
M.L. & I.A.
MÁS ALLÁ DE LOS PERFILES
Técnicas de monitoreo de comportamiento
CPLDFT
13
CPLDAMERICAS.COM

Disclaimer

- El taller **no** busca convertir a los participantes en científicos de datos.
- Se busca, a través de ejemplos sencillos, que los participantes puedan comprender los conceptos fundamentales de IA y ML y ver su aplicabilidad en monitoreo transaccional.
- Se busca que, un OC pueda entender cómo se aplican los modelos, entienda las salidas y pueda tener una conversación con los equipos de ML de su organización.

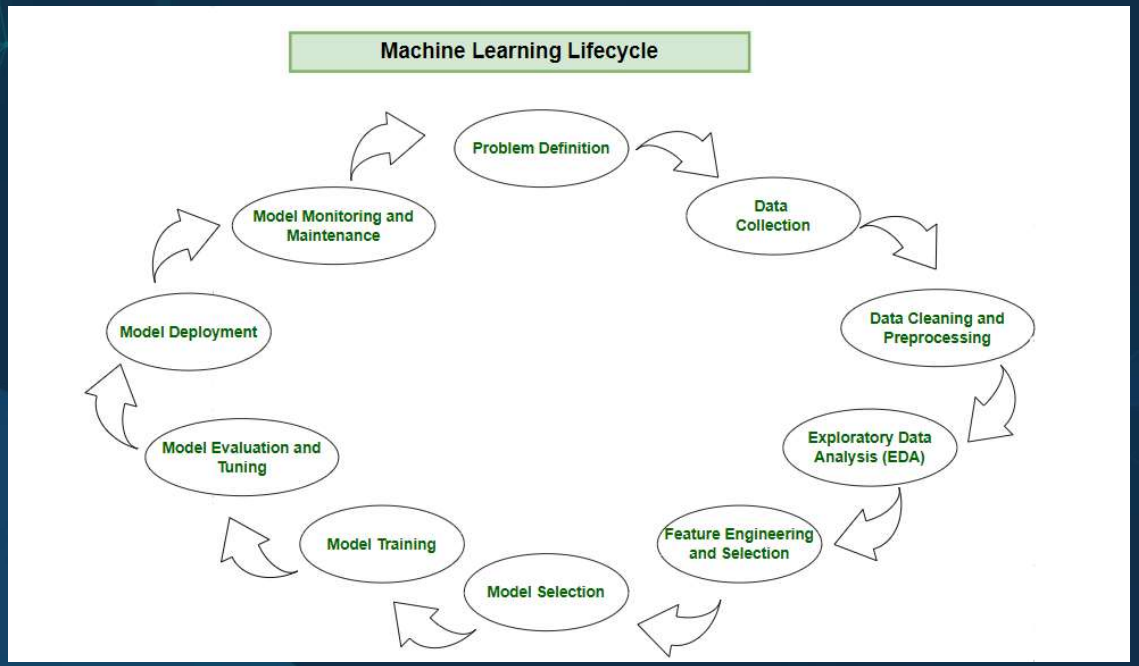
1. Conceptos de M.L.

¿Qué es machine learning y para qué sirve?

Definición técnica

Machine Learning es una rama de la **inteligencia artificial** que desarrolla **algoritmos** capaces de **aprender de los datos** para **reconocer patrones**, **identificar anomalías** y realizar **predicciones de manera automática**.

