Web Scraping & K-means

Juan Felipe Acevedo Pérez

Monitor Junior

UIFCE
UNIDAD DE INFORMÁTICA

Curso Libre

Correo:
uniic_bog@unal.edu.co
Tel: 3165000 ***Ext:*** 12301

Facultad de Ciencias Económicas
Unidad de Informática
Sede Bogotá

UNIVERSIDAD NACIONAL DE COLOMBIA

Contenido

Web Scraping

Inercia

K-means

Correo: uniic_bog@unal.edu.co
Teléfono: 3165000 **ext** 12301

1

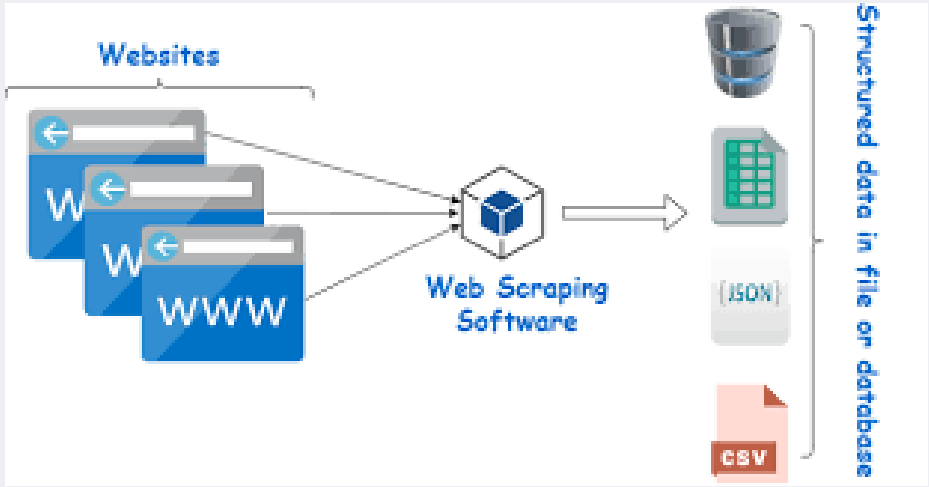
Web Scraping

Correo: uniic_bog@unal.edu.co
Teléfono: 3165000 **ext** 12301

UIFCE
UNIDAD DE INFORMÁTICA

Web Scraping

Web scraping o raspado web, es una técnica utilizada mediante programas de software para extraer información de sitios web. Usualmente, estos programas simulan la navegación de un humano en la World Wide Web ya sea utilizando el protocolo HTTP manualmente, o incrustando un navegador en una aplicación.



Correo: uniic_bog@unal.edu.co
Teléfono: 3165000 **ext** 12301

2

Inercia

Correo: uniic_bog@unal.edu.co
Teléfono: 3165000 **ext** 12301

UIFCE
UNIDAD DE INFORMÁTICA

Inercia

El algoritmo K-means tiene como objetivo elegir centroides que minimicen la inercia, o el criterio de suma de cuadrados dentro del grupo:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Problemas

- La inercia se puede reconocer como una medida de la coherencia interna de los clústeres. Adolece de varios inconvenientes:

  - **Supone** que los cúmulos son convexos e isotrópicos, lo que no siempre es así. Responde mal a racimos alargados o variedades con formas irregulares.

  - **No** es una métrica **normalizada**: solo sabemos que los valores más bajos son mejores y cero es óptimo. Pero en espacios de dimensiones muy altas, las distancias euclidianas tienden a inflarse (este es un ejemplo de la llamada "maldición de la dimensionalidad"). Ejecutar un algoritmo de reducción de dimensionalidad como el análisis de componentes principales (PCA) antes del agrupamiento de k-medias puede aliviar este problema y acelerar los cálculos.

