



CAPACITACIÓN
PROFESIONAL

Especialización en Machine Learning con Python

Sesión 12

Docente: Jose de Lama Zegarra

Reglas



Se requiere **puntualidad** para un mejor desarrollo del curso.



Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.



Las preguntas se realizarán **a través del chat** y en caso de que lo requieran **podrán activar el micrófono**.



Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.



Identificarse en la sala Zoom con el primer nombre y primer apellido.

MALLA CURRICULAR



Contenido – Módulo 15

- Métodos no supervisados
- Clustering: K – means y PAM
- Clustering Jerárquico
- Clustering basado en densidades
- Aplicación: Segmentación de clientes
- Market Basket Analysis: Análisis de Asociación
- Introducción a los sistemas de recomendación

b — number of replicates
DINC CAPACITACION CORPORATIVA

Training set S

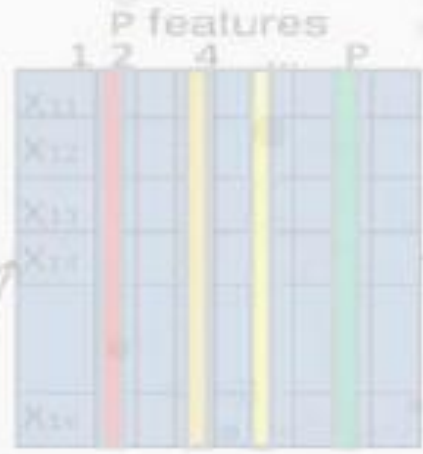


S_1

S_2

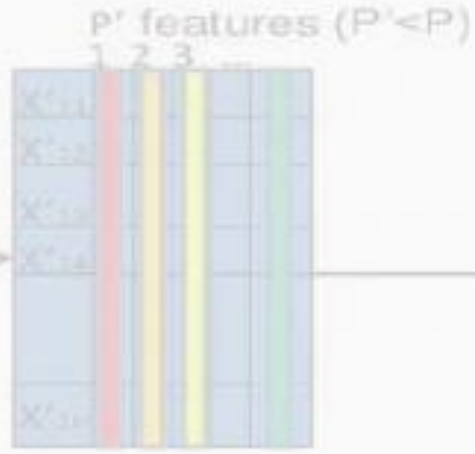
S_b

Bootstrap sampling with replacement



Random Subspace selection

S'_1



Learning algorithm

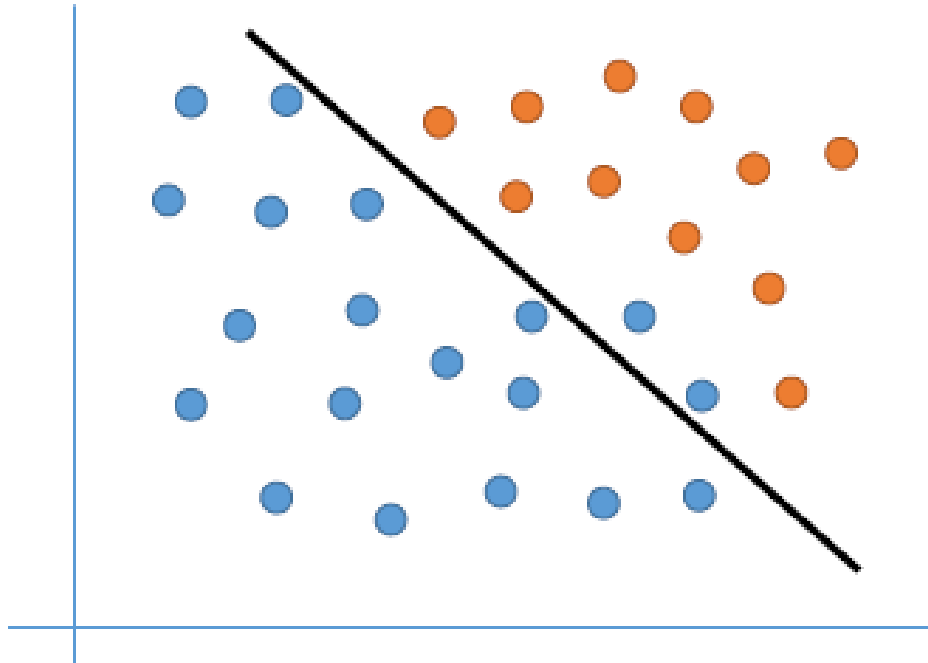
Classifier C_1

Classifier C_2

Classifier C_b

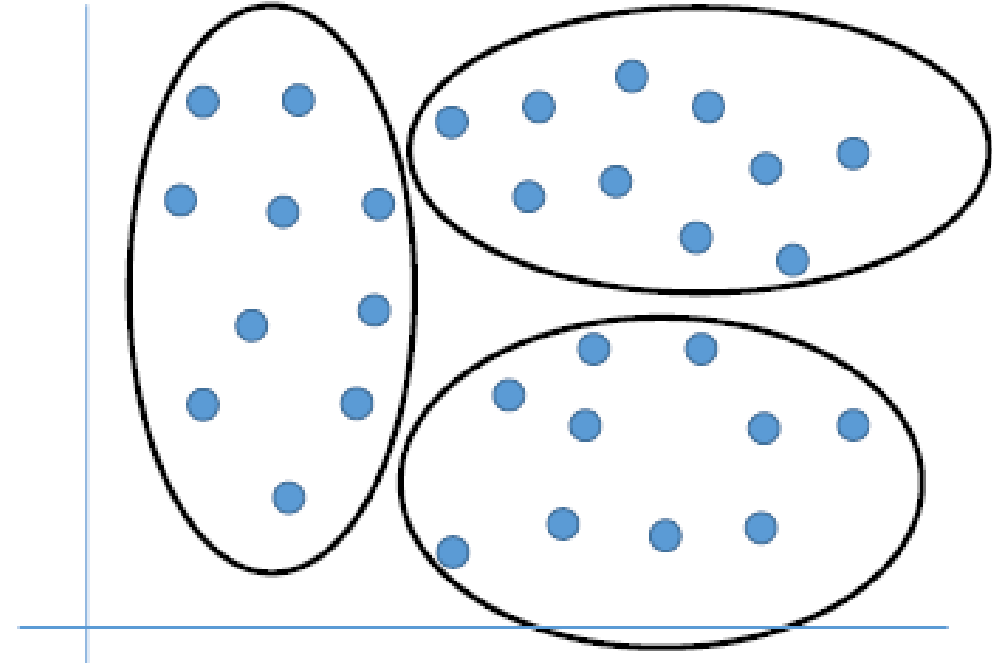
Aprendizaje No Supervisado
“ante cualquier acción de negocio, primero conoce bien a tu cliente”

Métodos no supervisados



Aprendizaje Supervisado

Conocemos las clases del dataset



Aprendizaje No Supervisado

No conocemos las clases del dataset

Métodos no supervisados

- ✓ En el **Aprendizaje NO Supervisado**, los datos de entrenamiento no están etiquetados con una **salida Y** (variable objetivo, target, etc.).
- ✓ A diferencia del aprendizaje supervisado, en el no supervisado **no hay forma determinística de verificar el performance del modelo**. Sólo se puede evaluar con conocimiento del negocio.
- ✓ Algunas aplicaciones típicas del aprendizaje no supervisado son:
 - Segmentación de clientes.
 - Detección de fraude o anomalías.
 - Análisis de asociación
- ✓ Otra aplicación importante es la **reducción de dimensionalidad**.

Métodos no supervisados: Características

Aprendizaje
no
supervisado

Dataset de Entrenamiento

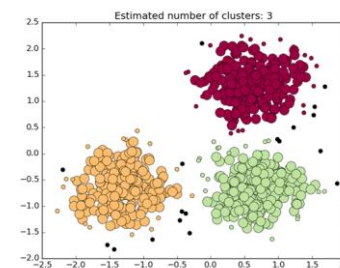
		Predictores (Features)			
		x_1	x_2	x_p
Muestras	1				
	n				

n muestras de
entrenamiento
o instancias

p
predictores

Resolución de problemas:

- Clustering
- Reducción de dimensionalidad.
- Detection de anomalías.

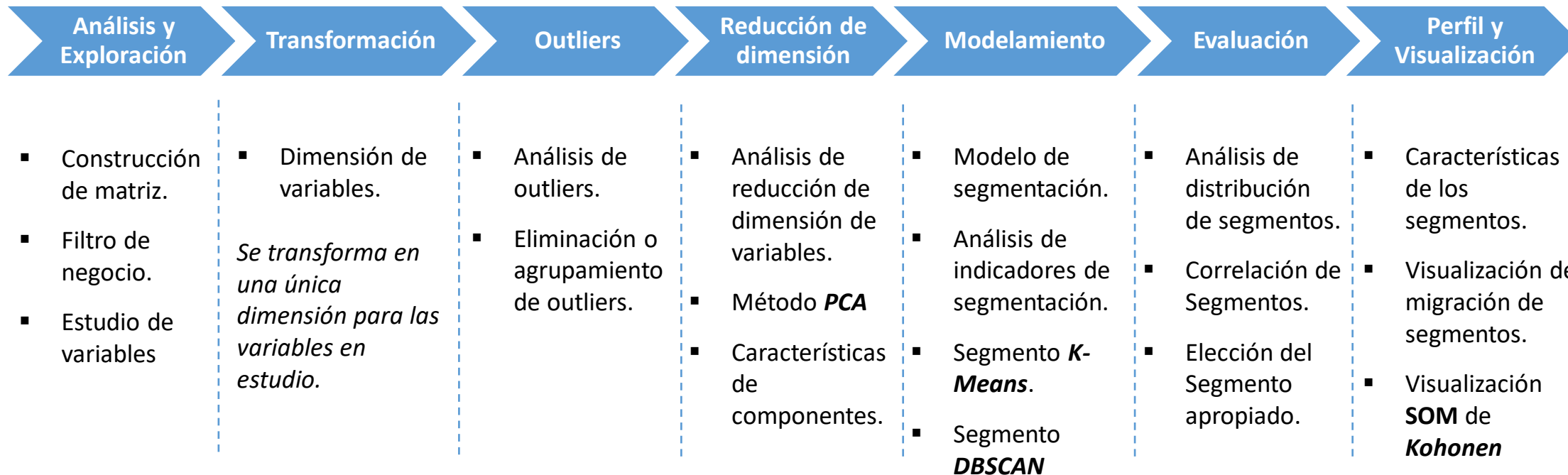


Datos Transformados y
Agrupados en Clústeres

No tenemos etiquetas
en los datos (Y)

Métodos no supervisados: Metodología

Proceso de Segmentación



Retroalimentación

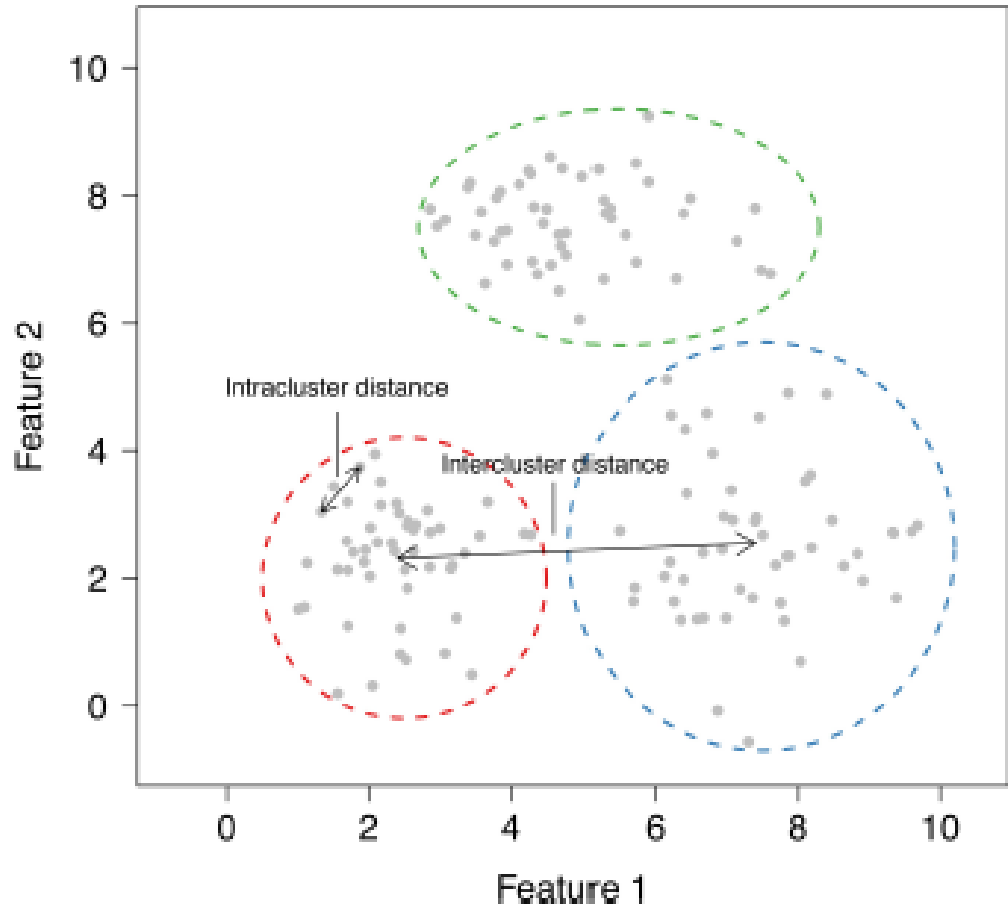
Métodos no supervisados: Clustering

Clustering

Es la tarea de encontrar agrupamientos (clusters) o grupos homogéneos dentro de un conjunto de datos

Se busca optimizar dos objetivos a la vez:

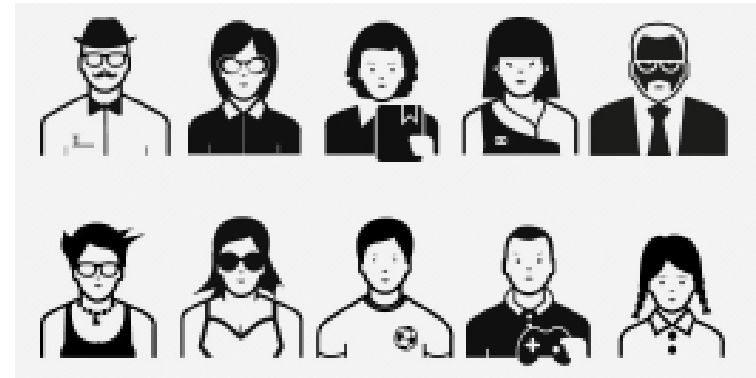
- Que los datos dentro de un mismo clusters sean muy **similares** entre sí.
- Que los datos de clusters **distintos** muy diferentes entre sí



Métodos no supervisados: Clustering

Ejemplos de aplicaciones de clustering

- Dado un conjunto de clientes, encontrar segmentos de mercado para aplicar estrategias de comunicación diferenciadas
- Dado un conjunto de noticias, identificar tópicos de información y agruparlas por su contenido
- Dentro de un conjunto de correos, identificar aquellos relacionados a spam o correos no deseados



Segmentación de clientes

Métodos no supervisados: Clustering

¿Cómo agrupar los datos?



Métodos no supervisados: Clustering

La agrupación es subjetiva



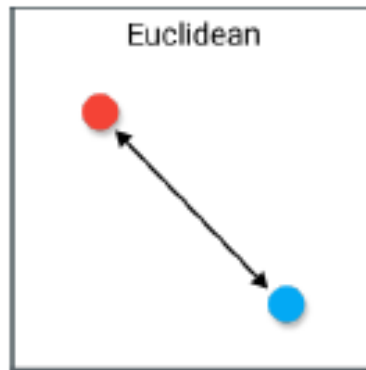
Etapa de evolución



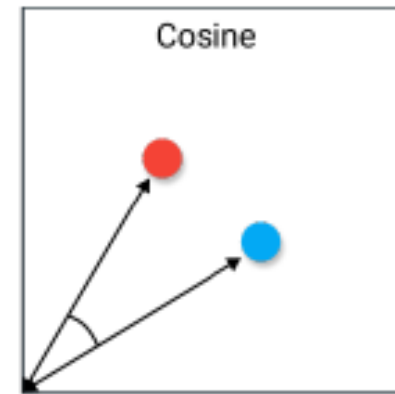
Tipo de pokémon

Métodos no supervisados: Clustering

Definición de distancia



$$D(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$



$$D(x, y) = \cos(\theta) = \frac{x \cdot y}{\|x\| \|y\|}$$

Validación del Clustering

- El **método de la silueta** suele ser un indicador para medir cómo de bien una observación se adapta a su cluster.
- Varía entre -1 y +1. Un valor cercano a +1 significa que la instancia está bien ubicada en su cluster y alejada de otros clusters. Un valor de 0 significa que está cerca a una frontera. Un valor de -1 significa que la instancia está en el cluster equivocado.
- El ancho de la silueta (**silhouette width**) de la i-ésima observación es definida por:

$$sil_i = (b_i - a_i) / \max(a_i, b_i)$$
- Donde, a_i denota la distancia promedio entre la observación i y todas las otras que están en el mismo cluster de i y b_i denota la distancia promedio mínima de i a las observaciones que están en otros clusters

