

# Clustering k-means

Gerardo Luis Muriel Lopez

Universidad del Valle  
Escuela de Ingeniería Sistemas y Computación  
Maestría en Ingeniería con énfasis en sistemas y computación  
Minería de datos

# Contenido

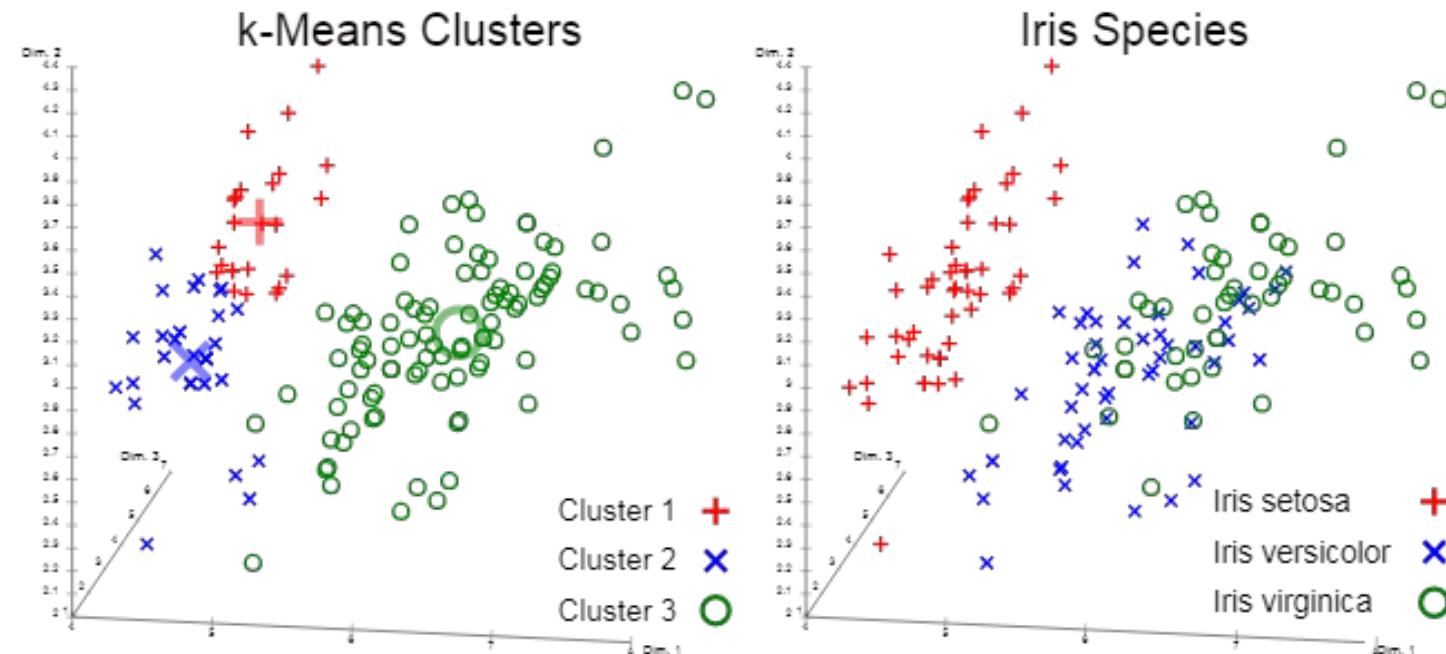
1. Algoritmo de clasificación no supervisada Clustering k-means
2. Reseña histórica
3. Precondiciones
4. Implementación en Jupyter Notebook Python 3.8
5. Casos de uso
6. Algoritmos de clustering
7. Referencias

# Contenido

1. Algoritmo de clasificación no supervisada Clustering k-means
2. Reseña histórica
3. Precondiciones
4. Implementación en Jupyter Notebook Python 3.8
5. Casos de uso
6. Algoritmos de clustering
7. Referencias

## 1. Algoritmo de clasificación no supervisada Clustering k-means

Este algoritmo de clasificación, sirve cuando se tiene un conjunto de datos de entrenamiento **no etiquetado**, es decir, que la clase a la cual pertenece un evento es desconocida. Lo que se busca es agrupar el conjunto de datos, de acuerdo a la similitud de sus características.



## *1. Algoritmo de clasificación no supervisada Clustering k-means*

Este algoritmo de clasificación, sirve cuando se tiene un conjunto de datos de entrenamiento **no etiquetado**, es decir, que la clase a la cual pertenece un evento es desconocida. Lo que se busca es agrupar el conjunto de datos, de acuerdo a la similitud de sus características.

“Automatic  
Classification”

## *1. Algoritmo de clasificación no supervisada Clustering k-means*

Este algoritmo de clasificación, sirve cuando se tiene un conjunto de datos de entrenamiento **no etiquetado**, es decir, que la clase a la cual pertenece un evento es desconocida. Lo que se busca es agrupar el conjunto de datos, de acuerdo a la similitud de sus características.

“Automatic  
Classification”

“Data  
segmentation”

## *1. Algoritmo de clasificación no supervisada Clustering k-means*

Este algoritmo de clasificación, sirve cuando se tiene un conjunto de datos de entrenamiento **no etiquetado**, es decir, que la clase a la cual pertenece un evento es desconocida. Lo que se busca es agrupar el conjunto de datos, de acuerdo a la similitud de sus características.

“Automatic  
Classification”

“Data  
segmentation”

“Outlier detection”

# *1. Algoritmo de clasificación no supervisada Clustering k-means*

Este algoritmo de clasificación, sirve cuando se tiene un conjunto de datos de entrenamiento **no etiquetado**, es decir, que la clase a la cual pertenece un evento es desconocida. Lo que se busca es agrupar el conjunto de datos, de acuerdo a la similitud de sus características.

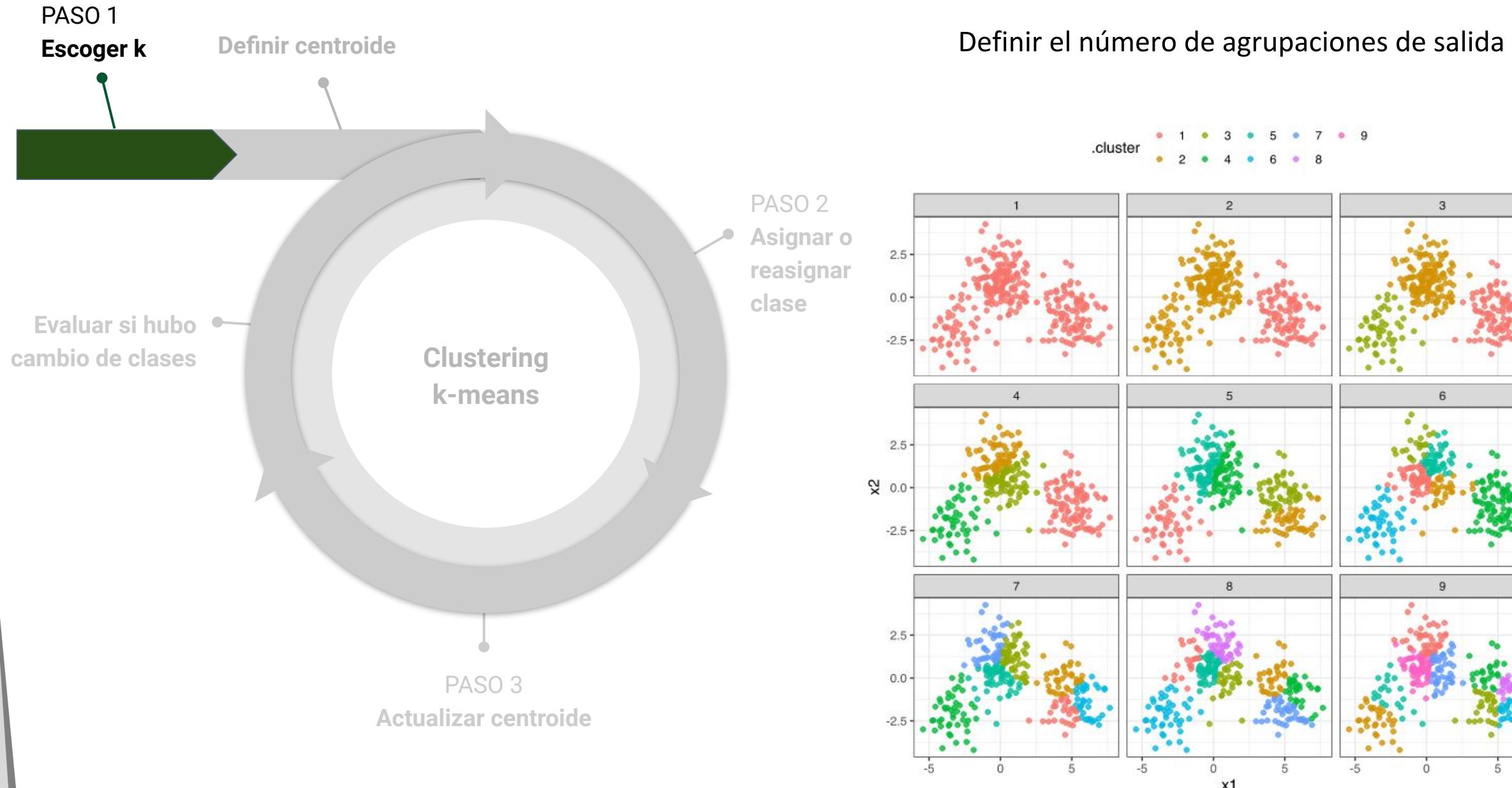
“Automatic  
Classification”

