Ciencia de la
Computación

Clustering

Cristian López Del Alamo
clopezd@utec.edu.pe
IPRODAM3D - Research group

2022

Programa

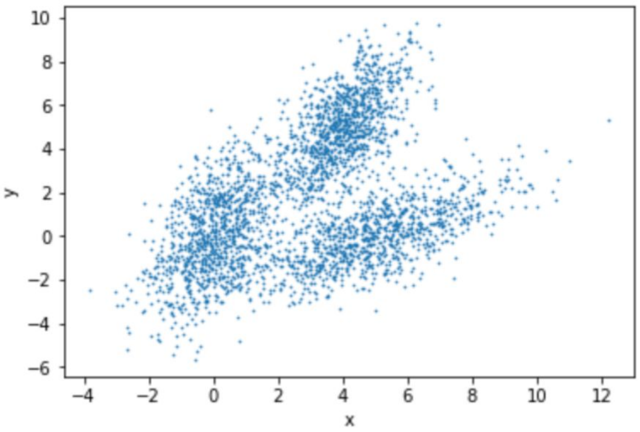
1. Kmenas
2. Mean Shift
3. DBSCAN

1 Clustering

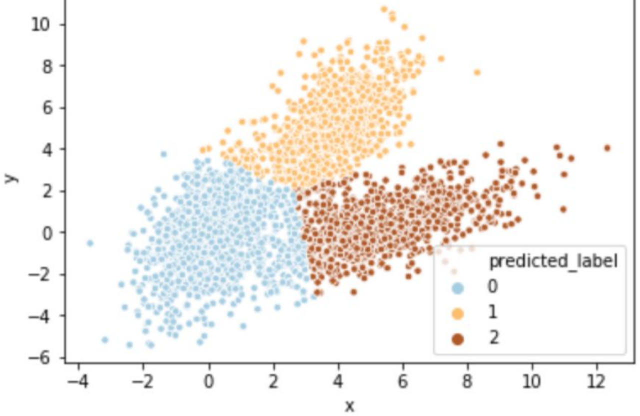
Clustering

- Técnica de ***machine learning*** que permite agrupar, de manera **no supervisada** un conjunto de datos de acuerdo a su estructura o **características similares**.

Before



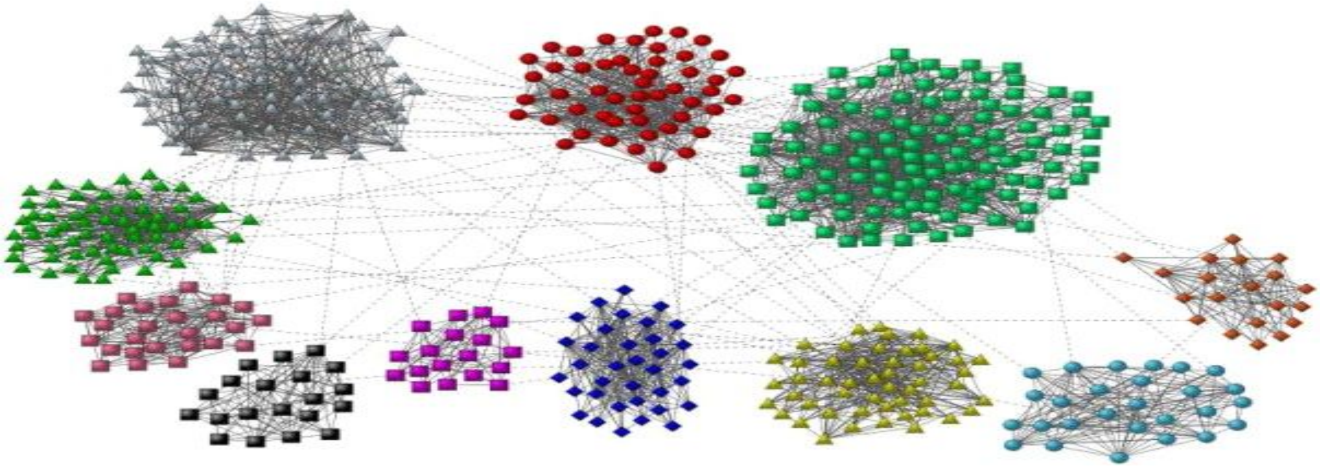
After



[Fuente: Click]

