**Explicación de cómo funciona PCA (Álgebra Lineal)**

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica de reducción de dimensionalidad ampliamente utilizada en aprendizaje automático y análisis de datos. En esencia, transforma un conjunto de variables posiblemente correlacionadas en un conjunto más pequeño de variables no correlacionadas llamadas componentes principales, conservando la mayor cantidad de información posible.

**Fundamentos Matemáticos de PCA**

**1 Representación del Dataset como una Matriz:** El dataset se representa como una matriz 𝑋

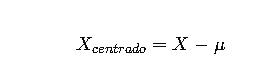
de tamaño n×m, donde:

n: Número de observaciones (filas).

m: Número de características (columnas).

**Ejemplo:** En el dataset Titanic, las columnas como Age, Fare, Sex, etc., forman las características de X.

**2 Centrar los Datos:** Para garantizar que PCA sea efectivo, cada columna de X debe centrarse restando la media:



Donde μ es el vector de medias de cada característica.

**3 Cálculo de la Matriz de Covarianza:** Se construye la matriz de covarianza C para medir cómo varían las características conjuntamente:



​

C es una matriz simétrica de tamaño m×m que captura las relaciones entre las columnas de X

**4 Descomposición en Valores Propios: Se** realiza una descomposición espectral en C, descomponiéndola en sus valores propios (λ) y vectores propios (v):



Los valores propios λ indican cuánta varianza explica cada componente principal.

Los vectores propios v representan las direcciones de los nuevos ejes (componentes principales).

**5 Ordenar los Componentes:** Los valores propios se ordenan en orden descendente. Los vectores propios correspondientes forman una matriz de proyección V.

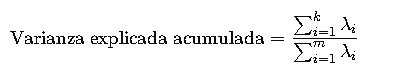
**6 Transformación de los Datos:** Los datos originales se proyectan en el espacio de los componentes principales:



Donde 𝑉𝑘 contiene los primeros k vectores propios correspondientes a los k valores propios más grandes.

**Selección de la Cantidad Óptima de Componentes:**

**Varianza Explicada Acumulativa:** Se calcula la proporción de varianza explicada por cada componente:



Elegimos k componentes que expliquen un porcentaje significativo (por ejemplo, 95%) de la varianza total.

**Intuición de PCA**

PCA puede entenderse como encontrar un nuevo sistema de coordenadas en el cual los datos tengan la mayor dispersión posible a lo largo de los primeros ejes. Esto permite reducir la dimensionalidad del dataset mientras se conserva la mayor parte de la información.

**Aplicación al Titanic:**

En el caso del dataset Titanic, PCA puede:

* Reducir el número de características como Age, Fare, Sex, etc., a unas pocas componentes principales.
* Facilitar la visualización de los datos en dos o tres dimensiones.
* Mejorar el rendimiento computacional de los modelos al reducir la dimensionalidad.