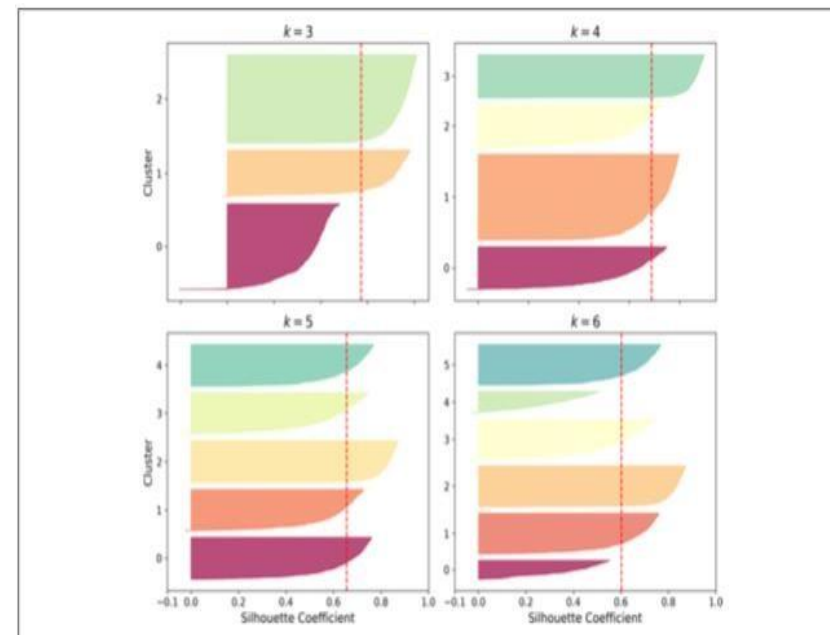
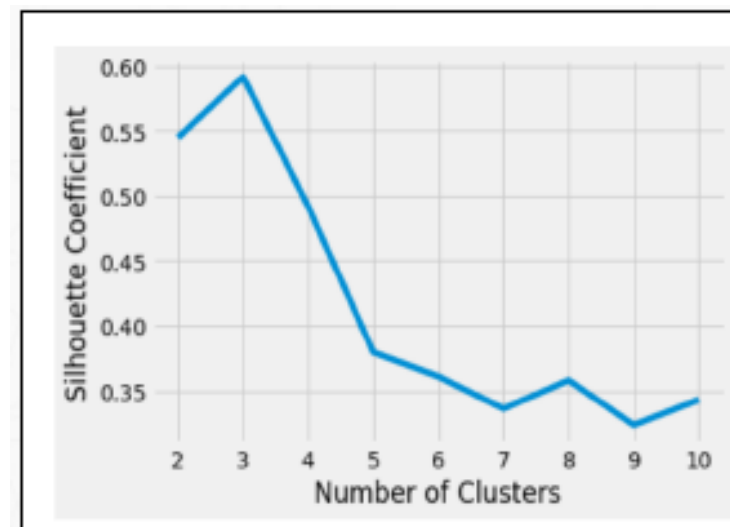
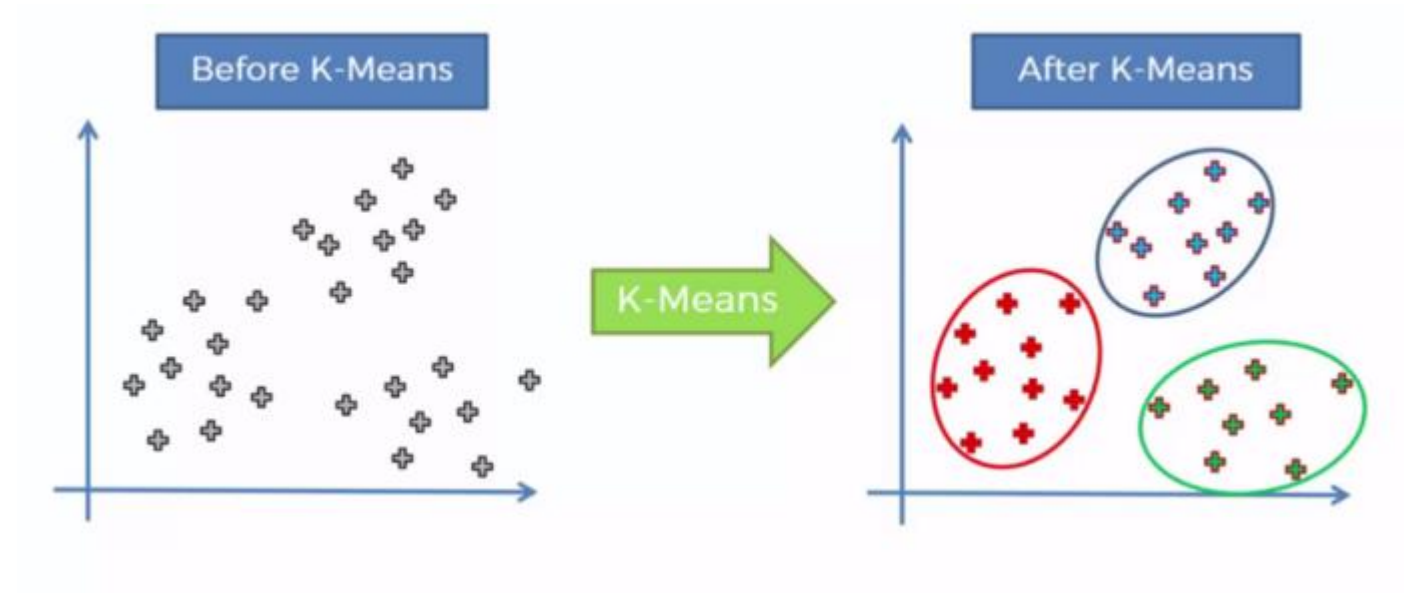


Figure 9-10. Silhouette analysis: comparing the silhouette diagrams for various values of k



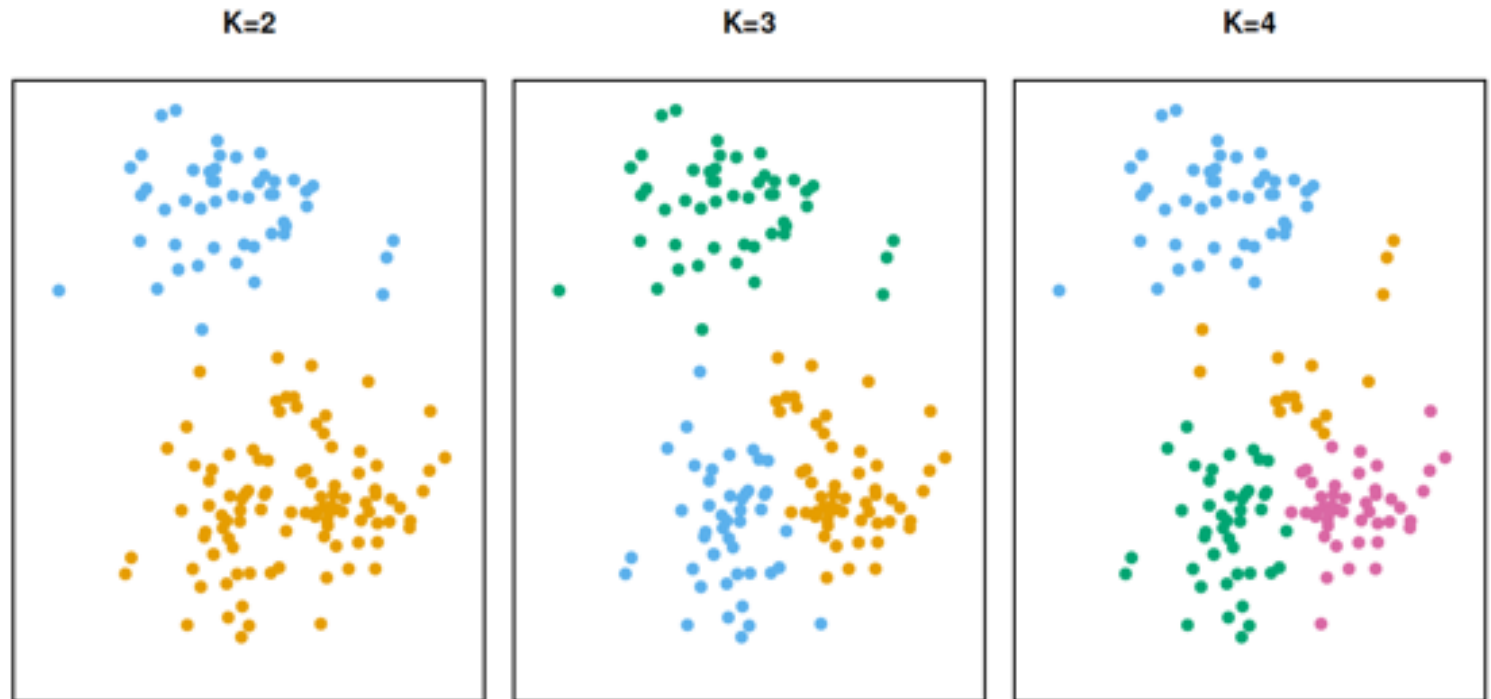
K-Means



Segmentación K-Means

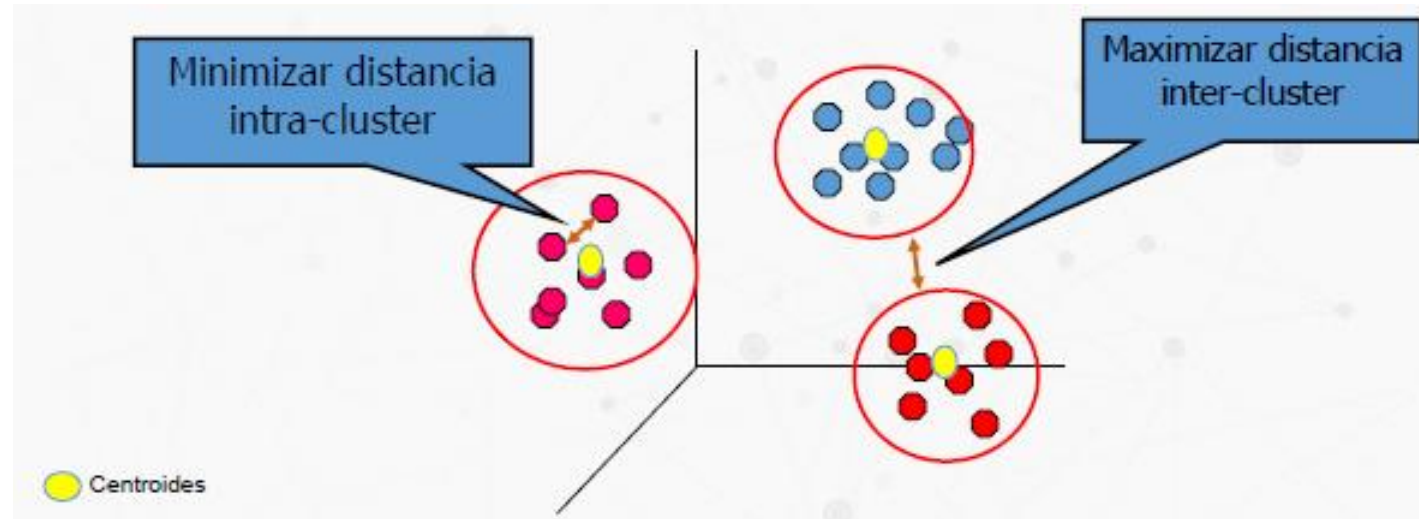
- ✓ En K-means el objetivo es agrupar las observaciones de un dataset en un número K de clústeres.
- ✓ En número K es **hiperparámetro** que hay que brindarle al algoritmo.
- ✓ Puede definirse este valor por conocimiento a priori o por indicaciones del negocio.
- ✓ En otro caso, se puede seguir la regla: $k = \text{raíz}(n/2)$

Ejemplo: para el caso de dos predictores para distintos valores de K



Segmentación K-Means

- ✓ Al igual que el caso del PCA, las variables predictoras deben ser normalizadas antes de hacer el clustering.
- ✓ Los atributos han de ser numericos (pues se computa la media de los mismos para reasignar los **centroides**).
- ✓ Es muy sensible a los valores anómalos (outliers).



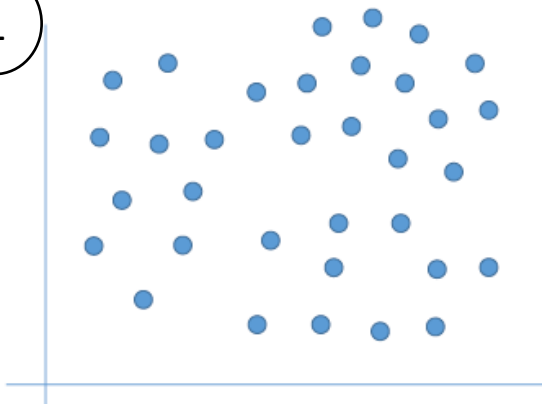
Segmentación K-Means: Pasos

El algoritmo K-means se explica de la siguiente manera:

- ✓ De manera aleatoria asignar un número de 1 a K a cada observación. Esto será la asignación inicial a los clústeres de cada observación.
- ✓ Iterar sobre los siguientes pasos hasta que las asignaciones a los clústeres deje de cambiar:
 - a. Para cada clúster, calcule el centroide será un vector compuesto por la media de los p predictores de las observaciones del mismo cluster.
 - b. Reasigne cada observación al clúster cuyo centroide esté más cercano a la observación.

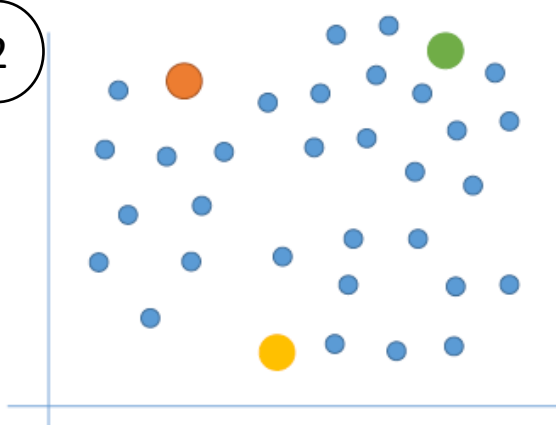
Segmentación K-Means: Pasos

1



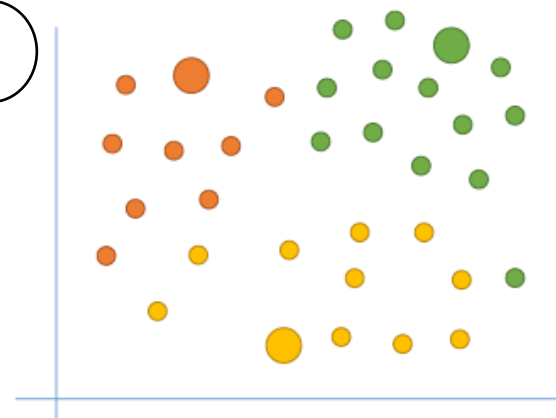
Seleccionar k puntos como centroides iniciales

2



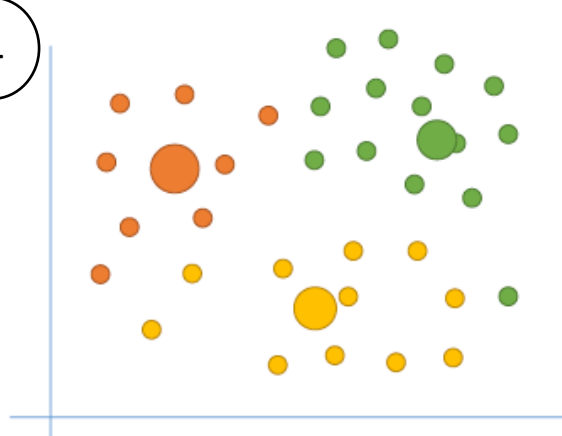
Asignar los puntos a cada cluster

3



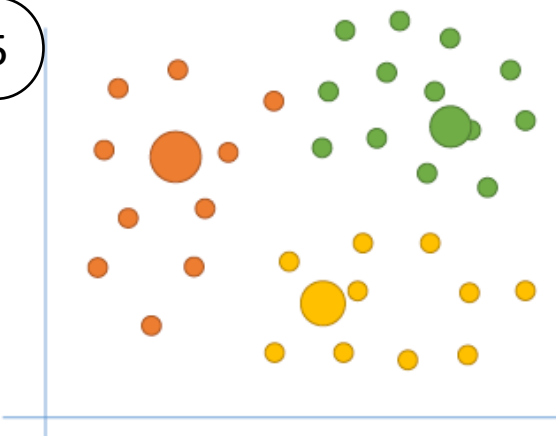
Recalcular los centros de cada cluster

4

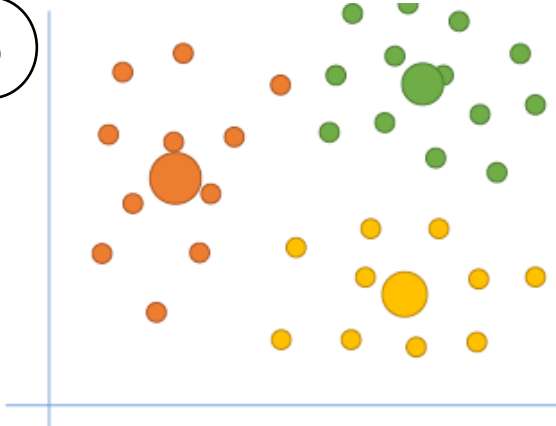


Reasignar los elementos a cada cluster hasta que el centroide no cambie

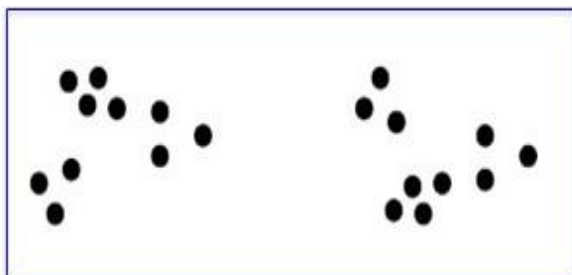
5



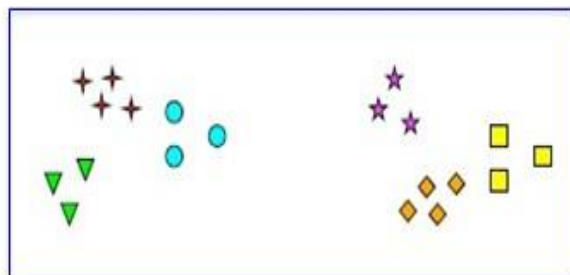
6



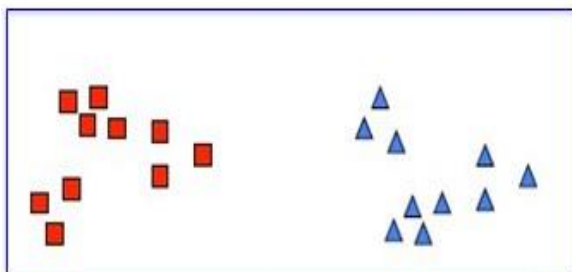
Segmentación K-Means



Datos originales



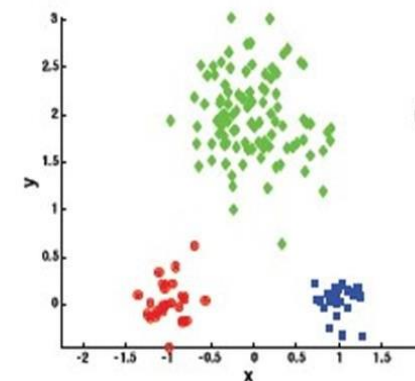
6 clústeres



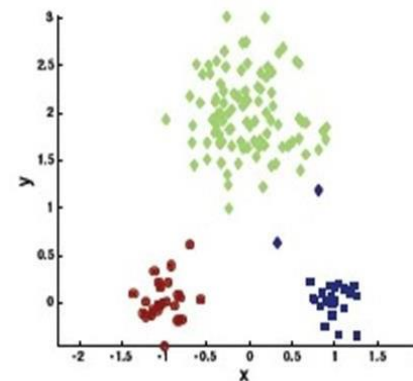
2 clústeres



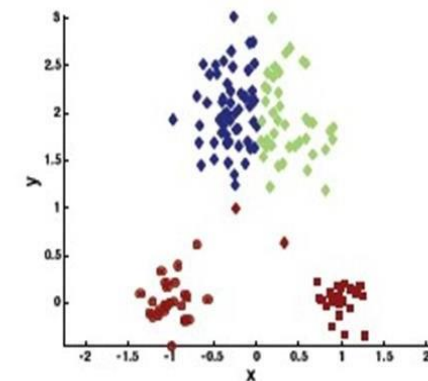
4 clústeres



Original Points



Optimal Clustering



Sub-optimal Clustering

Segmentación K-Means

Ventajas

- Simple, entendible
- Los elementos son asignados automáticamente a los clústers

