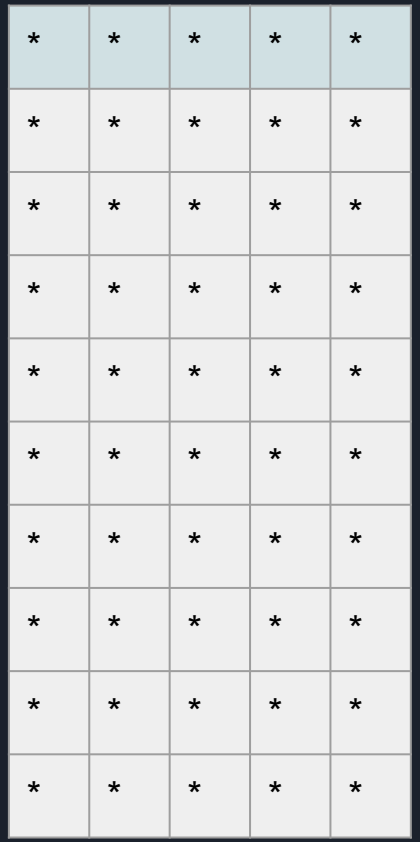
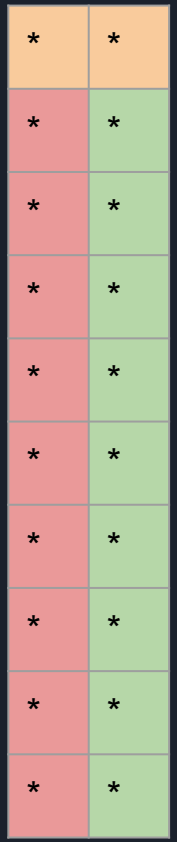
PCA