3

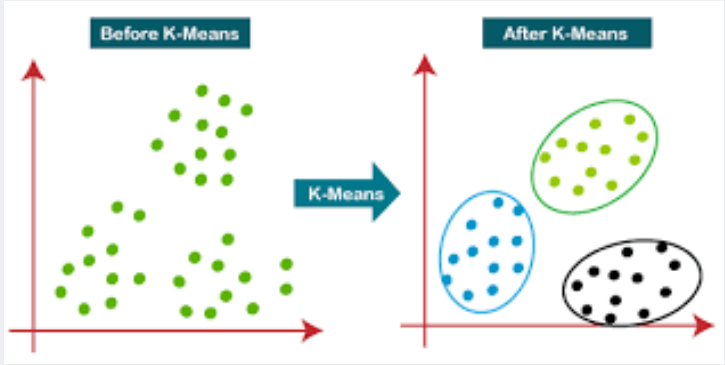
K-means

Correo: uniic_bog@unal.edu.co
Teléfono: 3165000 **ext** 12301

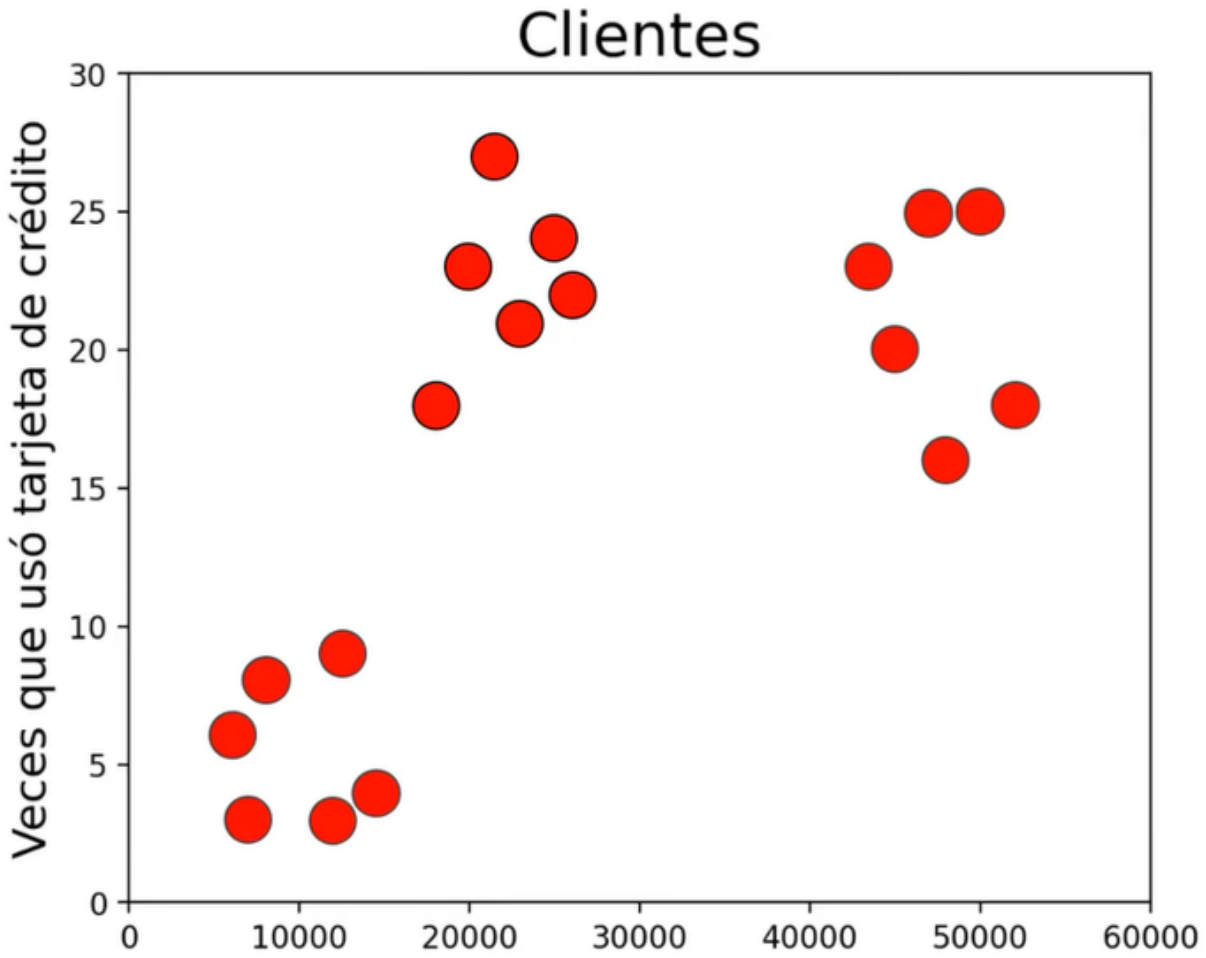
UIFCE
UNIDAD DE INFORMÁTICA

¿Qué es?

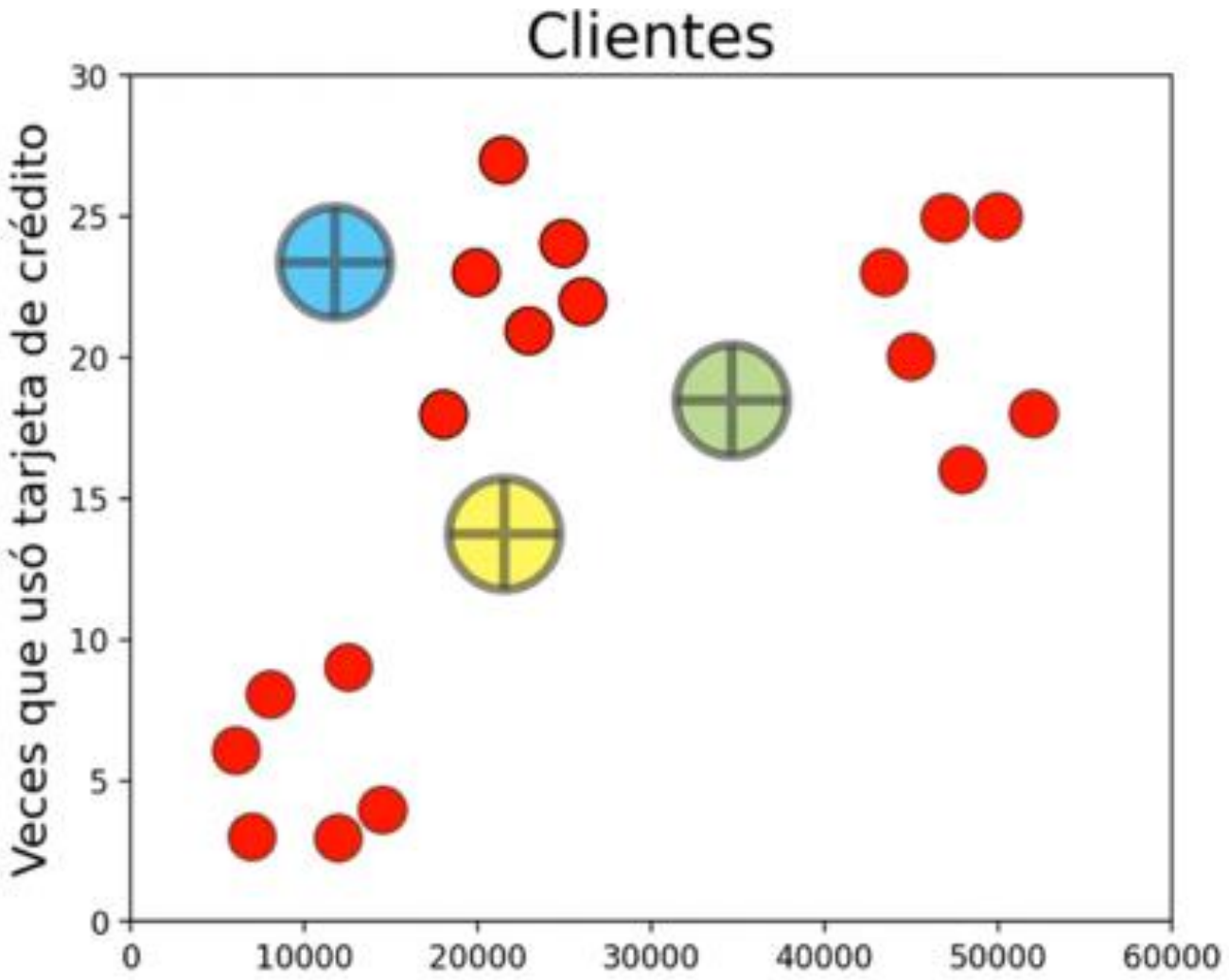
El algoritmo K-Means agrupa los datos tratando de **separar muestras en n grupos** de igual *varianza*, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo. Este algoritmo requiere que se especifique el número de **clusters**. Se adapta bien a una gran cantidad de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.



