

Aprendizaje no supervisado

Aprendizaje de Máquina Aplicado

Juan David Martínez Vargas
jdmartinev@eafit.edu.co

César Leandro Higueta
clhiguitap@eafit.edu.co

2023

Agenda

- Introducción
- Agrupamiento (clustering)
 - K-medias
- Reducción de dimensionalidad
 - PCA

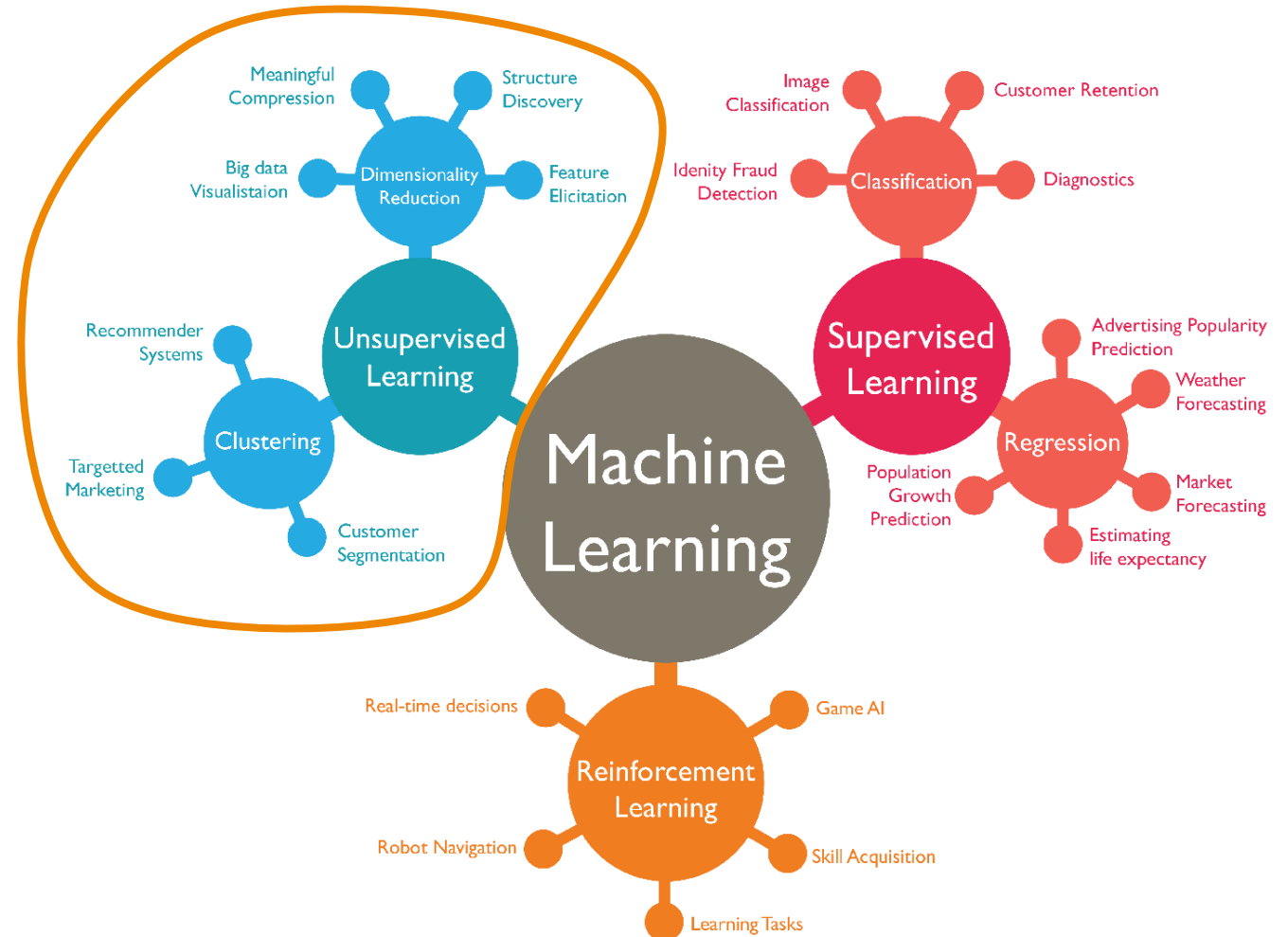


Introducción

Tipos de Aprendizaje en ML

Existen tres tipos principales de aprendizaje en ML:

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje por refuerzo



Aprendizaje no supervisado

En el aprendizaje no supervisado no contamos con etiquetas a predecir, pero podemos tomar algunas decisiones sobre los datos mediante aprendizaje automático. Las tareas más comunes de aprendizaje no supervisado son:

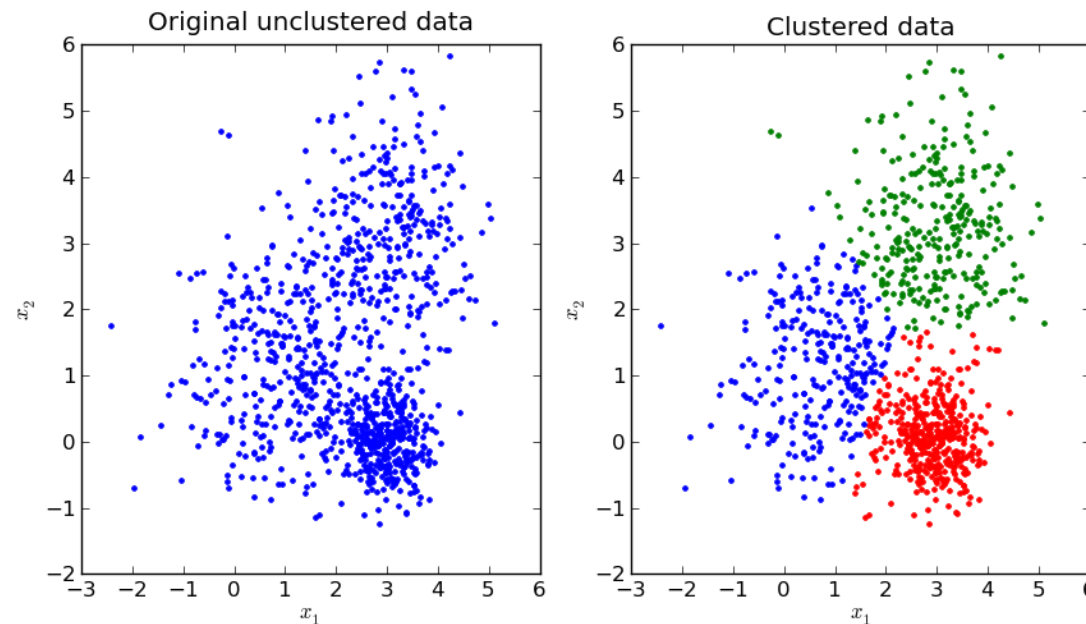
- Agrupamiento (clustering)
- Reducción de dimensionalidad
- Detección de anomalías



Agrupamiento (Clustering)

Agrupamiento (Clustering)

- Encontrar etiquetas (labels) a objetos sin etiquetas.
- Es el proceso de particionar un conjunto de objetos en subconjuntos.