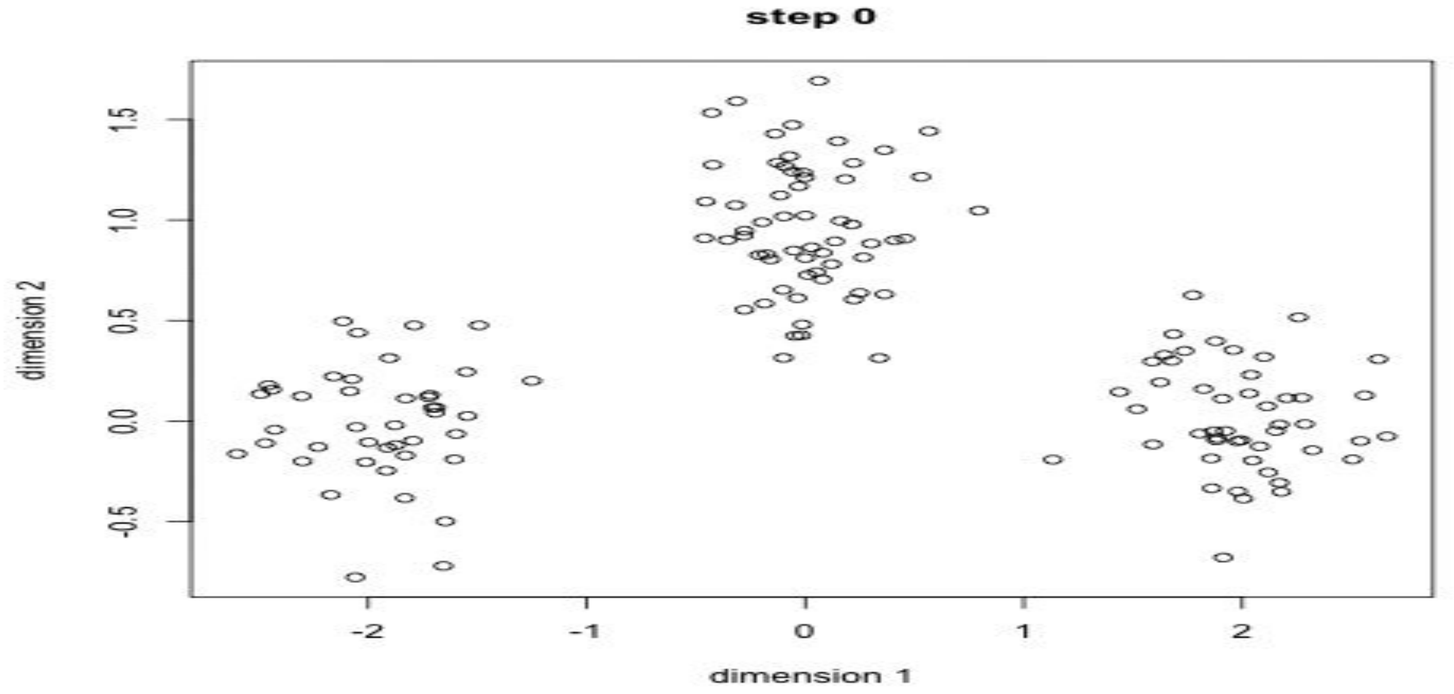
Clustering

- El conjunto de datos que pertenecen a **un mismo grupo** deben tener características **propiedades similares**, y a la vez, características **muy disímiles** respecto a elementos de **otros grupos**



