraremos resultados con estas técnicas para identificar patrones y relaciones en los datos y qué interpretación se puede hacer a partir de los resultados visuales y numéricos generados.

1.1. Dataset Wine

Del inglés: .Este es uno de los datasets más tempranos usado en la literatura en métodos de clasificación y ampliamente usado en estadística y aprendizaje máquina. El set de datos contiene 3 clases de 50 instancias cada uno, donde cada clase corresponde a un tipo de planta de iris. Una clase es linealmente independiente de otros 2; los otros no son linealmente independientes entre sí".

Este dataset se incluye en los datasets de práctica de la biblioteca Sckit y, originalmente, en el banco de datasets de la UCI.

1.2. Dataset Glass Identification

Del inglés: "Del Servicio de Ciencia Forense de los EEUU; 6 tipos de vidrio; definidos en términos de su contenido de óxido (p.e. Na, Fe, K, etc)".

Este dataset se encuentra disponible en el banco de datasets de la UCI.

2. Principal Components Analysis

En el código se hizo uso tanto del cálculo manual de los PCA como de los métodos incluidos en la biblioteca Sckit.

2.1. Método manual

Para el método manual se siguieron los siguientes pasos:

1. Se normaliza el dataset
    1.1. Cálculo de la media
    1.2. Cálculo de le desviación estándar
    1.3. Cálculo de valores estandarizados
2. Cálculo de la covarianza
3. Cálculo de los eigenvalores y eigenvectores
    3.1. Ordenamiento de los eigenvalores y eigenvectores
    3.2. Suma acumulativa de eigenvalores y eigenvectores
4. Cálculo de componentes principales
    4.1. Cálculo de componentes mínimos
    4.2. Cálculo de los PCA a partir de los cálculos anteriores
5. Transformación del espacio original del dataset al de los PCA (bidimensional)

2.1.1. valores

A continuación se detallarán los resultados en los pasos más relevantes:

- **Media**: Indican el valor representativo o el valor ubicado a la mitad del rango de valores donde se encuentran la mayoría de datos.
- **Desviación estándar**: Este valor se puede interpretar como el número que indica qué tan alejado está el punto a analizar con respecto a la media.

- **Valor estandarizado**: Este valor corresponde a su reflejo en una distribución normal estandarizada.
- **Covarianza**: Este número indica la tendencia a variar de manera paralela, en este caso, de los distintos datos del dataset.
- **Eigenvalores y eigenvectores**: Estos valores nos indica la relación matemática entre las dimensiones del dataset y nos ayuda, posteriormente, al cálculo de los PCA.

2.1.2. Diagrama de relación de Componentes Principales

A continuación se muestra el diagrama de los PCA calculados y su relación con las dimensiones originales:

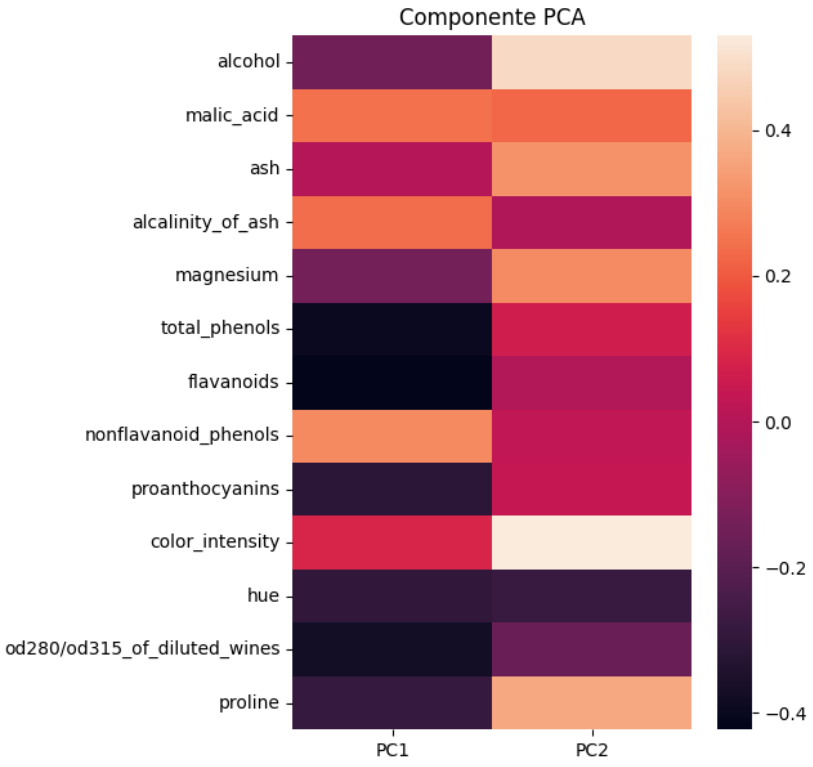

Figura 1: Diagrama de relación de los componentes calculados con las dimensiones originales

En ese diagrama se observa que las columnas representan a los 2 componentes calculados anteriormente y, por la escala, aquellos con el tono correspondiente al valor 0.0 o cercanos son aquellos que no tienen una relación directa con el componente. Así, para el PCA 1 se mantienen relacionadas las siguientes dimensiones:

- Positivas
  - malic_acid
  - alcalinity_of_ash
  - nonflavanoid_phenols
- Negativas
  - alcohol
  - magnesium
  - total_phenols
  - flavanoids
  - proanthocyanins
  - hue
  - od280/od315_of_diluted_wines
  - proline
- Cercanas ninguna
  - ash
  - color_intensity

Y para el PCA 2:

- Positivas
  - alcohol
  - malic_acid
  - ash

        - magnesium
        - color_intensity
        - proline
    - Negativas
        - hue
        - od280/od315_of_diluted_wines
    - Cercanas ninguna
        - alcalinity_of_ash
        - total_phenols
        - flavanoids
        - nonflavanoid_phenols
        - proanthocyanins

Con estos resultados podemos decretar que con estos dos componentes se mantiene la relación entre las dimensiones originales y los calculados gracias a que los componentes cumplen complementariamente con la corespondencia.

2.1.3. Transformación

Una vez calculados los PCA, los datos se llevan al espacio bidimensional. La gráfica correspondiente se muestra a continuación:

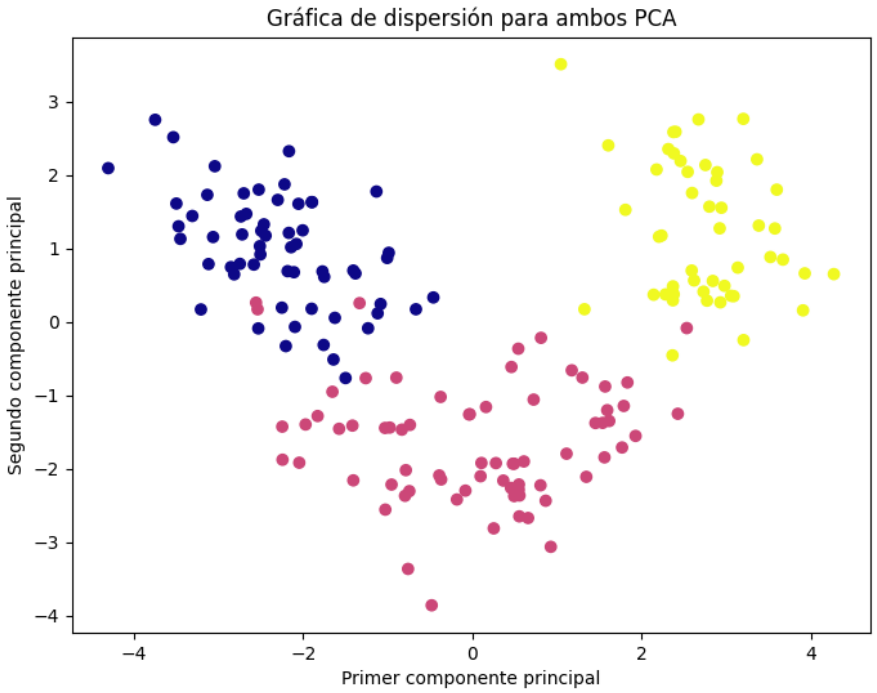

Figura 2: Diagrama de dispersión de los datos originales con las PCA calculadas

2.2. Método automático

Este método está incluido en la biblioteca de Scikit y se tiene que importar explícitamente.

Los resultados que arroja son los siguientes:

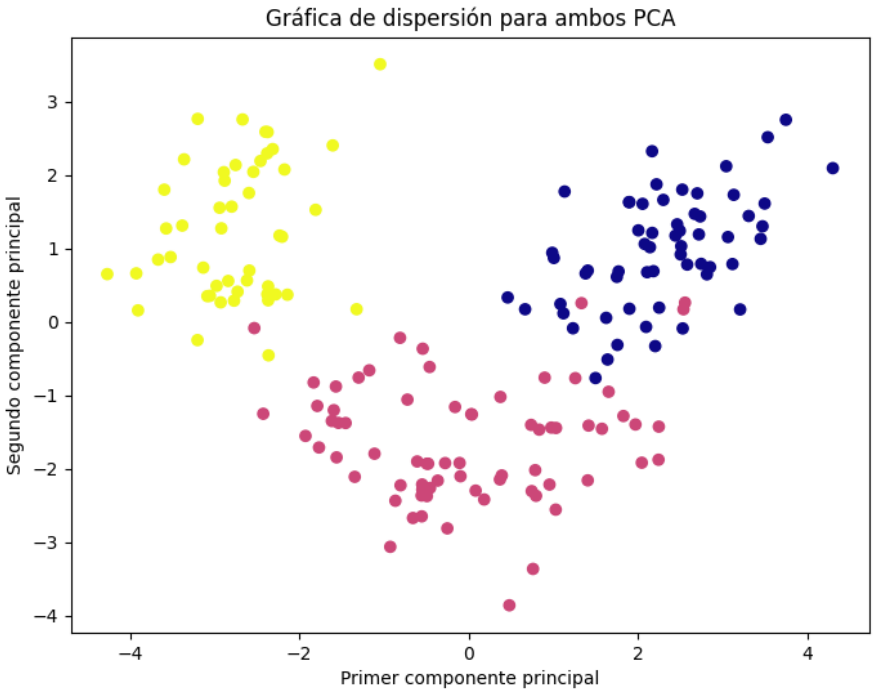

Figura 3: Diagrama de dispersión de los datos originales con los PCA generados por Sckit

2.3. Comparación

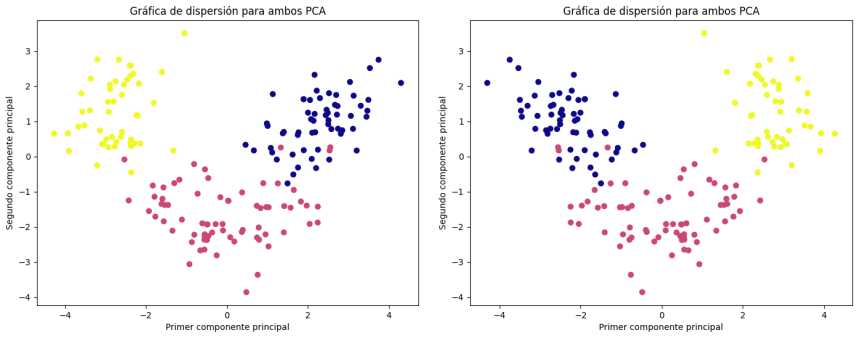

Figura 4: Comparación de diagramas de dispersión calculados (izquierda) y generados (derecha)

Como se puede ver, gran parte de los datos mantienen su posición en ambas gráficas, siendo que aquellos que difieren lo hacen a causa de los distintos pesos asignados a cada dimensión en los componentes generados o calculados.s

3. Self Organized Maps

El código se basó en la documentación cargada de la página de Scikit-learn de Python.