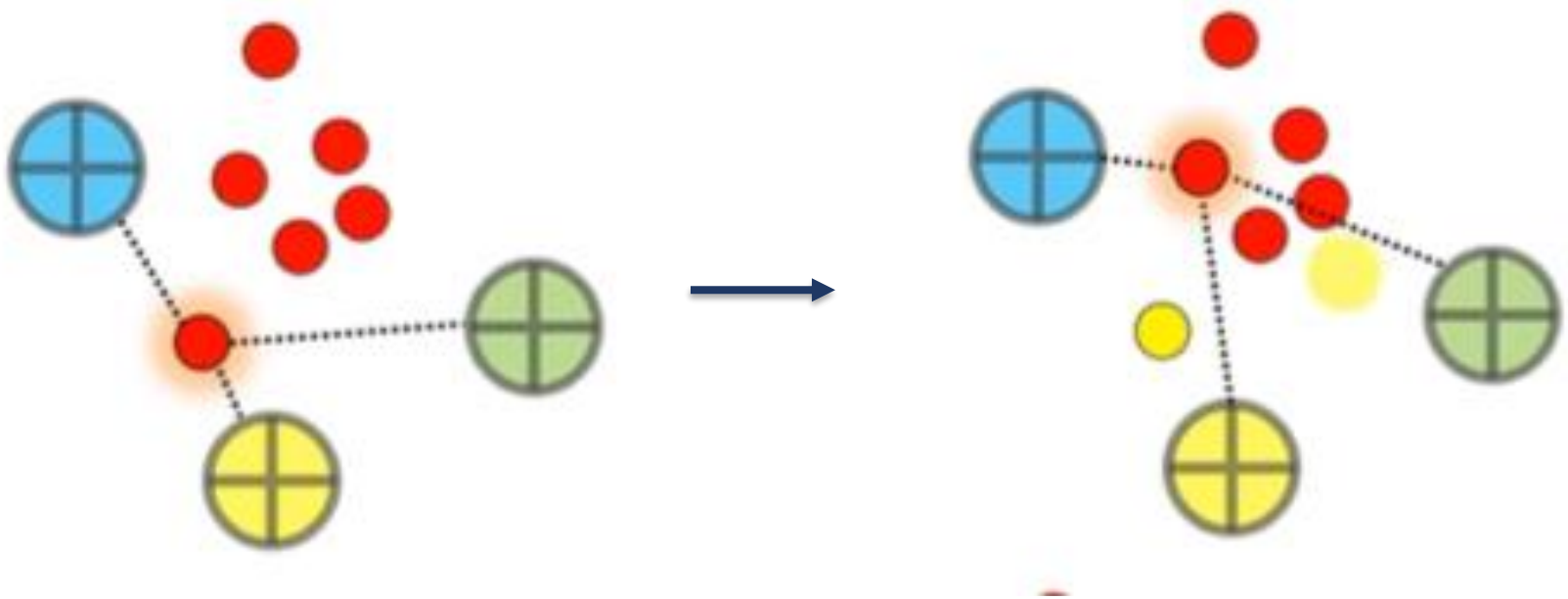
El algoritmo k-means divide un conjunto de N muestras X en K grupos separados C, cada uno descrito por la media de las muestras en el grupo. Los medios se denominan comúnmente **"centroides"** del grupo; tenga en cuenta que no son, en general, puntos de X, aunque viven en el mismo espacio.