“Data  
segmentation”

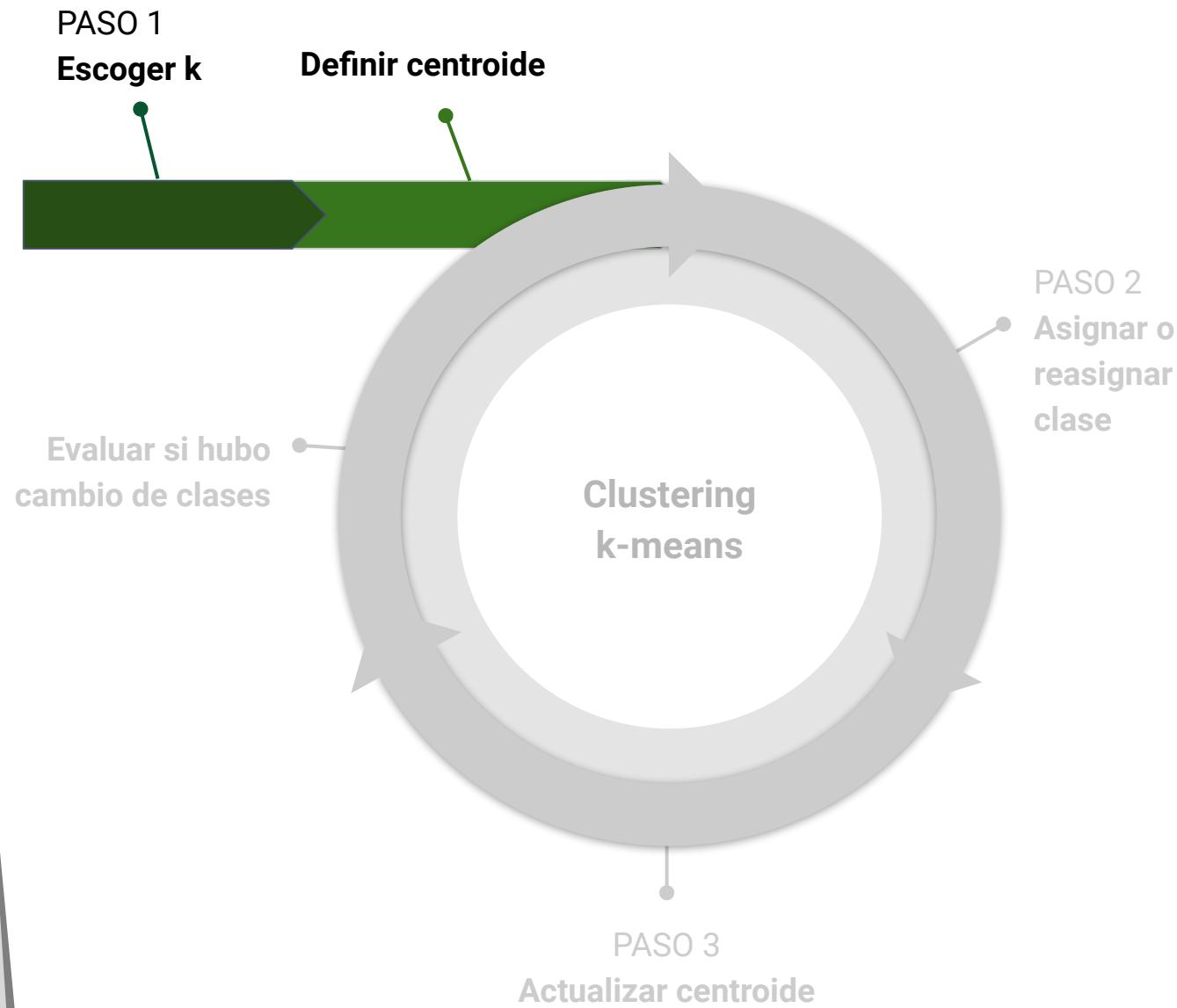
“Outlier detection”

“learning by  
observation”

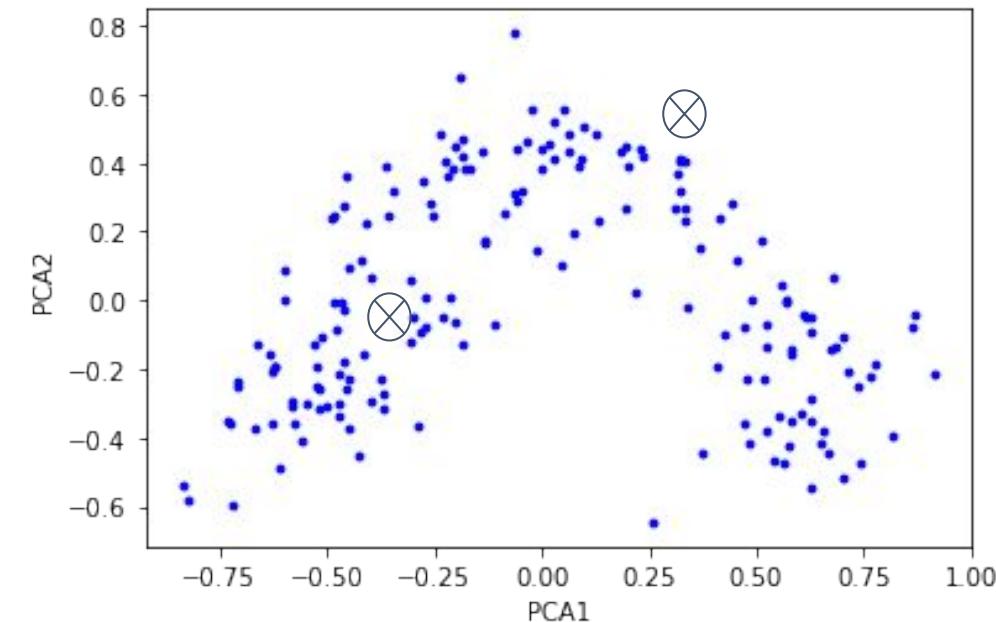
# 1. Algoritmo de clasificación no supervisada Clustering k-means



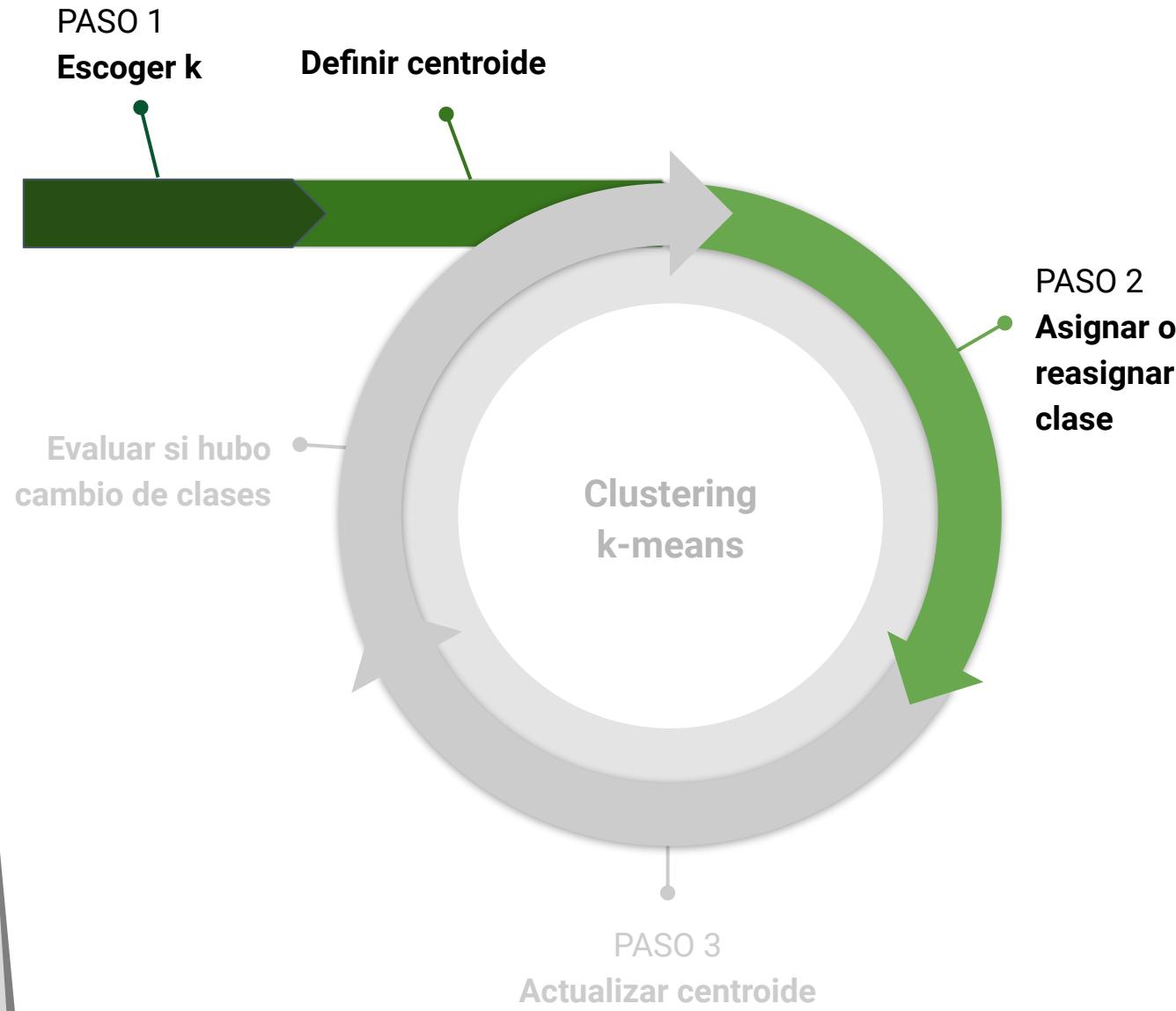
# 1. Algoritmo de clasificación no supervisada Clustering k-means