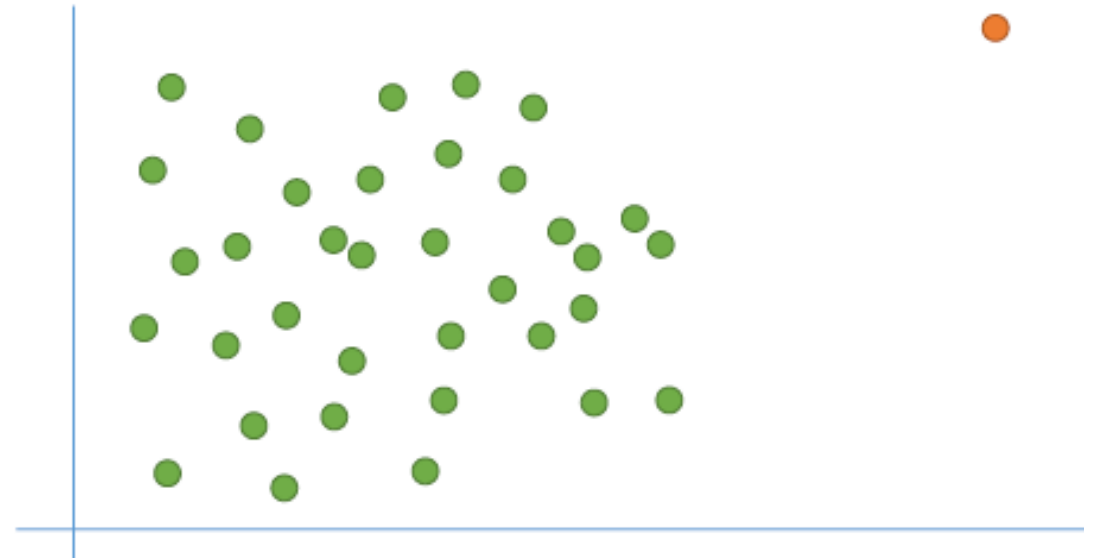
Desventajas

- No se sabe a priori el número de clusters
- Todos los elementos se deben asignar a un cluster
- Los resultados pueden variar de acuerdo a la asignación inicial de los centroides
- Es muy sensible a valores extremos

Variaciones:

- Kmedoids
- Kmedians
- Kmeans ++

Kmeans: Valores extremos



Visualización de Kmeans:

<https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>

PAM (Partitioning Around Medoids)

- ✓ El algoritmo PAM se basa en la búsqueda de k objetos representativos o medoides entre las observaciones del conjunto de datos
- ✓ Un Medoide es un objeto de un cluster cuya disimilaridad media al resto de objetos del cluster es mínima. Es el punto ubicado más hacia el centro en todo el grupo.
- ✓ Trabaja, como el K Means con particiones, dividiendo el conjunto de datos en grupos:
 - Ambos intentan minimizar la distancia entre puntos que se añadirían a un grupo y otro punto designado como el centro de ese grupo
 - k medoids escoge datapoints como centros y trabaja con una métrica arbitraria de distancias entre datapoints
 - Minimiza suma de disimilaridades (entre pares de puntos) en vez de una suma de distancias euclidianas cuadradas

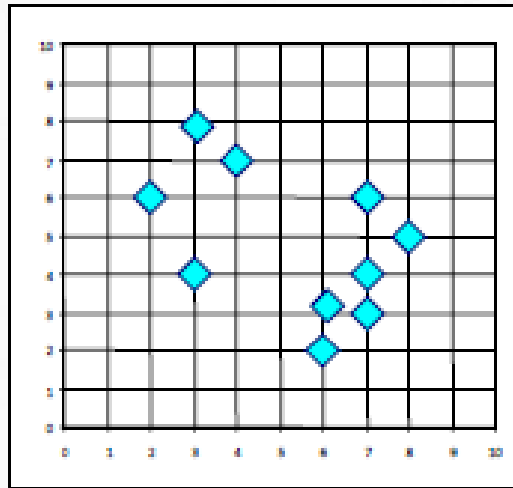
Ventajas:

- Es más robusto ante el ruido
- Muy flexible

Desventajas

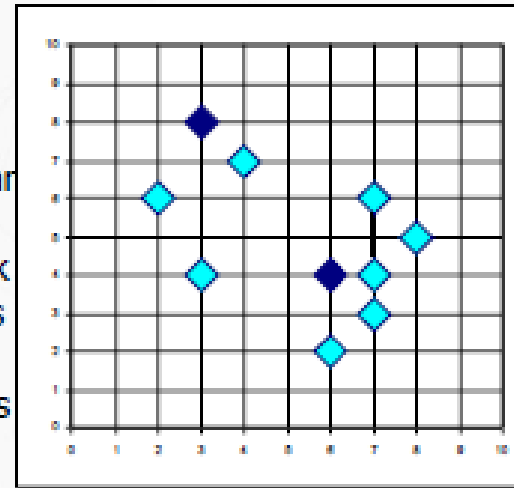
- No muy sofisticado
- No está garantizado encontrar en número de clusters óptimo

PAM (Partitioning Around Medoids)



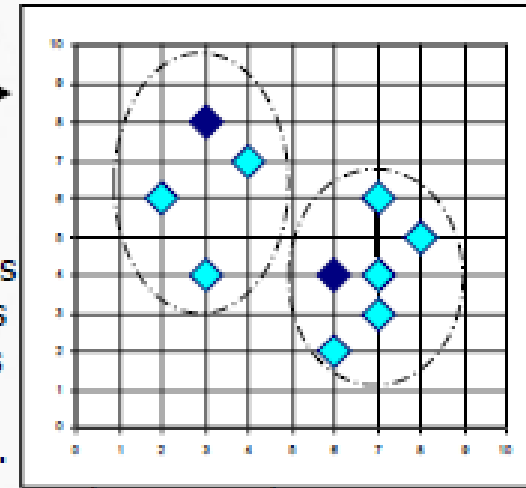
K=2

Arbitrariamente
escoger k
instancias
como
medoides



Total Cost = 26

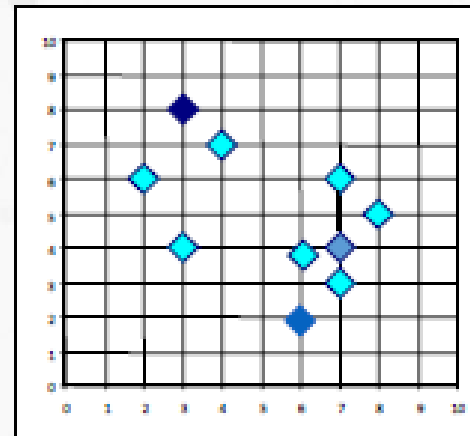
Asignar
las
instancias
restantes
a su mas
cercano
medoide.



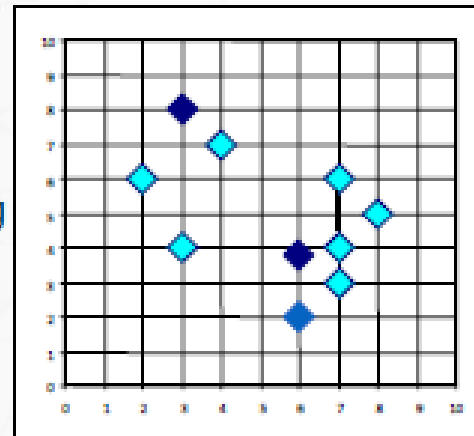
Total Cost = 20

Seleccionar al azar una
instancia
nonmedoide, O_{random}

Calcular el
costo total
de swapping

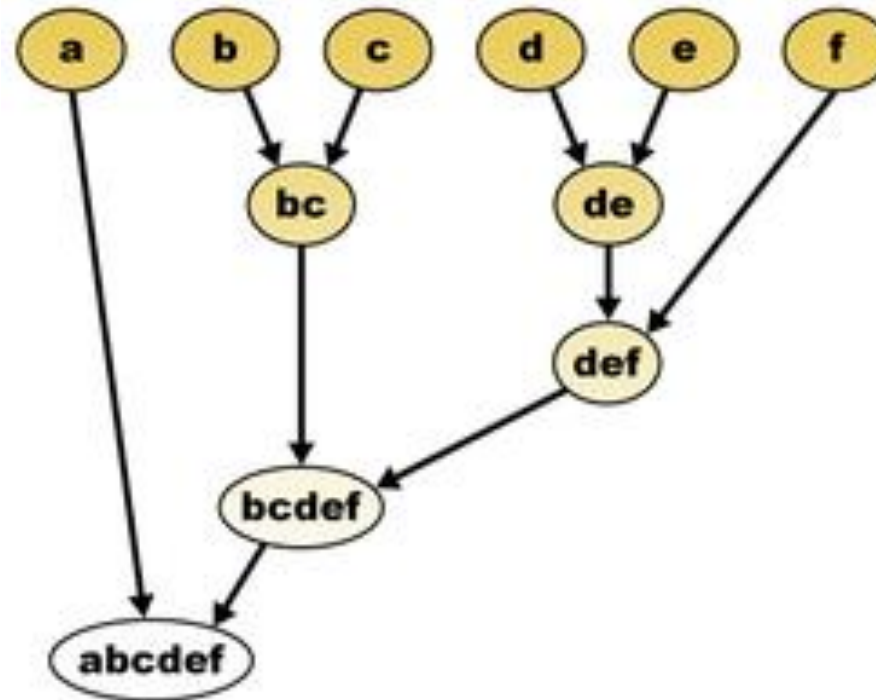


Swapping O
y O_{random}
Si se mejora
la calidad



**Hacer el loop
hasta que no
haya cambios**

Clustering Jerárquico



Clustering Jerárquico

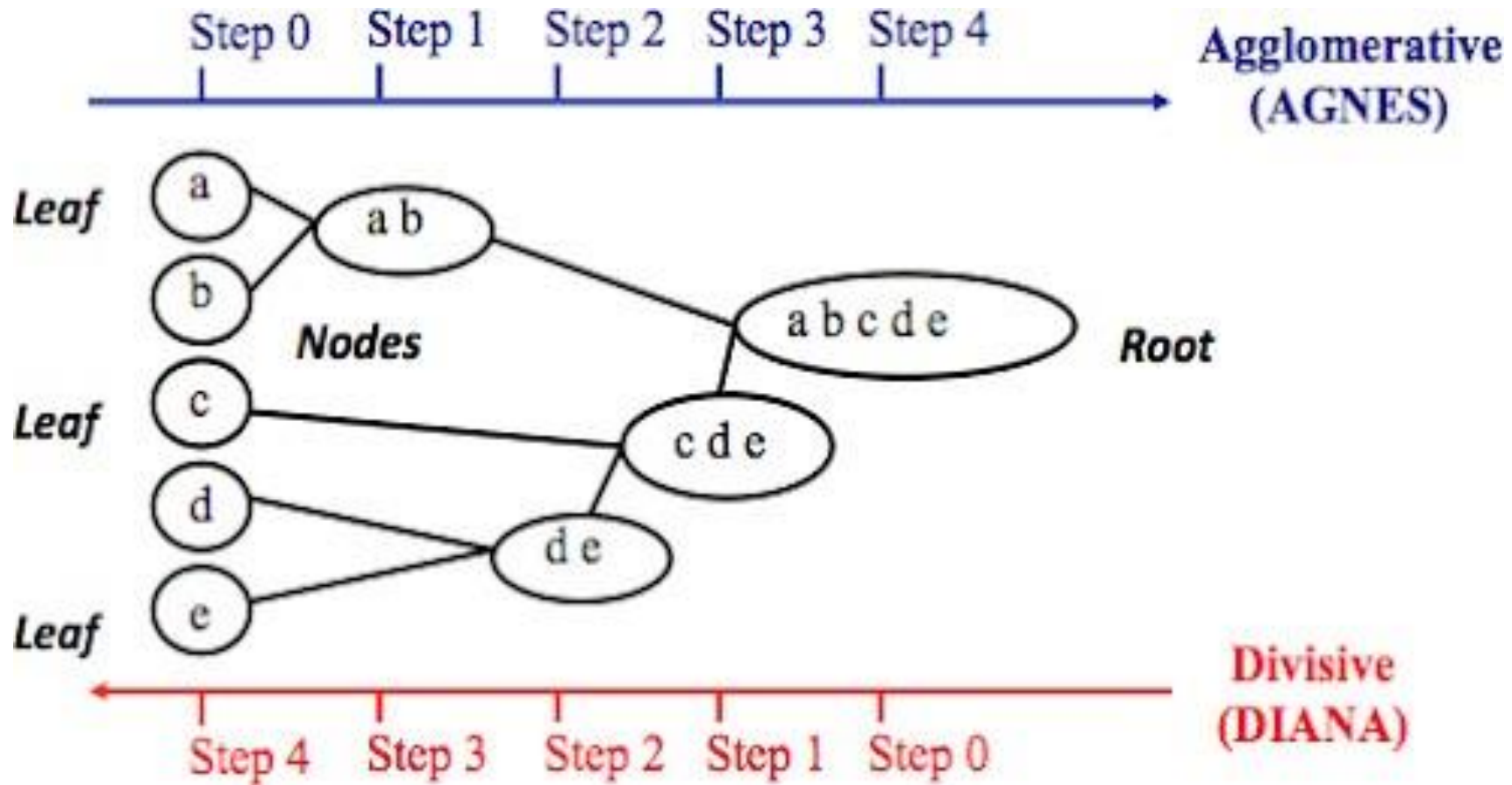
En estos algoritmos se generan sucesiones ordenadas (jerarquías) de conglomerados. Puede ser juntando cluster pequeños en mas grande o dividiendo grandes clusters en otros mas pequeños. La estructura jerárquica es representada en forma de un árbol y es llamada **Dendograma**. Cada corte del árbol da una partición. Cuanto más arriba se corte el árbol se obtendrá un menor número de clases y clases menos homogéneas

Se dividen en dos tipos:

Algoritmos jerárquicos aglomerativos (bottom-up, inicialmente cada instancia es un cluster).
AGNES

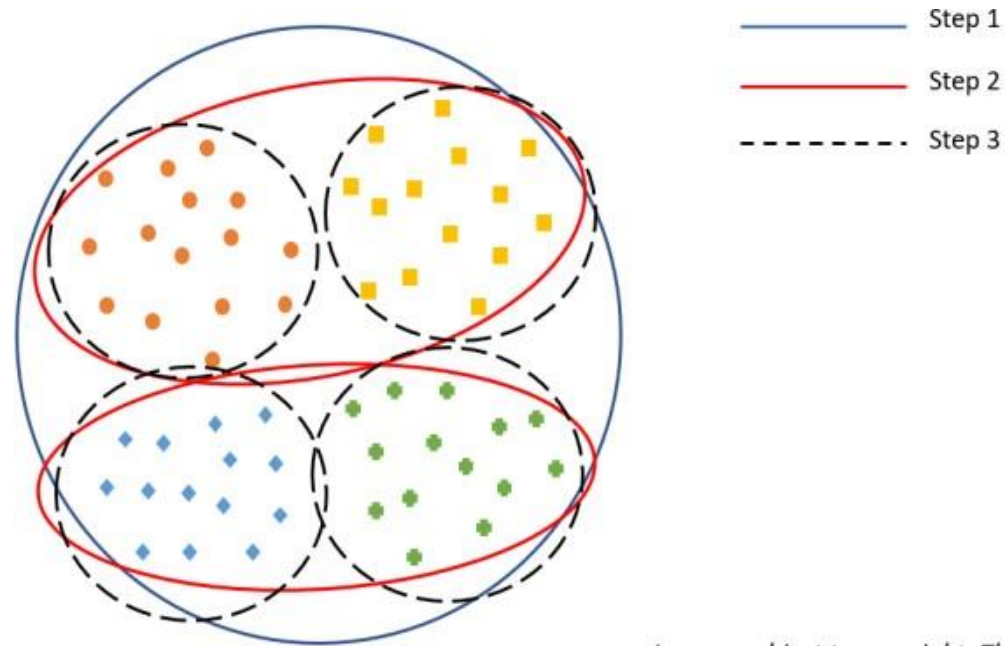
Algoritmos jerárquicos divisivos (top-down, inicialmente todas las instancias están en un solo cluster). **DIANA**.