El ciclo de implementación

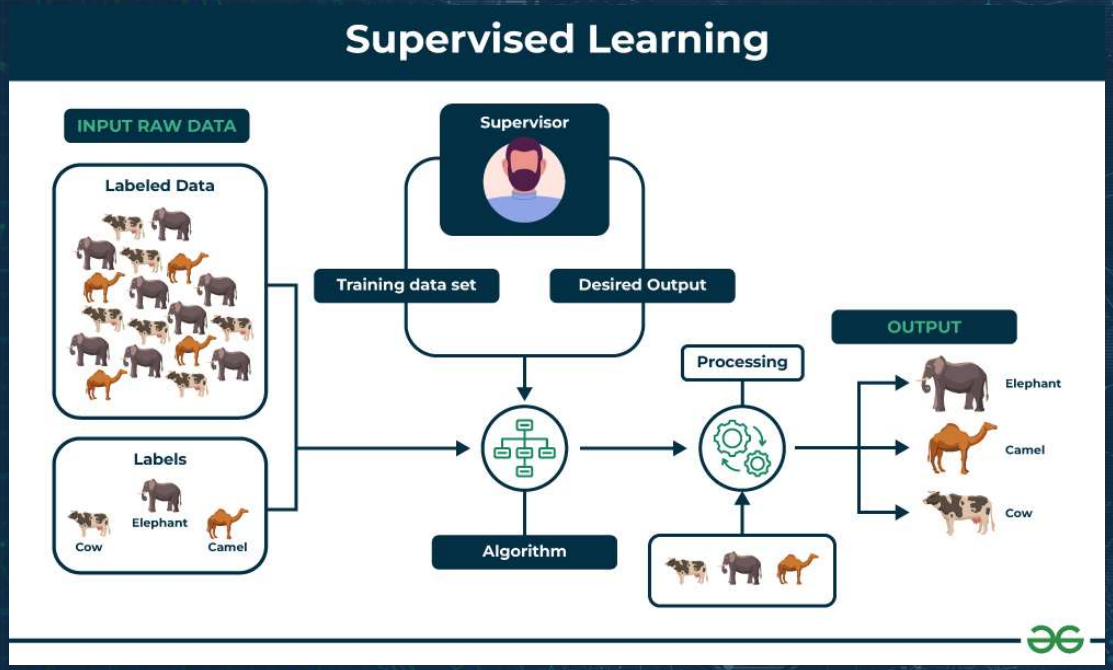


Modelos de aprendizaje automático

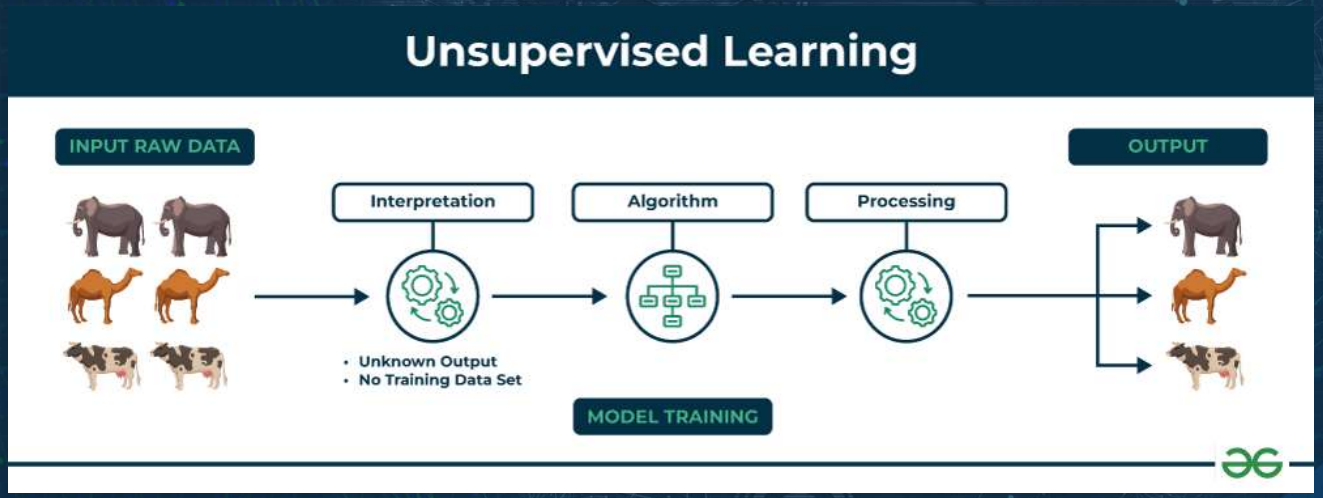
Supervisado

No Supervisado

Aprendizaje supervisado



Aprendizaje no supervisado

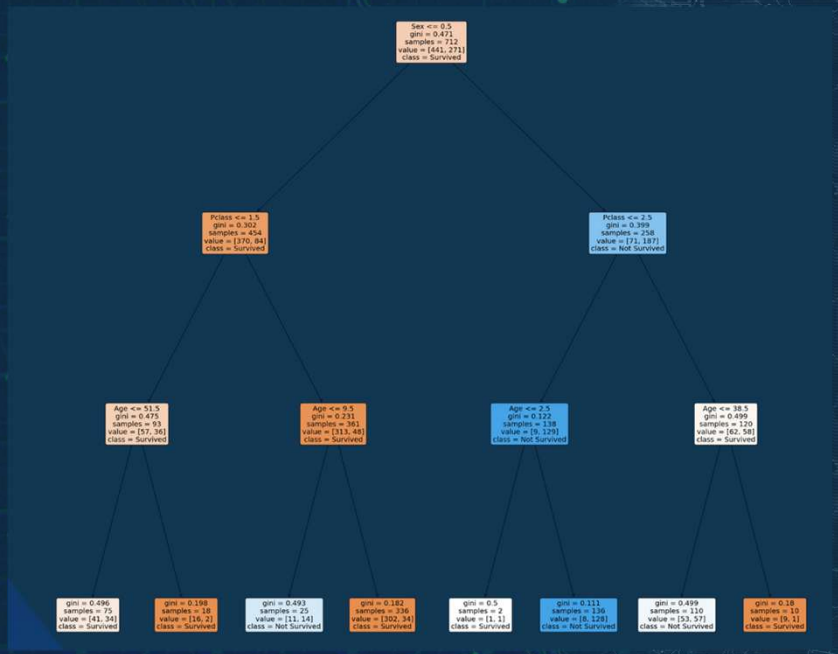


¿Qué algoritmos usamos para M.L.?

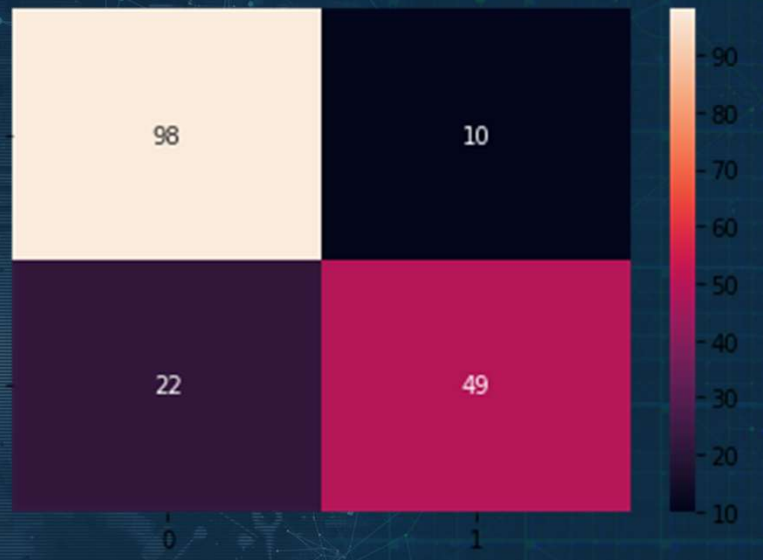
- Regresión Lineal
- Decision Tree
- Random Forest
- K-Means
- DBSCAN
- Gaussian Mixture Model

PERO NO HACE FALTA QUE SEA TAN COMPLICADO

Ejemplo: Supervivencia del Titanic



Ejemplo:
https://github.com/alicevillar/titanic-kaggle/blob/main/Titanic_DecisionTree.ipynb



b) Precision

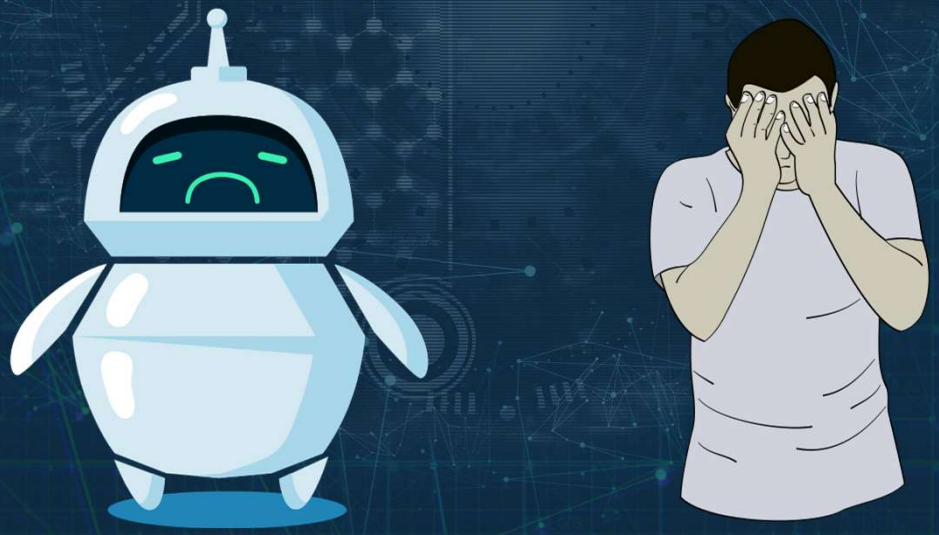
```
# Precision = true positive / true positive + false positive
precision_dt = metrics.precision_score(Y_real,Y_pred_dt)
precision_dt
```

```
0.8305084745762712
```

La pregunta es fácil

¿Por qué no creamos un modelo de M.L. con la capacidad de medir la probabilidad de que un perfil o comportamiento esten ligados a lavado de activos?

La respuesta no es igual de fácil



Limitaciones

- Base de Datos no balanceada
- El impacto de los falsos negativos
- Falta de datos para un “labeling” real
- No hay un solo patrón - Más de 30 delitos precedentes
- Los datos claves no siempre están

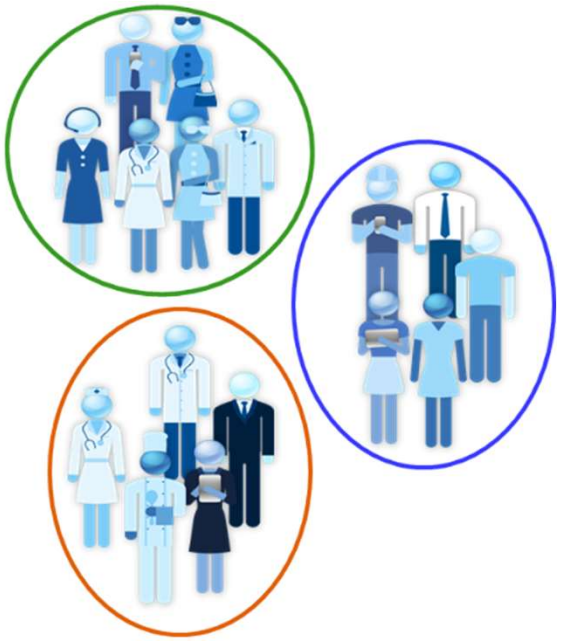
La Buena Noticia

Que no nos podamos saltar al final no significa que no nos pueda ayudar durante el proceso

¿De qué manera se podría aplicar M.L. en el proceso de determinación de perfiles y monitoreo de comportamiento?