Algoritmo K-Medias

Inicializar los centros $k = \{\vec{c}_1, \vec{c}_2, \dots, \vec{c}_k\}$

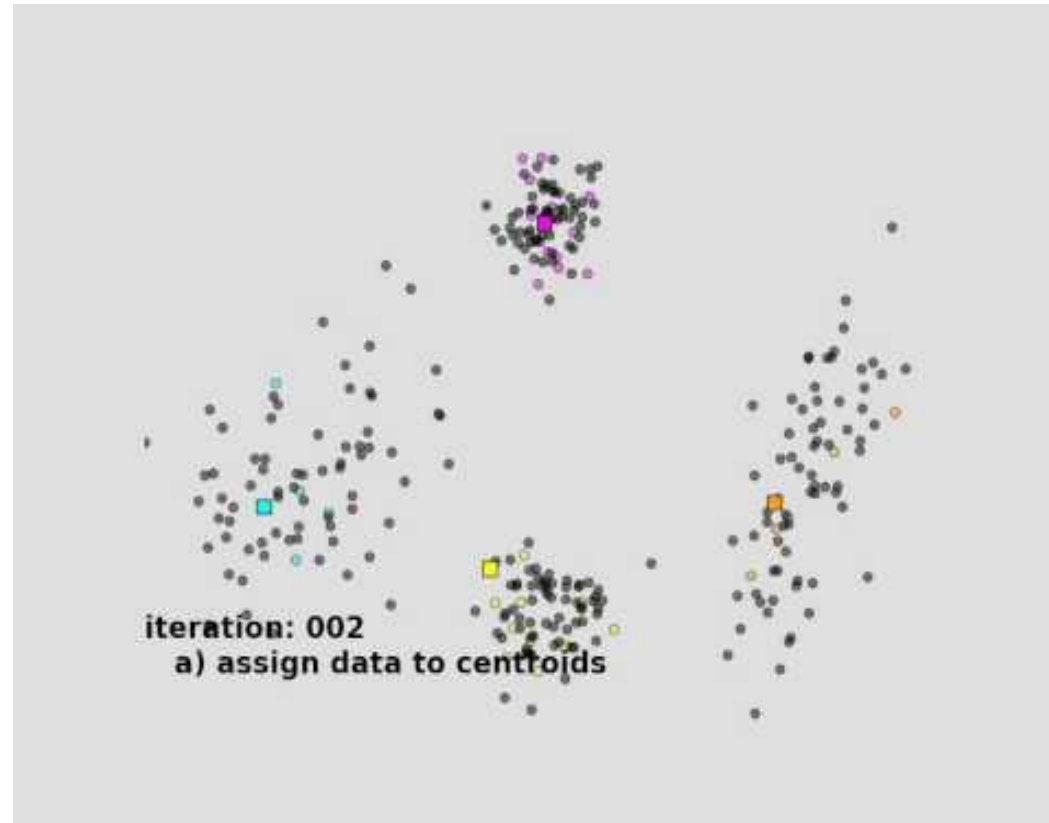
1. Para cada centro encontrar datos más cercanos.
2. Se recalcula el centro sacando la media de los puntos en el cluster.
3. Repetir desde el paso 1 hasta que los centros dejen de cambiar.

K-Medias: Función Objetivo

El procedimiento de K-medias equivale a intentar minimizar la **inercia**, desviación o dispersión del conjunto de datos con respecto a los centros.

