

CLUSTERING

Equipo

Patricia Arvizu 1823604

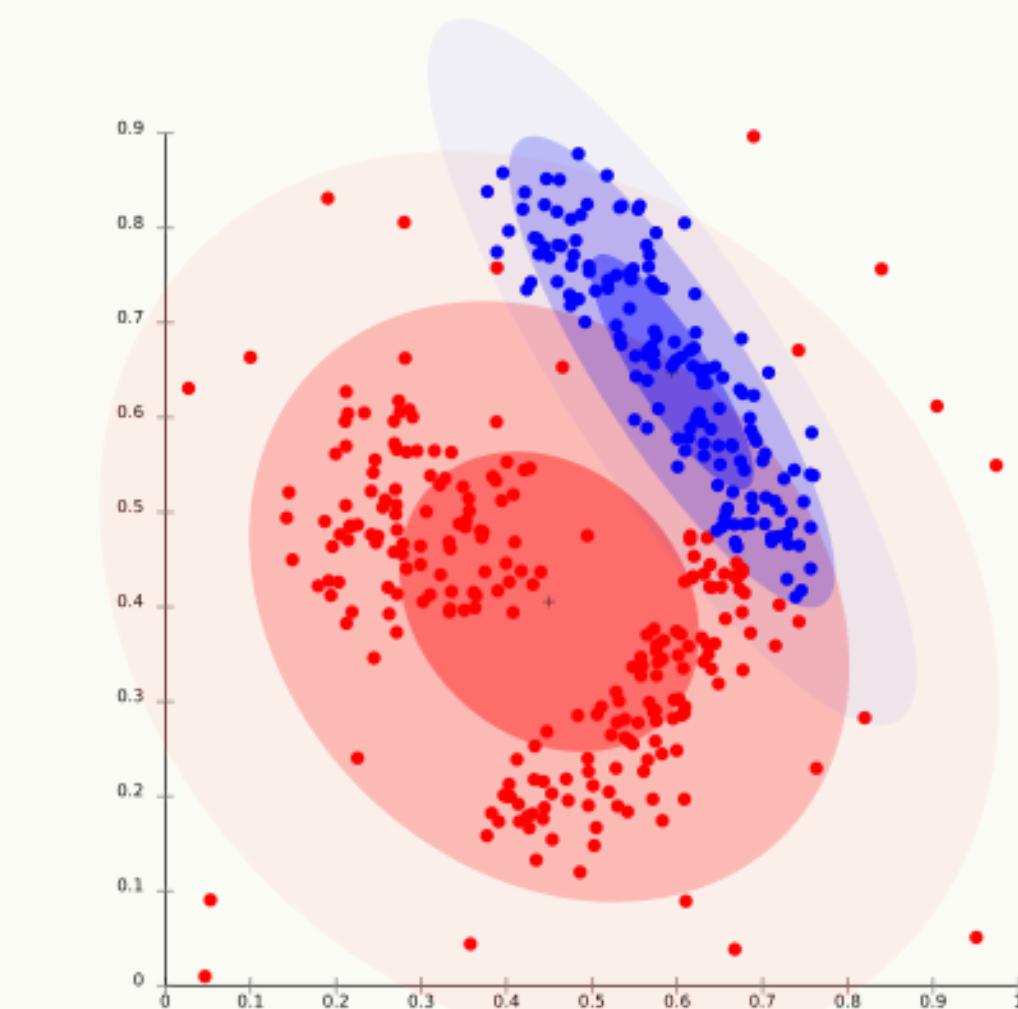
Helena Carrillo 1725370

Noé López 1812678

Vanessa Ortiz 1810699

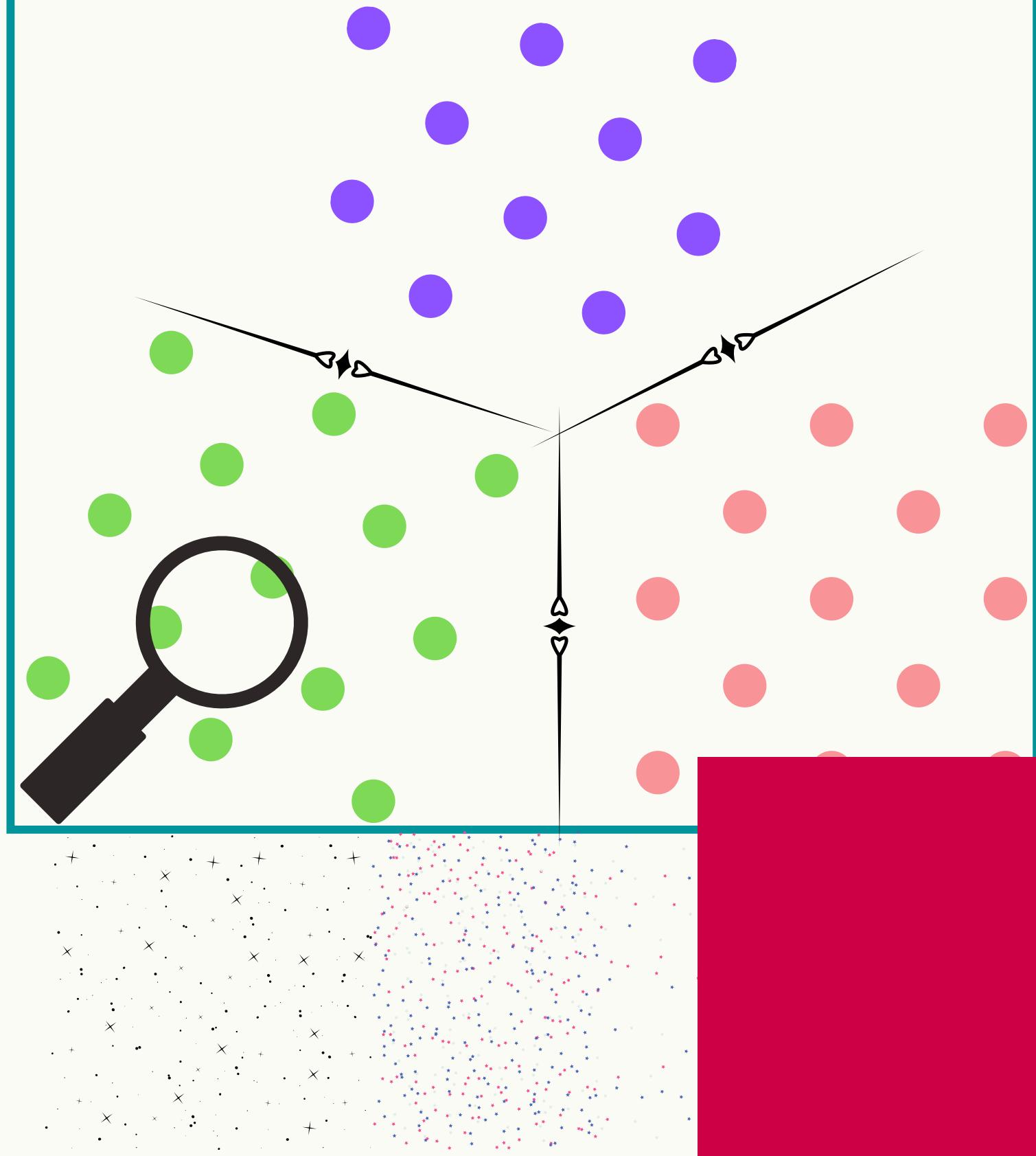
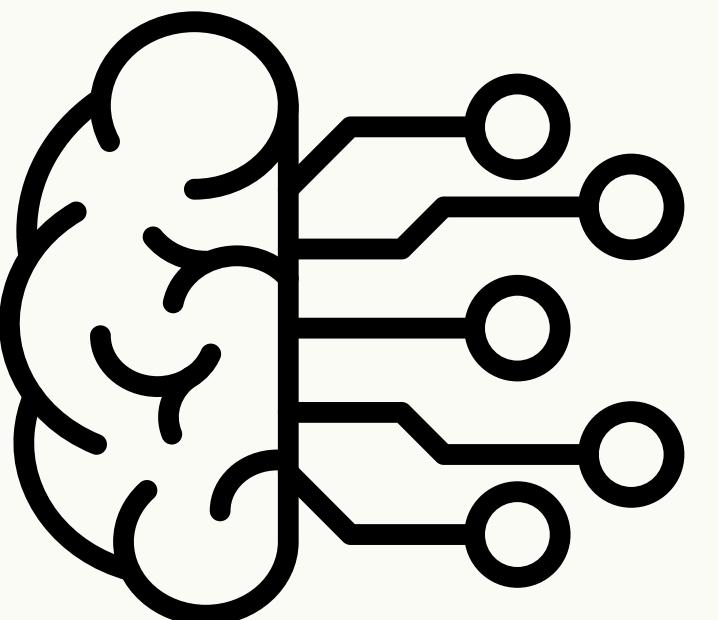
Miguel Ovalle 1801990

Keila Puente 1807864



¿QUÉ ES?

Es una técnica de aprendizaje de máquina no supervisada que consiste en **agrupar puntos de datos** y de esta forma crear particiones basándonos en similitudes.