En un espacio multidimensional se define los centroides aleatorios o se escoge uno aleatoriamente del conjunto de datos, por cada agrupamiento

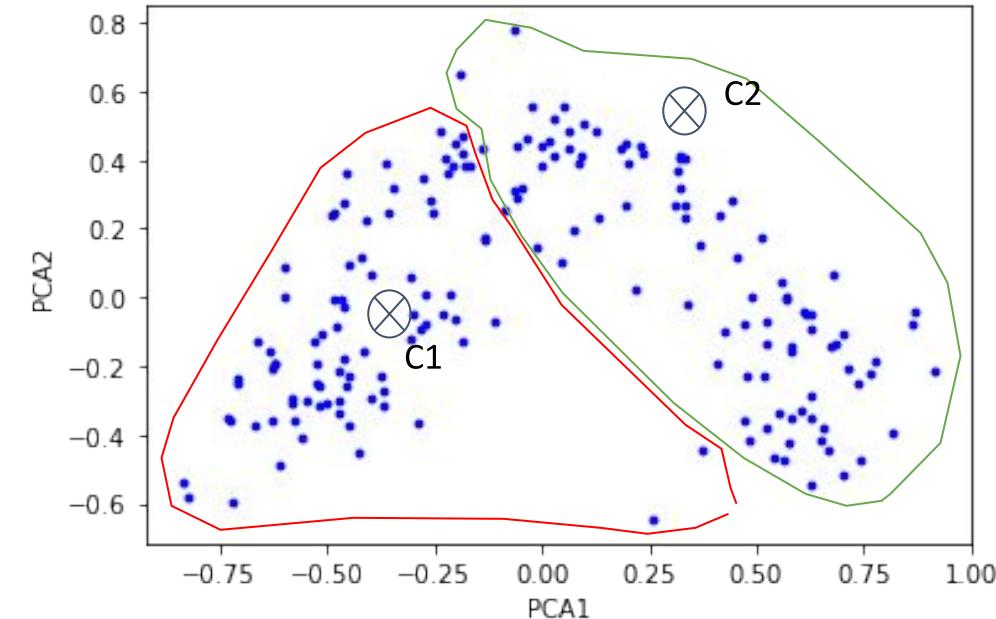


# 1. Algoritmo de clasificación no supervisada Clustering k-means

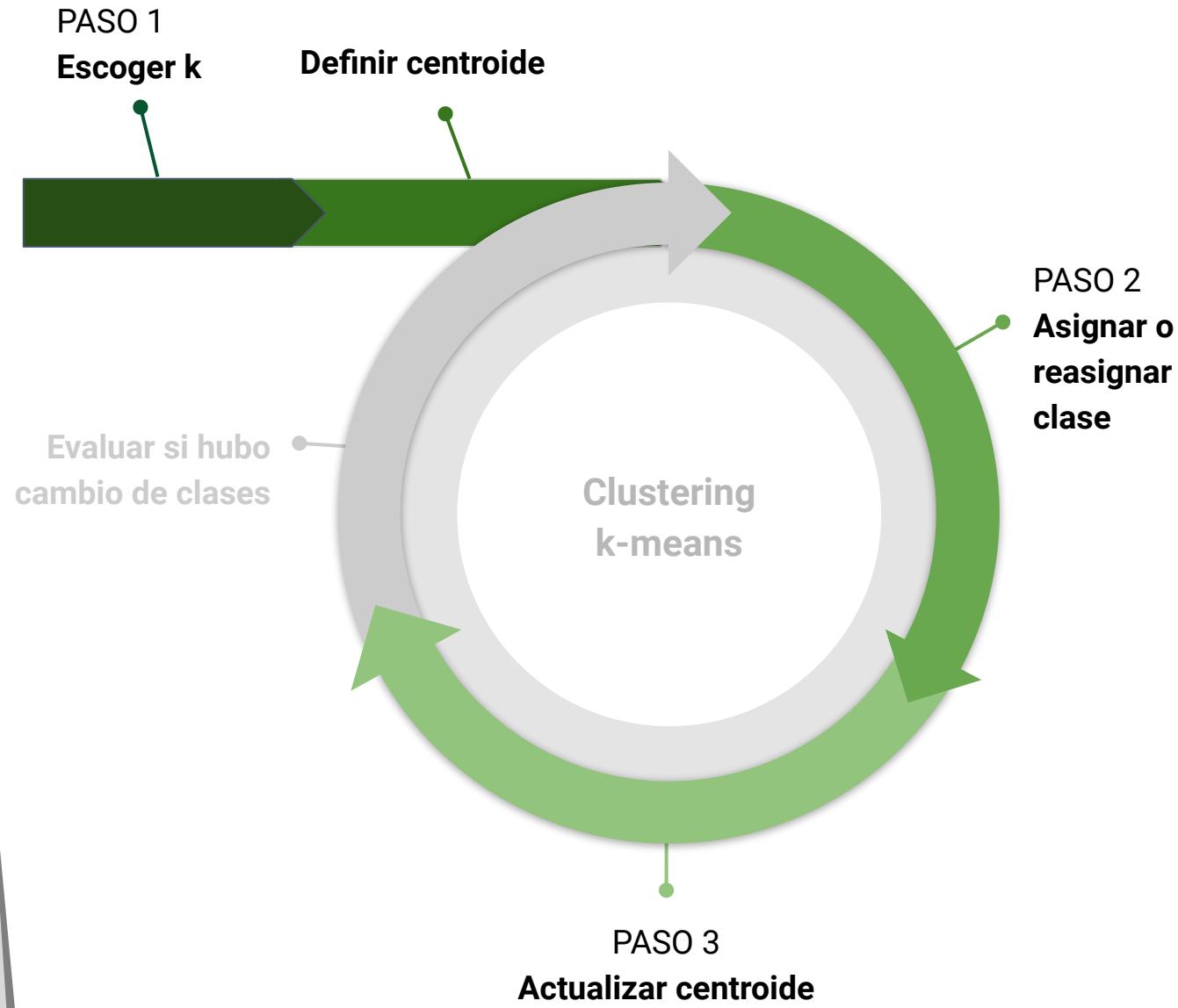


Se asigna cada dato al cluster que representa el centroide más cercano

$$dist(p, c_i) = \sqrt{(p_x - c_{ix})^2 + (p_y - c_{iy})^2}$$



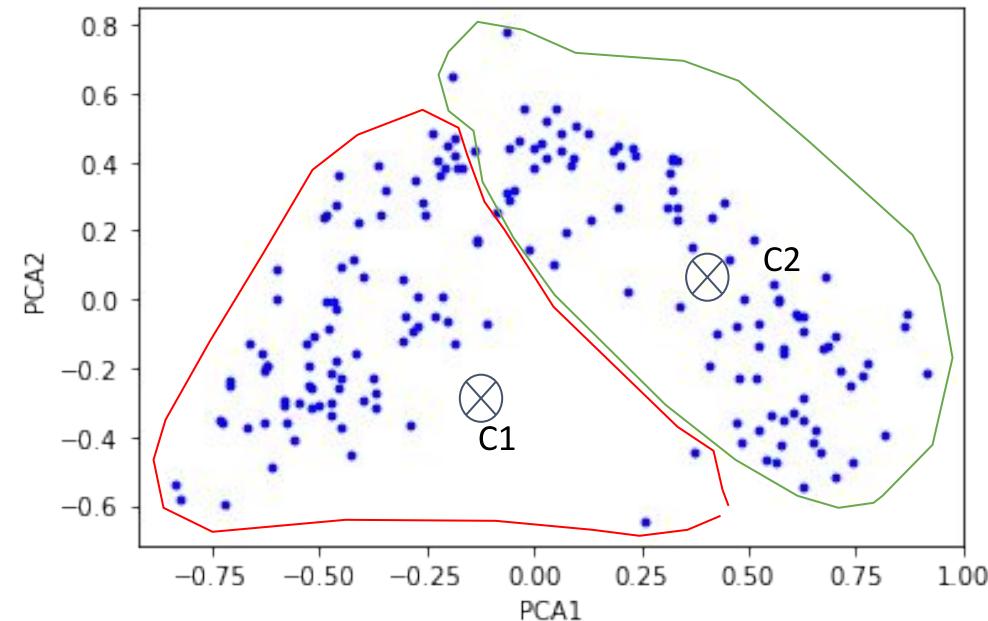
# 1. Algoritmo de clasificación no supervisada Clustering k-means