$$\arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

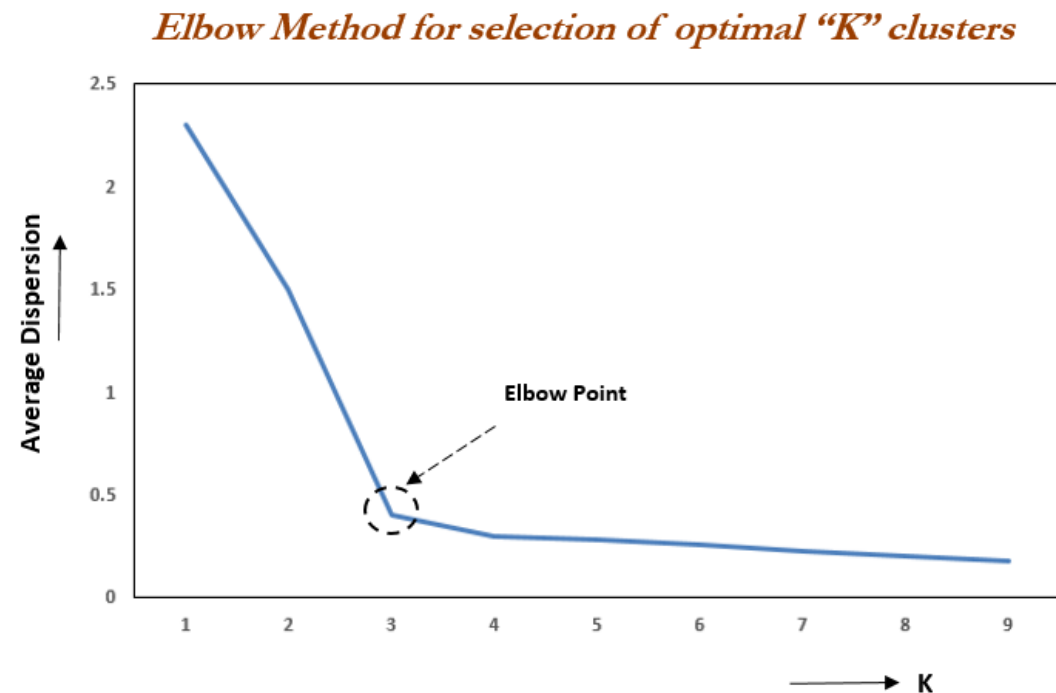
K-Medias: Animación



<https://www.youtube.com/watch?v=5l3Ei69l40s>

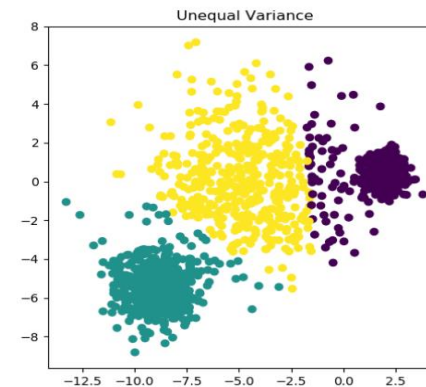
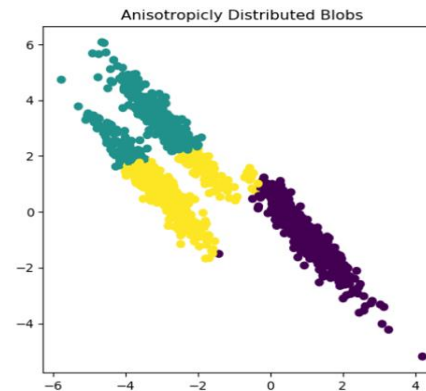
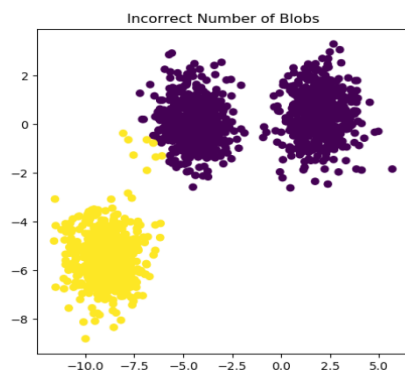
K-Medias: Criterio del Codo

Para decidir el número de clusters se puede utilizar el criterio del codo:



K-Medias: Desventajas

- No garantiza encontrar el agrupamiento óptimo.
- Depende de la distancia.
- Sensible a outliers.
- Problemas con muchas dimensiones.