2. Uso de cluster como herramienta de monitoreo

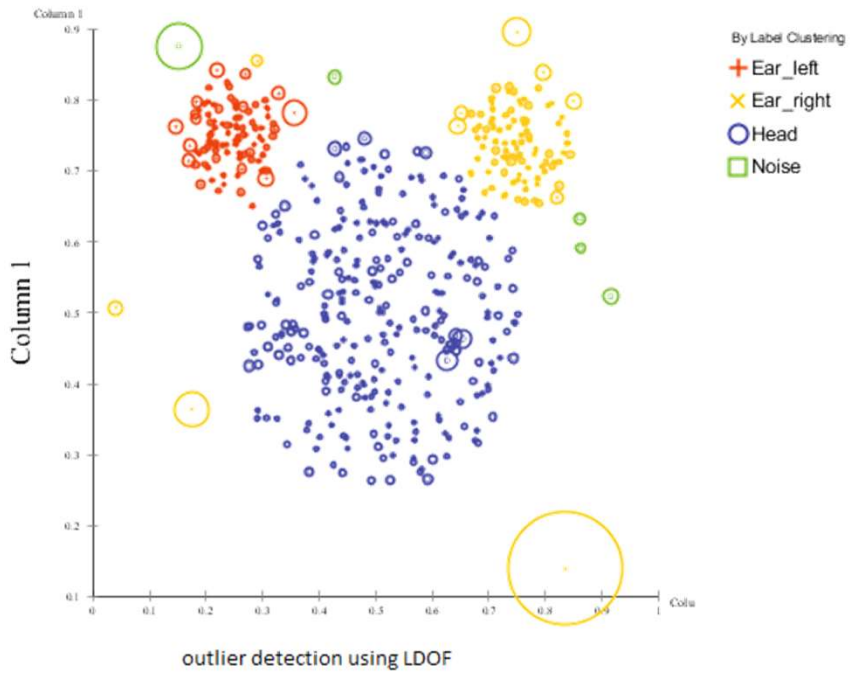
Qué es un cluster?



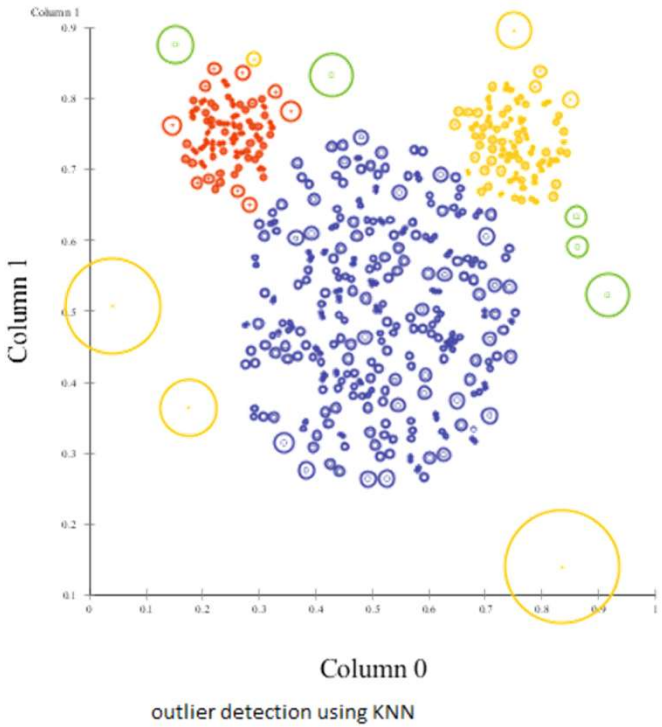
- Técnica que busca encontrar grupos (clusters) de puntos de datos de tal forma que los puntos de un mismo grupo sean **similares** (o parecidos) entre si y al mismo tiempo, **diferentes** (o no relacionados) con puntos de datos en otros grupos.