USOS DEL CLUSTERING

Investigación de mercado



Identificar comunidades



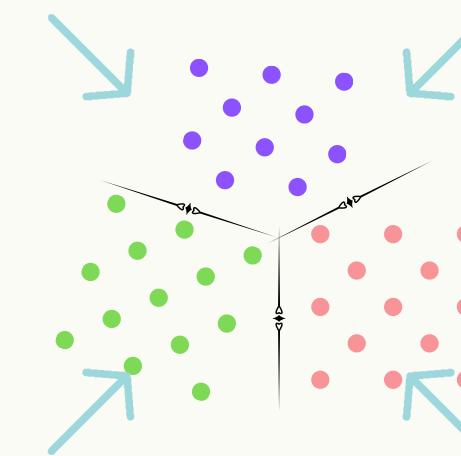
Prevención de crimen



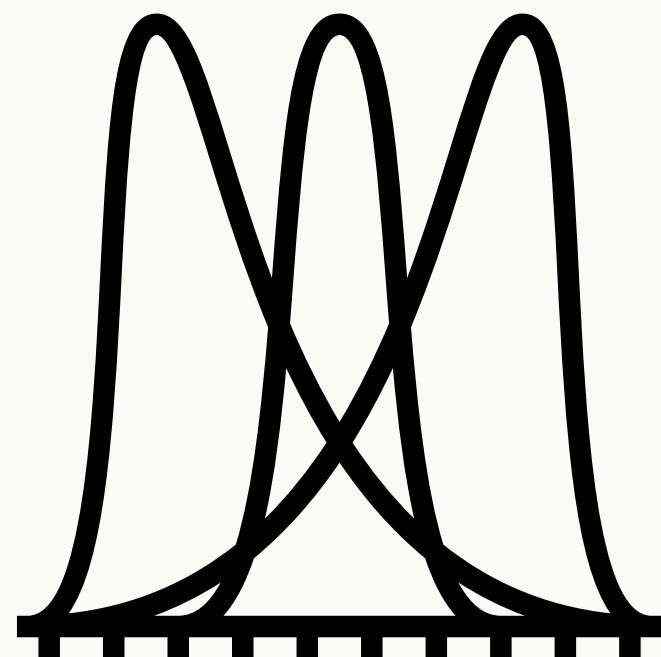
Procesamiento de imágenes



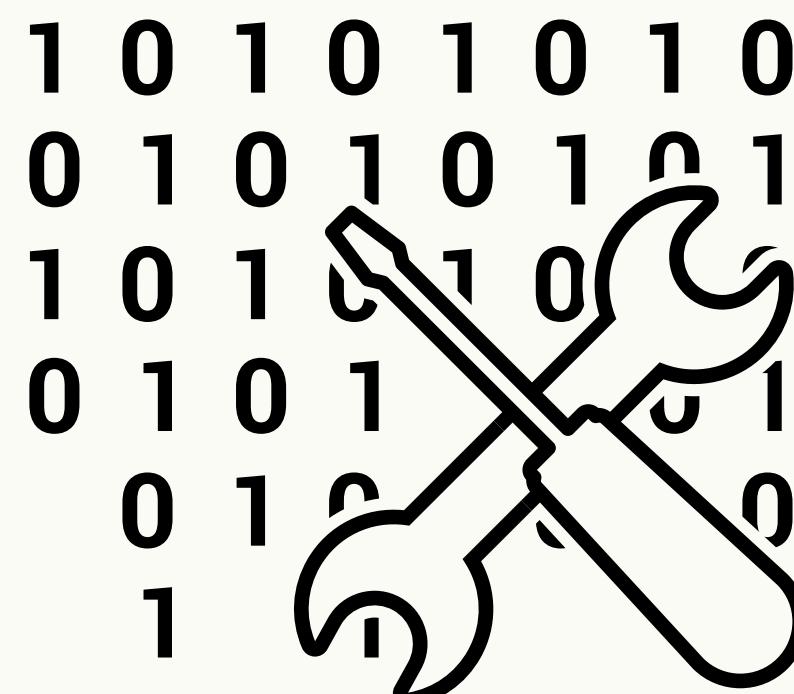
TRANSFORMACIÓN DE DATOS



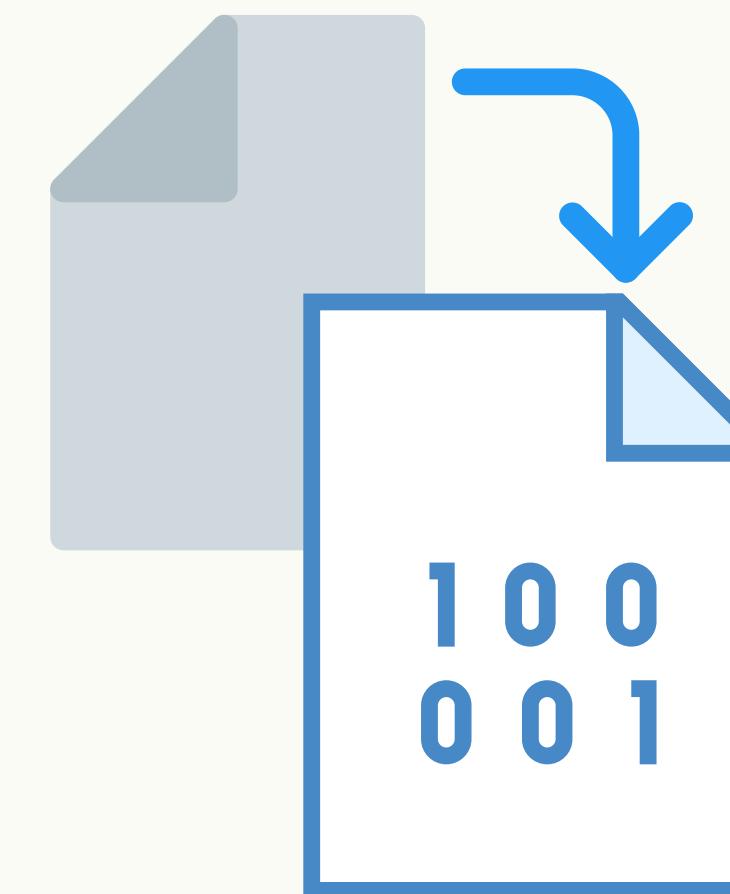
Variables cuantitativas



Variables Binarias



Variables categóricas



TIPOS BÁSICOS DE ANÁLISIS