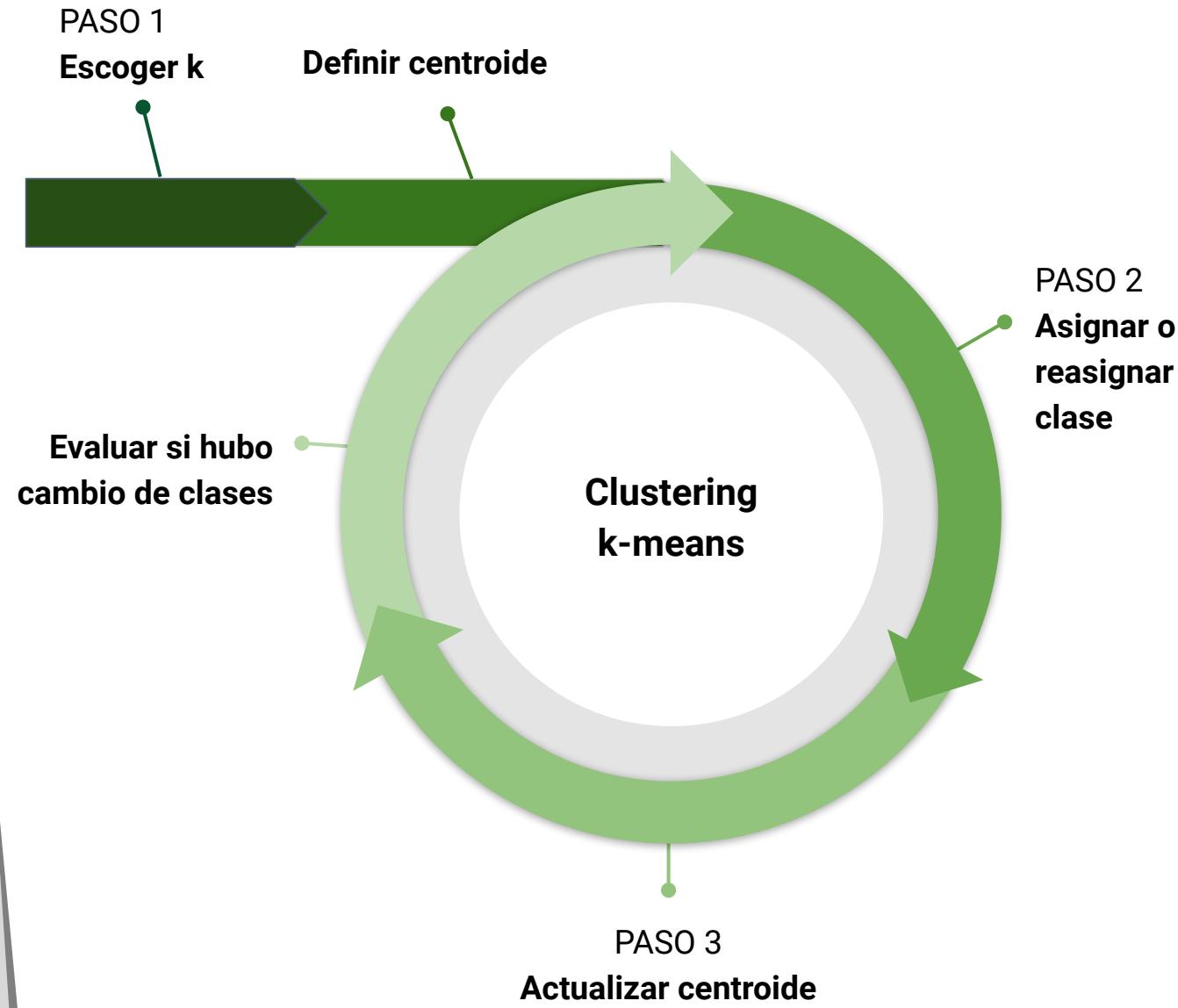
Recalcula los centroides de cada grupo como el promedio de los puntos que pertenecen

$$C_{ix} := \frac{1}{j} \sum_{n=1}^j p_{nx}$$

$$C_{iy} := \frac{1}{j} \sum_{n=1}^j p_{ny}$$



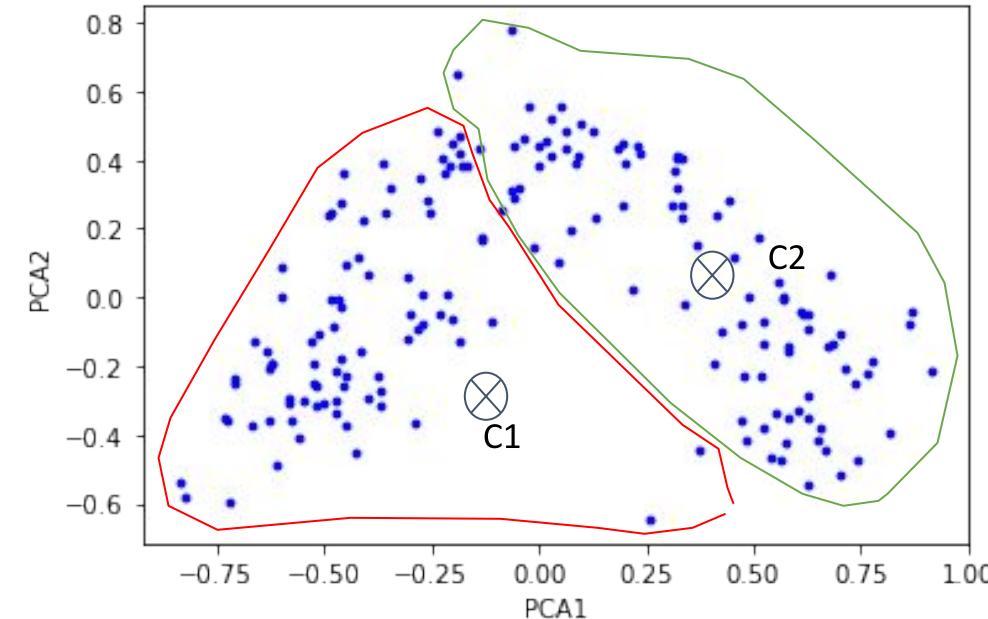
# 1. Algoritmo de clasificación no supervisada Clustering k-means