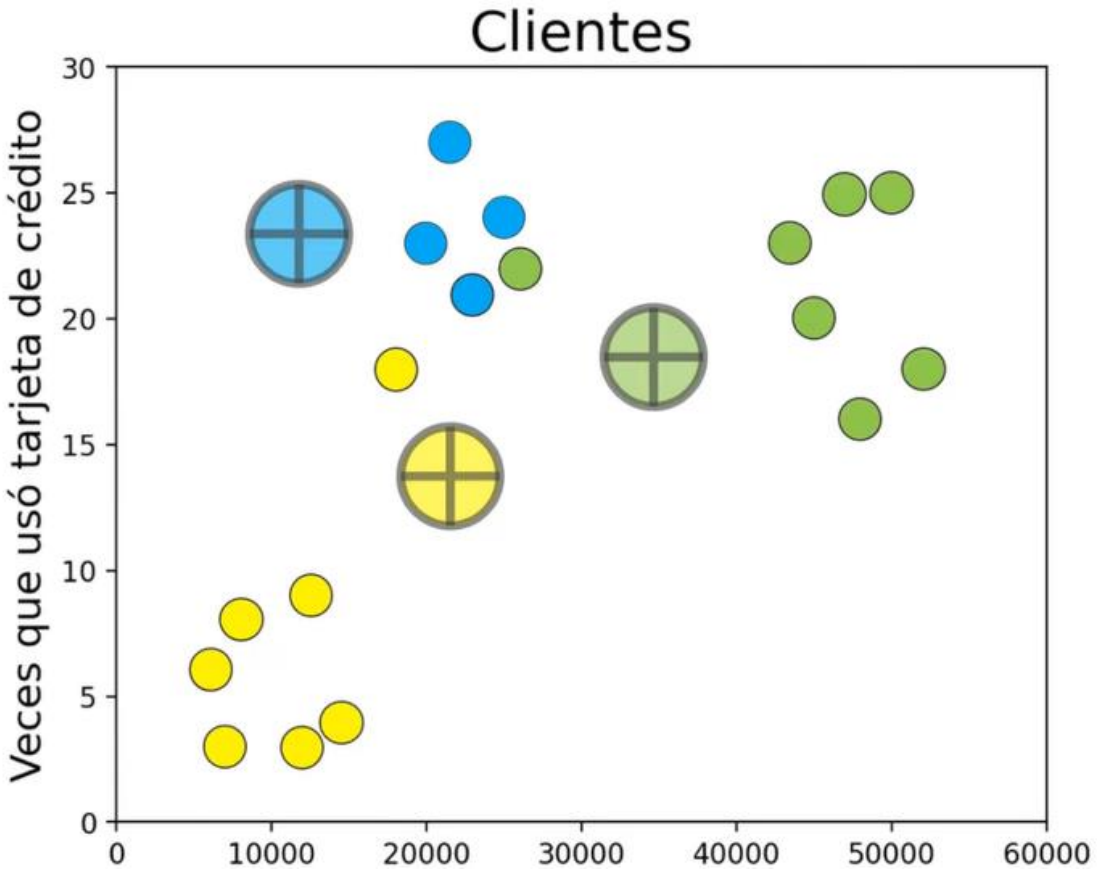
Suponemos el valor k de clusters (o lo encontramos con métodos estadísticos)