[Fuente: Click](#)

K-Means



[Fuente: Click]()

Algorithm: K-Means

Input —Dataset
—number of clusters

Output —K clusters

Step-1: —Initialize K centers of the cluster

Step-2: —Repeat

—Calculate the mean of all the objects belonging to that cluster

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q$$

where μ_k is the mean of cluster k and N_k is the number of p
belonging to that cluster

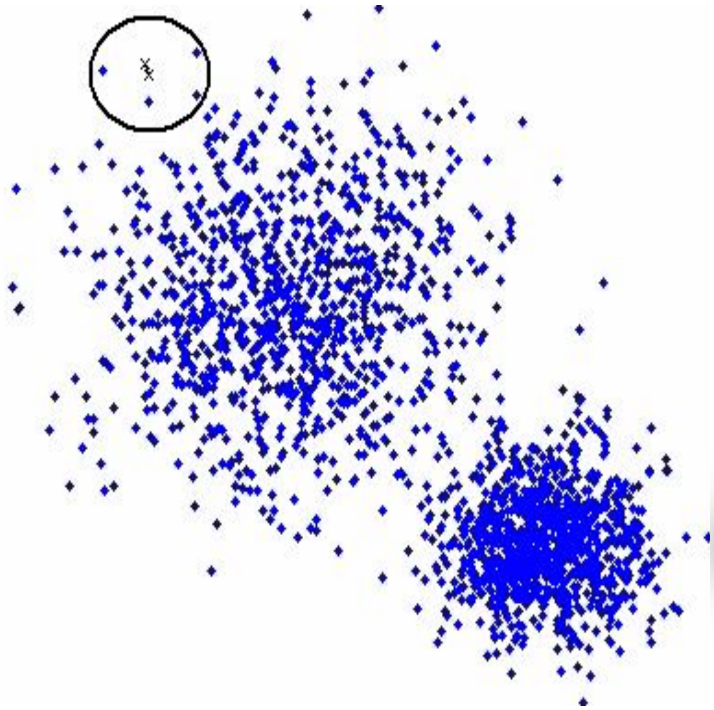
—Assign objects to the closest cluster centroid