Reducción de Dimensionalidad

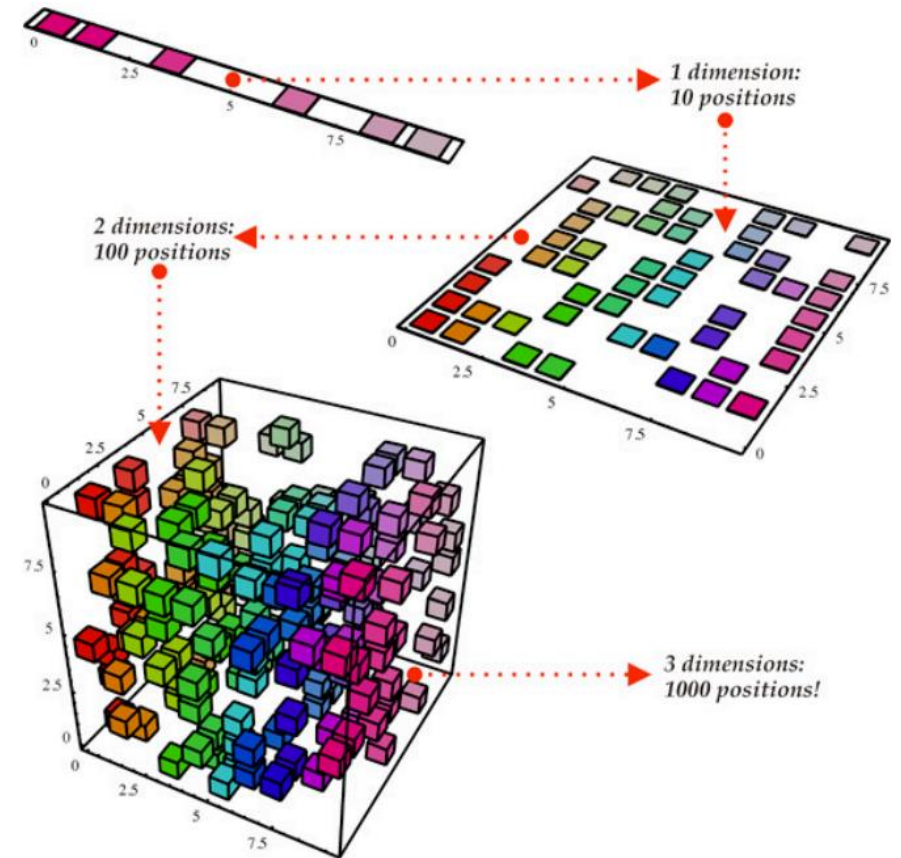
Reducción de Dimensionalidad

Como su nombre lo indica, busca llevar un conjunto de variables :

$$x_1, \dots, x_n$$

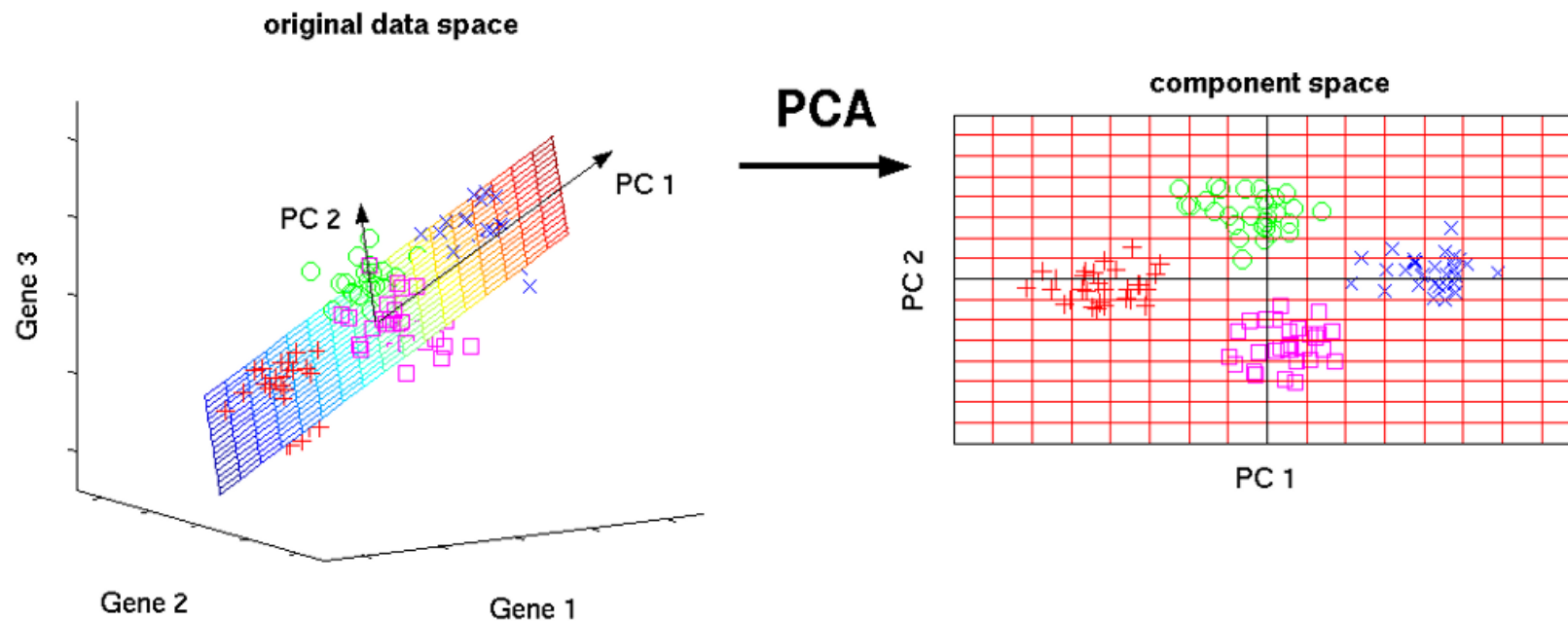
a una representación más pequeña de ellas:

$$z_1, \dots, z_k, \quad k < n$$



Principal Component Analysis (PCA)

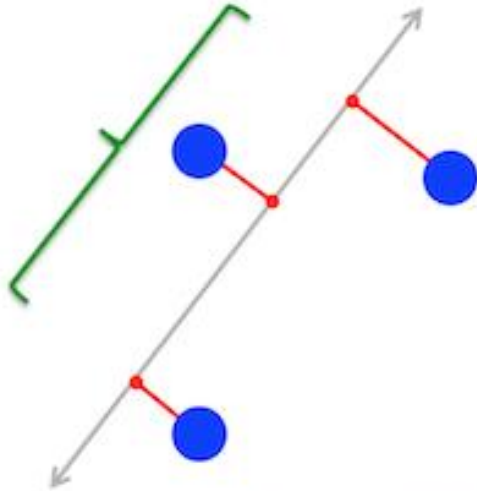
Es un método para reducir la dimensionalidad de los datos conservando las direcciones que mayor varianza tienen.



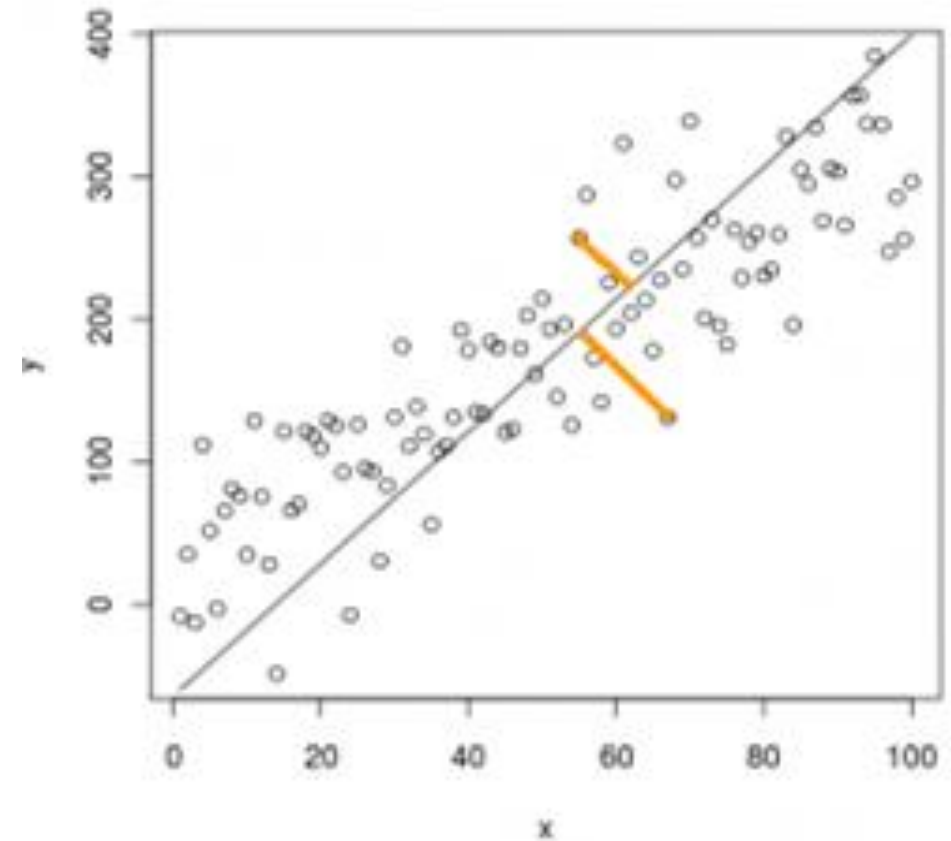
PCA: Algoritmo

1. Calcular la media de cada variable.
2. Centrar los datos para que tengan media cero.
3. Calcular la matriz de covarianza.
4. Encontrar la descomposición en valores singulares.