Clustering Jerárquico



Clustering Jerárquico divisivo (DIANA)

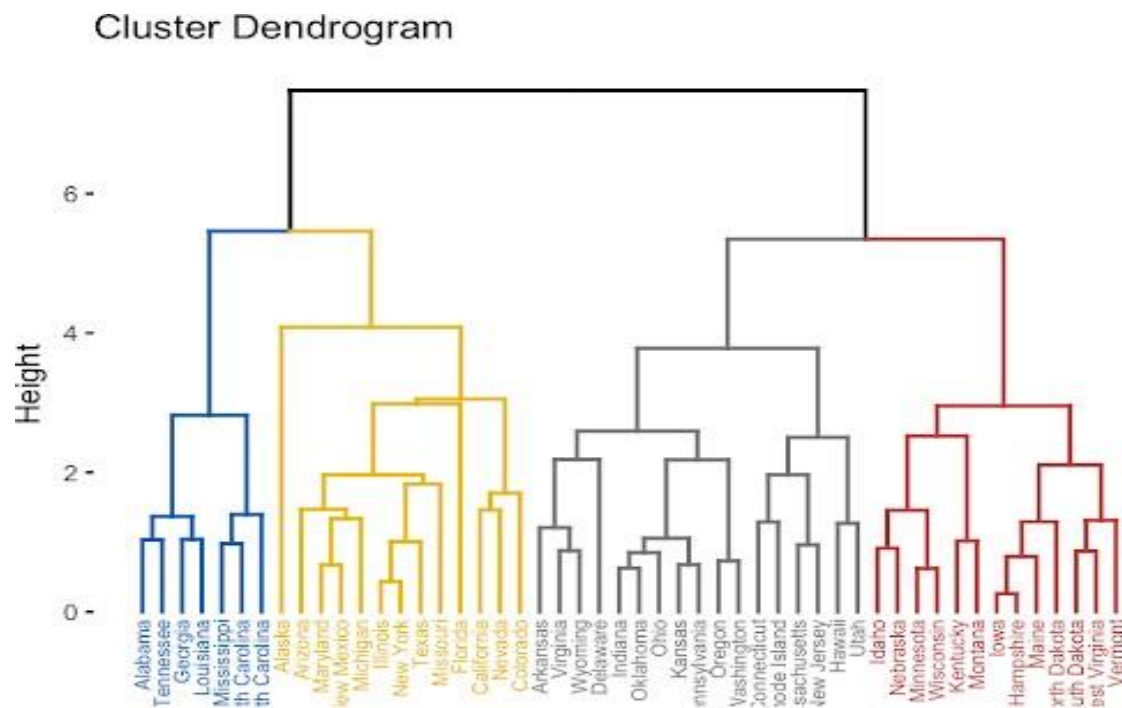
En el método Divisivo suponemos que todas las observaciones pertenecen a un único grupo y luego dividimos el clúster en dos grupos menos similares. Esto se repite recursivamente en cada grupo hasta que haya un grupo para cada observación.



Images subject to copyright: The Datum

Clustering Jerárquico Aglomerativo (AGNES)

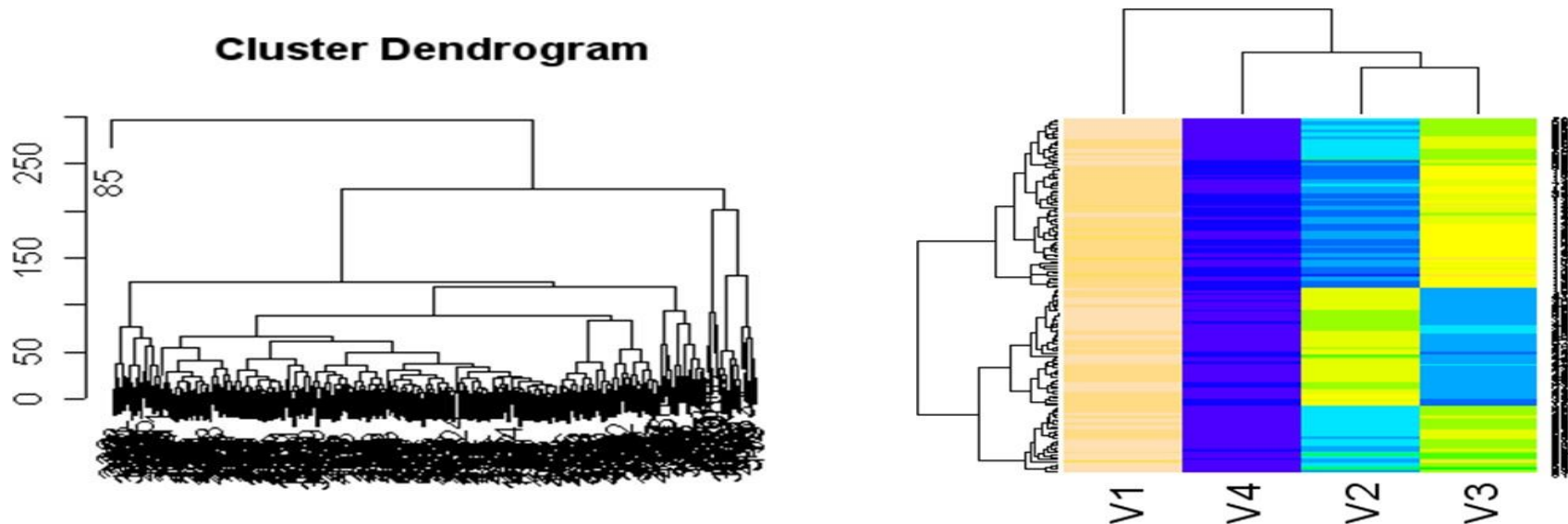
En este método, cada observación se asigna a su propio clúster. Luego, se calcula la similitud (o distancia) entre cada uno de los clusters y los dos clusters más similares se fusionan en uno. Finalmente, los pasos 2 y 3 se repiten hasta que solo quede un grupo.



Desventajas de los dendogramas

Los dendogramas son fáciles de interpretar pero pueden conducir a falsas conclusiones por las siguientes razones:

- El dendograma correspondiente a un conglomerado jerárquico no es único, puesto que por cada junte de clusters uno necesita especificar que sub-árbol va a la derecha y cuál a la izquierda.
- La estructura jerárquica del Dendograma no representa con certeza las verdaderas distancias entre los objetos distintos del conjunto de datos.













Clustering Jerárquico Aglomerativo: Pasos

Matriz de distancia

Se inicia con una matriz de distancia que contiene las distancias entre cada par de objetos en la base de datos

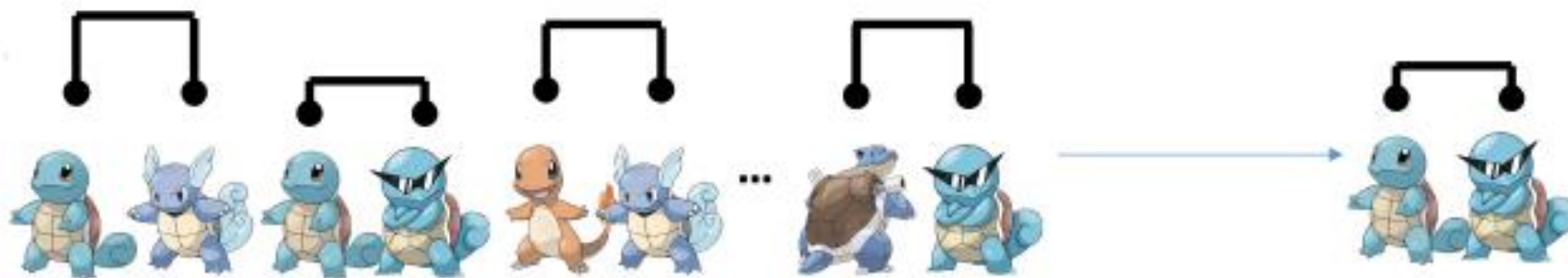
$$D(\text{Charmander}, \text{Blastoise}) = 8$$

$$D(\text{Squirtle}, \text{Blastoise}) = 1$$

					
	0	8	8	7	7
		0	2	4	4
			0	3	3
				0	1
					0

Clustering Jerárquico Aglomerativo: Pasos

Empezar con cada elemento en su propio cluster, encontrar el mejor par y unirlo en un nuevo cluster. Repetir hasta que todos los clusters estén unidos

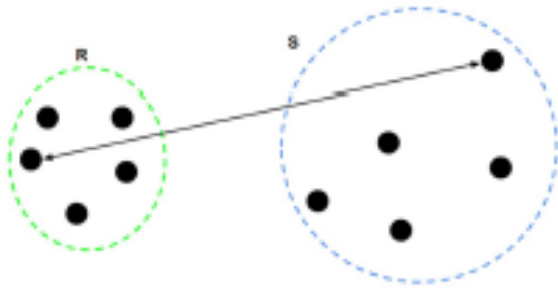


Considerar todas las posibles uniones y elegir la mejor

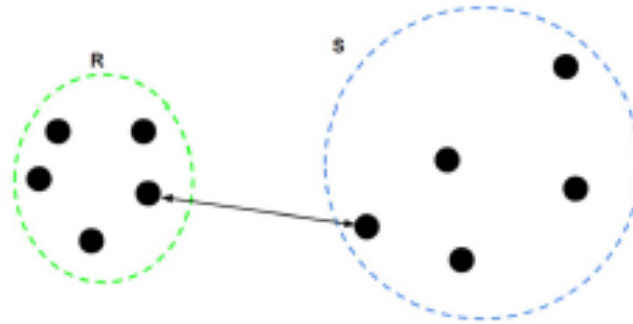


Clustering Jerárquico Aglomerativo: Pasos

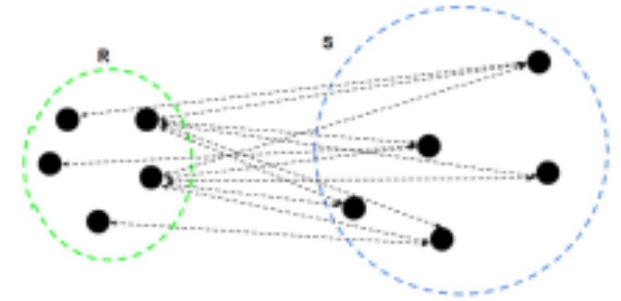
Criterios de similitud



Complete Linkage
Distancia máxima entre dos
puntos de distintos clusters



Single Linkage
Distancia mínima entre dos
puntos de distintos clusters



Average Linkage
Promedio de las distancias entre
los puntos de distintos clusters

Clustering Jerárquico

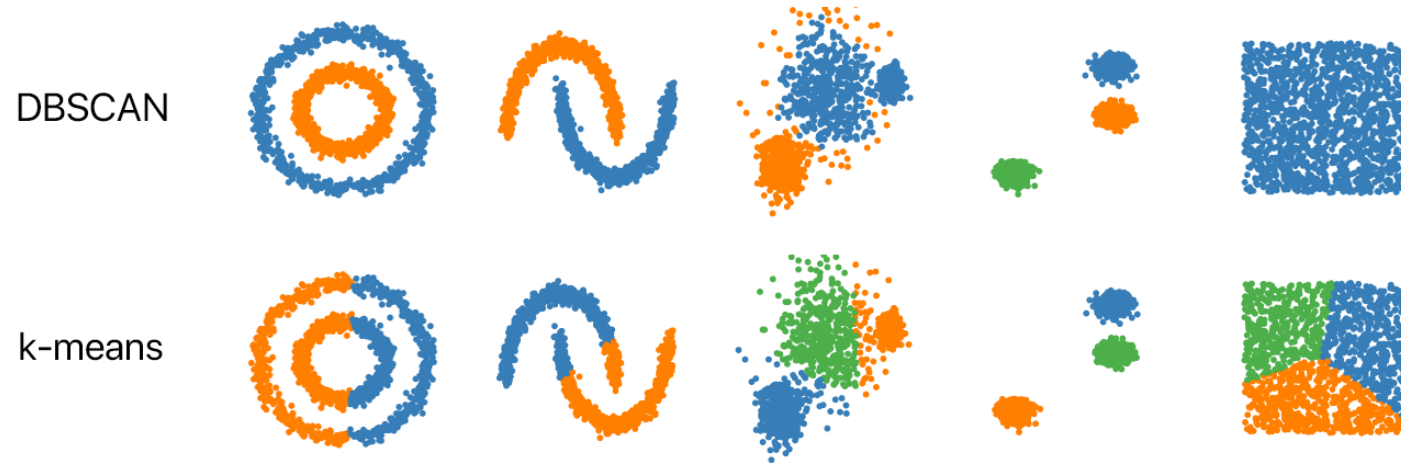
✓ Ventajas:

- No requiere un número de *clusters* predefinido.
- Permite establecer jerarquías entre *clusters* y elementos.
- Un dendrograma es fácilmente interpretable

✓ Desventajas:

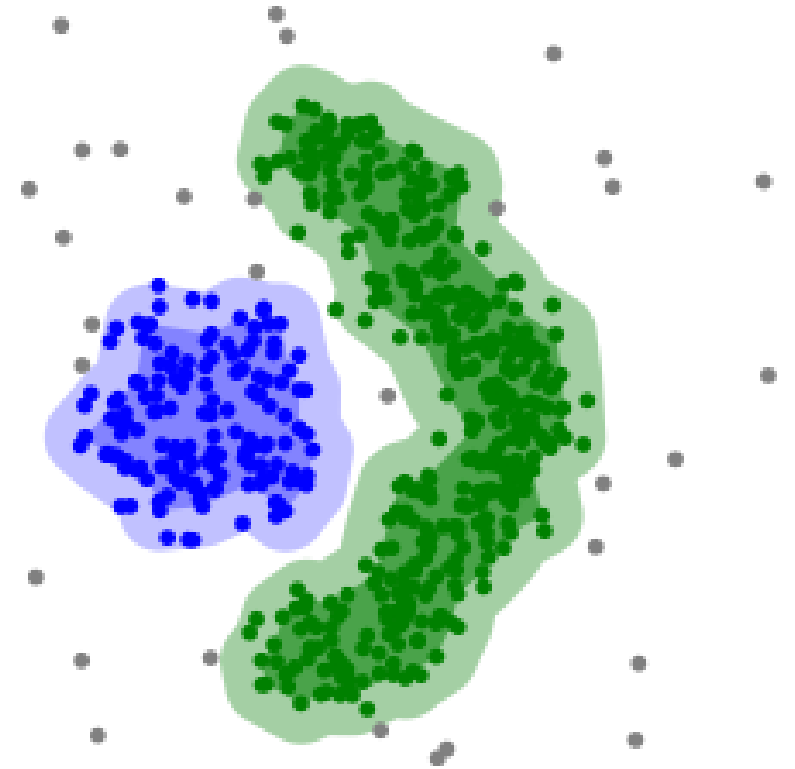
- Carece de una función objetivo global. No siempre es fácil definir los nivel de selección de los clústers.
- Es costoso en tiempo y espacio de almacenamiento
- La decisión de mezclar clusters es irreversible.
- Lo anterior representa un problema en conjuntos de datos con mucho ruido y de muchas dimensiones

Agrupamiento basado en densidad: Segmentación DBSCAN



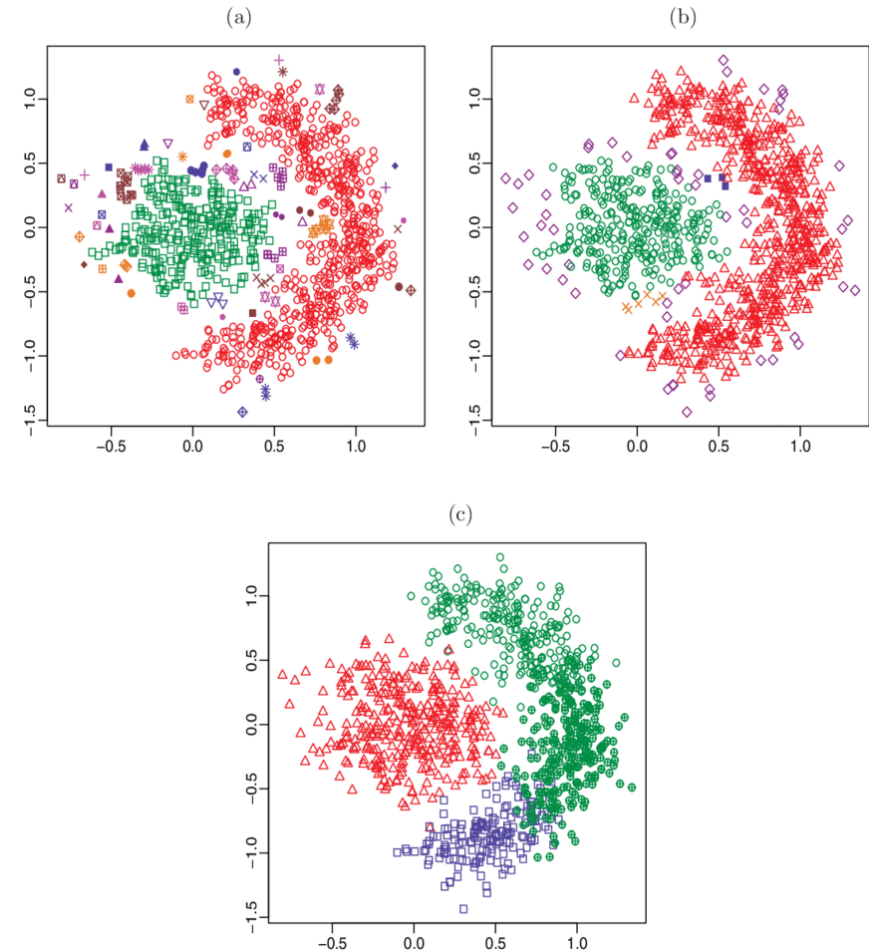
Agrupamiento basado en densidad

- ✓ Los métodos basados en la distancia tienden a funcionar bien con clusteres esféricos y mal con clusteres con otras formas.
- ✓ Para solucionar este problema otros métodos han desarrollado el concepto de densidad, el cual permite descubrir clusteres con ***formas arbitrarias*** (conjunto de datos que determinan un volúmen).
- ✓ La idea es hacer crecer un cluster siempre cuando la densidad en el entorno del objeto exceda de un ***umbral***.



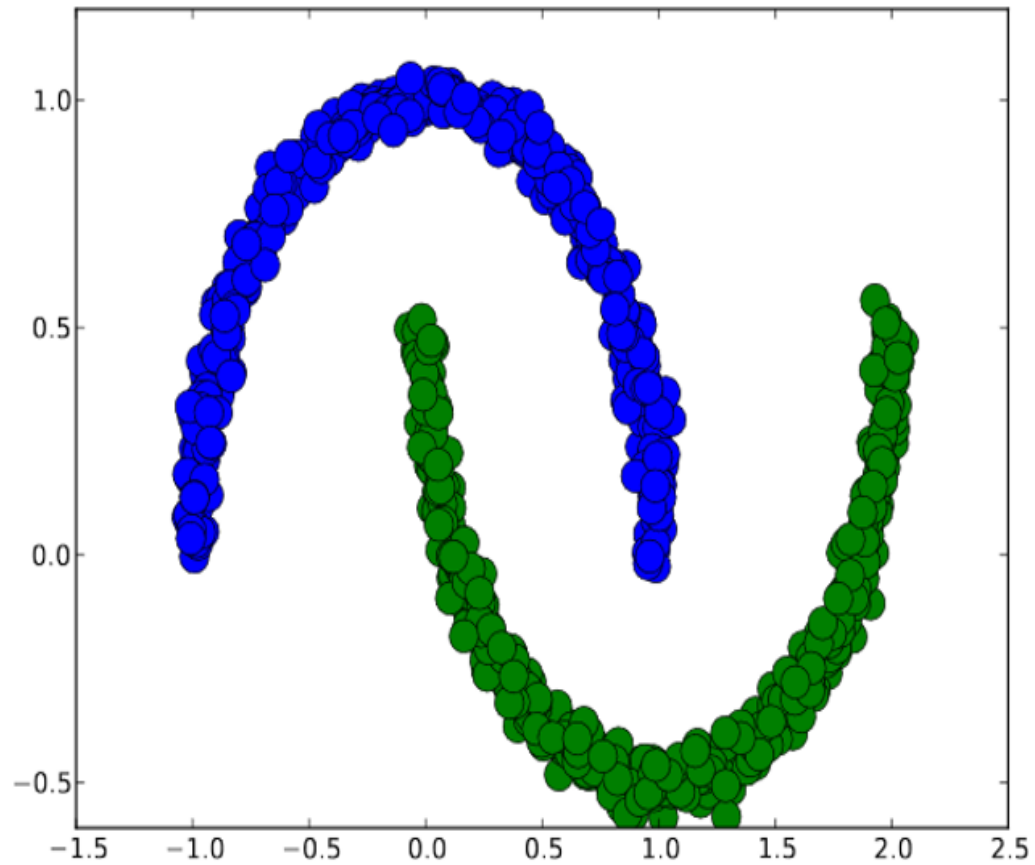
DBSCAN

- ✓ DBSCAN: *Density-Based Spatial Clustering of Application Noise*.
- ✓ Asume que la densidad alrededor de los datos normales es similar a la densidad alrededor de sus vecinos.
- ✓ La densidad alrededor de los valores atípicos es considerablemente diferente a la densidad alrededor de sus vecinos.
- ✓ Hace crecer regiones con suficiente alta densidad en grupos
- ✓ Estos grupos están separados por regiones de baja densidad de objetos (ruidos).