Centroid Based Clustering

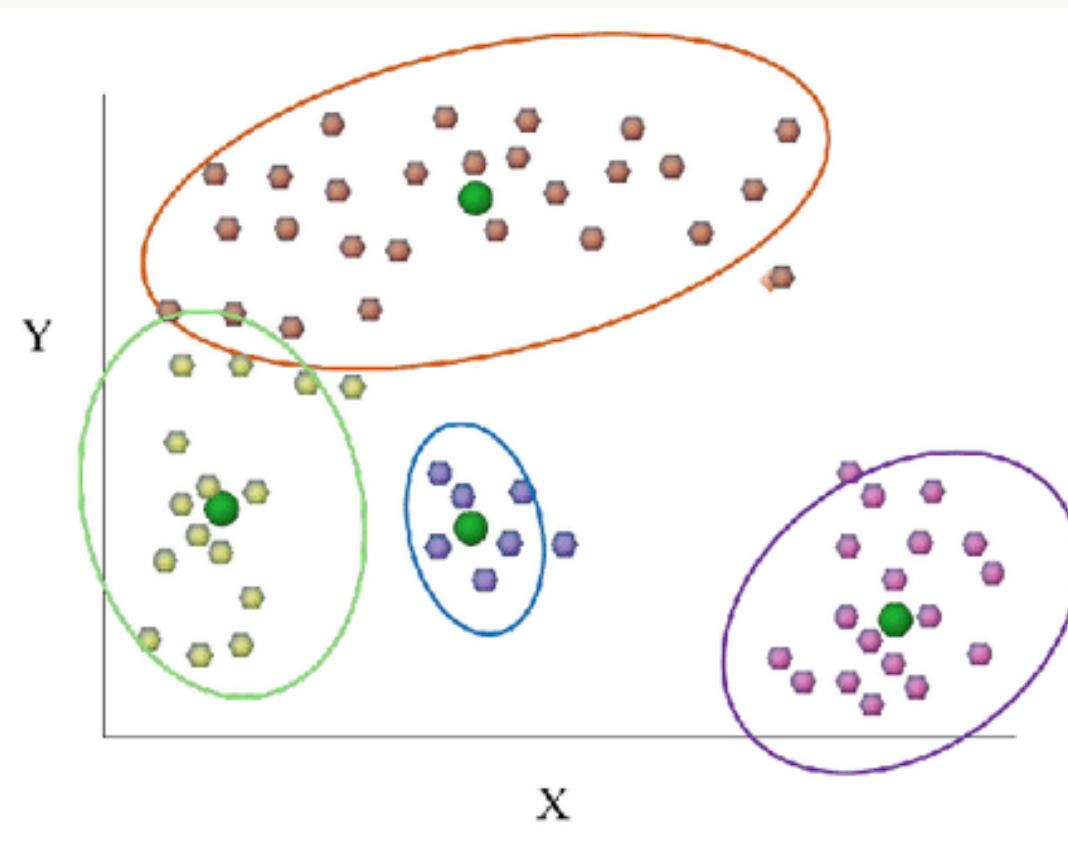
Connectivity Based Clustering

Distribution Based Clustering

Density Based Clustering



CENTROID BASED CLUSTERING



Cada cluster es representado por un **centroide**.

Los clusters se construyen basados en la **distancia** de punto de los datos hasta el centroide. Se realizan varias iteraciones hasta llegar al mejor resultado.