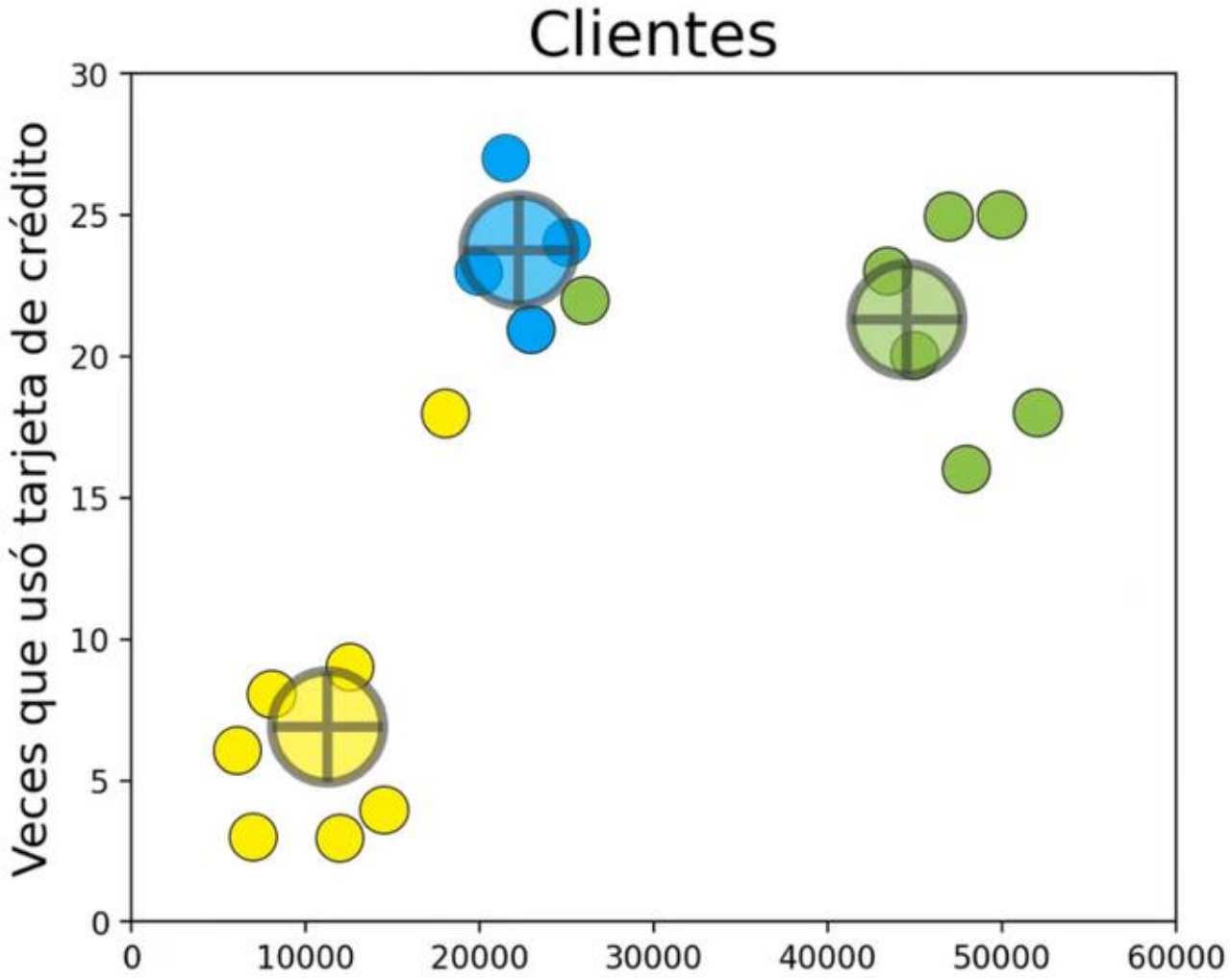
Suponiendo **K = 3** tenemos tres centroides (eventualmente se deben ubicar en el centro de cada agrupación de grupos) ubicados de forma aleatoria