—Update cluster centroids based on the assignment

—**Until** centroids do not change

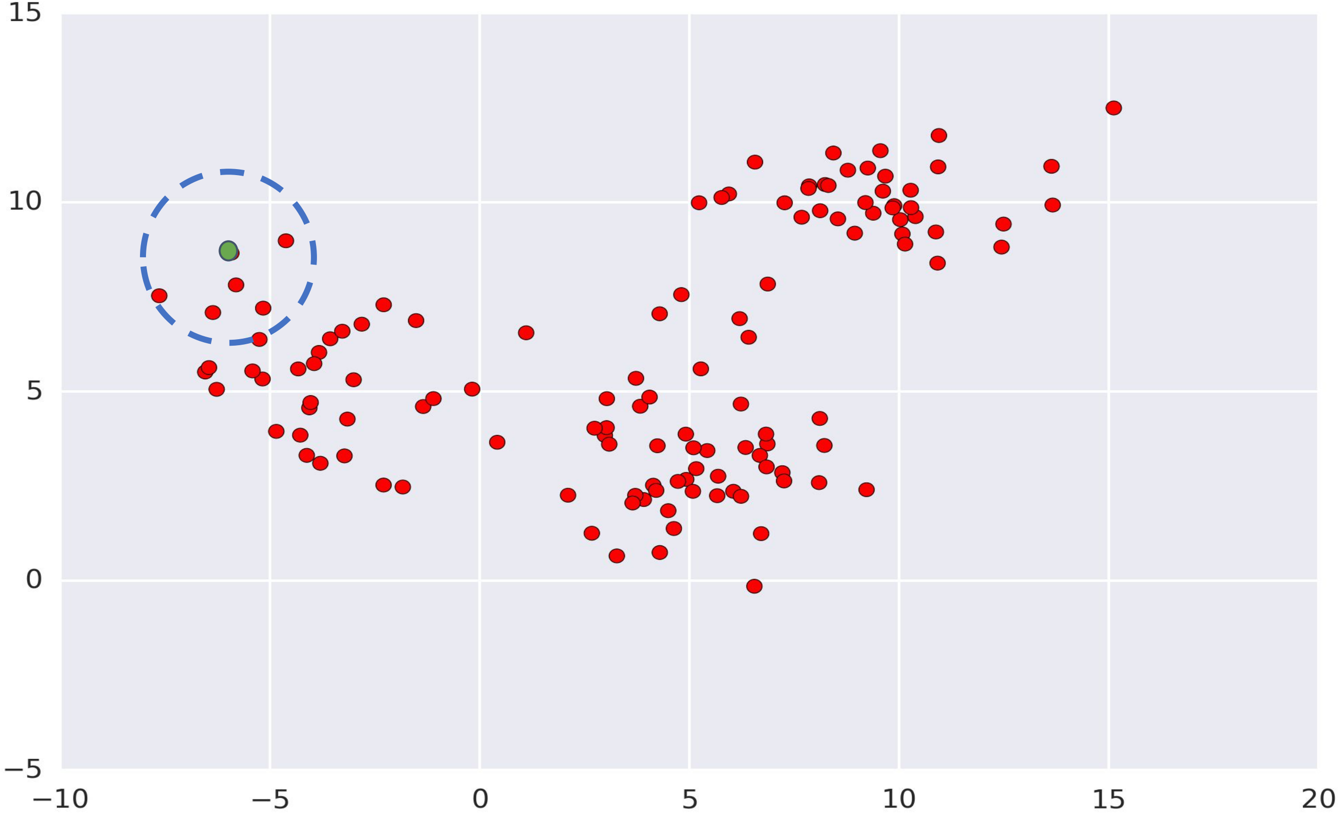
Mean-Shift

Fuente: Click



Fuente: Click

Mean-Shift



Mean-Shift

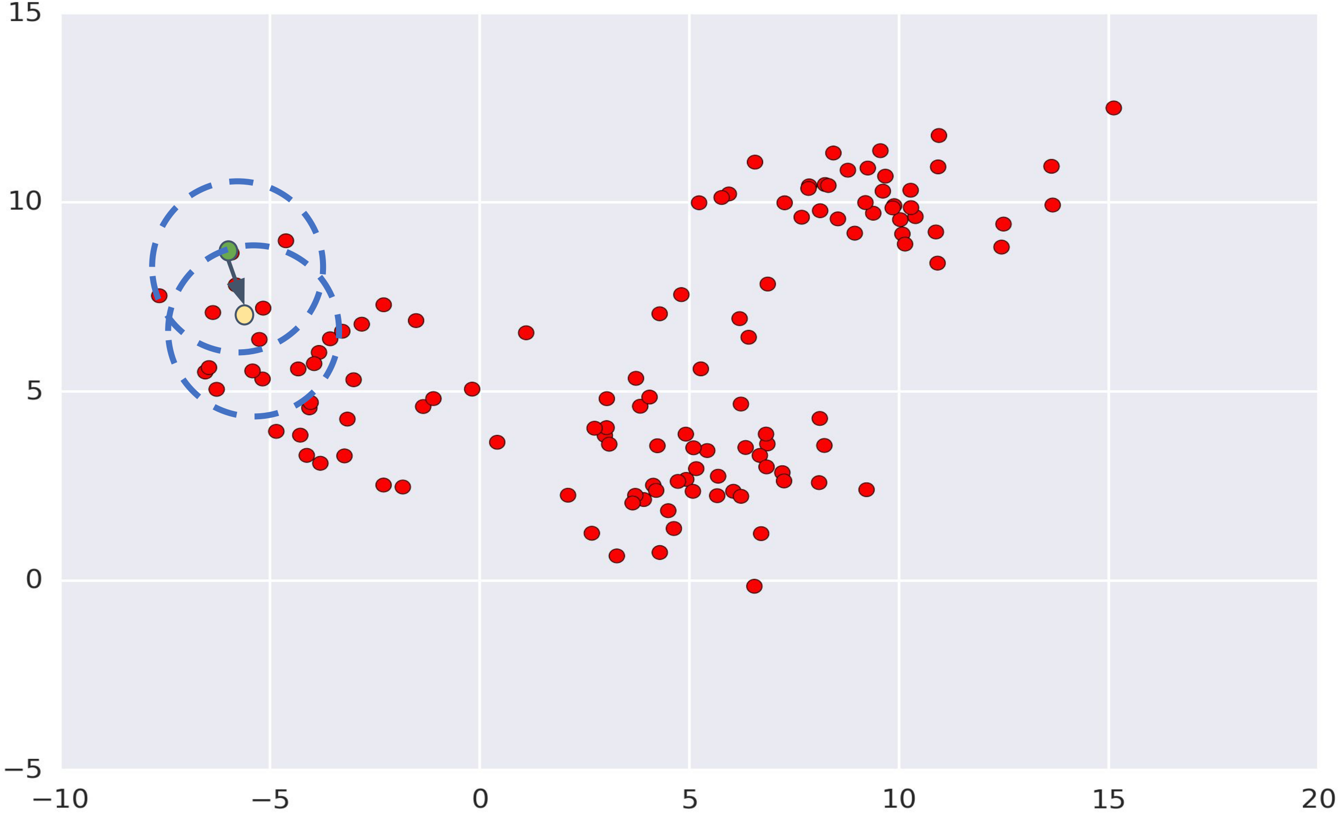
- Smoothing kernel

$$f(x) = \sum_m K(x - x_m) = \sum_m k(\frac{x - x_m}{h})$$