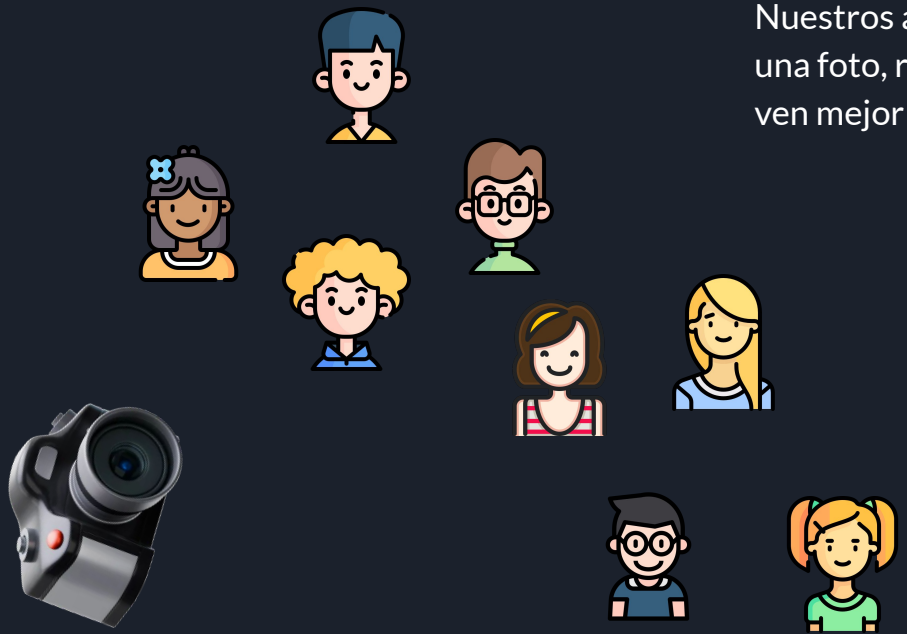
Modelizado de Minería de Datos - Q22025

PCA





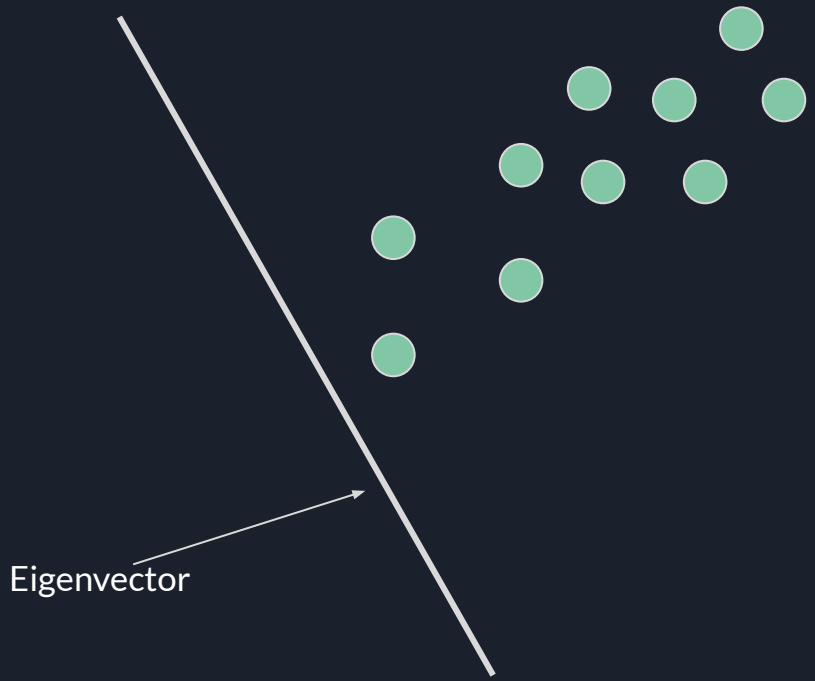
PCA



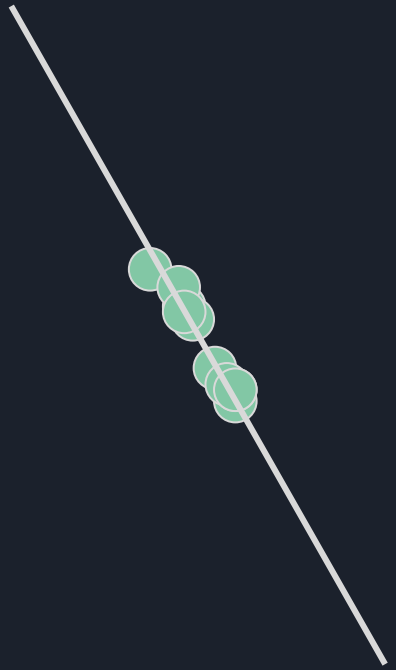
Nuestros amigos viven en 3 dimensiones, pero si sacamos una foto, reduciremos a 2 dicha cantidad. De qué ángulo se ven mejor todos ellos?

PCA

Probamos una primer proyección…



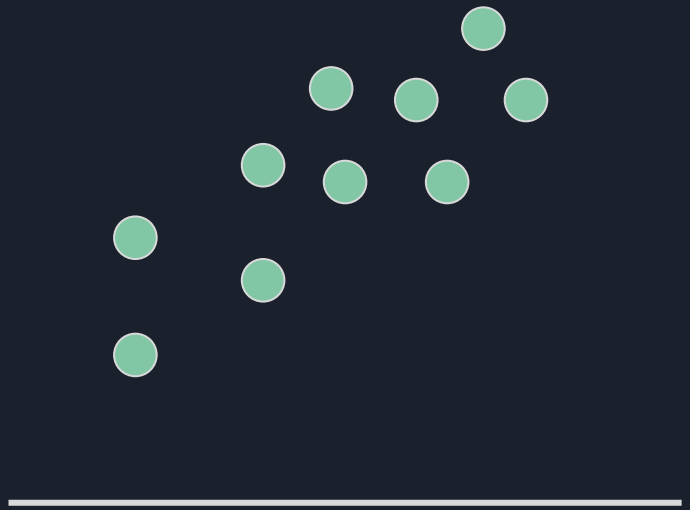
PCA

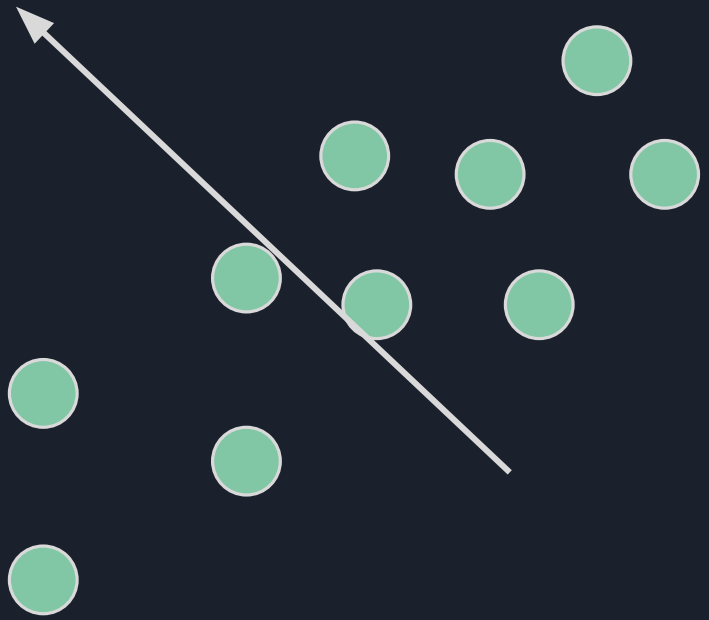


Nuestros amigos viven en 3 dimensiones, pero si sacamos una foto, reduciremos a 2 dicha cantidad. De qué ángulo se ven mejor todos ellos?