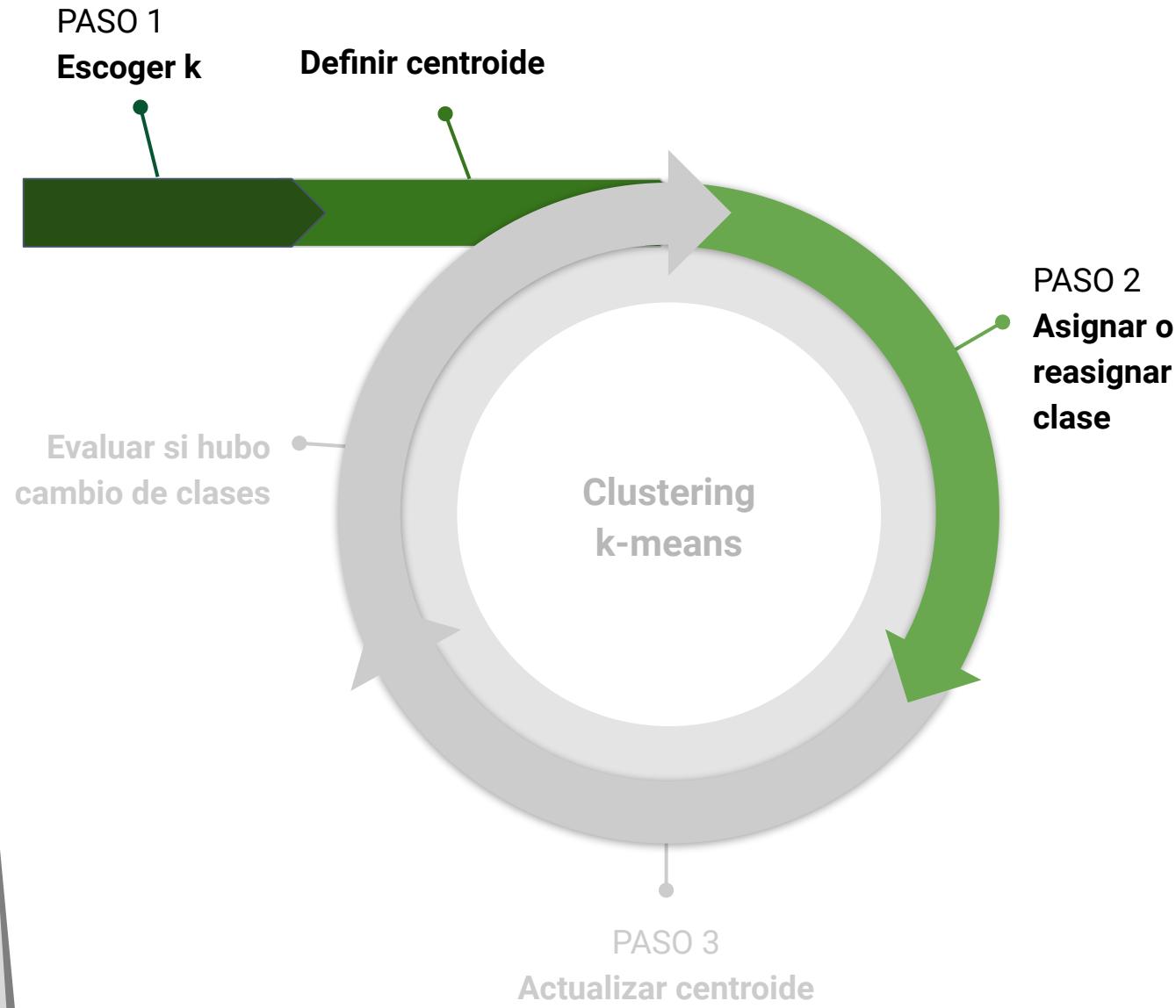
Recalcula los centroides de cada grupo como el promedio de los puntos que pertenecen