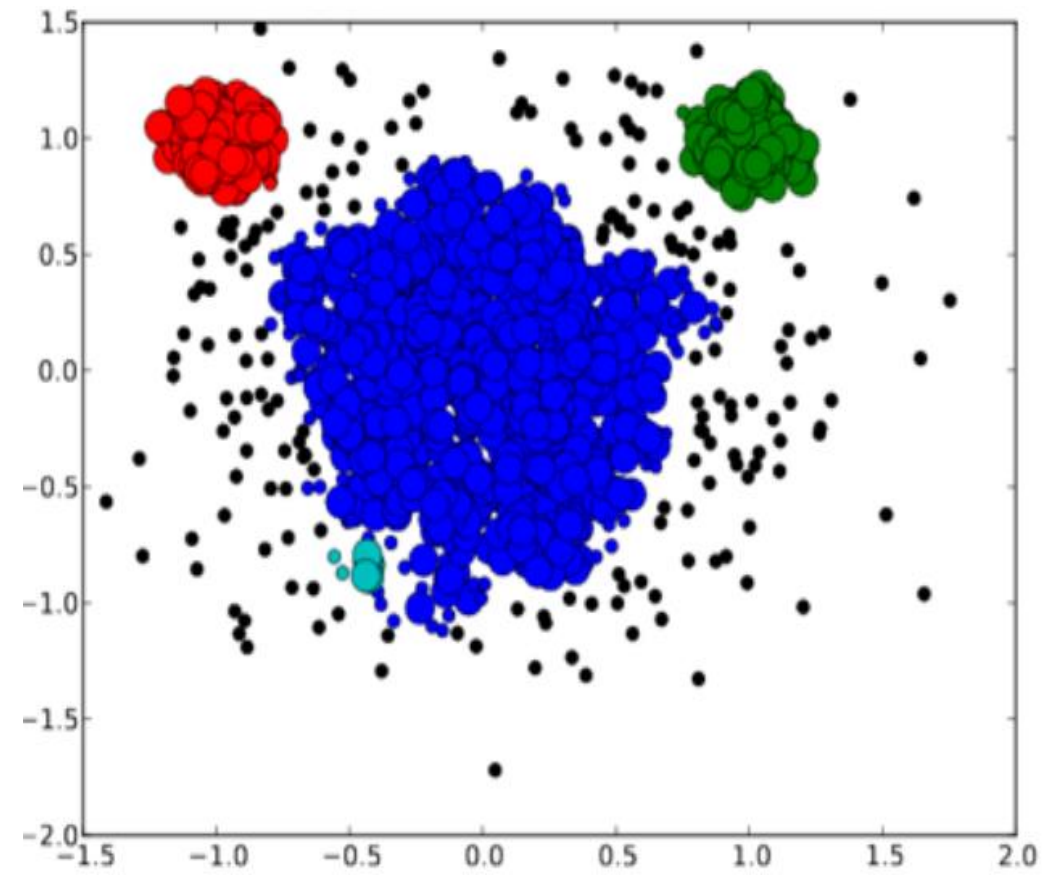


DBSCAN

Número estimado de cluster: 2



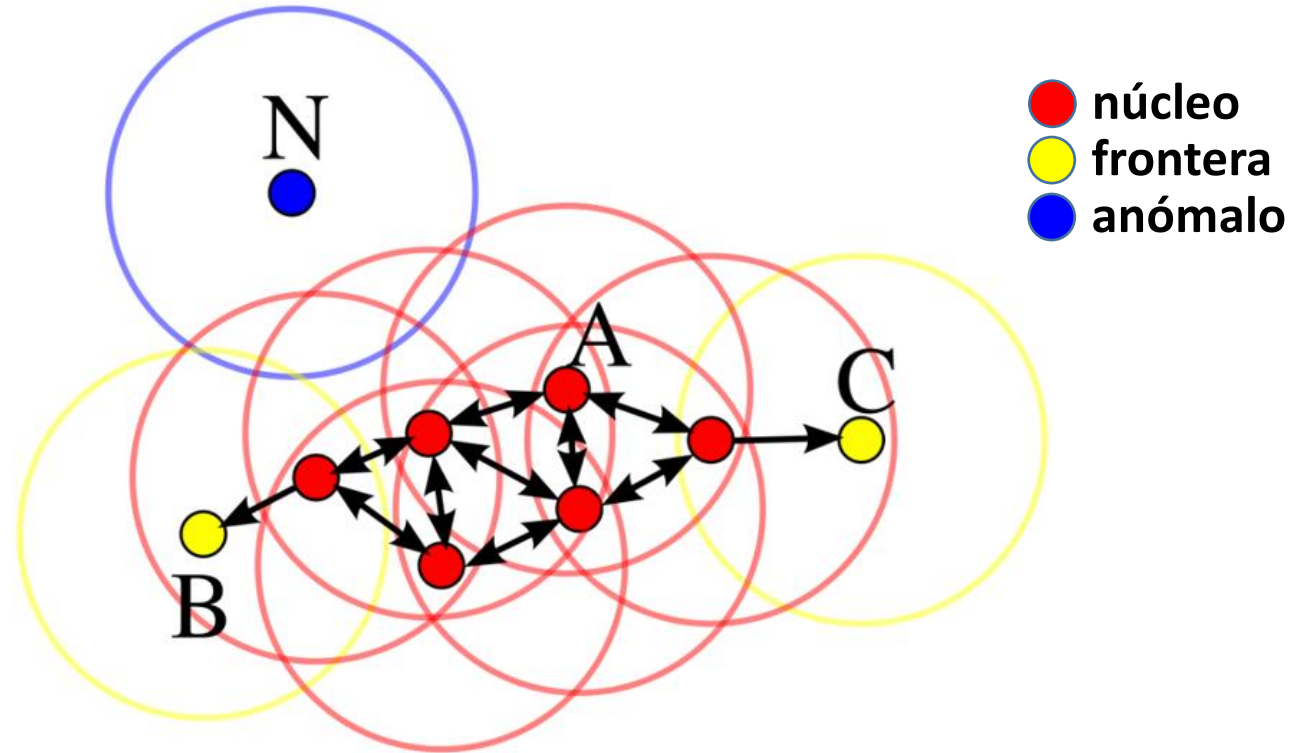
Número estimado de cluster: 4



DBSCAN: Algoritmo

- ✓ El algoritmo requiere dos parámetros principales:
 - ***El parámetro epsilon (eps)***, define el radio de vecindad alrededor de un punto.
 - ***El número mínimo de puntos (MinPts)*** de vecinos en un radio **eps**.
- ✓ Cualquier punto en el set de datos, con un número mayor o igual que ***MinPts*** se considera un **punto núcleo**.
- ✓ Un punto se considera **punto frontera** si tiene menos de MinPts vecinos pero el es vecino de un punto núcleo.
- ✓ Un punto que no es ni núcleo ni frontera, se considera un **punto de ruido** o valor atípico (outlier).

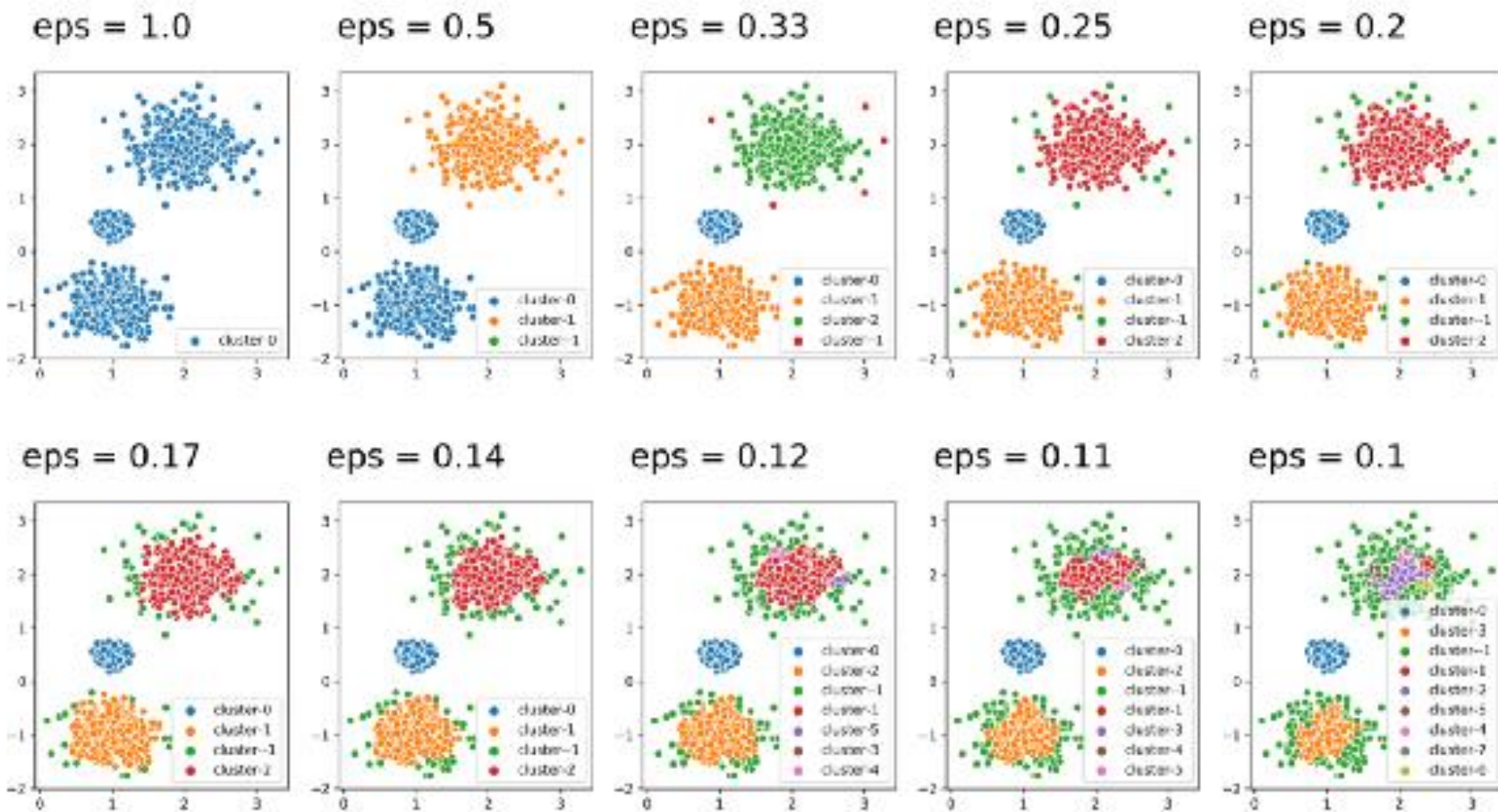
DBSCAN: Algoritmo



¿Cómo funciona?

- 1) Busca clústeres comprobando el vecindario (eps) de cada punto
- 2) Si existe MinPts, un nuevo clúster es creado
- 3) Busca iterativamente puntos que son directamente alcanzables
- 4) El proceso termina cuando no se pueden añadir nuevos puntos a ningún clúster

DBSCAN: Algoritmo



DBSCAN

✓ **Ventajas:**

- No necesita asumir un número fijo de clusteres.
- No depende de las condiciones de inicio.
- Encuentra clústeres no separables linealmente

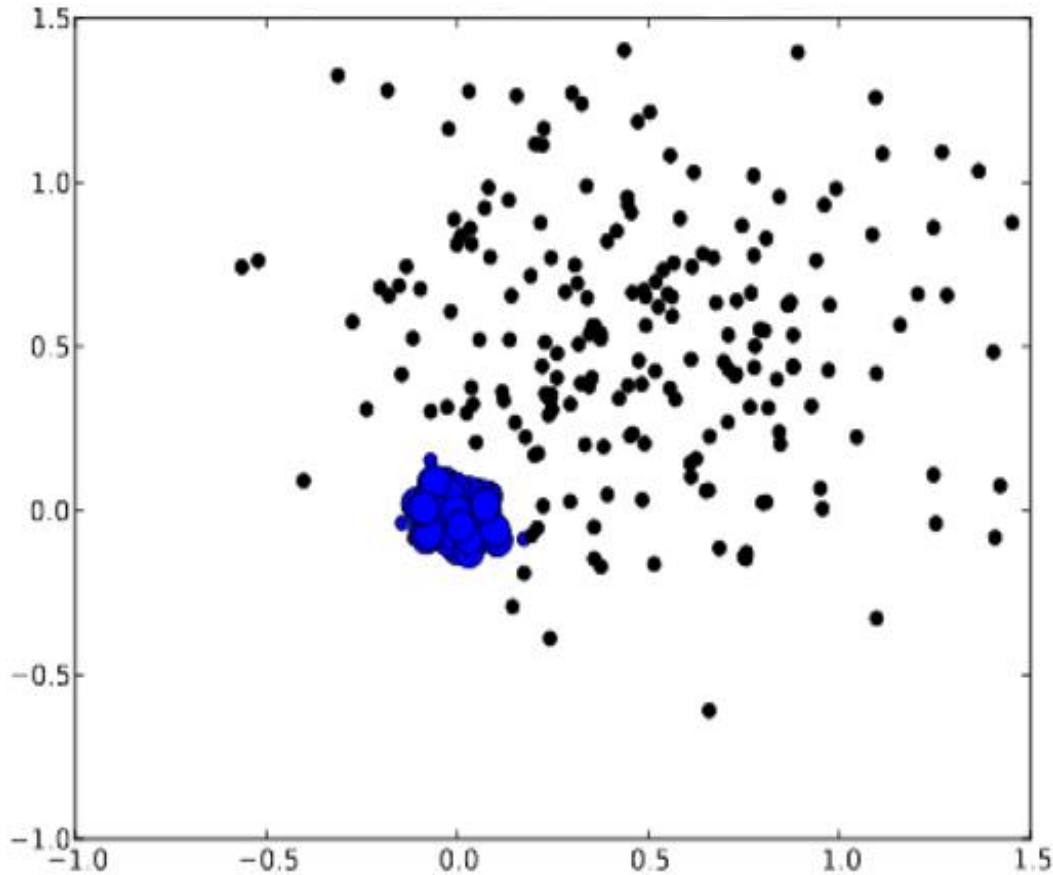
✓ **Desventajas:**

- Asume densidades similares en todos los clusteres.
- Puede tener problemas al separar clusteres.
- No siempre es fácil definir los nivel de selección de los clústers.

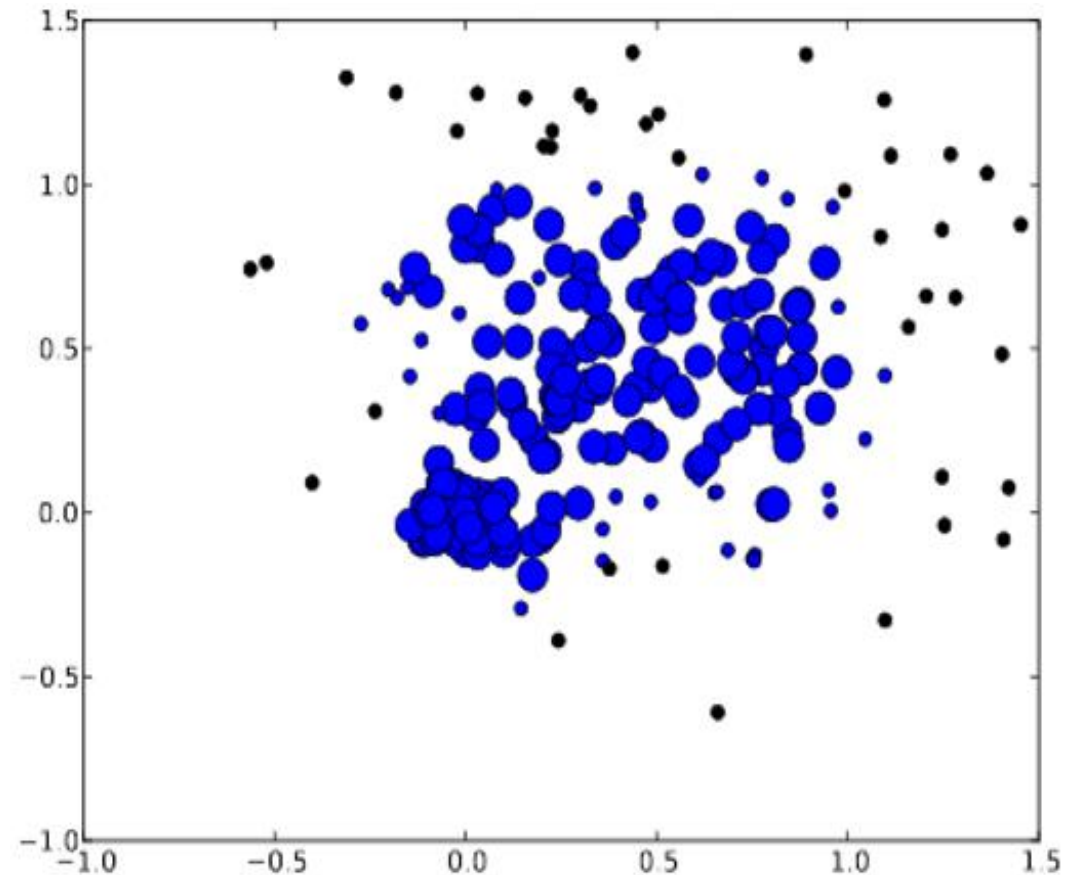
DBSCAN

Para el caso de un cluster muy disperso y otro compacto uno cercano al otro ¿qué pasaría?