3.1. Proceso

Para el método manual se siguieron los siguientes pasos:

1. Carga de los datos de Iris de `sklearn.datasets`, de la cual se toman únicamente los datos de longitud y ancho del sépalo.

2. Se crea una instancia de un SOM con una matriz 3x1 y una dimensión de 2, que corresponde a las características seleccionadas de los datos de Iris.

3. Después del entrenamiento, se generan predicciones con el SOM para los datos de entrada utilizando `predict()`. Gracias al entrenamiento previo, el SOM los clasifica automáticamente.

3.2. Resultados obtenidos

El diagrama obtenido previo al entrenamiento de SOM:

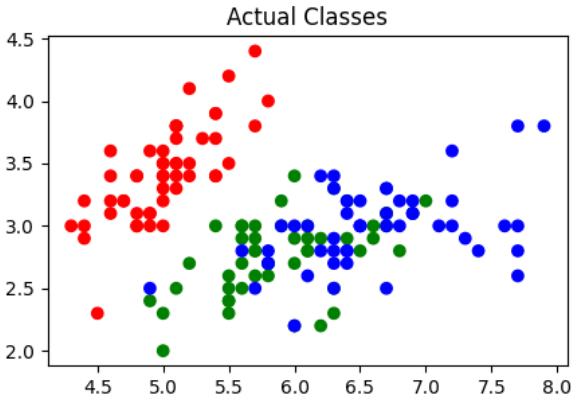

Figura 5: Datos previos al entrenamiento de SOM

Datos después del entrenamiento del modelo y utilizando el modelo predictivo:

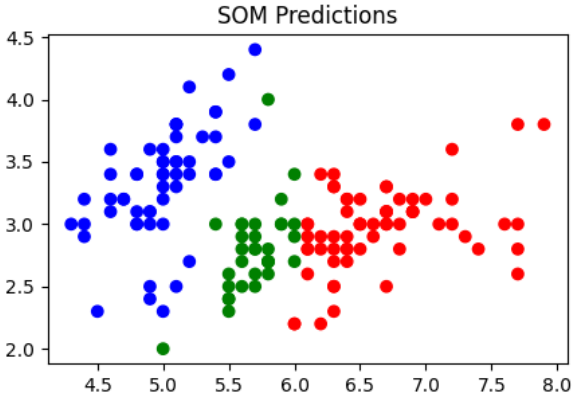

Figura 6: Datos después del entrenamiento de SOM

4. Conclusiones

4.1. Samuel Patiño Flores

Con esta práctica vimos lo útil que es utilizar los métodos de reducción de dimensionalidad, como PCA y SOM. Esto nos ayuda a visualizar los datos de una forma más intuitiva, en forma de plano, y mostrando la información de manera más clara. En el caso de SOM, al capturar relaciones no lineales, lo hace más versátil en aplicaciones para clasificación.

4.2. Hernández Jiménez Erick Yael

Con ambos métodos, si bien se demostró que no se puede llegar a representar los datos originales con una fidelidad total, se puede decir que se llega a una representación considerablemente buena de los datos en n dimensiones y, en el proceso, a una reducción conveniente para su análisis para posteriores procesos. Cabe mencionar que, por los resultados, SOM es más conveniente para clasificar o reducir dimensiones de datos de manera no lineal y PCA para datos con relaciones más lineales. Además, dado que PCA no está hecho para clasificar, SOM es mejor para este objetivo aún cuando PCA pueda .ªccidentalmenteçumplir el cometido