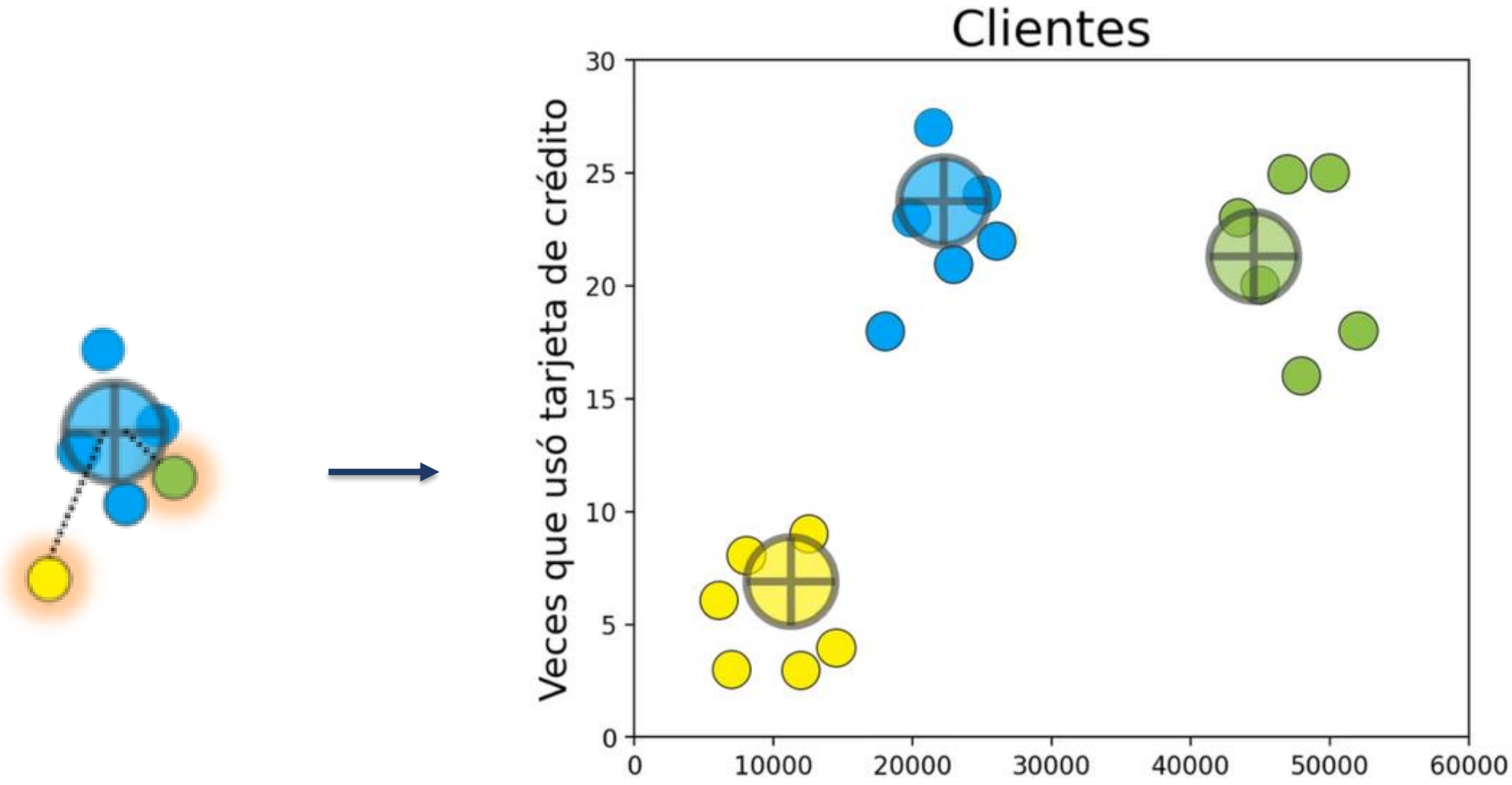
Para que cada centroide llegue al centro de los potenciales clusters se calcula la **distancia euclidiana** (hipotenusa) de cada punto al centroide. En este caso lo asignamos momentáneamente al color amarillo, pues es el más cercano. Continuamos con la misma lógica para cada punto.