Número de cluster estimado: 1



Número de cluster estimado: 1



Aplicación: Segmentación de Clientes



Segmentación de clientes

Estudios de Mercado



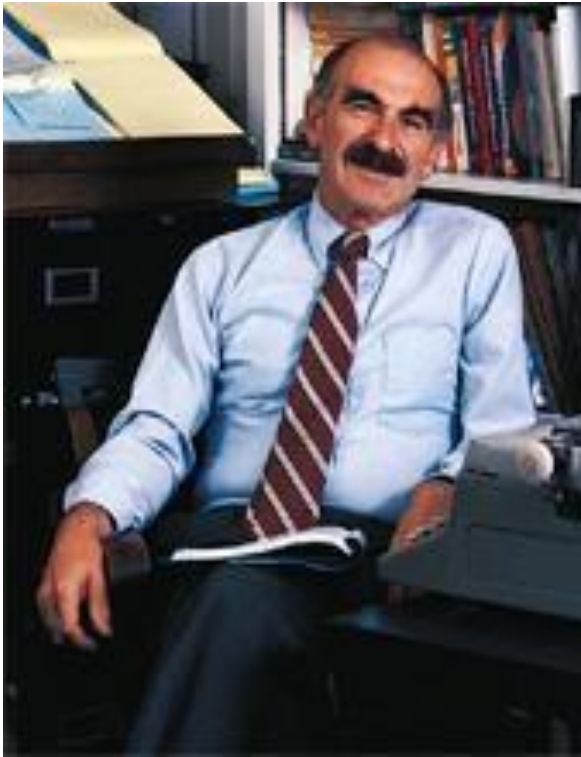
Se pueden implementar soluciones Machine Learning para la detección del fraude, incumplimiento de normativas, o asistencia virtual para clientes por perfiles.

La aplicación del Machine Learning permite la segmentación de clientes para identificar mejor a los grupos de clientes y personalizar mejor la oferta.

Sistema Financiero



Segmentación de clientes



*“If you’re not thinking segments,
you’re not thinking.
To think segments means you have
to think about what drives
customers, customer groups, and
the choices that are or might be
available to them.”*

- Ted Levitt, Marketing Imagination

Importancia de la segmentación de clientes

- ❑ **Tenemos que manejar cientos, miles, o millones de clientes / prospectos cada día.**
- ❑ **No podemos tratarlos todos de forma idéntica.**
 - ✓ Tienen otras necesidades, deseos, perfiles, potencial de rentabilidad.
 - ✓ La era del “one fits all” se fue hace mucho tiempo.
 - ✓ En Marketing, se trata de “customization”.
- ❑ **No podemos tratar los todos de forma diferente.**
 - ✓ Son demasiados.
 - ✓ Una empresa generalmente no tiene los recursos, / la capacidad de adaptar todo (producto, oferta, precio, comunicación) para cada cliente 50.

Características de un buen segmento

❑ Distinto de los otros segmentos.

- ✓ Si dos segmentos tienen los mismos clientes, **cual es el punto?**

❑ Homogéneo

- ✓ De lo contrario, clientes en un segmento no se pueden considerar como idénticos y eso niega el objetivo de la segmentación.

❑ Identificable

- ✓ Un segmento no es solo conceptual. Tiene que ser operacional .
- ✓ Si no se puede identificar/predecir los miembros de un segmento, tiene muy poco valor.

❑ Sustancial

- ✓ Mas en términos económicos que en cantidad de clientes.
- ✓ Se justifica el trato preferencial y/o diferente

❑ Interesante (useful),operacional, y informativo

- ✓ Demasiado segmentos mata la segmentación. Quienes son? Que me ensenaña? Ayuda a definir o mejorar mi estrategia?

Pasos para implementar un modelo de segmentación

Realizar el análisis exploratorio de datos, con la finalidad de tener completitud de los datos además de una correcta estandarización de datos.

AED DE DATOS

Definir la metodología de segmentación a utilizar, sea un cluster jerárquico o no jerárquico el que deseemos implementar.

SEGMENTACIÓN

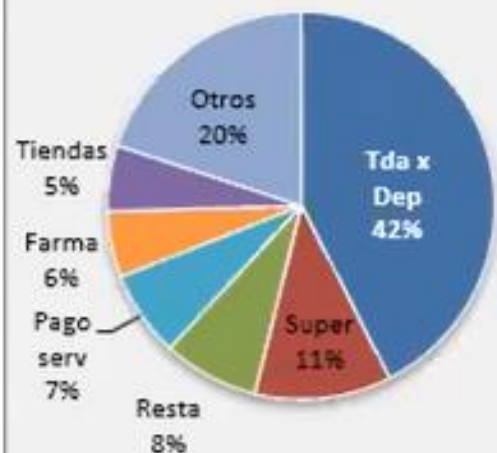
Examinar el centroide y perfilar los clúster o grupos de acuerdo a éstos, añadiendo otras variables relevantes.

PERFILAMIENTO

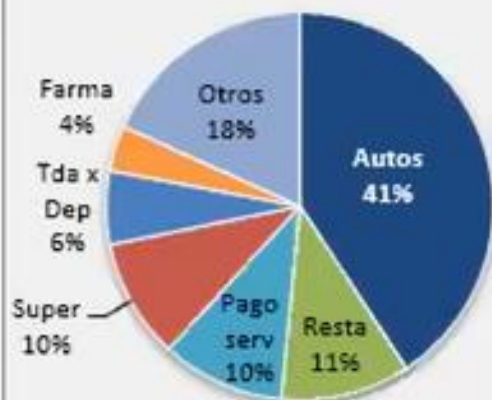
Revisar algunos indicadores de segmentación o agrupamiento.

VALIDACIÓN

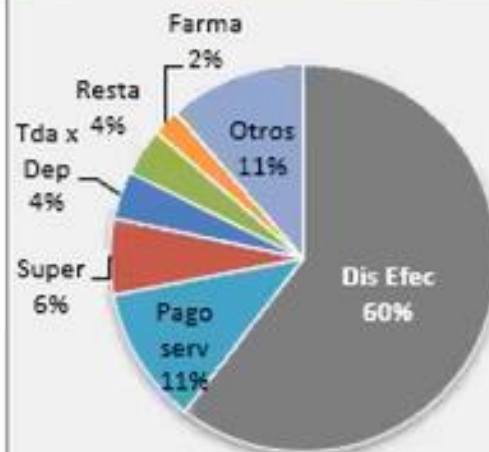
Amantes de la moda (66K, 14%)



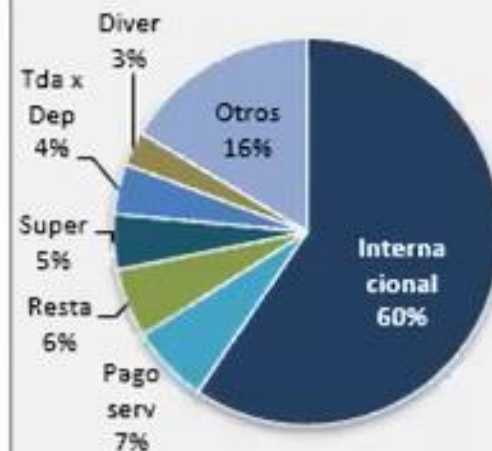
Amantes de los autos (26K, 5%)



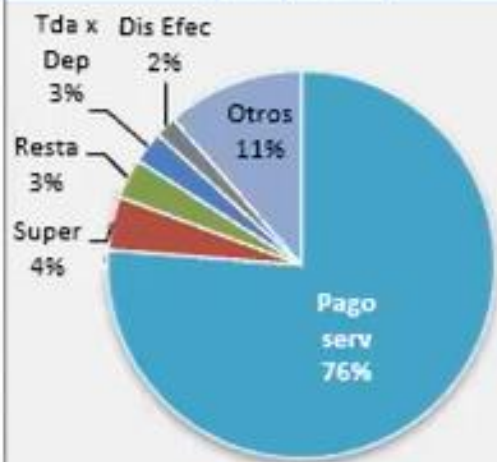
Ajustados (32K, 7%)



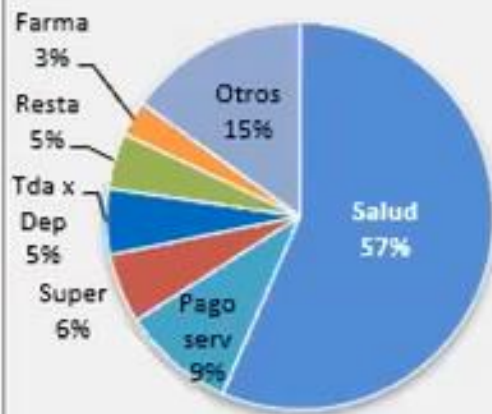
Cosmopolita (39K, 8%)



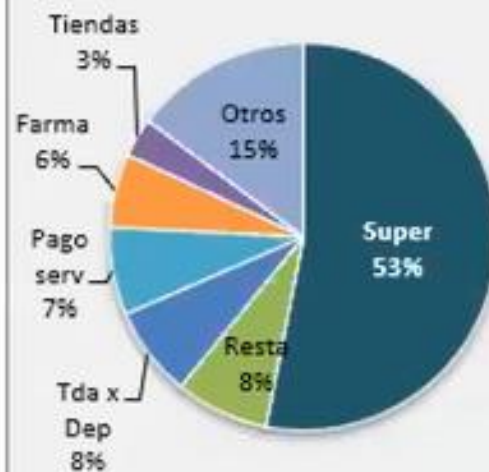
Planificados (52K, 11%)



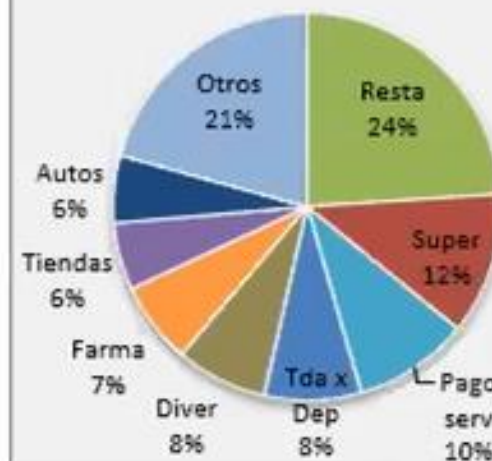
Saludables (28K, 6%)



Tradicionales (68, 14%)



Diversificados (159K, 34%)





Amantes de los autos

(26K, 5%)



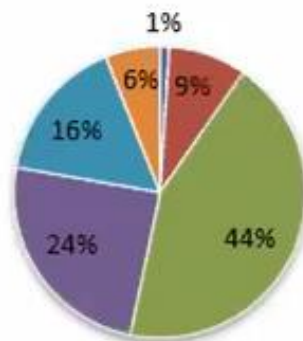
- 41% trx en autos
- 5 trx x cliente mensual (5% del total)
- S/ 821 mto x cliente mensual (5% del tot)



24%

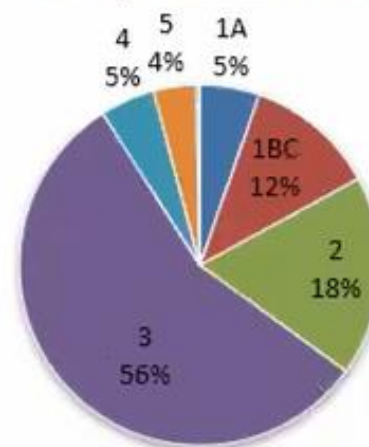


76%

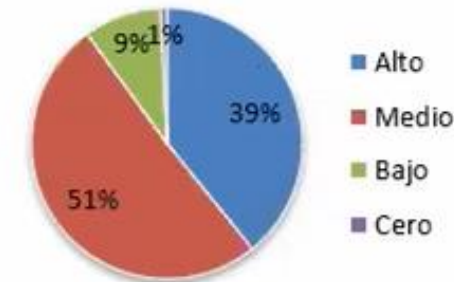


- Hasta 25 años
- De 26 a 30 años
- De 31 a 45 años
- De 46 a 55 años
- De 56 a 65 años
- Más de 65 años

56% pertenecen al S3



39% tienen rentabilidad alta

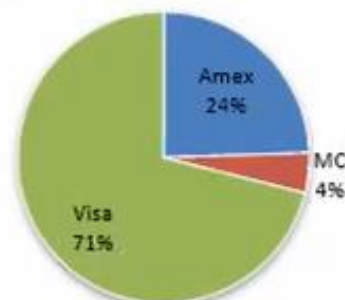


Clientes de TC con 2 años de antigüedad

Residencia del cliente



Tipo TC



Segmento / Total

Ingreso	S/ 2.9K / S/ 2.7K
SOW	57% / 56%
Saldo últ. mes	S/ 2.1K / S/ 1.6K
% Trx en cuotas	26% / 21%
% Afiliación PA	3.0% / 3.9%

Comercios Top



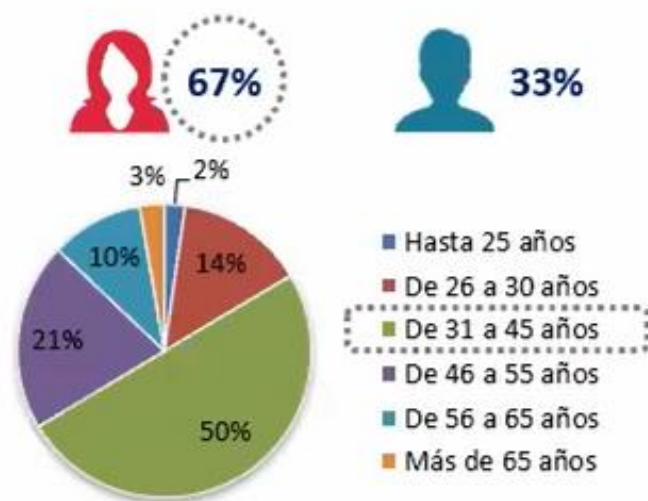


Amantes de la moda

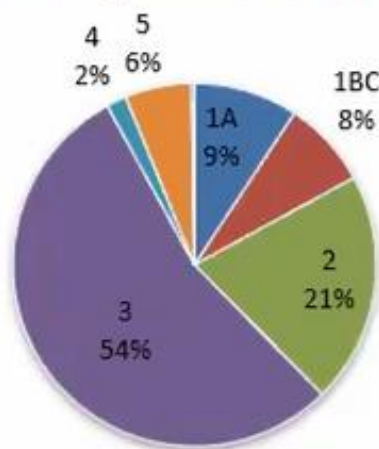
(66K, 14%)



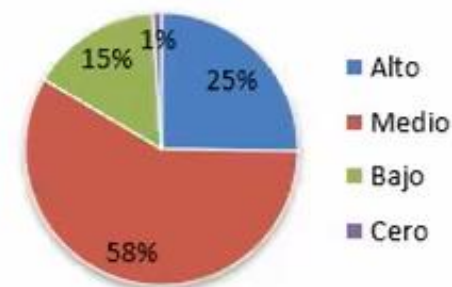
- 42% trx tiendas por departamento
- 3 trx x cliente mensual (9% del total)
- S/ 608 mto x cliente mensual (10% del to)



54% pertenecen al **S3**



74% tienen rentabilidad media y baja

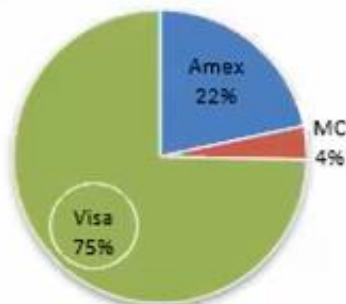


Clientes de **TC** con **2 años** de antigüedad

Residencia del cliente



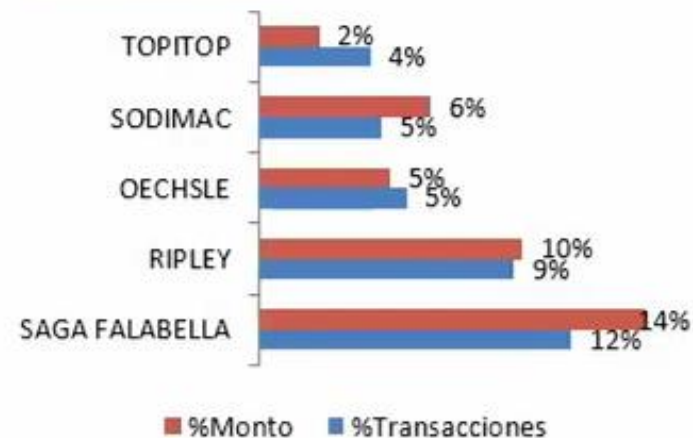
Tipo TC



Segmento / Total

Ingreso	S/ 2.3K / S/ 2.7K
SOW	60% / 56%
Saldo últ. mes	S/ 1.3K / S/ 1.6K
% Trx en cuotas	37% / 21%
% Afiliación PA	2.2% / 3.9%

Comercios Top



Market Basket Analysis



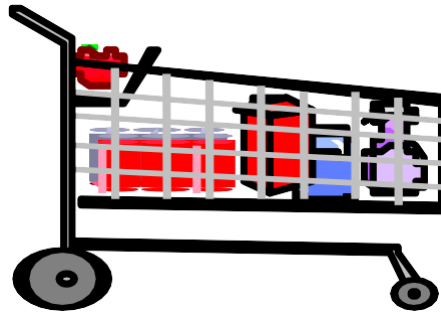
Market Basket Analysis

- ✓ Los hábitos de compra de los clientes pueden ser representados a través de asociaciones o correlaciones entre los diferentes productos que compran en sus “canastas”.



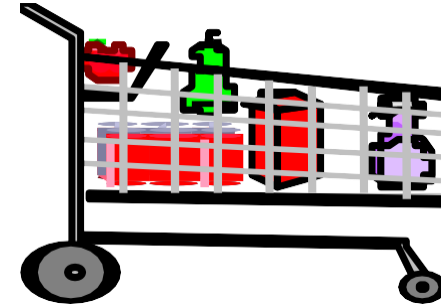
Cliente 1:

Arroz, puré, bebida



Cliente 2:

Arroz, helado, pan



Cliente 3:

Arroz, bebida, cerveza

- ✓ En base a la información obtenida, el análisis de Cesta de Canasta de Mercado nos permite conocer:
 - Las Necesidades de los Consumidores.
 - Los productos que suelen comprarse al mismo tiempo.
 - El perfil del consumidor que adquiere determinado tipo de producto y el momento en que lo hace.