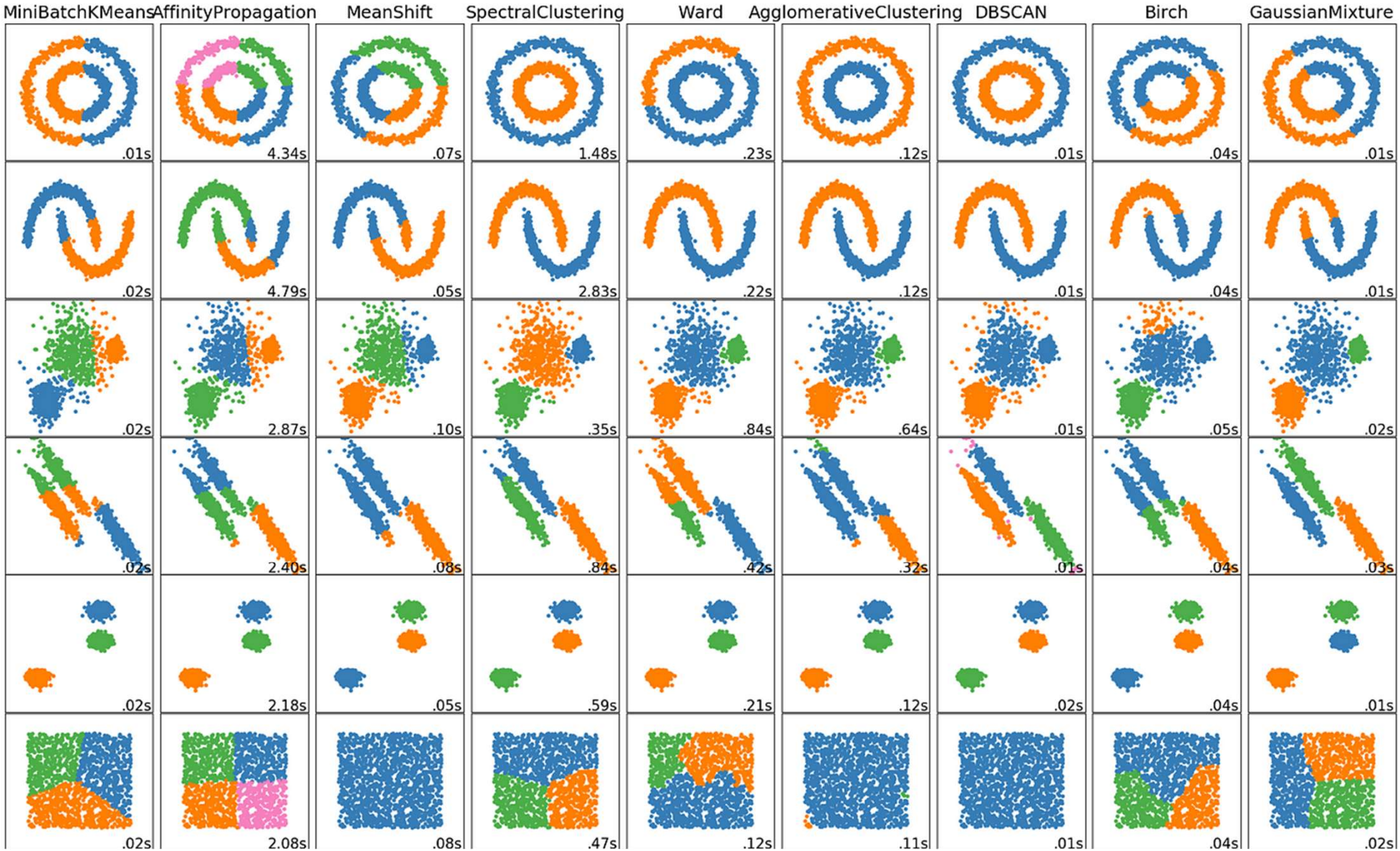
Otros usos de cluster

Detección de anomalias



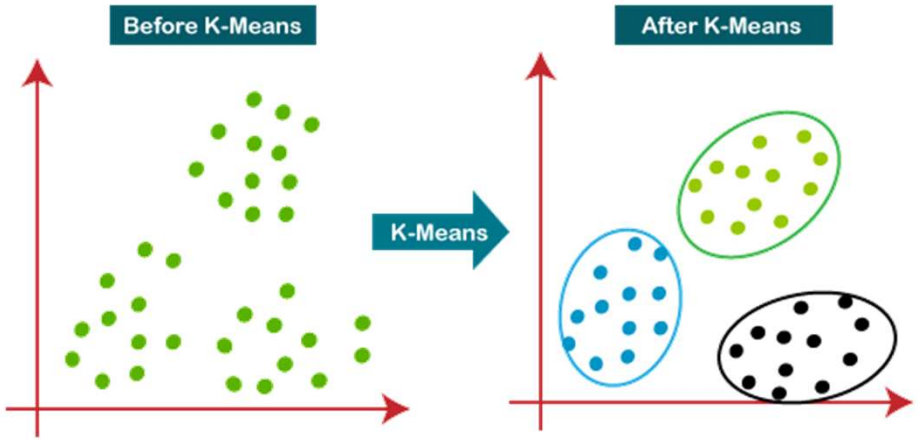
EDA



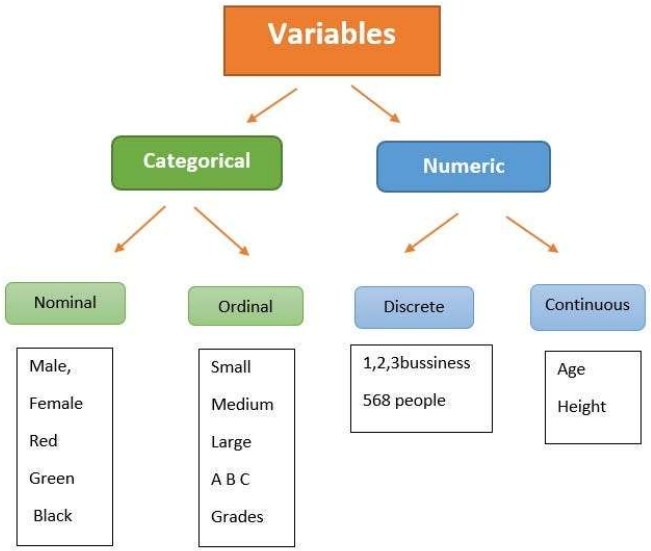

MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s
4.34s
.07s
1.48s
.23s
.12s
.01s
.04s
.01s
.02s
4.79s
.05s
2.83s
.22s
.12s
.01s
.04s
.01s
.02s
2.87s
.10s
.35s
.84s
.64s
.01s
.05s
.02s
.02s
2.40s
.08s
.84s
.42s
.32s
.01s
.04s
.03s
.02s
2.18s
.05s
.59s
.21s
.12s
.02s
.04s
.01s
.02s
2.08s
.08s
.47s
.12s
.11s
.01s
.04s
.02s

Kmeans cluster - Requisitos

Grupos hiperesféricos



Variables numéricas



Ejercicio

- Tenemos la siguiente data de KYC y deseamos disenar un modelo de cluster para segmentar clientes desde la perspectiva de riesgo AML. Cuáles variables podríamos utilizar?
    - Nombre
    - Direccion
    - Teléfono
    - Tipo (Natural, Jurídico)
    - Nacionalidad
    - Actividad Ecónomica
    - Ocupación
    - Nivel de estudios
    - Productos / servicios
    - Canales que utiliza
    - Vol Txs money in
    - Q Txs money in
    - Vol Txs money out
    - Q Txs money out

Ejercicio

- Más variables no siempre es mejor. Cuáles podriamos “agrupar”?

- Nombre
- Direccion
- Teléfono
- Tipo (Natural, Jurídico)
- Nacionalidad
- Actividad Ecónomica
- Ocupación
- Nivel de estudios

- Productos / servicios
- Canales que utiliza
- Vol Txs money in
- Q Txs money in
- Vol Txs money out
- Q Txs money out

Kmeans: Transformacion de datos

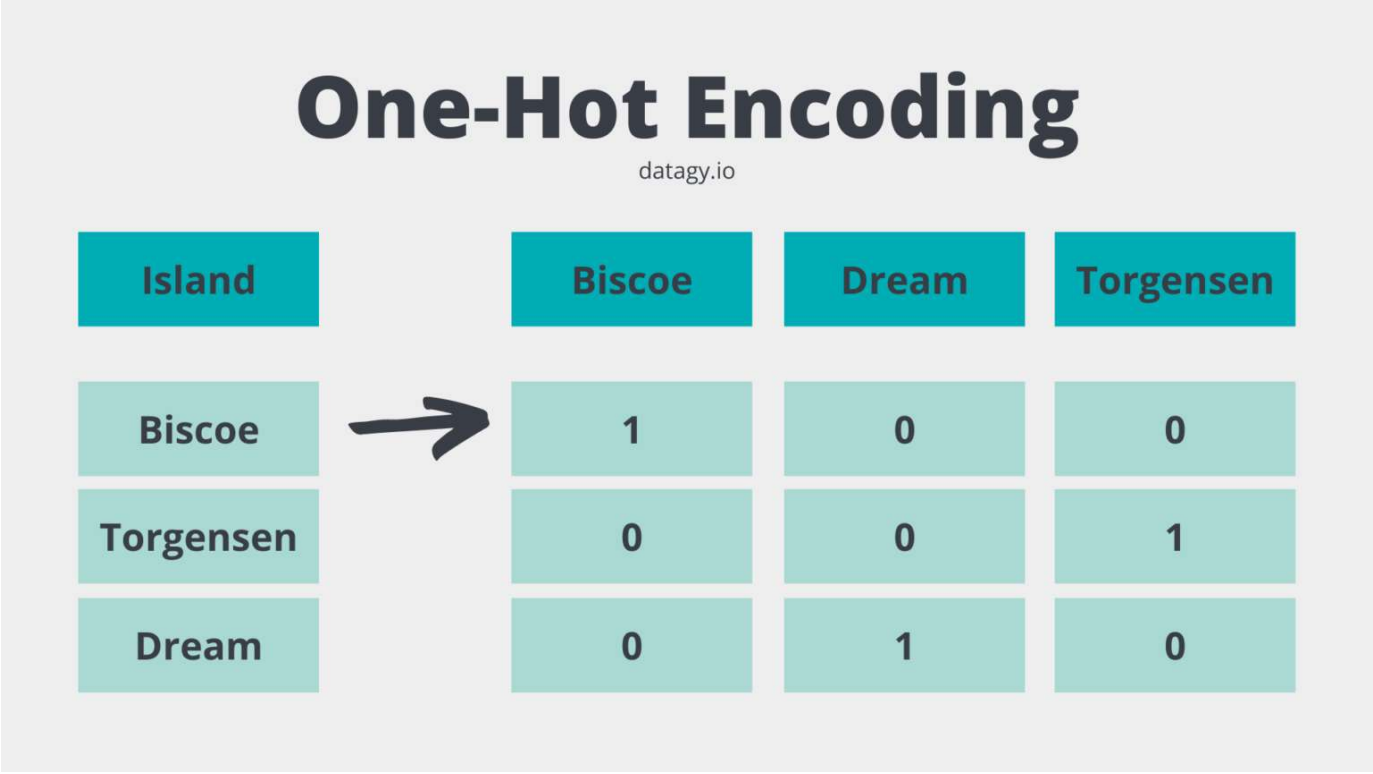
Normalización

Z Score Formula

$$z = \frac{x - \mu}{\sigma}$$

Kmeans: Transformacion de datos

Transformacion de variables categoricas



Ejercicio Transformacion de datos

Ejercicio Transformacion de datos

- Se tiene la siguiente base de datos, cómo se podría transformar?

IdCliente	NR LAFT	Nacional	PEP	Vol Txs money in	Q Txs money in	Vol Txs money out	Q Txs money out
1	Alto	Nacional	Si	12000	20	11500	14
2	Medio	Nacional	No	132000	67	113450	56
3	Medio	Extranjero	No	6400	45	2400	32
4	Bajo	Nacional	No	500	28	500	16
5	Alto	Nacional	No	276000	3	267000	1