$$C_{ix} := \frac{1}{j} \sum_{n=1}^j p_{nx}$$

$$C_{iy} := \frac{1}{j} \sum_{n=1}^j p_{ny}$$

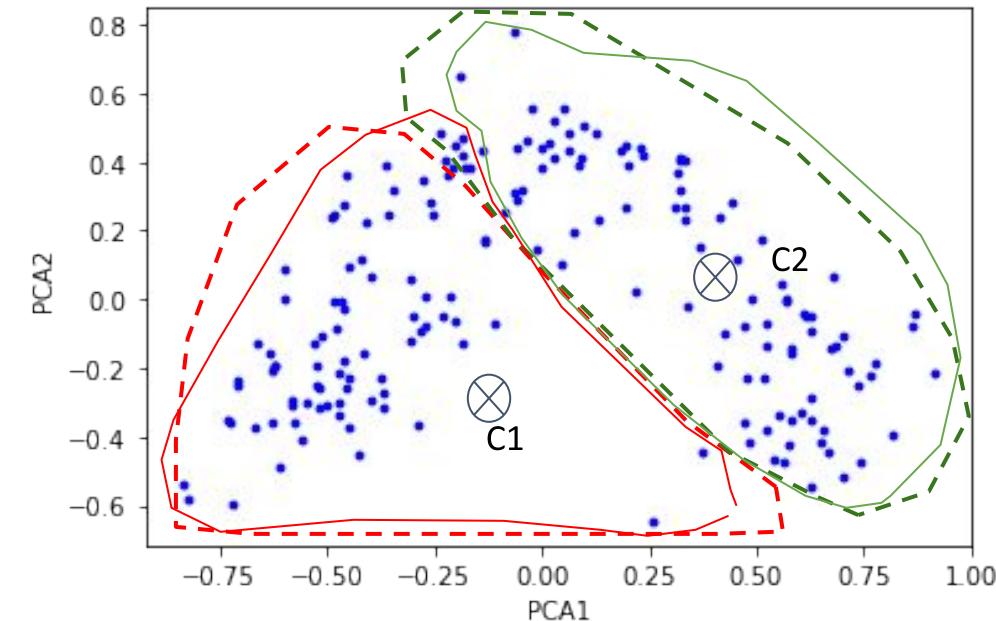


# 1. Algoritmo de clasificación no supervisada Clustering k-means

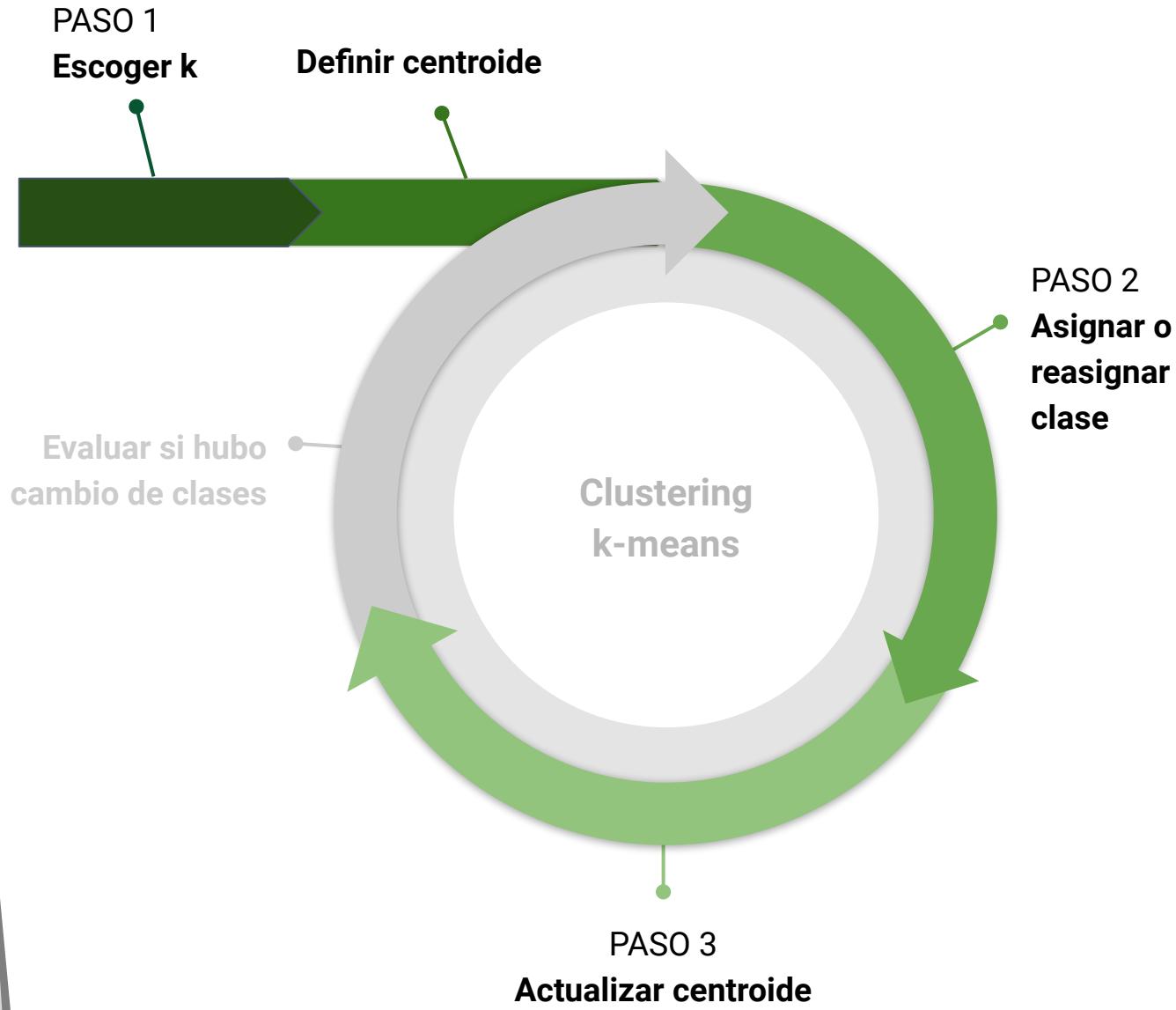


Cambio de clase de todos los datos de acuerdo a este nuevo centroide

$$dist(p, c_i) = \sqrt{(p_x - c_{ix})^2 + (p_y - c_{iy})^2}$$



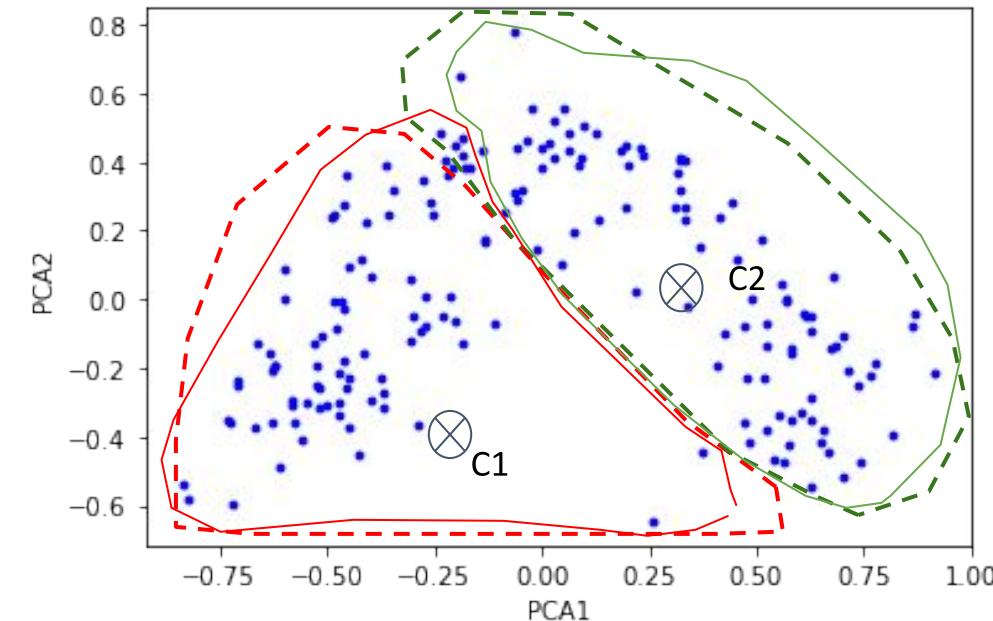
# 1. Algoritmo de clasificación no supervisada Clustering k-means



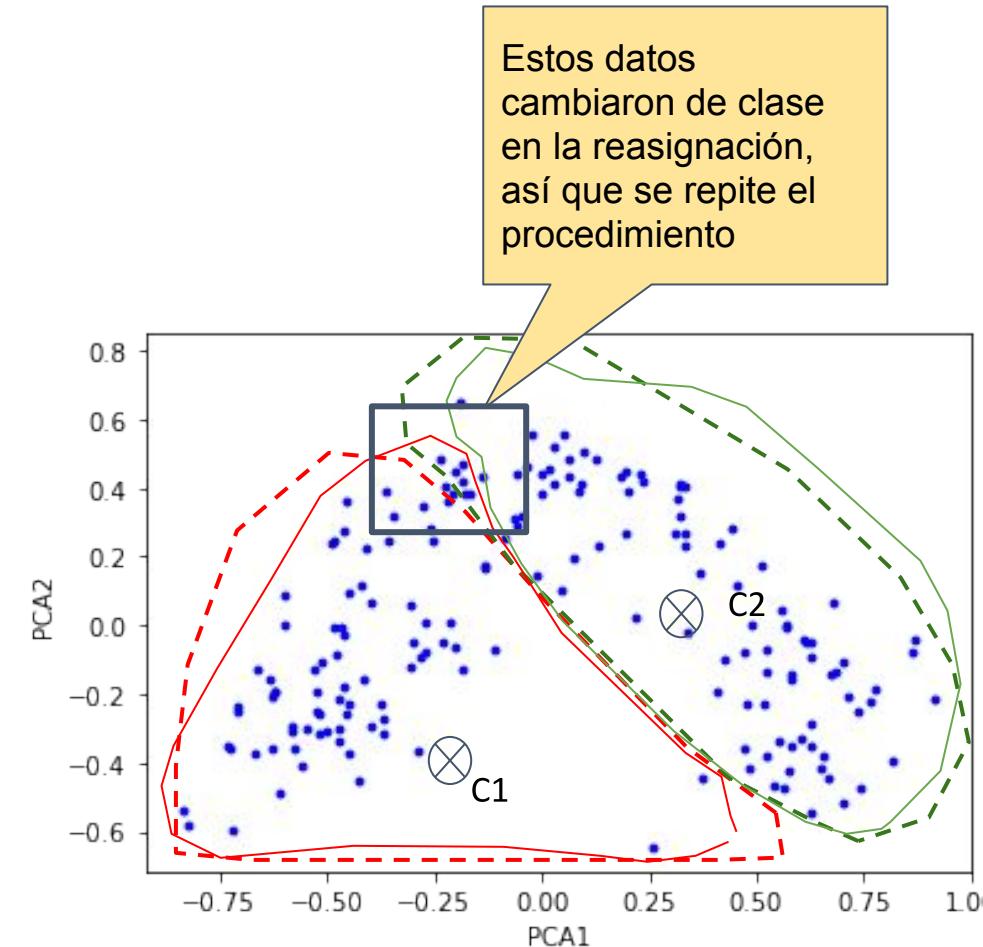
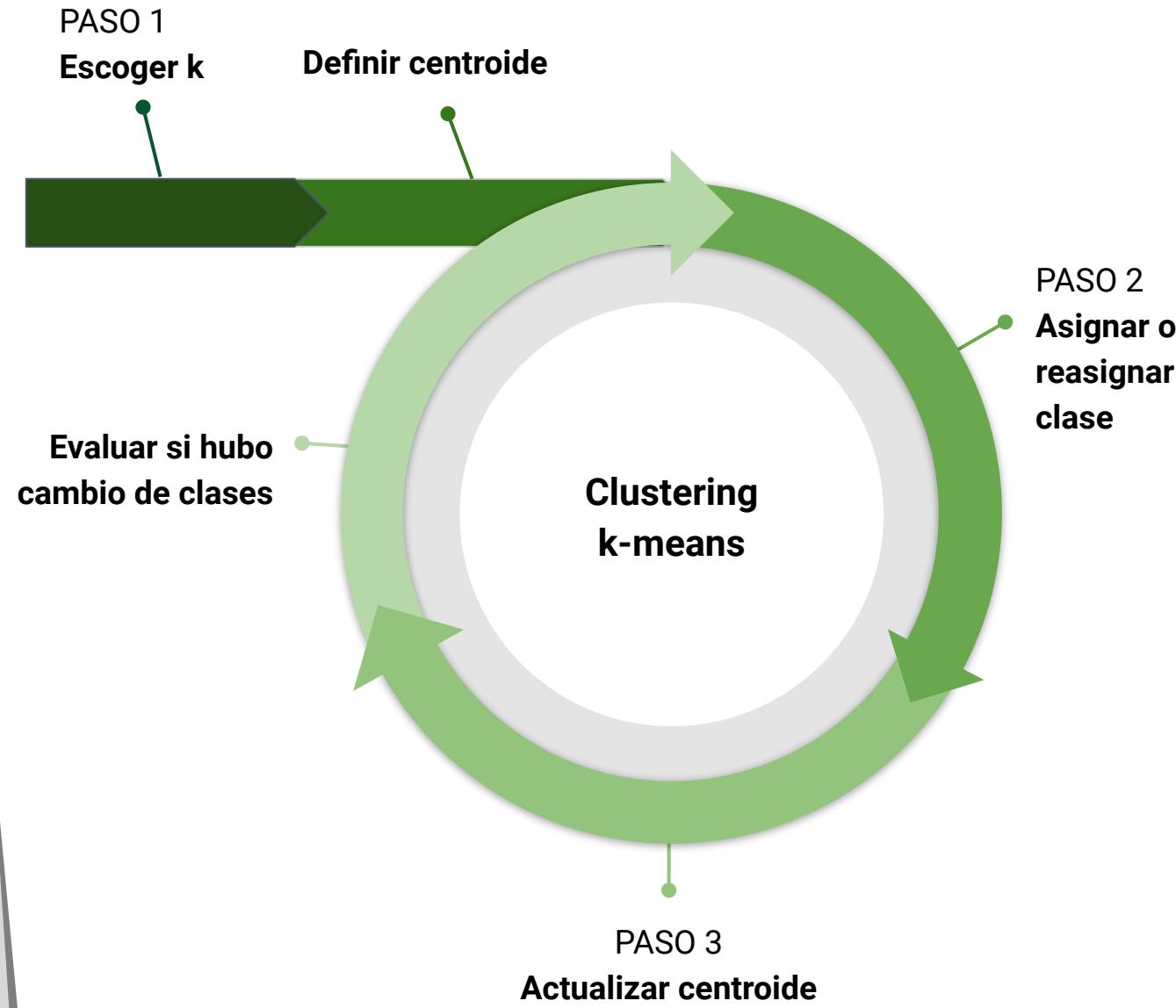
Recalcula los centroides de cada grupo como el promedio de los puntos que pertenecen

