MÉTODO PCA
IMPLEMENTACIÓN EN EL DATASET IRIS

En el mundo de la computación, el *Análisis de componentes principales* o *PCA*, es una técnica estadística ampliamente utilizada para la *reducción de dimensionalidad*, su objetivo principal es transformar un conjunto de datos de alta dimensionalidad en un espacio de menor número de dimensiones, preservando la mayor cantidad de información posible; dentro de sus implementaciones más comunes se encuentran los casos en los que se trabaja con información con muchas variables muy redundantes o correlacionadas entre sí.

Elementos matemáticos

Como uno de los elementos principales en la implementación del algoritmo *PCA*, se encuentran algunos principios matemáticos del álgebra vectorial, como matrices o vectores; en este caso se encuentran los *autovectores*, que representan las direcciones en las que los datos varían más, y los *autovalores*, que miden la magnitud de la varianza a lo largo de estas direcciones; ambos, son utilizados después en la implementación del algoritmo (posteriormente al cálculo de la matriz de covarianza), y posteriormente las direcciones de mayor varianza (autovectores) y utilizarlas como nuevas *componentes principales.*

Como primer paso concreto del análisis de componentes, se encuentra la construcción de la matriz de covarianza, ésta mide la relación entre las variables del conjunto de datos y captura cómo varían las características juntas: si dos características aumentan o disminuyen de manera conjunta, su covarianza será alta, mientras que sí varían de manera opuesta, la covarianza será negativa.

Componentes principales

En el contexto del algoritmo PCA, los componentes principales son combinaciones lineales de las variables originales que maximizan la varianza en los datos; el primer componente es la dirección que explica la mayor parte de la varianza, el segundo explica la mayor cantidad de varianza restante, y así sucesivamente; son ortogonales entre sí para garantizar que no haya redundancia en la información.

Como producto de la reducción de la dimensionalidad se proyectan los datos originales en estas nuevas direcciones, y por tanto se proyectan los datos originales en estas nuevas direcciones. Para medir la varianza explicada por cada componente se utiliza un porcentaje de varianza del total, con ello es posible llegar a

determinar cuántos componentes se deben conservar para garantizar que se mantenga la mayor cantidad de información posible.

Aplicaciones Ventajas

El método PCA es utilizado en una amplia variedad de campos diversos, incluyendo el reconocimiento de patrones, procesamiento digital de imágenes, análisis de datos complejos, financieros, genómica, entre otros. Dentro de sus de sus principales aplicaciones y ventajas se encuentran las siguientes:

- **Reducción de dimensionalidad**: permite reducir la cantidad de variables necesarias para representar los datos.
- **Eliminación de ruido**: Al seleccionar solo las componentes principales más relevantes, el PCA ayuda a eliminar variaciones insignificantes o ruido en los datos, mejorando la precisión de los modelos de aprendizaje automático.
- **Mejora del rendimiento computacional**: Al trabajar con un número reducido de variables, también se reduce el costo computacional.

Limitaciones

Como todos los diferentes métodos de procesamiento de datos, el *Análisi de componentes principales* también posee algunas limitaciones; entre ellas se encuentran su incapacidad para capturar adecuadamente relaciones complejas y no lineales entre las variables debido a su naturaleza lineal, su alta sensibilidad a la escala de las variables, su dificultad para interpretar componentes principales, y la necesidad de estandarizar los datos antes de aplicar la técnica.

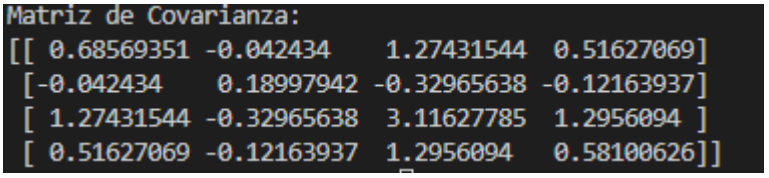
Dataset Iris

Para la implementación del método *PCA*, en un programa en Python, se utilizó el conjunto de datos denominado *Iris*, el cual es uno de los métodos más utilizados en el campo del aprendizaje automático. Este *dataset* consta de 50 muestras de flores de iris, divididas en tres especies diferentes: *Iris setosa*, *Iris versicolor*, *Iris virginica,* cada una con una representación de 50 muestras, de igual manera cada muestra cuenta con diferentes características relacionadas a la especie.

Este conjunto de datos es continuamente utilizado en este tipo de procesos (aprendizaje automático) debido a su simpleza y su capacidad para ser visualizado en 2D o 3D. Para el algoritmo previamente mencionado (implementación Python), el dataset *Iris* fue procesado mediante el método *PCA* para obtener los componentes.

Análisis de matriz de covarianza

Una vez ejecutado el programa que implementa el método *PCA*, la matriz de covarianza que se obtuvo fue la siguiente:



Se pueden observar las siguientes tendencias al analizar la matriz:

- **Varianzas Elevadas**: Las varianzas en la diagonal (que indican la variabilidad de cada característica) pueden revelar cuáles características son más dispersas.
- **Covarianzas Positivas y Negativas**: Los elementos fuera de la diagonal indican cómo varían las características en relación entre sí. Covarianzas positivas sugieren que, a medida que una característica aumenta, la otra también tiende a aumentar. Covarianzas negativas indican que a medida que una característica aumenta, la otra tiende a disminuir.

De igual manera se obtuvieron las siguientes correlaciones significativas, las cuales fue posible identificar gracias al análisis de éste elemento:

- Si se encuentra una covarianza alta entre dos características, sugiere que estas están fuertemente relacionadas y contribuyen a la misma varianza
- Las características altamente correlacionadas pueden ser combinadas en componentes principales, lo que reduce la redundancia en los datos.

Finalmente es posible concluir que la matriz de covarianza es fundamental para calcular los autovalores y autovectores. Los autovectores derivados de la matriz de covarianza son las direcciones (componentes principales) en las que los datos varían más. Por lo tanto, una matriz de covarianza que muestre características interrelacionadas puede resultar en componentes principales que capturan eficientemente la estructura de los datos.