- Gaussian Kernel

$$k(x) = (2\pi)^{-d/2} exp(-\frac{1}{2}||x||^2)$$

Mean-Shift

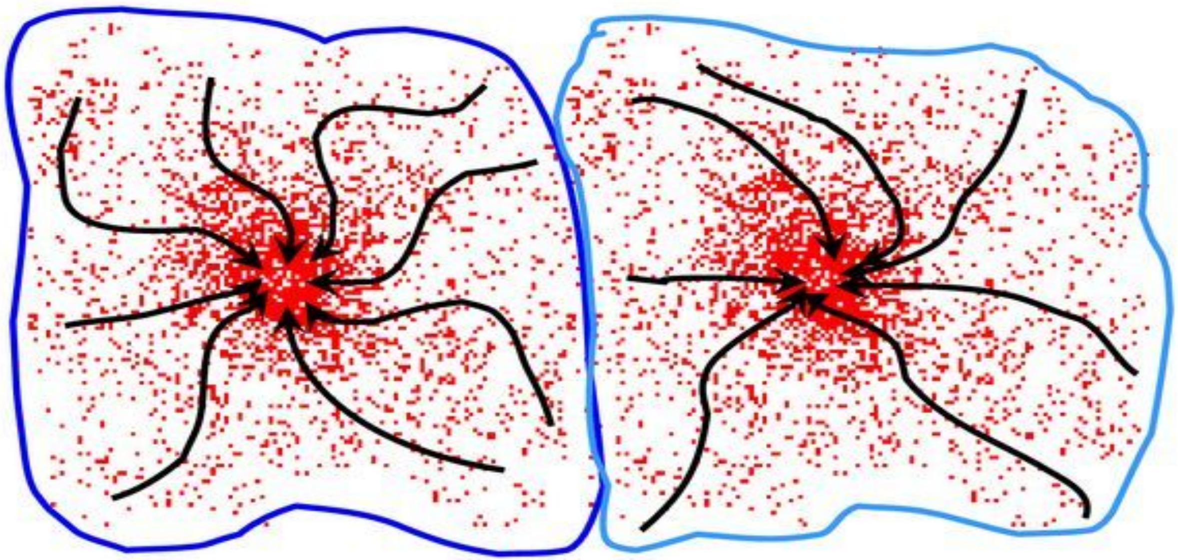


Mean-Shift

- Calculamos el mean shift vector
- Movemos el kernel windows

$$\hat{x} = \frac{\sum_m x_m exp(-\frac{1}{2}||\frac{x-x_m}{h}||^2)}{\sum_m exp(-\frac{1}{2}||\frac{x-x_m}{h}||^2)} - x$$