Kmedias: Ventajas y desventajas

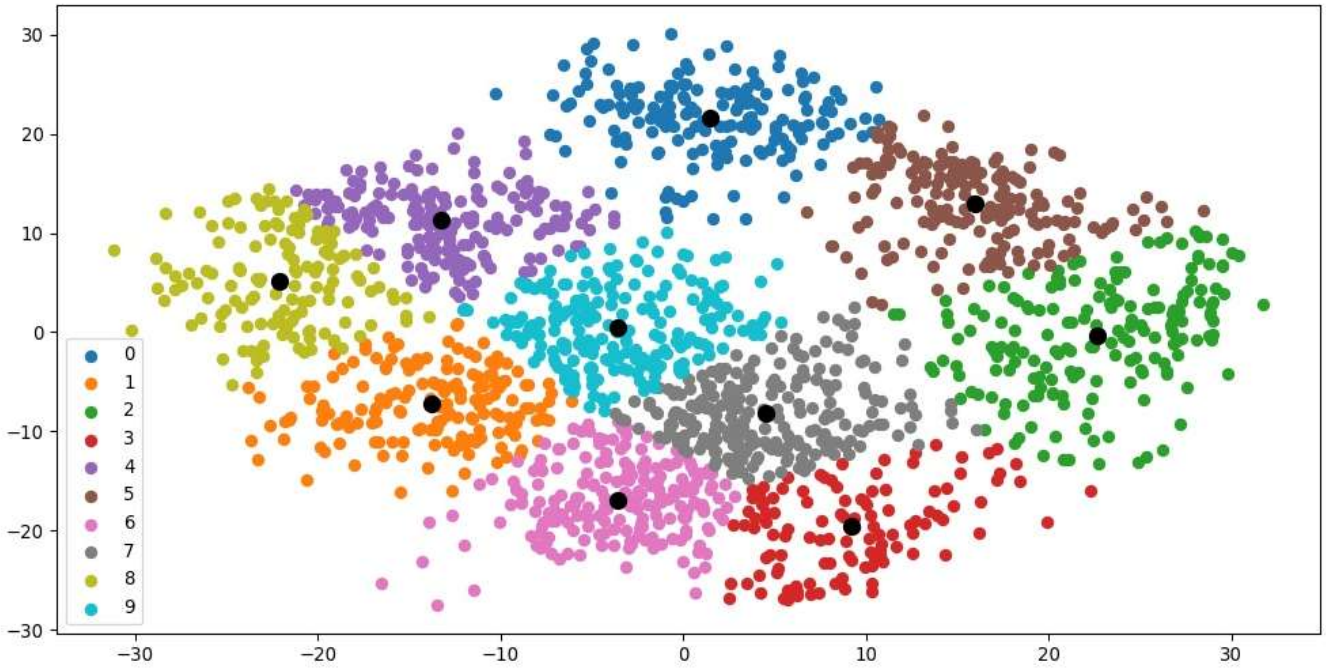
Ventajas

- Simple y fácil de entender
- Rápido y escalable (útil en datasets grandes)
- Puede manejar diferentes tipos de datos*
- Puede ser utilizado en una amplia gama de aplicaciones

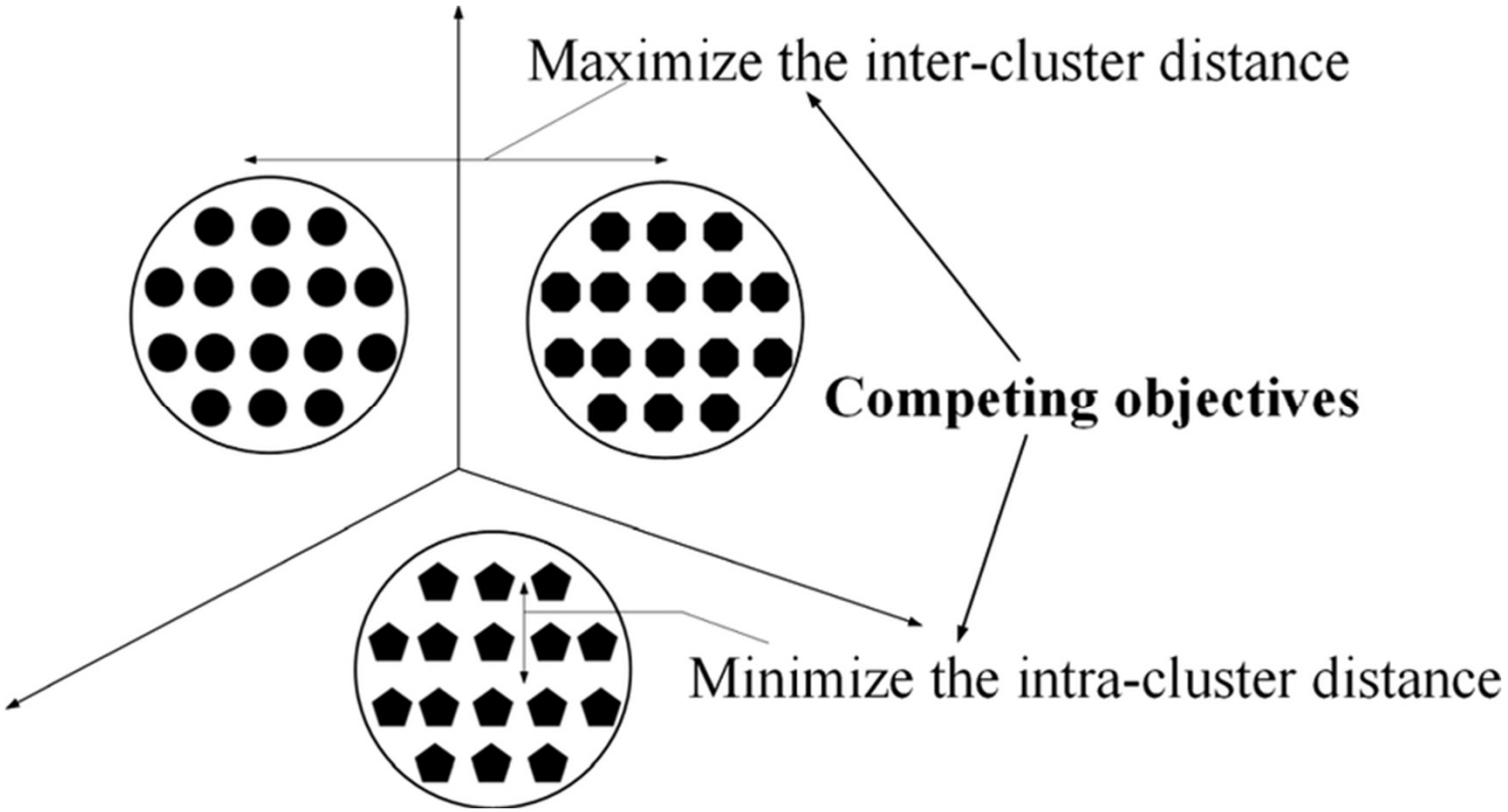
Limitantes

- Requiere que se especifique el numero de segmentos con antelación.
- Sensible a la inicialización de centroides
- Problemas con data con estructura compleja
- Puede verse afectado por outliers

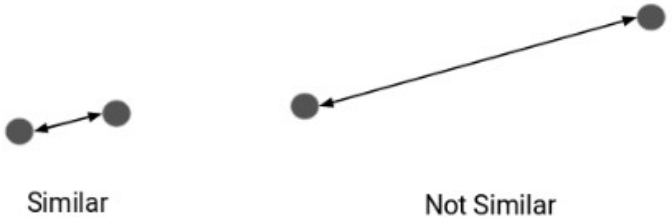
Kmedias y centroides



Cómo se mide la similaridad?



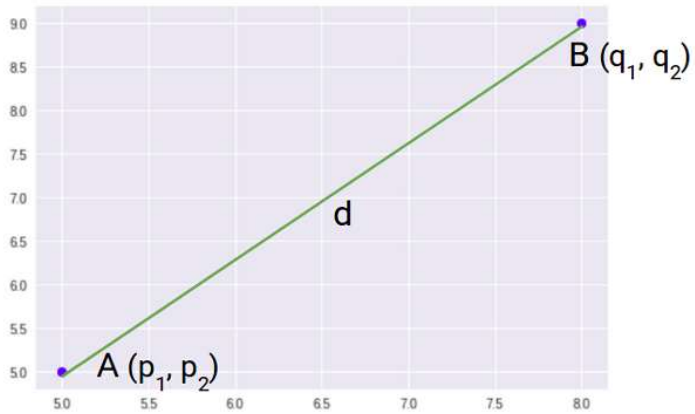
Similaridad (homogeneidad)