$$C_{ix} := \frac{1}{j} \sum_{n=1}^j p_{nx}$$

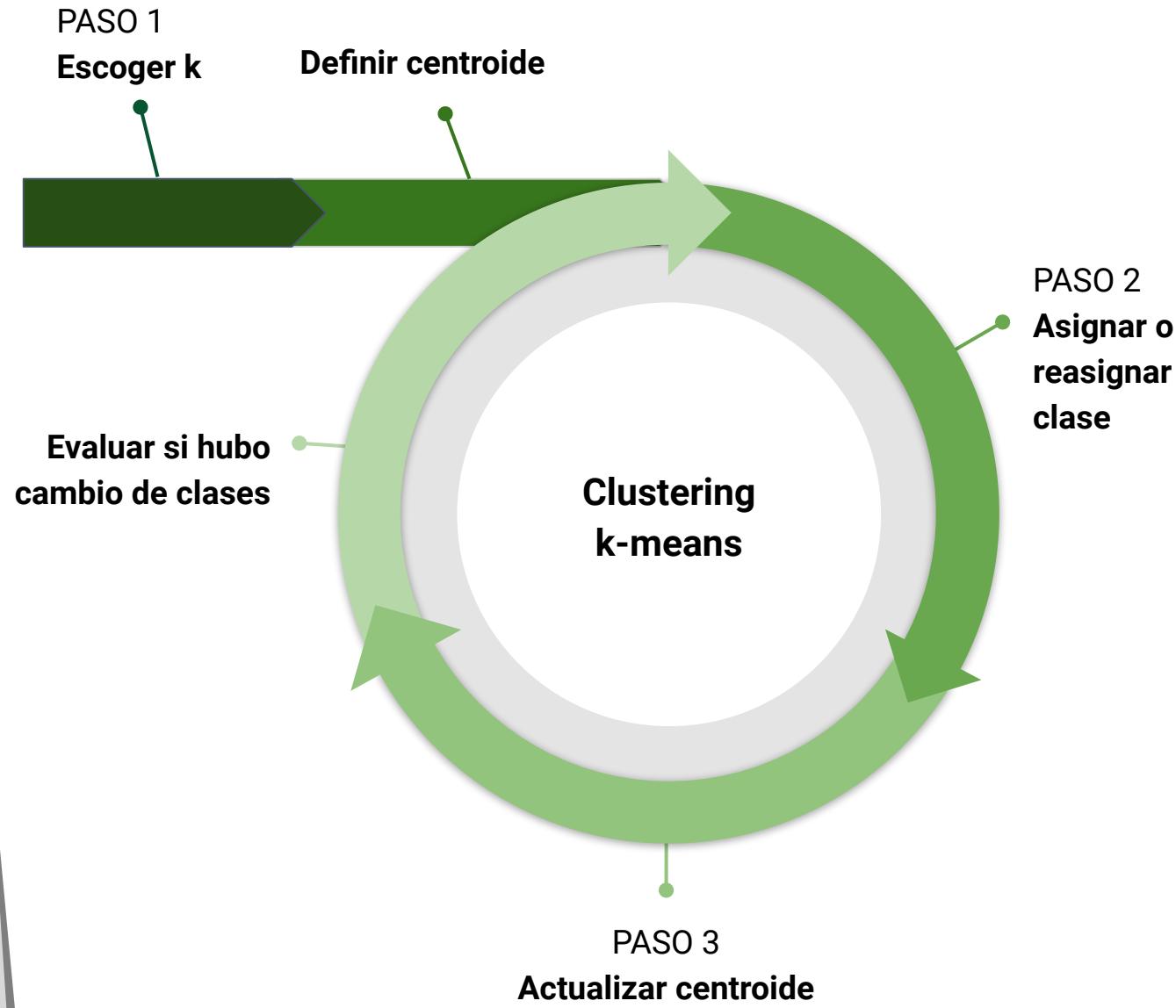
$$C_{iy} := \frac{1}{j} \sum_{n=1}^j p_{ny}$$



# 1. Algoritmo de clasificación no supervisada Clustering k-means



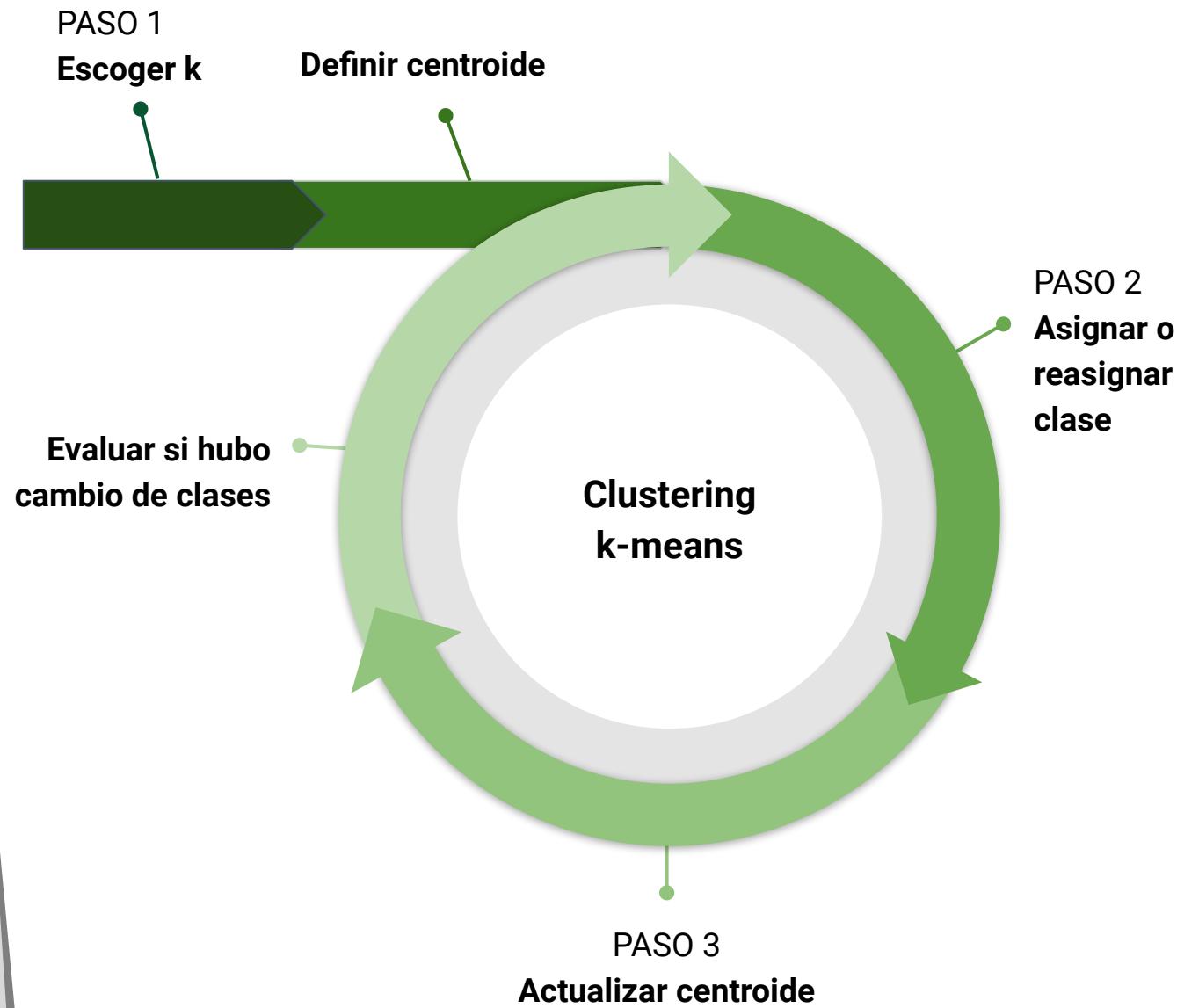
# 1. Algoritmo de clasificación no supervisada Clustering k-means



## Convergencia:

- 1- Hasta que los datos no cambien de clase en iteraciones consecutivas
- 2- Hasta un número de iteraciones definidas
- 3- Hasta que el grado de separabilidad de los datos alcance un mínimo aceptable