Market Basket Analysis

- ✓ Es una técnica utilizada para descubrir relaciones entre los productos que compran los usuarios.
- ✓ Se observan las combinaciones de productos que son compradas conjuntamente en las transacciones.
- ✓ Utilizando esta información, ¿es posible que una tienda pueda tomar decisiones para incrementar sus ventas?



Market Basket Analysis: Aplicaciones en el Negocio

- ☐ El **posicionamiento de los productos en el lineal**. Colocar juntas la cerveza y las patatas fritas si se ha determinado que se compran simultáneamente, maximiza la venta de ambos productos.
- ☐ Las **mejoras de las ofertas “trade” o combos**. En el sector del retail habitualmente se hacen ofertas de paquetes de compra de 2 o más productos.
- ☐ La **selección del surtido** para las tiendas. En las tiendas no caben todos los productos de los que dispone el supermercado.
- ☐ La **venta cruzada** (o “cross-selling”) de productos complementarios en caja, ¿qué pasaría si pudiéramos hacer una oferta personalizada en función del ticket de compra del cliente, en lugar de ofrecer siempre el mismo producto, sea cual sea el cliente? .

Market Basket Analysis

- ✓ Si tengo conocimiento de que dos productos se compran juntos frecuentemente:
 - Puedo colocar ambos productos cerca del otro en un supermercado.
 - Puedo aplicar descuentos para uno de los dos productos.
 - Se puede ofrecer publicidad de un producto a compradores del otro producto.
 - Puedo generar nuevos productos o bundles a partir de los productos originales.

Frequently bought together



i These items are shipped from and sold by different sellers. [Show details](#)

- ✓ **This item:** Nikon D850 FX-Format Digital SLR Camera Body **\$2,996.95**
- ✓ Sony Professional XQD G Series 64GB Memory Card (QDG64E/J) **\$129.95**
- ✓ EN-EL15a Rechargeable Li-ion Battery **\$54.95**

**¿Cómo descubrimos asociaciones
entre productos?**

Análisis de Asociación

- ✓ El Análisis de Asociación, o Análisis de Reglas de Asociación, se define como la tarea de encontrar relaciones.
- ✓ interesantes/relevantes en un largo conjunto de datos.
- ✓ Dicho de otro modo, se trata de descubrir cómo diferentes elementos se encuentran asociados entre sí.

$$\{X \rightarrow Y\}$$





{Antecedente → Consecuente}

Regla de asociación

Y sucede si es que ha sucedido X
(el sentido inverso no es igual)

Análisis de Asociación: Métricas e Indicadores

- ✓ Imaginemos que se realizan las siguientes compras en un retail :

Transaction ID				
TRX - 001	1	1	1	1
TRX - 002	1	0	1	1
TRX - 003	0	0	1	1
TRX - 004	0	1	0	0
TRX - 005	1	1	1	1
TRX - 006	1	1	0	1

Análisis de Asociación: Métricas e Indicadores

- ✓ Observando las compras anteriores de las 6 transacciones podemos encontrar o buscar reglas de asociación o patrones de compras interesantes. Revisemos los principales indicadores:

1.- Support: Es la popularidad predeterminada, frecuencia de ocurrencia o impacto en ventas de un artículo. En términos matemáticos, el soporte del elemento A no es más que la relación entre las transacciones que involucran a A y el número total de transacciones.

$$\text{Soporte (Uvas)} = \frac{(\text{Transacciones que involucran uvas})}{(\text{Total de transacciones})}$$

$$\text{Soporte (Uvas)} = 0.666$$



Análisis de Asociación: Métricas e Indicadores

2.- Confidence: Probabilidad condicional de que el cliente que compró el producto A, compre B. Divide el número de transacciones que involucran tanto a A como a B, por el número de transacciones que involucran a A. **Desventaja:** Puede representar erróneamente la importancia de una asociación.

Confidence (A => B) = (Transacciones que involucran A y B) / (Total de transacciones que involucran A) = Support (A, B) / Support (A)

Confidence({Uvas, Manzanas} => {Mango}) =
(Transacciones que involucran uvas, manzanas y mango)
/ (Transacciones que involucran uvas y manzanas)

Confidence({Uvas, Manzanas} => {Mango}) = ((2/6) / (3/6)) = **0.666**

Análisis de Asociación: Métricas e Indicadores

3.- Lift: Esta medida indica que tan usual sería que un ítem B sea comprado si es que el ítem A también fue comprado, tomando en cuenta que tan popular es el ítem B.

$$\begin{aligned}
 \text{Lift (A} \Rightarrow \text{B)} &= \text{Confidence (A, B)} / \text{Support (B)} \\
 &= \text{Support (A, B)} / (\text{Support (A)} \times \text{Support (B)})
 \end{aligned}$$

$$\begin{aligned}
 \text{Lift (\{Uvas, Manzana\} } \Rightarrow \text{\{Mango\})} &= \text{Confidence(\{Uvas, Manzana\}, \{Mango\})} / \text{Support} \\
 (\text{Mango}) &= \text{Support(\{Uvas, Manzana\}, \{Mango\})} / (\text{Support (\{Uvas, Manzana\})} \times \text{Support} \\
 &\quad \text{\{Mango\}})
 \end{aligned}$$

$$\text{Lift (\{Uvas, Manzana\} } \Rightarrow \text{\{Mango\})} = 1$$



Análisis de Asociación: Métricas e Indicadores

Métrica 1: Support

$$\text{Support} \{\text{🍎}\} = \frac{4}{8}$$

Métrica 2: Confidence

$$\text{Confidence} \{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Support} \{\text{🍎, 🍺}\}}{\text{Support} \{\text{🍎}\}}$$

Métrica 3: Lift

$$\text{Lift} \{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Support} \{\text{🍎, 🍺}\}}{\text{Support} \{\text{🍎}\} \times \text{Support} \{\text{🍺}\}}$$

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍏
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍏

- Si el valor fuera mayor que 1, el item Y sería usualmente comprado si X es comprado.
- Si el valor fuera menor que 1, el item Y no sería usualmente comprado si el item X es comprado.

Análisis de Asociación: Procedimiento

- ✓ Comenzar con itemsets de tamaño $k = 1$ ir incrementando el valor de k de 1 en 1 descartando aquellos itemsets que no cumplan un soporte mínimo:

Items (1-itemsets)

Item	Count
Bread	4
Peanuts	4
Milk	6
Fruit	6
Jam	5
Soda	6
Chips	4
Steak	1
Cheese	1
Yogurt	1

Minimum Support = 4

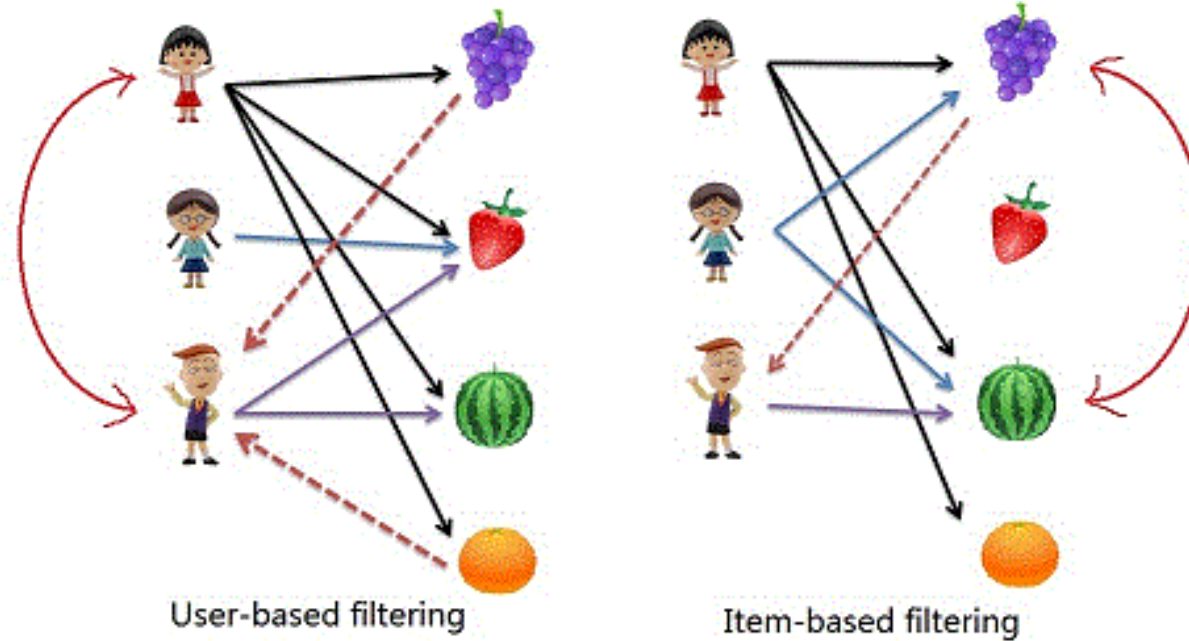
2-itemsets

2-Itemset	Count
Bread, Jam	4
Peanuts, Fruit	4
Milk, Fruit	5
Milk, Jam	4
Milk, Soda	5
Fruit, Soda	4
Jam, Soda	4
Soda, Chips	4

3-itemsets

3-Itemset	Count
Milk, Fruit, Soda	4

Introducción a los sistemas de recomendación



Introducción a los sistemas de recomendación

La era del ecommerce ha permitido que las opciones disponibles para los usuarios incrementen de forma considerable





Ventaja

Muchas opciones para los consumidores

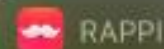
Desventaja

Demasiadas opciones para los consumidores

People You May Know

	Alonso Huiman Salarrayán 1 mutual friend	Add Friend	Remove
	Carlos Huisa 22 mutual friends	Add Friend	Remove
	Marianella Beatriz 40 mutual friends	Add Friend	Remove
	Victor Hugo Quilca 6 mutual friends	Add Friend	Remove

Because you watched Marvel's Daredevil



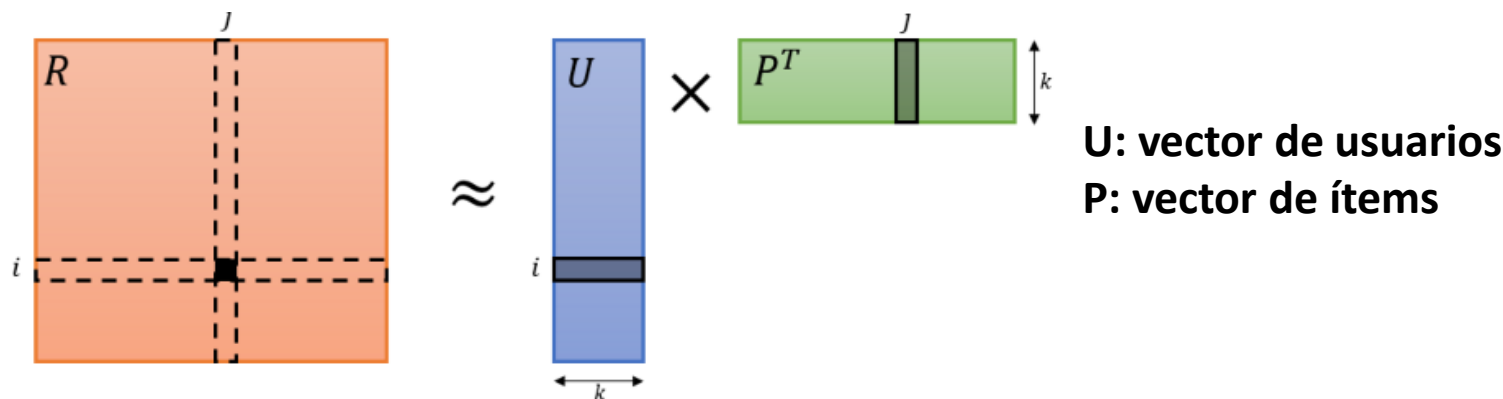
RAPPI

8m ago

Manuel, ¡no te quedes por fuera!
En los últimos días 62 personas como tú han pedido
a Roll-Star Sushi 🍱

Introducción a los sistemas de recomendación

- ✓ Es un sistema inteligente que proporciona a los usuarios una serie de sugerencias personalizadas (**recomendaciones**) sobre un determinado tipo de elementos (**items**).
- ✓ Está comprendida por la técnica Collaborative filtering (toma en cuenta el comportamiento pasado).
- ✓ Modelo de factores latentes (método de factorización de matrices).



- ✓ Se busca minimizar el error cuadrático medio (MSE)

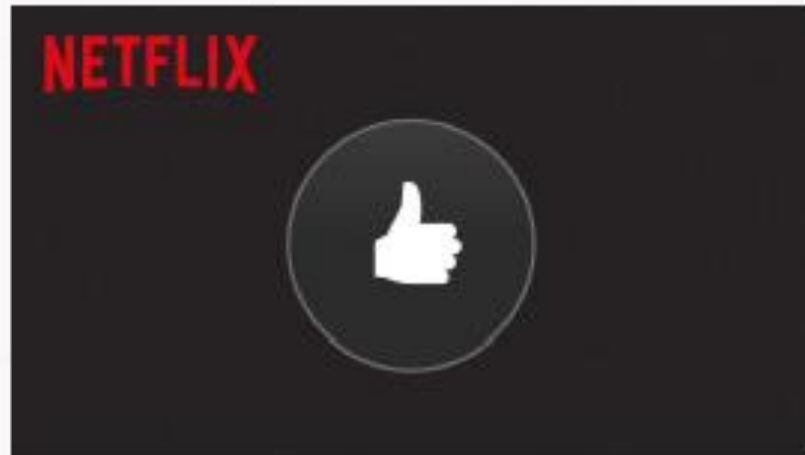
$$MSE = \frac{1}{n} \sum_{i,j} (r_{i,j} - u_i \cdot p_j)^2$$

$r_{i,j}$ es el rating observado y $u_i \cdot p_j$ representa el rating predicho del usuario i al ítem j

¿Cómo modelar lo que le gusta a mis usuarios?

Feedback explícito

El usuario me dice que le gusta y que no

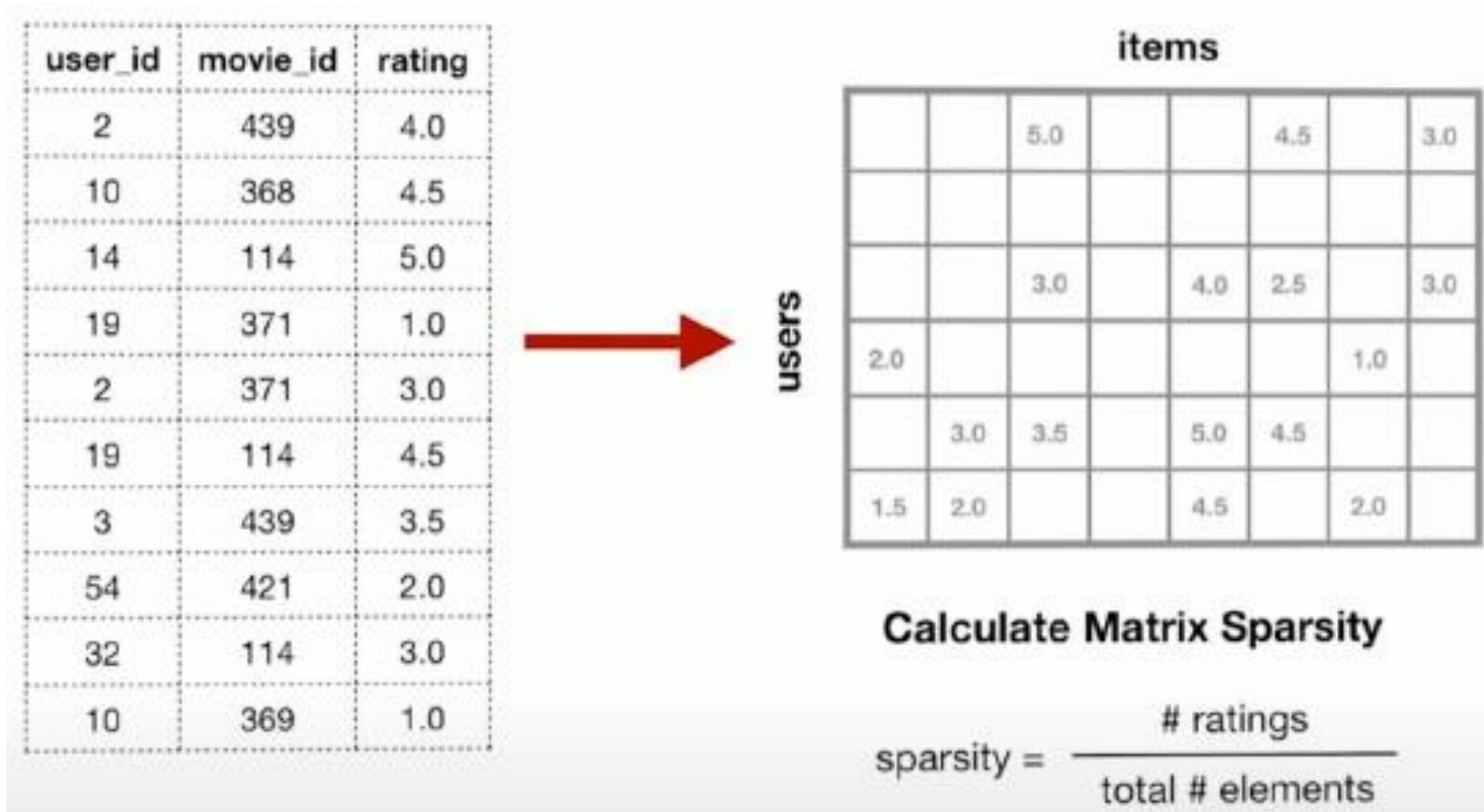


Feedback implícito

Infiero que le gusta al usuario según su comportamiento



Paradigma clásico en los sistemas de recomendación



Paradigma clásico en los sistemas de recomendación



U1	2	?	5	?	?
U2	5	?	?	?	2
U3	?	5	?	3.5	?
U4	?	?	4.5	?	3.5
U5	?	3.5	?	?	?
U6	3.5	?	5	5	?