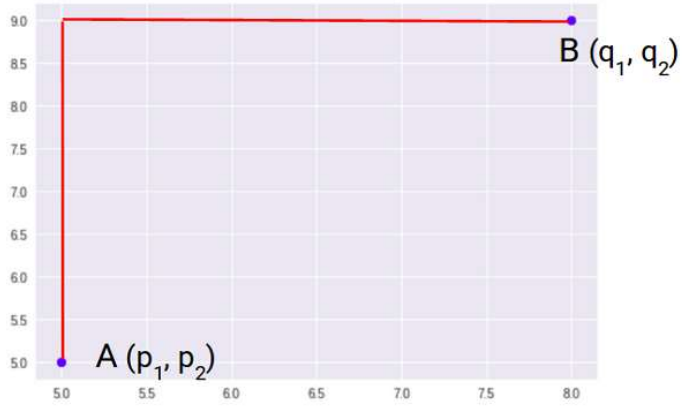
Minkowski

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

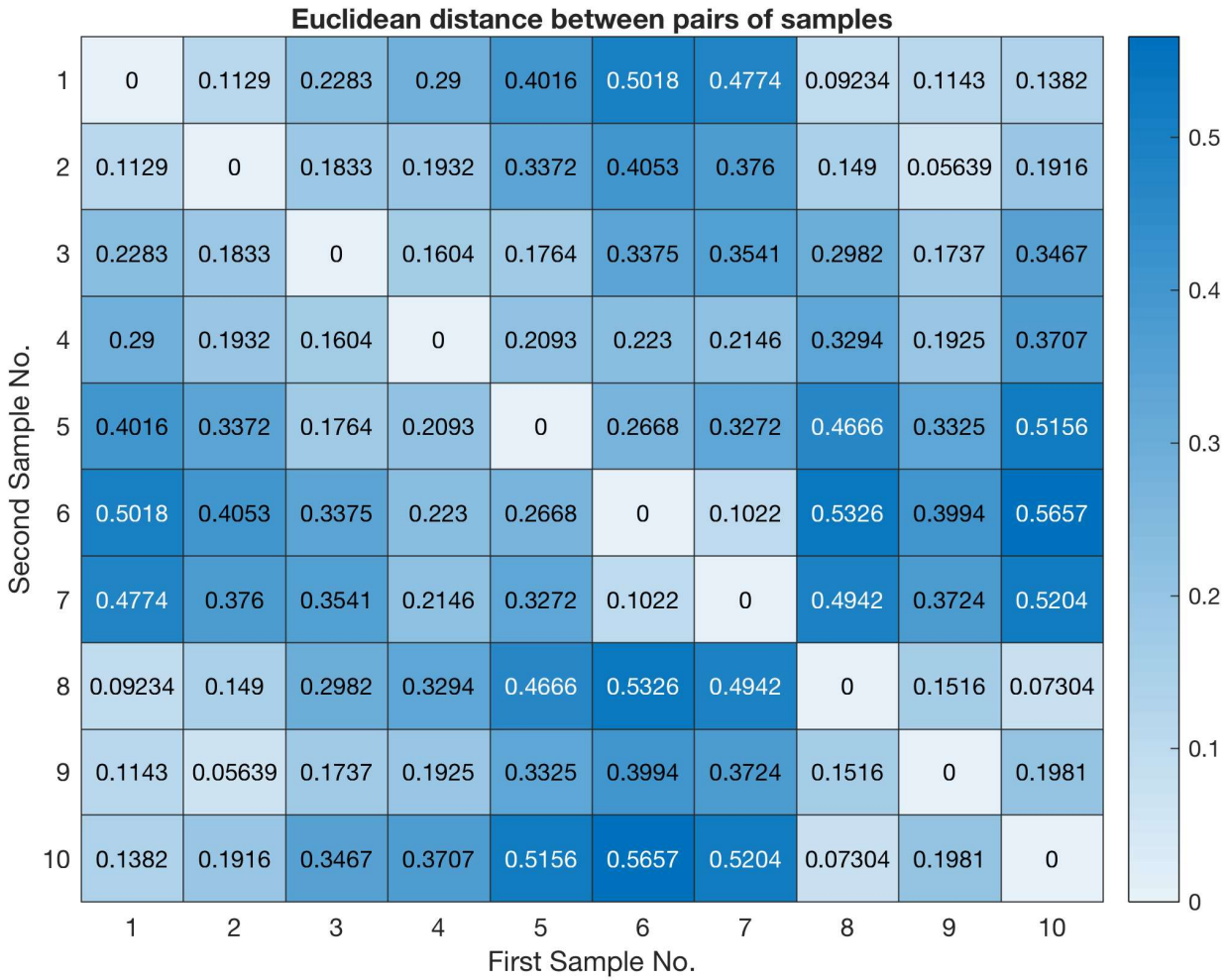
Euclidiana



Manhattan



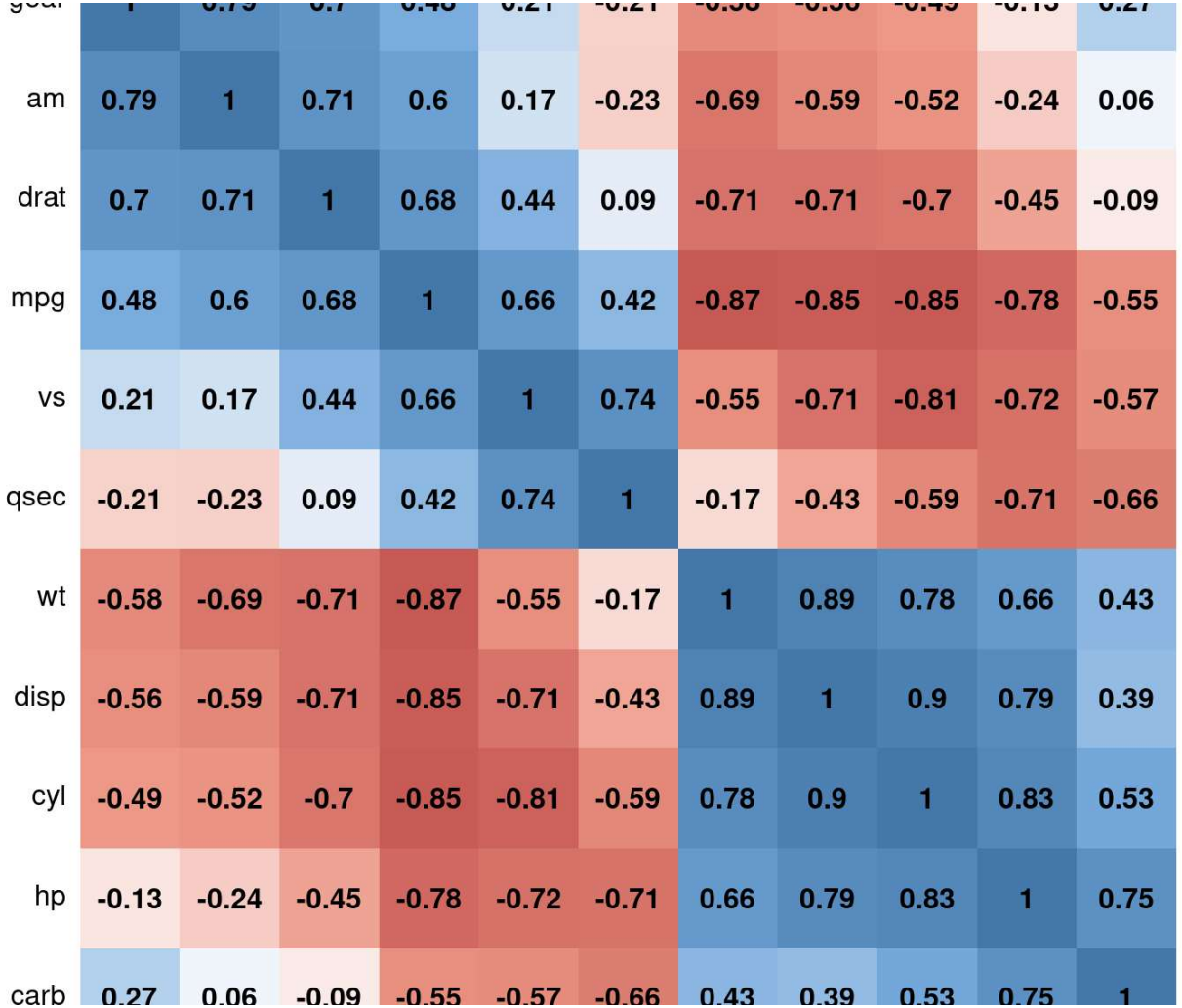
Matriz de distancias



3. Análisis preliminar

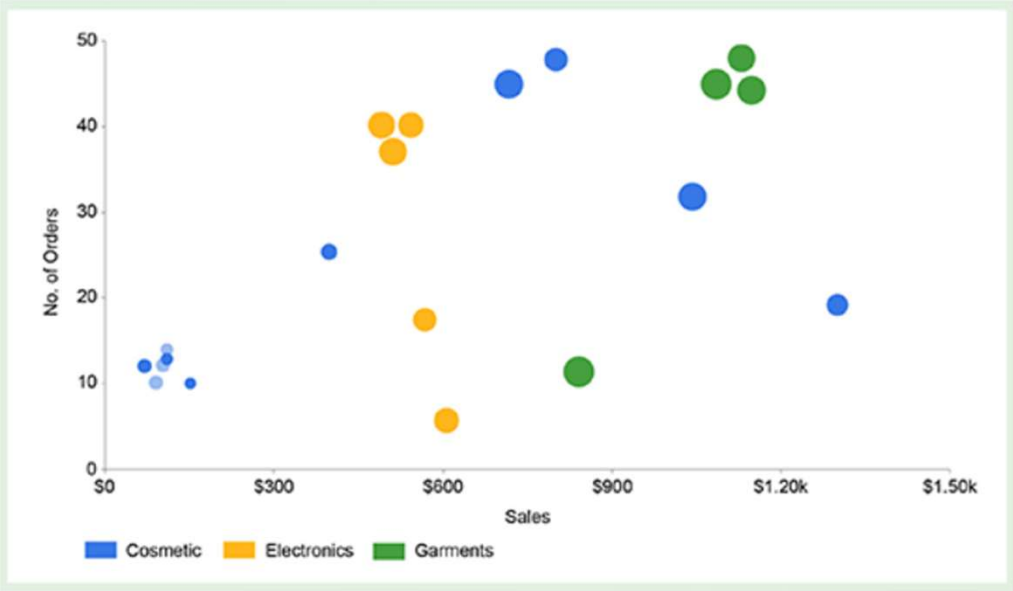