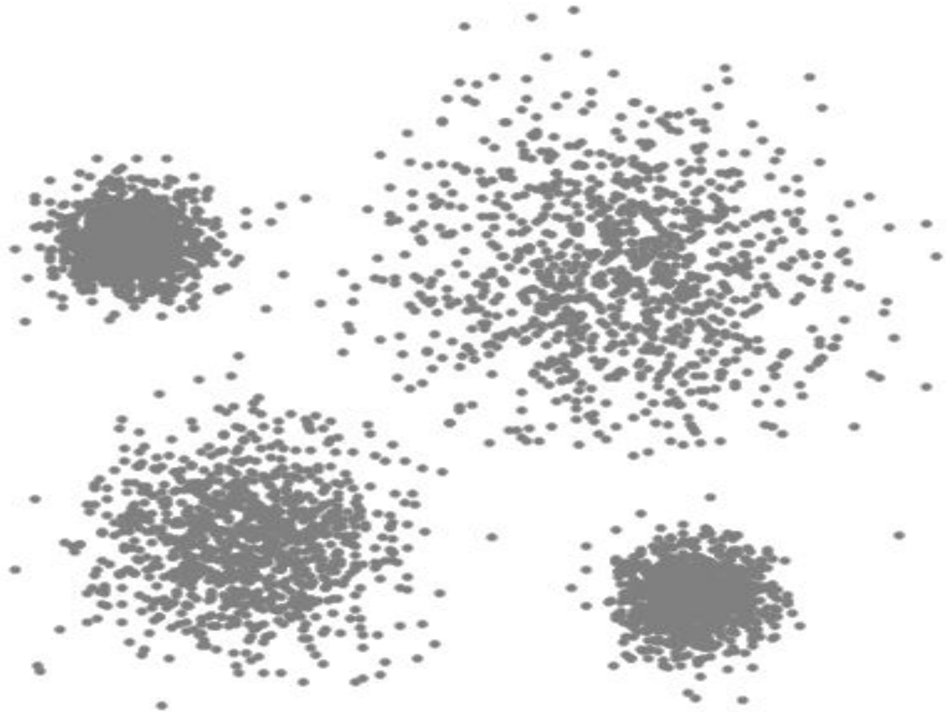
Mean-Shift



Mean-Shift

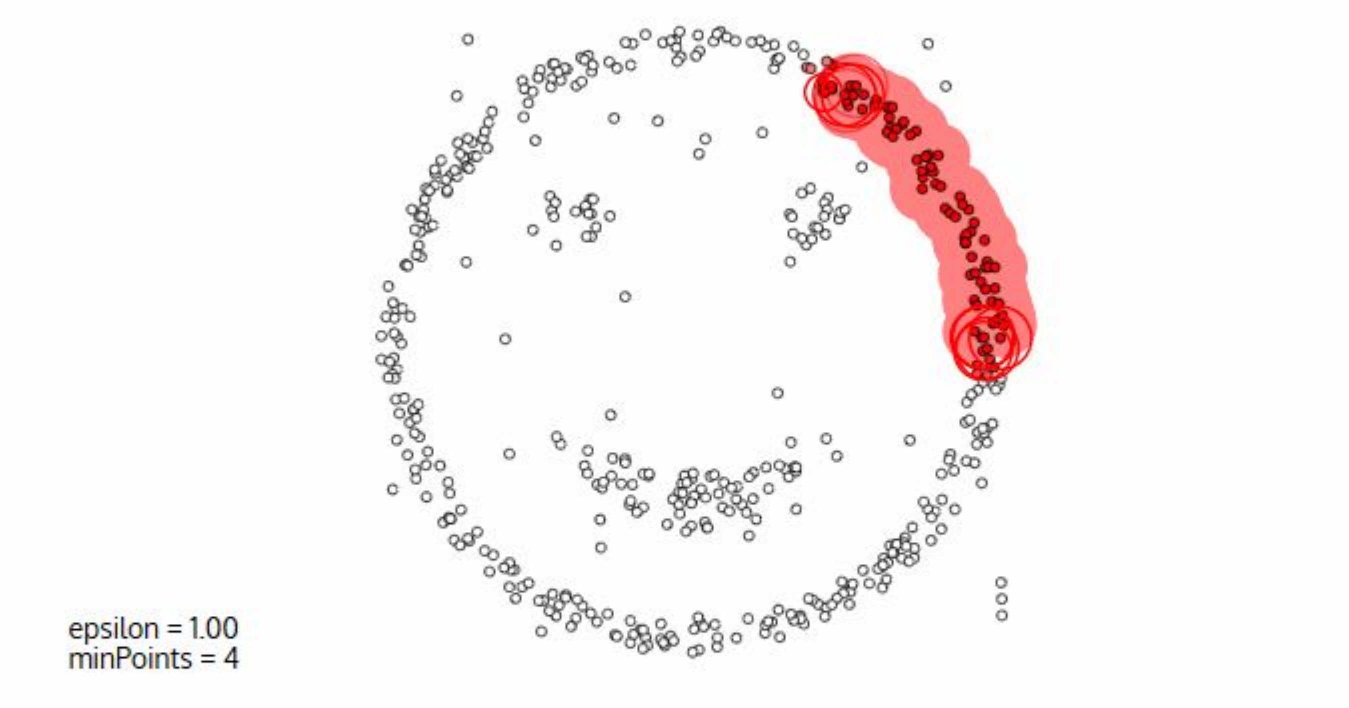
1. Elegir un kernel y un radio
2. Por cada punto p:
   a. Centrar una ventana en el punto p
   b. Computar la media de los datos en la ventana de búsqueda
   c. Centrar la ventana en la nueva localización
   d. Repetir (b,c) hasta que converga
3. Asignar el cluster el centroide a todos los nodos que le dieron orígen.

Mean-Shift



Fuente: Click

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)



Fuente: Click

ALGORITHM 1: Pseudocode of Original Sequential DBSCAN Algorithm

```
Input: DB: Database
Input: ε: Radius
Input: minPts: Density threshold
Input: dist: Distance function
Data: label: Point labels, initially undefined
foreach point p in database DB do                       // Iterate over every point
    if label(p) ≠ undefined then continue               // Skip processed points
    Neighbors N ← RANGEQUERY(DB, dist, p, ε)            // Find initial neighbors
    if |N| < minPts then                                // Non-core points are noise
        label(p) ← Noise
        continue
    c ← next cluster label                              // Start a new cluster
    label(p) ← c
    Seed set S ← N \ {p}                                // Expand neighborhood
    foreach q in S do
        if label(q) = Noise then label(q) ← c
        if label(q) ≠ undefined then continue
        Neighbors N ← RANGEQUERY(DB, dist, q, ε)
        label(q) ← c
        if |N| < minPts then continue                   // Core-point check
        S ← S ∪ N
```

Fuente: Click

Gracias

Inserta texto adicional aquí

UTEC
UNIVERSIDAD DE INGENIERÍA
Y TECNOLOGÍA