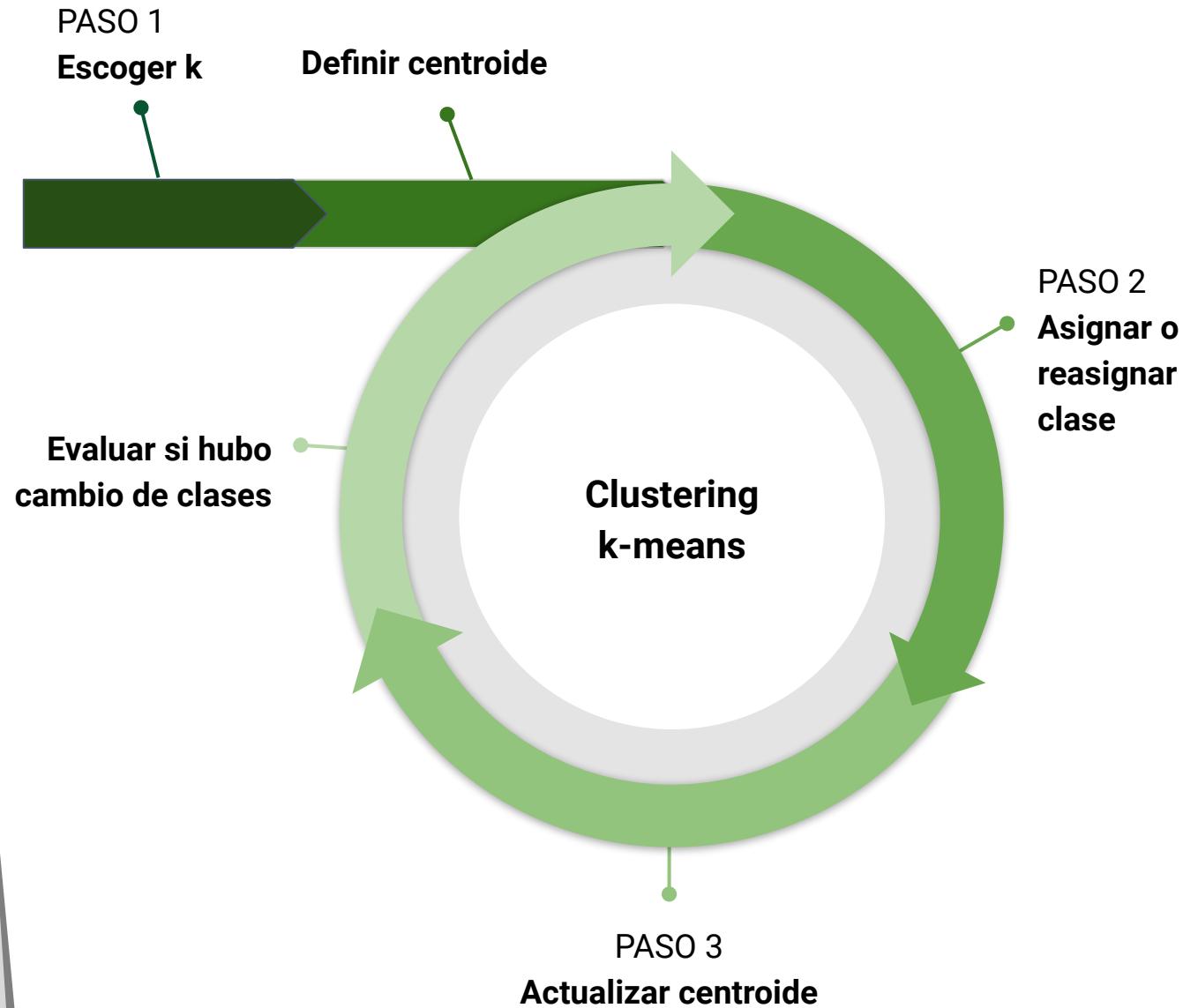
# 1. Algoritmo de clasificación no supervisada Clustering k-means



## Convergencia:

- 1- Hasta que los datos no cambien de clase en iteraciones consecutivas
- 2- Hasta un número de iteraciones definidas
- 3- Hasta que el grado de separabilidad de los datos alcance un mínimo aceptable

# 1. Algoritmo de clasificación no supervisada Clustering k-means



**Convergencia:**

Grado de separabilidad

$$E = \sum_{i=1}^k \sum_{p \in C_i} \text{dist}(p, c_i)^2$$

Donde :

$E$  : Suma del error cuadratico

$p$  : es el punto que representa un dato

$c_i$  es el centroide del agrupamiento  $C_i$

$k$  : es numero de agrupamientos

# 1. Algoritmo de clasificación no supervisada Clustering k-means

## Ventajas y “desventajas”

✓ Encuentra clases automáticamente

✓ No requiere de muestras etiquetadas

✓ Poco robusto y por lo tanto facil de implementar

! Limitaciones al agrupar formas no esfericas

! Es complicado escoger el valor optimo de k

! Muy sensible a los valores atipicos

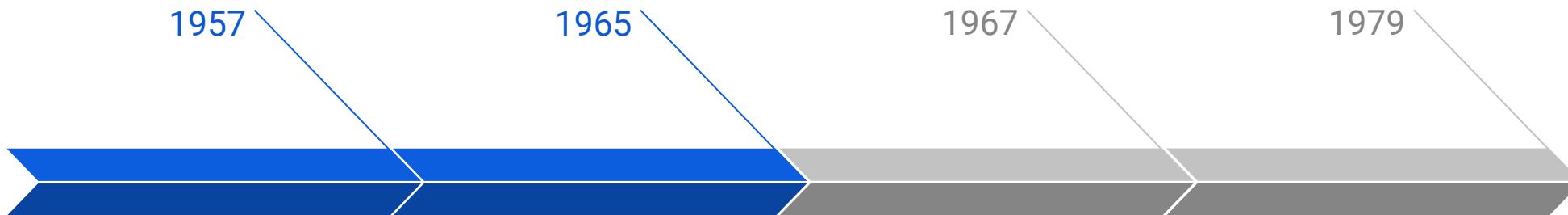
! Dificil de implementar para tipos de datos no numericos

! Resultados dependientes del orden de los datos de entrada

# Contenido

1. Algoritmo de clasificación no supervisada Clustering k-means
2. Reseña histórica
3. Precondiciones
4. Implementación en Jupyter Notebook Python 3.8
5. Casos de uso
6. Algoritmos de clustering
7. Referencias

## 2. Reseña histórica



Idea del algoritmo	Uso en clasificación	Término k-means	Versión Eficiente
Stuart Lloyd propone un método para modulación de impulsos codificados	E. W. Forgy, publica esencialmente el mismo algoritmo para clasificación	James MacQueen, le da el nombre de k-means como método de clasificación	Hartigan y Wong, publican una versión más eficiente del algoritmo en Fortran.
	“Cluster analysis of multivariate data: efficiency versus interpretability of classifications”	“Some Methods for classification and Analysis of Multivariate Observations”	“Algorithm AS 136: A k-medias Clustering Algorithm”

# Contenido

1. Algoritmo de clasificación no supervisada Clustering k-means
2. Reseña histórica
3. Precondiciones
4. Implementación en Jupyter Notebook Python 3.8
5. Casos de uso
6. Algoritmos de clustering
7. Referencias

### *3.Precondiciones*

*1- Definir el número de grupos o cluster (k)*

*2- Definir el espacio*

*3- Definir el numero de iteraciones maxima*

*4- Convertir a datos numericos*

*5- Normalizar las variables*

# Contenido

1. Algoritmo de clasificación no supervisada Clustering k-means
2. Reseña histórica
3. Precondiciones
- 4. Implementación en Jupyter Notebook Python 3.8**
5. Casos de uso
6. Algoritmos de clustering
7. Referencias

# Contenido

1. Algoritmo de clasificación no supervisada Clustering k-means
2. Reseña histórica
3. Precondiciones
4. Implementación en Jupyter Notebook Python 3.8
- 5. Casos de uso**
6. Algoritmos de clustering
7. Referencias

## 5. Casos de uso

Entender la  
distribucion de  
los datos

Observar  
caracteristicas  
de cada grupo

Herramienta de  
preprocesamiento  
para escoger  
datos

Detección de  
valores  
atípicos Ej:  
fraudes

Marketing

Busquedas  
web

Biología

Teledetección

# Contenido

1. Algoritmo de clasificación no supervisada Clustering k-means
2. Reseña histórica
3. Precondiciones
4. Implementación en Jupyter Notebook Python 3.8
5. Casos de uso
6. Algoritmos de clustering
7. Referencias

## 6. Algoritmos de clustering

01

### Métodos de partición

- Basado en la distancias
- Clases independientes
- Utiliza diferentes medidas de tendencia
- Ideales para conjuntos de datos pequeños a medianos (cientos a miles)

- **k-means**
- k-medoids

02

### Métodos Jerárquicos:

- Soporta diferentes niveles en cada clase
- Las fusiones o divisiones erróneas no se pueden corregir
- Se puede incluir otras técnicas como microclustering

- AGNES (AGglomerative NESting),
- DIANA (DIvisive ANAlysis)
- Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)
- Chameleon

## 6. Algoritmos de clustering

03

### Métodos basado en la densidad

- Agrupa formas aleatorias
- Puede filtrar valores atípicos

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- OPTICS: Ordering Points to Identify the Clustering Structure
- DENCLUE (DENsity-based CLUstEring)

04

### Métodos basado en grilla:

- Utiliza una grilla con múltiples resoluciones
- El tiempo de procesamiento es relativamente rápido, y no depende de la cantidad de los datos sino de la grilla

- STING: SStatistical INformation Grid
- CLIQUE (CLustering In QUEst)

## 6. Referencias

E.W. Forgy. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, 21, 768–769.

Han, J., Kamber, M., & Pei, J. (2012). 10—Cluster Analysis: Basic Concepts and Methods. In J. Han, M. Kamber, & J. Pei (Eds.), *Data Mining (Third Edition)* (Third Edition, pp. 443–495). Morgan Kaufmann.  
<https://doi.org/10.1016/B978-0-12-381479-1.00010-1>

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, 28(1), 100.  
<https://doi.org/10.2307/2346830>

MacQueen, J., & others. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281–297.

Clustering Método K-Means en Python: <https://www.youtube.com/watch?v=s6PSSzeUMFk>

Sklearn K-Means Python Example | Interpreting Clustering results:<https://www.youtube.com/watch?v=3Spa10-mwsW>

*Muchas Gracias*