PCA

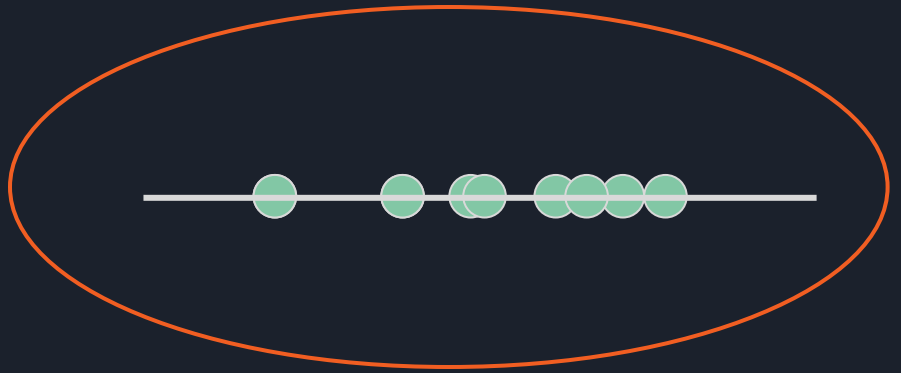
Probamos una segunda proyección…





$$PC1 = (0.7 \times \text{Edad}) + (0.5 \times \text{Ingresos}) + (0.3 \times \text{Puntuación de Crédito})$$

PCA



PCA

Datos de casas

Superficie

Cantidad de habitaciones

Escuelas cercanas

Tasa de criminalidad

Tamaño

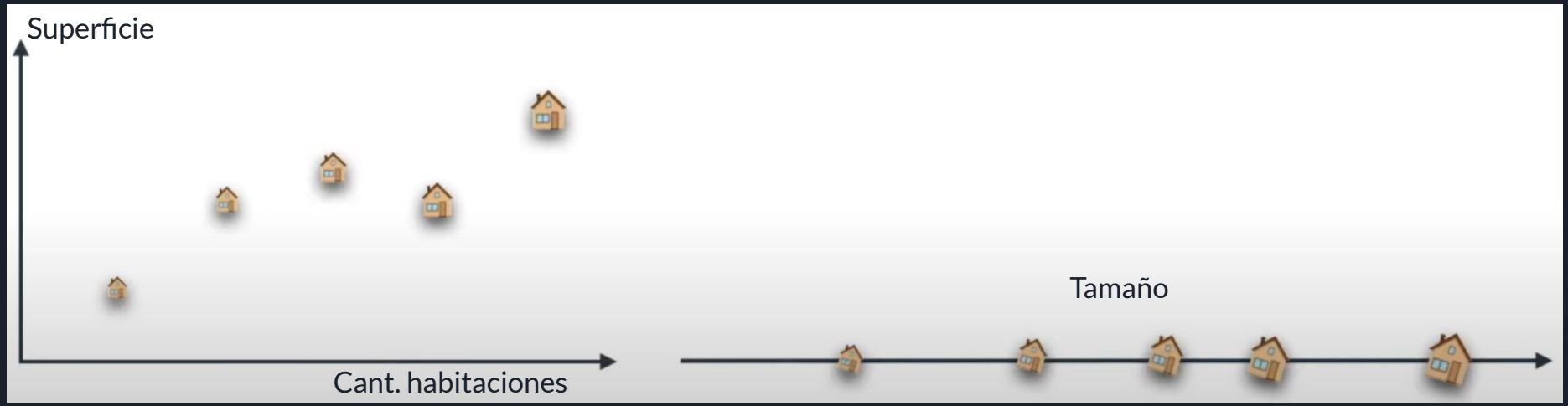
Ubicación

Eigenvalues

PCA

2 dimensiones

1 dimensión



PCA

El Análisis de Componentes Principales (PCA) no es un algoritmo de minería de datos en sí mismo, sino una poderosa técnica de reducción de dimensionalidad.

El PCA nos permite sintetizar esa información, encontrando nuevas variables, no correlacionadas, llamadas componentes principales

Estas nuevas variables son una combinación lineal de las originales. Son como el promedio ponderado de las variables originales

Imagina que tienes un conjunto de datos con muchas columnas (variables) y quieres simplificarlo sin perder la información más importante. Eso es esencialmente lo que hace PCA.

PCA - La Idea Central: Reducción de Dimensionalidad

Dimensionalidad: Un dataset con muchas columnas se dice que tiene alta dimensionalidad.

El Problema: Dificulta la visualización, impacto en eficiencia de algoritmos de ML

La Solución de PCA: PCA busca reducir el número de dimensiones (variables) de nuestros datos, creando nuevas variables artificiales **no correlacionadas** llamadas componentes principales.

PCA - Cómo funciona?

- Estandarizar datos.
- Calcula la matriz de covarianza para entender las relaciones entre variables.
- Encuentra las direcciones de mayor varianza (vectores propios) y la cantidad de varianza que explican (valores propios).
- Selecciona las direcciones más importantes (los primeros "k" vectores propios).
- Proyecta tus datos originales en estas nuevas direcciones, obteniendo un conjunto de datos de menor dimensión (los componentes principales) que aún conserva la mayor parte de la información importante.