Matriz de correlaciones

Exploratory Data Analysis

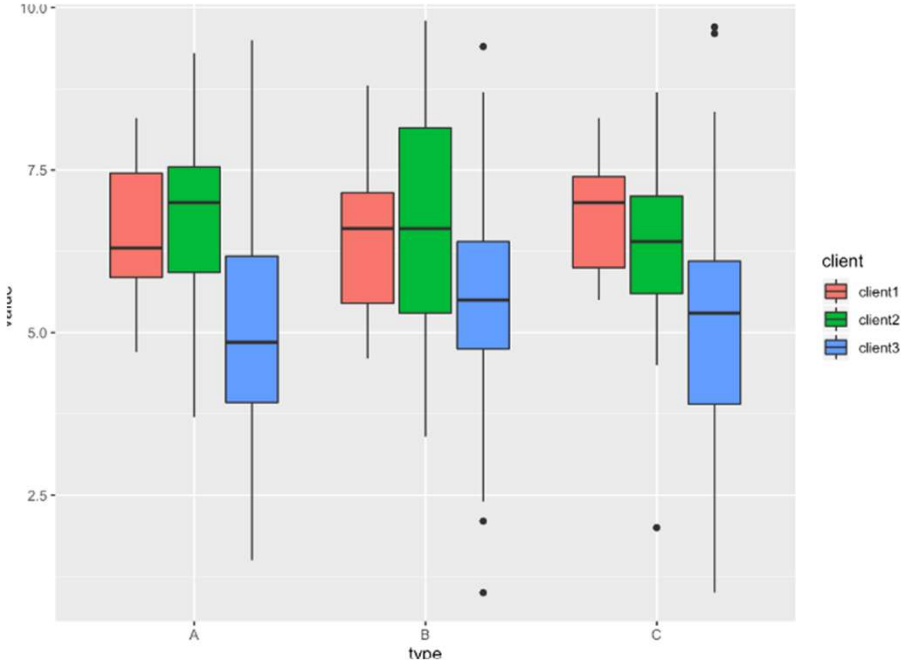


Técnicas de reducción de dimensiones

Qué pasa cuando se tienen más de 3 variables?
Cómo visualizarlas en conjunto?



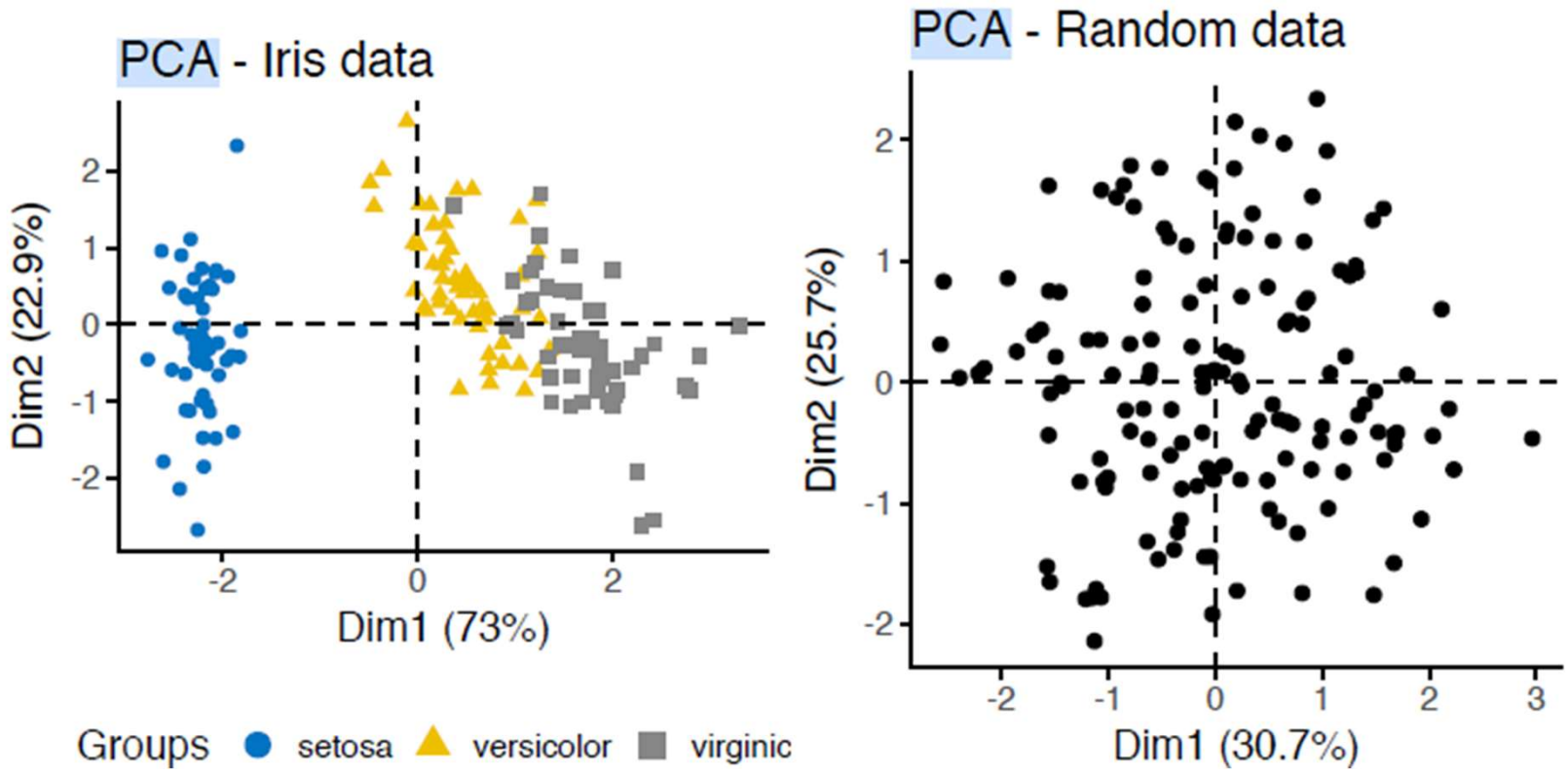
How to Create a Scatter Plot in Excel With 3 Variables?