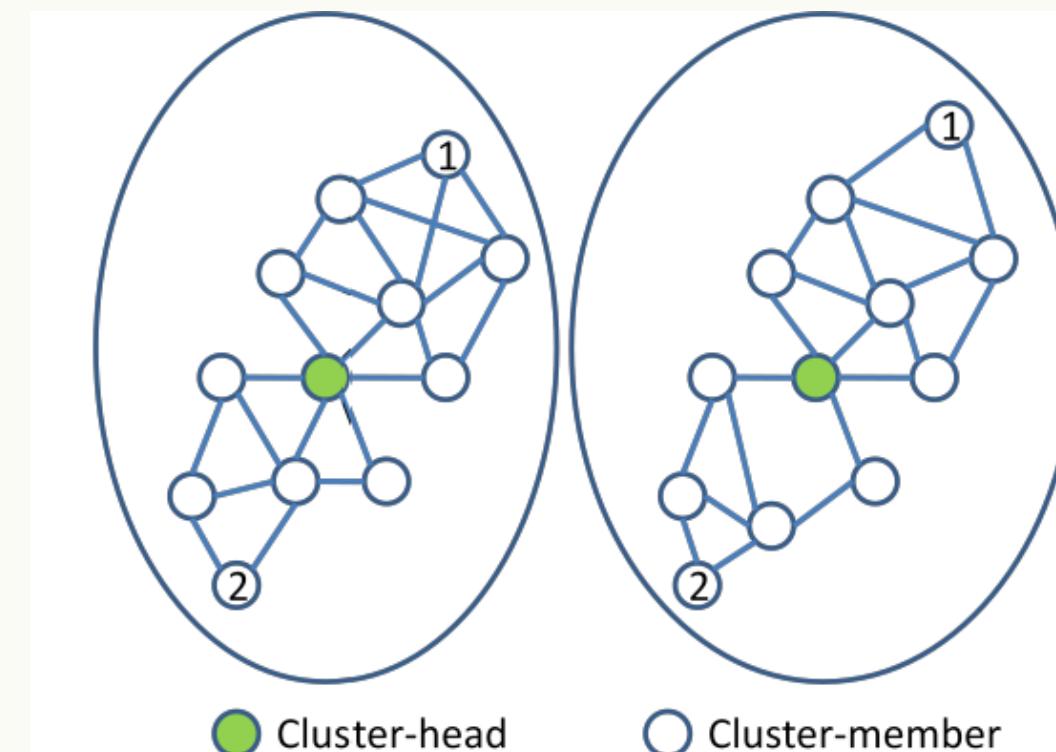
El algoritmo más usado de este tipo es el de **K-medias**.

CONNECTIVITY BASED CLUSTERING

Los clusters se definen agrupando a los datos más similares o **cercanos** (los puntos más cercanos están más relacionados que otros puntos más lejanos).

La característica principal es que un cluster contiene a otros clusters (representan una **jerarquía**).