Matriz de utilidad

Paradigma clásico en los sistemas de recomendación

Intuición

- Describir a cada **ítem** como una serie de características.
- Describir a cada **usuario** según qué tanto le gustan esas características



Qué tanto contiene

Acción	Romance
4.2	-1.2

Qué tanto les gusta

Acción	Romance
1.5	-0.8



Acción	Romance
-1.5	3.1



Paradigma clásico en los sistemas de recomendación

Factorización de Matrices

$$\begin{array}{c}
 \begin{array}{c} \text{User 1} \\ \text{Icon 1} \end{array} \times \begin{array}{c} \text{Movie 1} \\ \text{Icon 2} \end{array} = \begin{array}{c} \text{Qué tanto le gusta} \\ \begin{array}{|c|c|} \hline \text{Acción} & \text{Románticas} \\ \hline 1.5 & -0.8 \\ \hline \end{array} \end{array} \times \begin{array}{c} \text{Qué tanto contiene} \\ \begin{array}{|c|c|} \hline \text{Acción} & \text{Romántica} \\ \hline 4.2 & -1.2 \\ \hline \end{array} \end{array} = 6.72 \quad \begin{array}{c} \text{Like} \\ \text{Icon 3} \end{array}
 \end{array}$$

$$\begin{array}{c}
 \begin{array}{c} \text{User 2} \\ \text{Icon 4} \end{array} \times \begin{array}{c} \text{Movie 1} \\ \text{Icon 2} \end{array} = \begin{array}{c} \text{Qué tanto le gusta} \\ \begin{array}{|c|c|} \hline \text{Acción} & \text{Románticas} \\ \hline -1.5 & 3.1 \\ \hline \end{array} \end{array} \times \begin{array}{c} \text{Qué tanto contiene} \\ \begin{array}{|c|c|} \hline \text{Acción} & \text{Romántica} \\ \hline 4.2 & -1.2 \\ \hline \end{array} \end{array} = -9.48 \quad \begin{array}{c} \text{Dislike} \\ \text{Icon 5} \end{array}
 \end{array}$$

Paradigma clásico en los sistemas de recomendación

Factorización de Matrices

ITEMS

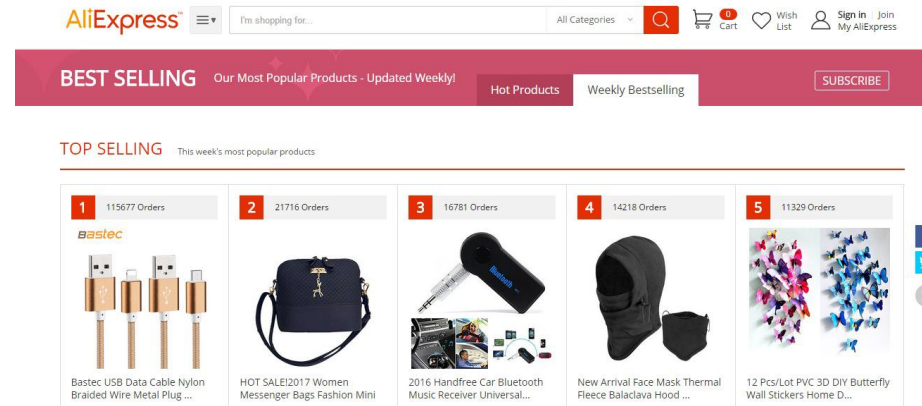
1.3	1.2	1.5	0.7	0.1
3.5	1.2	2.5	1.7	1.4

USUARIOS

1.2	2.1
0.5	0.7
0.2	1.3
1.7	1.1
1.2	1.3
1.5	1.1

2	3.96	5	4.41	3.06
5	1.44	2.5	1.54	2
4.81	5	3.55	3.5	1.84
5	3.36	4.5	3.06	3.5
5	3.5	5	3.05	1.94
3.5	3.12	5	5	1.69

Agregados por popularidad



Agregados por contenido



Año	1994	1997	1976	2003	2013
Género	Drama / Crime	Romantic / Drama	Drama / Sport	Drama / Crime	Sitcom
Director	Tarantino	Cameron	Avildsen	Tarantino	Roiland
Muerte	Si	Si	No	Si	No
Armas	Si	No	No	Si	No
Política	No	No	No	No	No
Actores	Travolta	Di Caprio	Stallone	Uma Thurman	Roiland

Agregados por contenido



Basados en contenido: Problemas



Agregados por contenido



Basados en contenido: Problemas

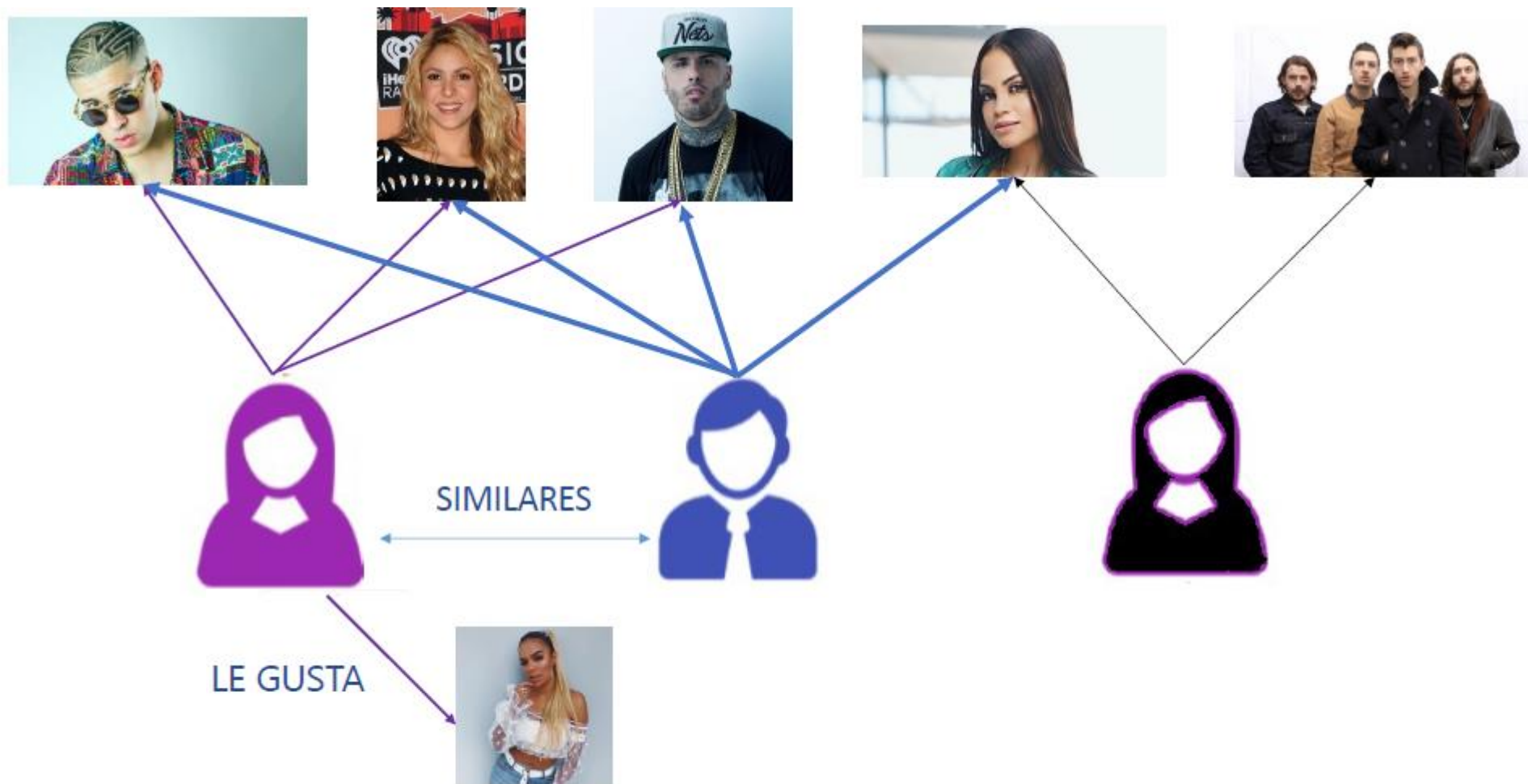


Modelos de vecinos más cercanos

Cada usuario es un vector de interacciones

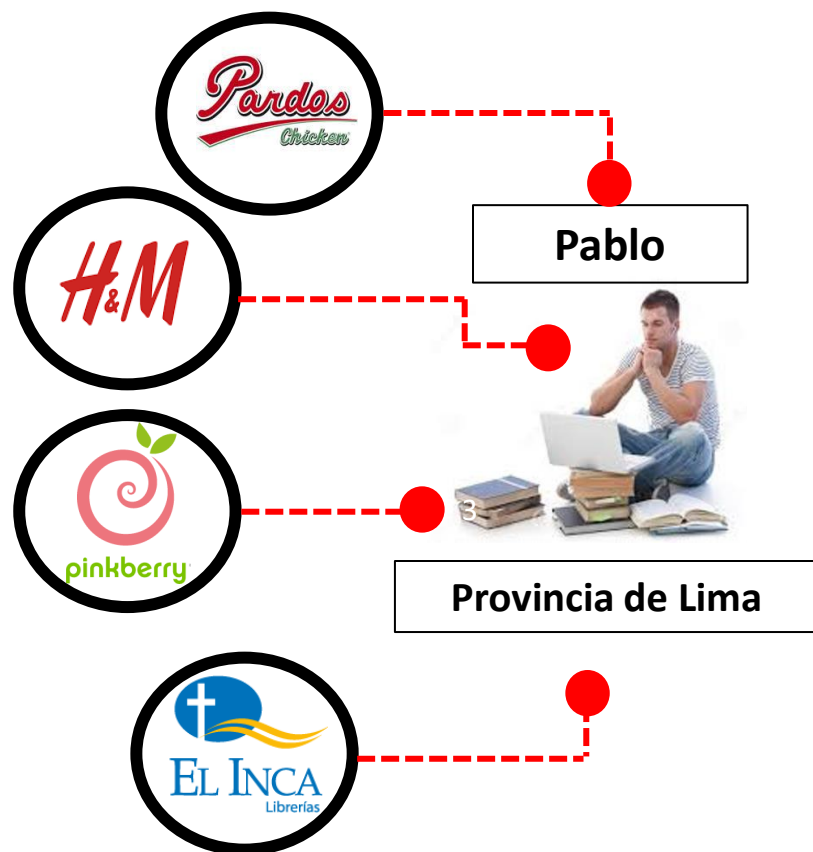
"A los usuarios parecidos a ti también le gustó ..."

"Los usuarios que compraron este producto también compraron..."

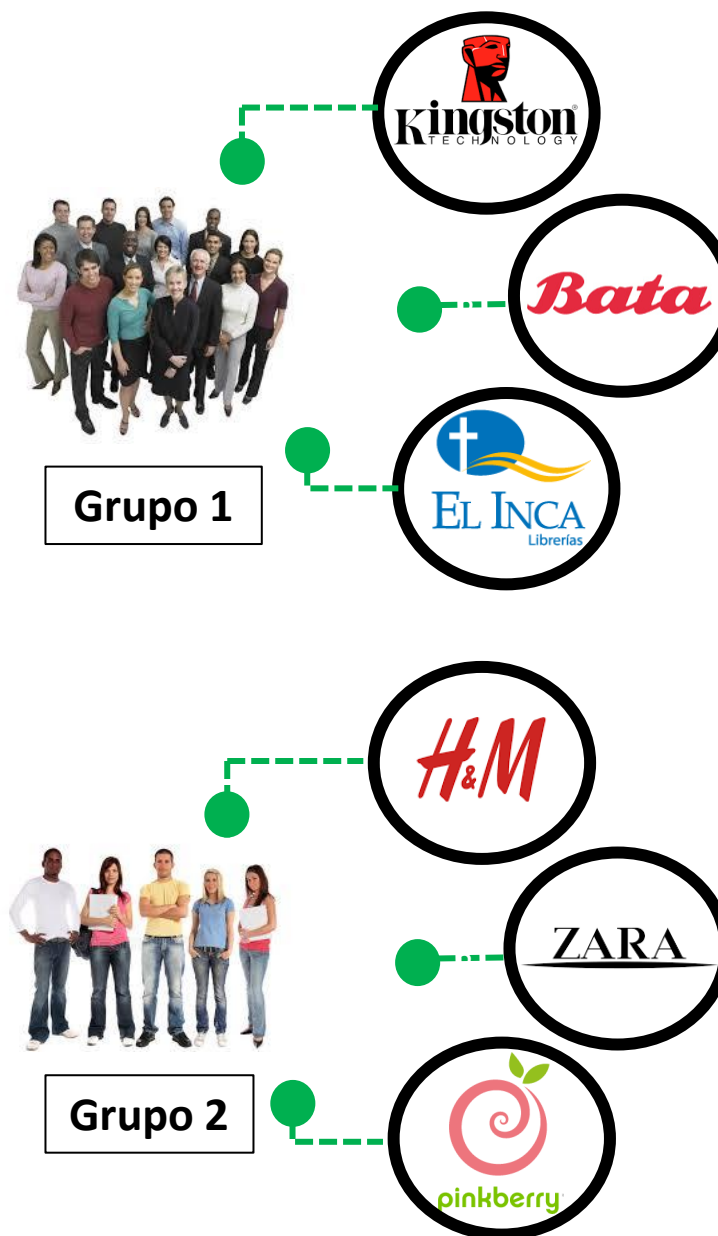


Caso práctico

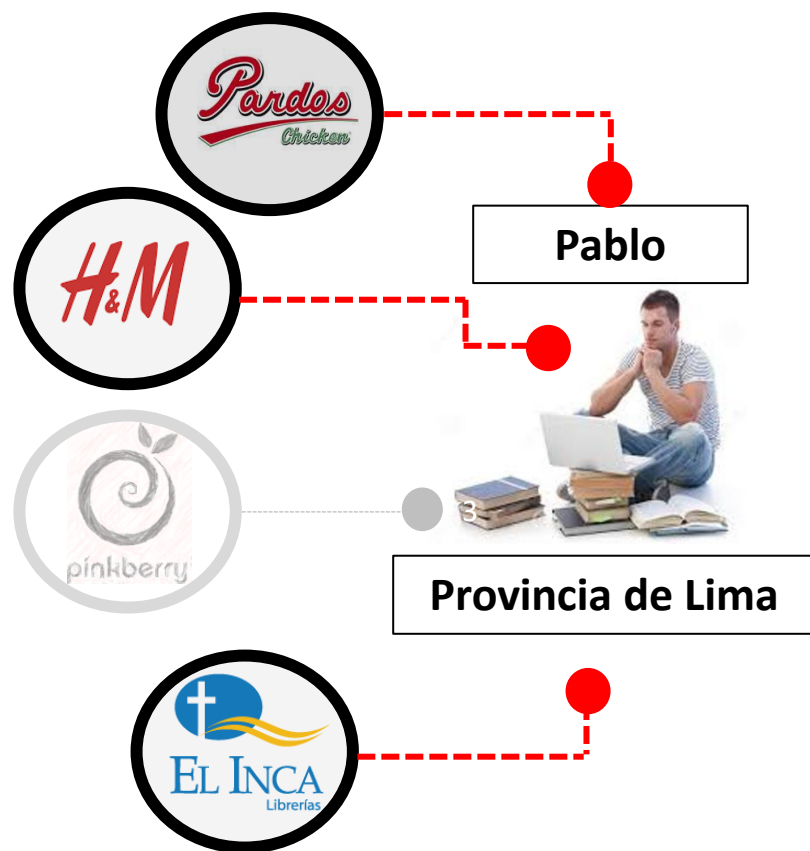
“ Consumo en diferentes
establecimiento (**comercios**) “



Los establecimientos están
mapeados georeferencialmente (X,Y)“



Caso práctico



“Existen algunos establecimientos comunes entre grupos”



Caso práctico

1

Comercios de
consumo histórico
(MBA)



Pardos
Chicken

H&M


pinkberry


EL INCA
Librerías

Pablo



Provincia de Lima

*“ En base al
Rating, se
priorizan los
comercios más
preferidos “*

2

Comercios a
recomendar
(sugerencia)




Kingston
TECHNOLOGY

Bata

ZARA

3

Rating de
recomendación
(prioridad)



Caso práctico

Cliente: Pablo

Domicilio: Lima

Comercio recomendado



Recomendado nuevo



Recomendado
histórico

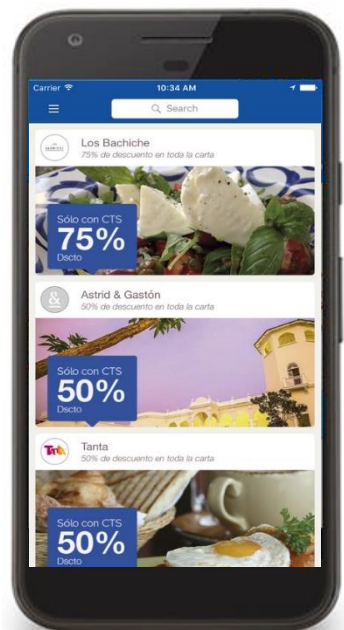
SE OBSERVA:

“ Comercios que se
le podría sugerir al
cliente, según
donde se encuentre
georreferenciado “

Caso práctico

Actualidad:

Tanto para los clientes
Marcos y Liza, reciben
las mismas ofertas.

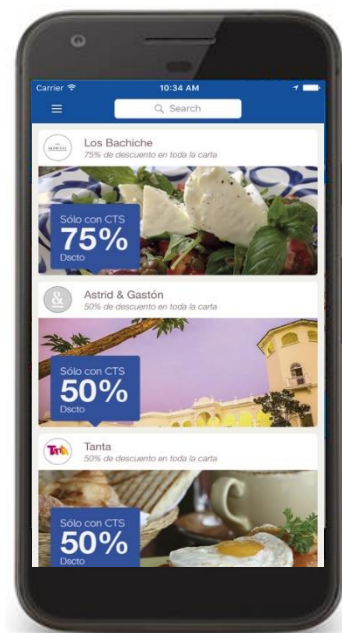


No se evalúa el perfil de
consumo de cada cliente.

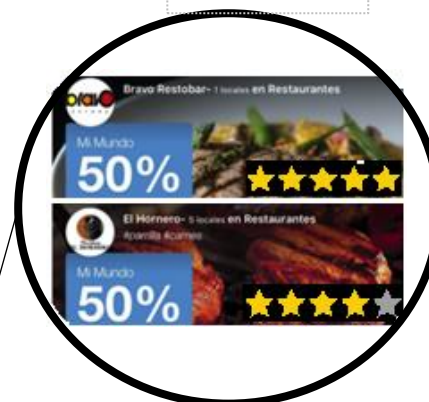
Ofertas **comunes** para todos los
clientes BBVA.

Recomendador:

En base al perfil de
consumo, se desarrolló
nuevas alternativas



Marcos



Liza



Ofertas **“personalizadas”**
recomendados por cliente.



CAPACITACIÓN
PROFESIONAL