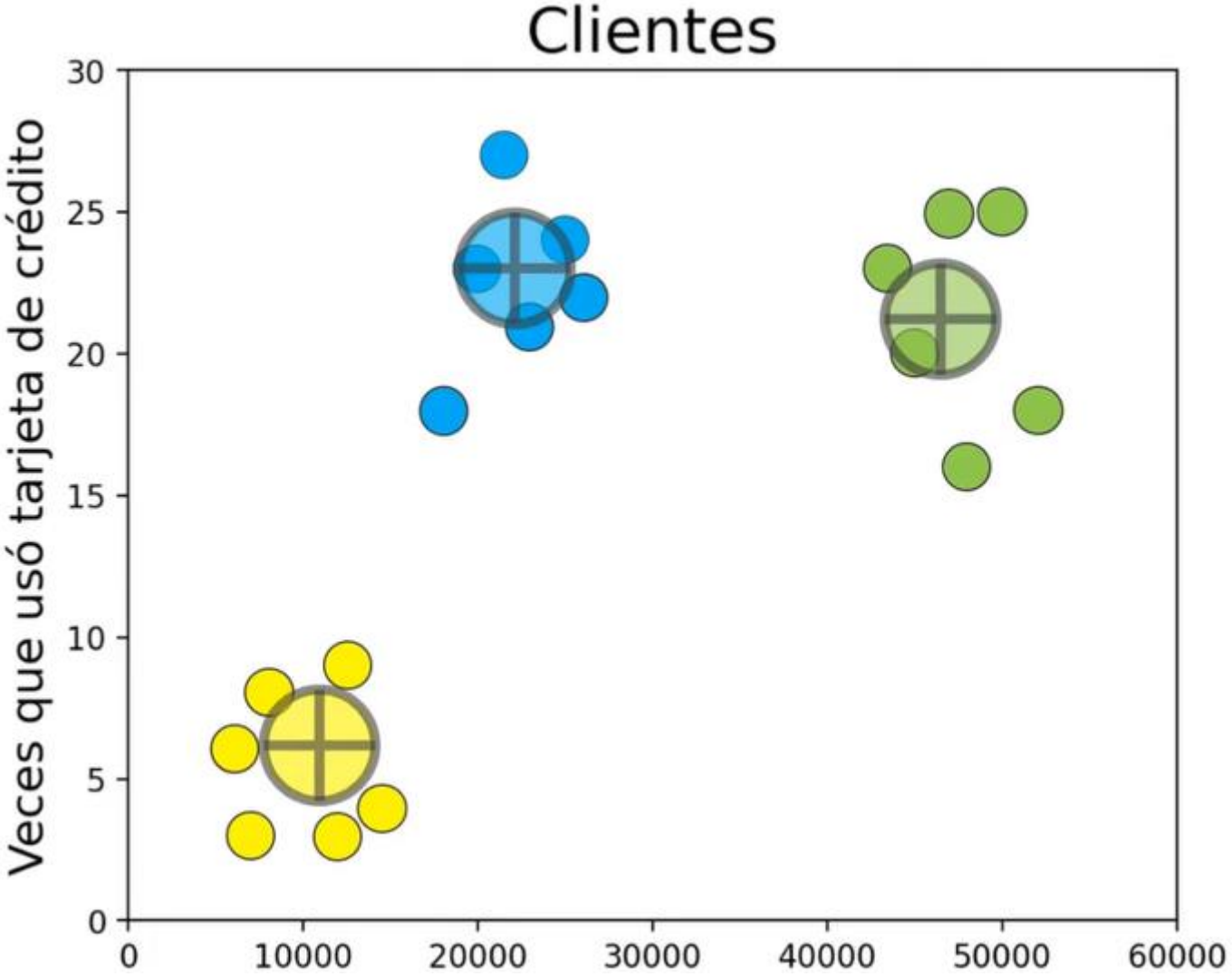
Tras la categorización de cada punto, se debe acercar el centroide a cada cluster mediante el **promedio** de los puntos (ver ejes).



Este es el resultado de la primera iteración; sin embargo, note que dos puntos (amarillo y verde) quedaron muy lejos de sus centroides. Volviendo a **repetir** el mismo proceso para todos los puntos…



Estos pequeños cambios afectan la posición de cada uno de los centroides, por ende, se deben reposicionar



Tras este último cambio, ¿qué pasa si volvemos a repetir el algoritmo?¿cambia alguna posición de los clusters?