Un algoritmo usado de este tipo es **Hierarchical clustering**

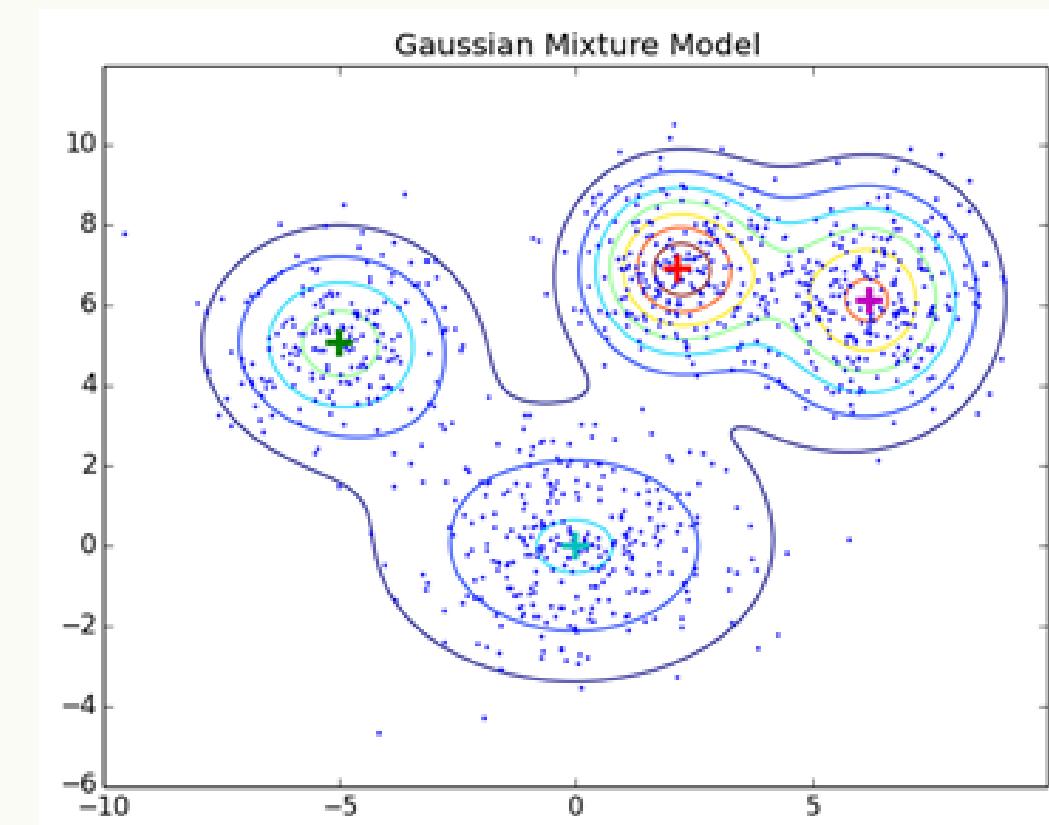


DISTRIBUTION BASED CLUSTERING

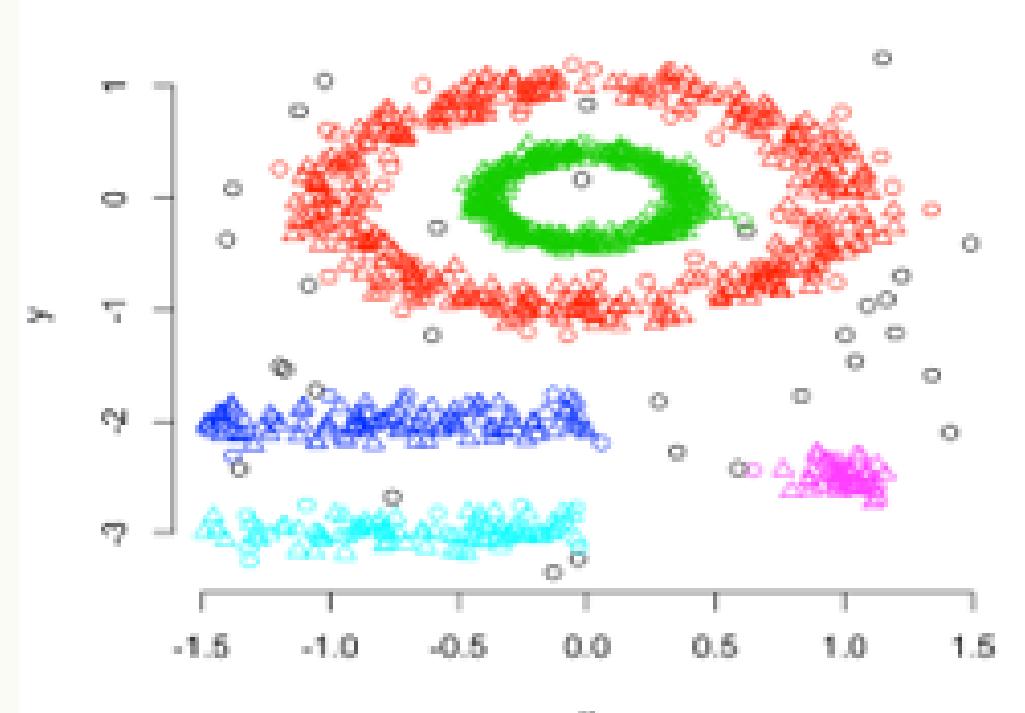
En este método cada cluster pertenece a una **distribución normal**,

La idea es que los puntos son divididos con base en la **probabilidad** de pertenecer a la misma distribución normal.

Un algoritmo de clustering perteneciente a este tipo es **Gaussian mixture models**.