5. Elegir el número de variables a conservar.
6. Proyectar.

PCA: Proyección

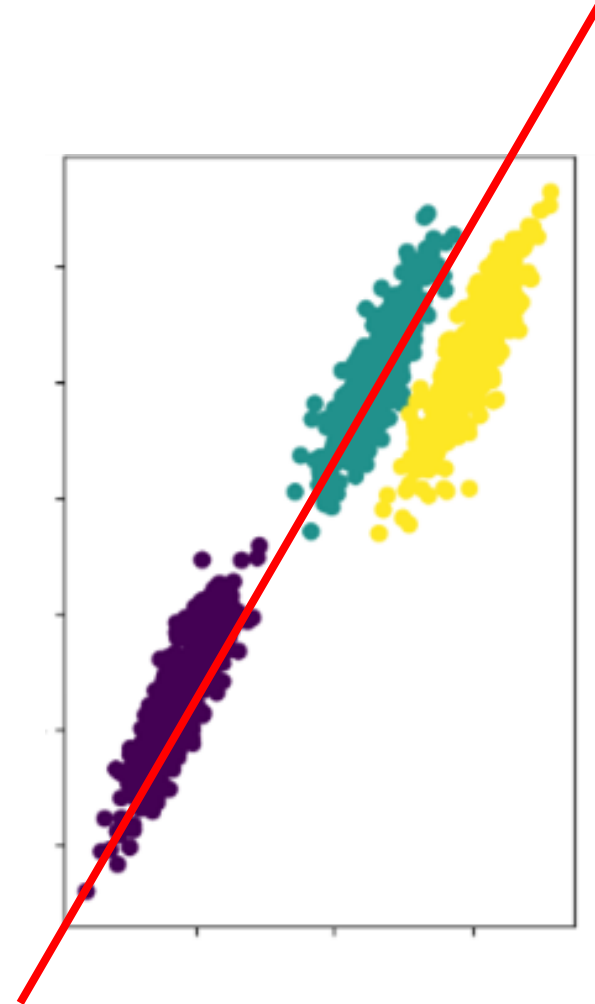


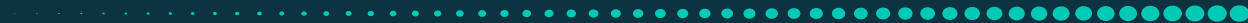
This axis produces the smallest total **projection error** and the largest **variance**



PCA: Clusters

En general, PCA no preserva clusters a la hora de reducir la dimensionalidad de los datos.





Ejemplos de agrupamiento y reducción de dimensionalidad



¡Muchas Gracias!