DENSITY BASED CLUSTERING



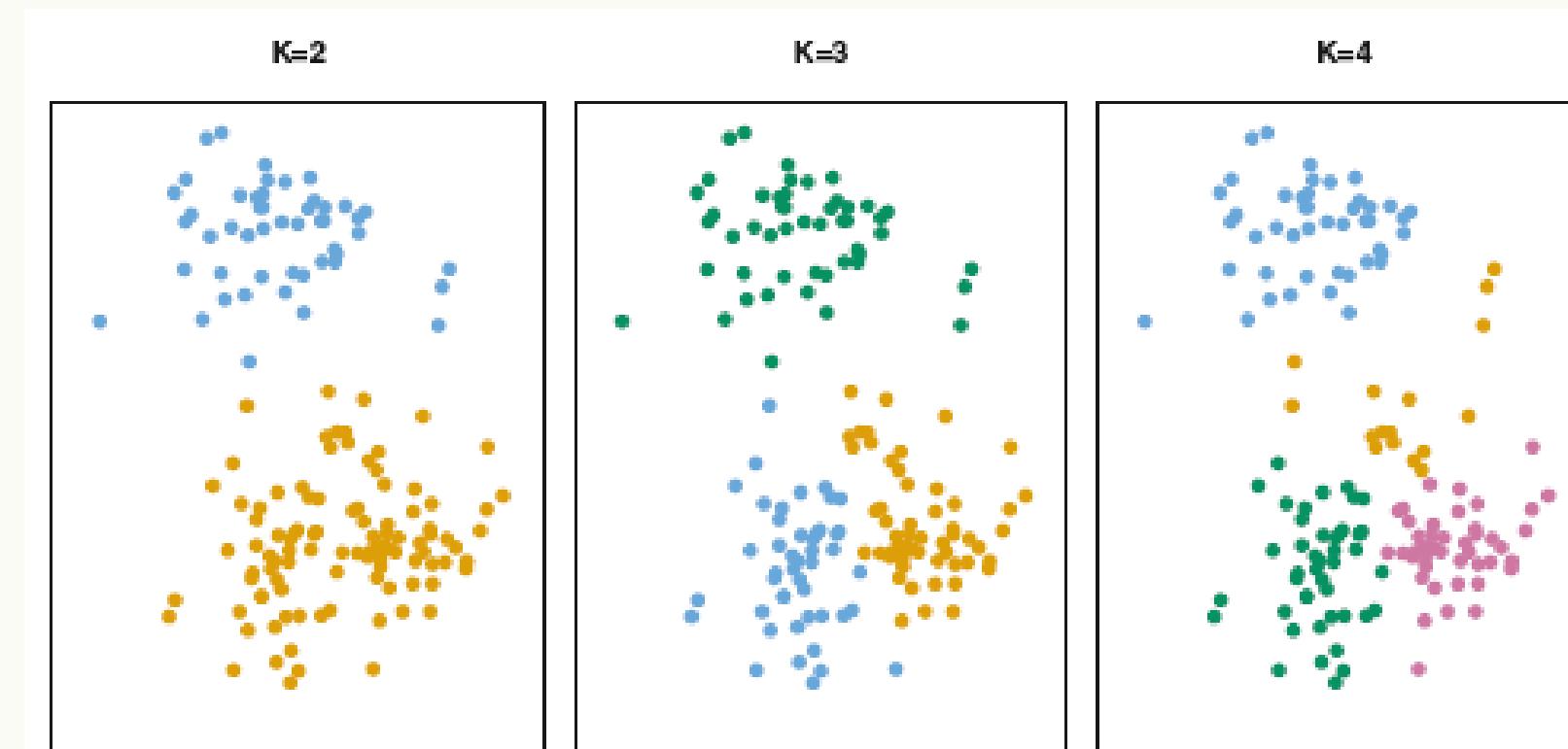
Los clusters son definidos por **áreas** de concentración.

Se trata de **conectar** puntos cuya distancia entre sí es considerada pequeña.

Un cluster contiene a todos los puntos relacionados dentro de una distancia limitada y considera como irregular a las áreas esparcidas entre clusters.

MÉTODO K-MEDIAS

Algoritmo de clustering basado en centroides. K representa el número de clusters y es definido por el usuario.



Una vez que escogemos el valor de k:

PASOS K-MEDIAS

1

CENTROIDES

Elegimos k datos aleatorios que pasarán a ser los centroides representativos de cada cluster

2

DISTANCIAS

Analizamos la distancia de cada dato al centroide más cercano, perteneciendo a su cluster.

3

MEDIA

Obtener media de cada cluster y este será el nuevo centro.

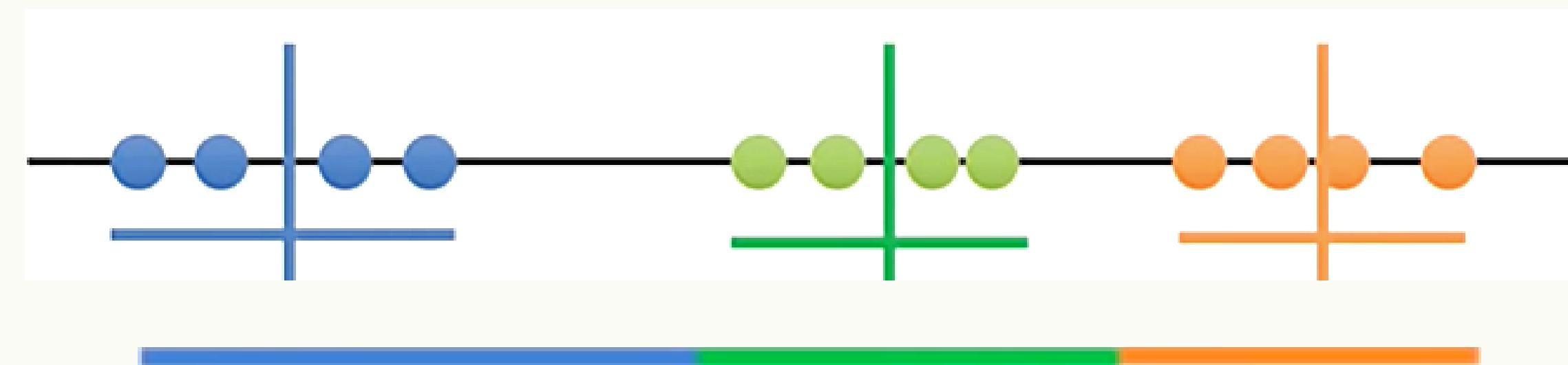
4

ITERAR

Repetimos el proceso hasta que los clusters no cambien

VARIANZA DE LOS CLUSTERS

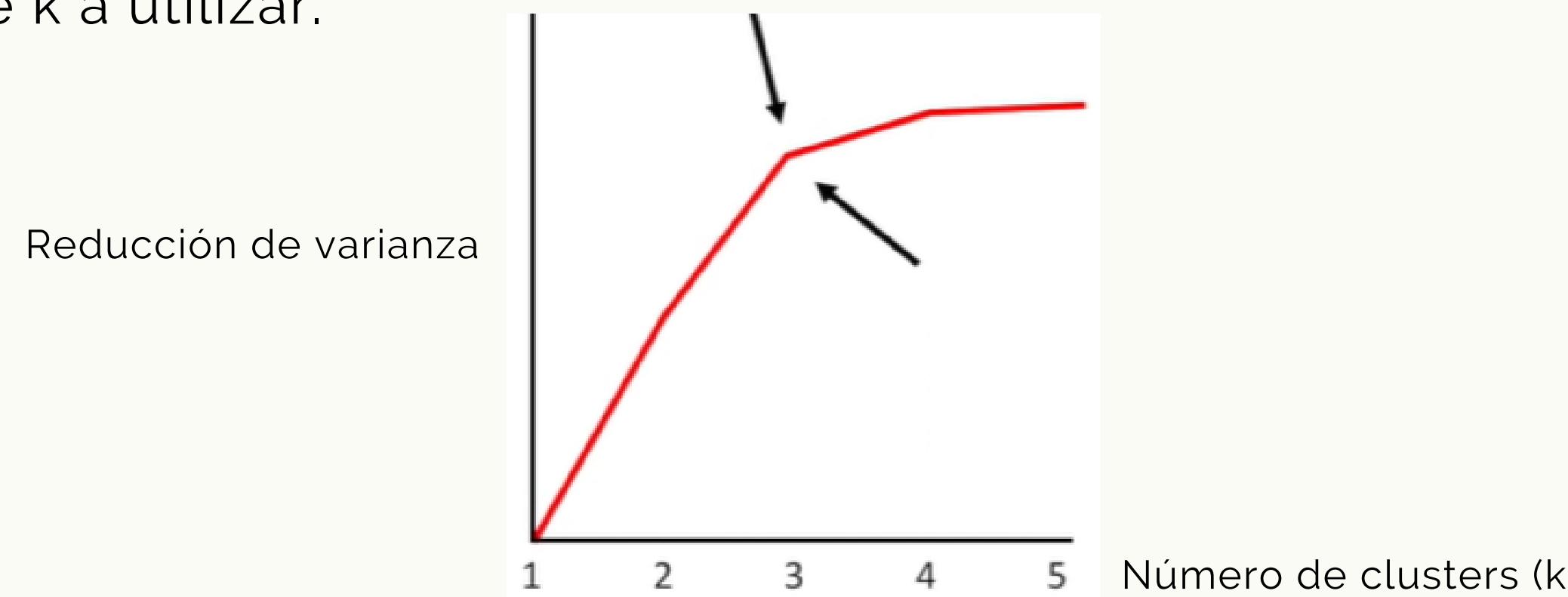
La varianza de cada cluster disminuye al aumentar k. Si sólo hay un elemento en el cluster, la varianza es de 0. Entre menor sea la suma de las varianzas de los clusters, mejor es nuestro clustering.



MÉTODO DEL CODO

Consiste en **graficar la reducción de la varianza total** a medida que k aumenta.

En un punto la reducción de la varianza no disminuirá de forma significativa entre un valor k y otro. Este punto es llamado **elbow plot** o codo y representa el numero de k a utilizar.



REFERENCIAS

http://www.exa.unicen.edu.ar/catedras/optia/public_html/clustering.pdf
<https://www.youtube.com/watch?v=4b5d3muPQmA>
<https://youtu.be/4cxVDUybHrI>
<https://developers.google.com/machine-learning/clustering/clustering-algorithms>
https://conceptosclaros.com/que-es-clustering/#La_caja_de_tomates_o_que_significa_clustering
<https://jarroba.com/que-es-el-clustering/>
<https://www.ugr.es/~mvargas/1.acluster.pdf>