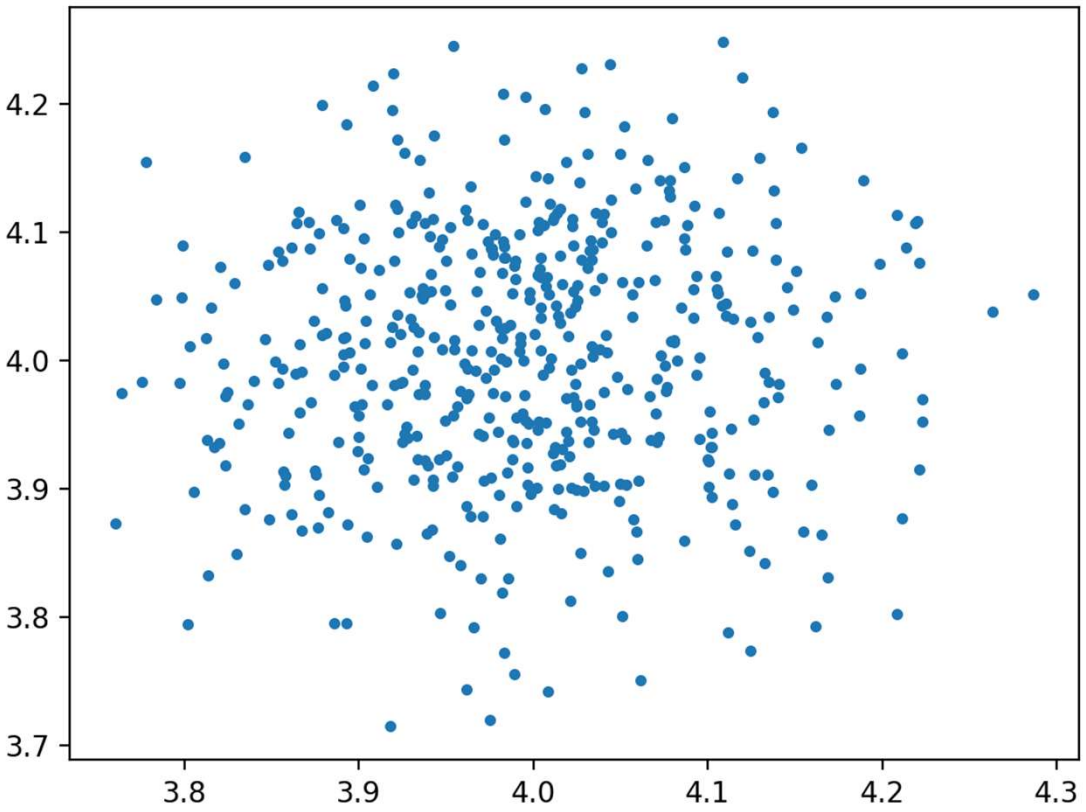
PCA

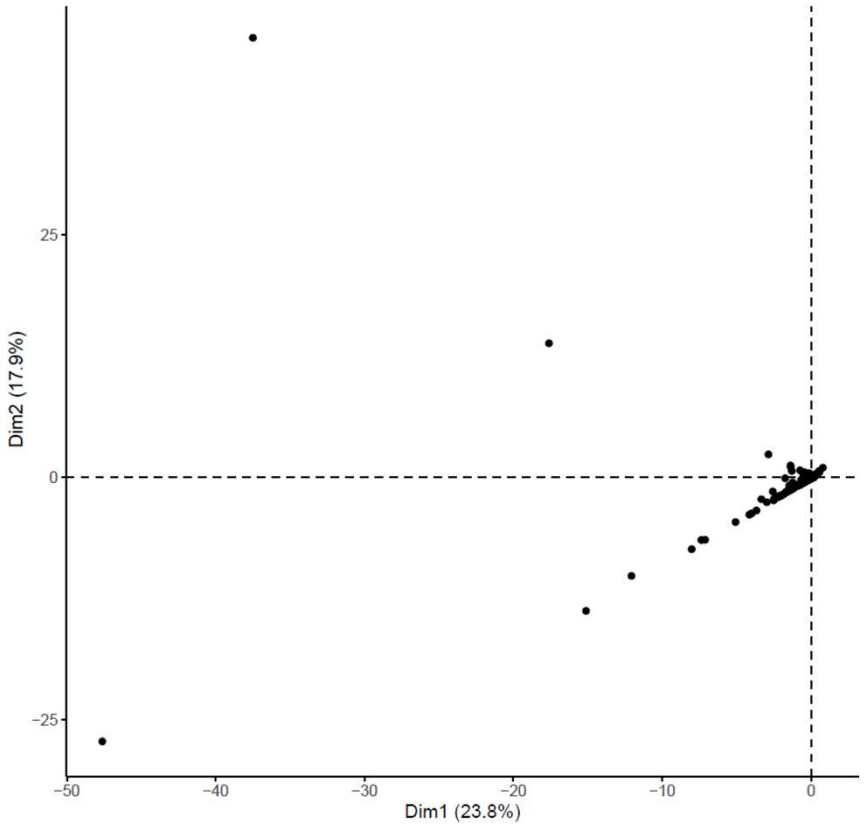
- Identifica “direcciones” que maximizan varianza
- Sobre ejes ortogonales entre si (correlacion 0)
- Las “direcciones” o nuevas variables, son una combinación lineal de las variables originales
- Retorna la “mejor reconstrucción”, es decir, minimiza la Perdida de informacion al moverse de N (dimensiones PCA) a M (dimensiones originales)
- Los valores propios son monotonicamente decrecientes (se reducen en cada nueva dimension) y representan el nivel de varianza que acumulan.
- Bien estudiado → algoritmos rapidos

PCA: Inspección visual de los datos para determinar “clustereabilidad”



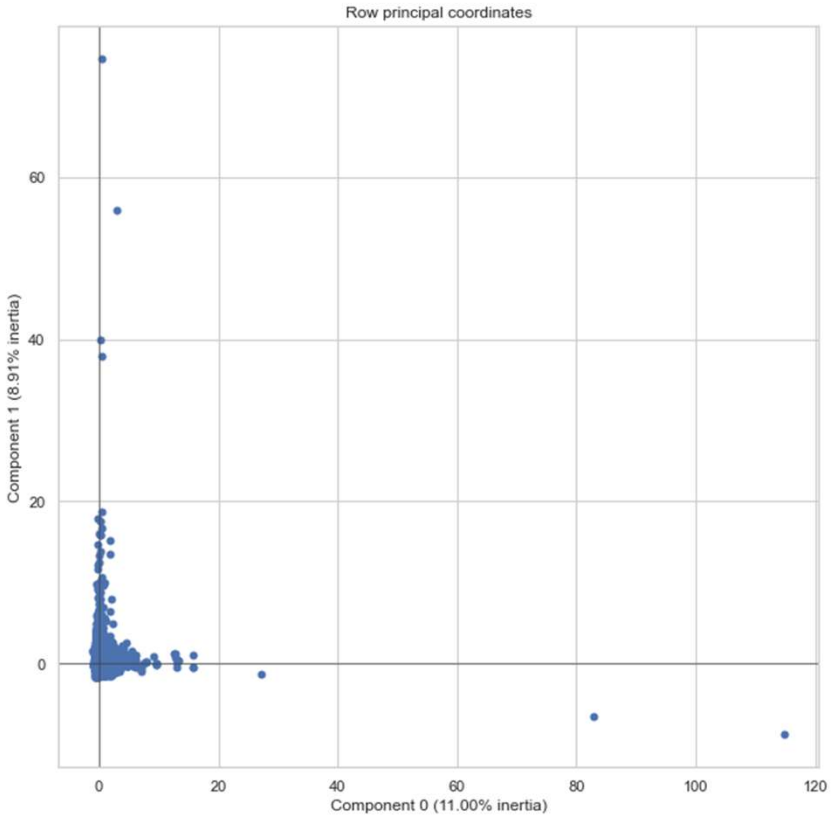
PCA: Ejercicio – Se podría aplicar cluster?

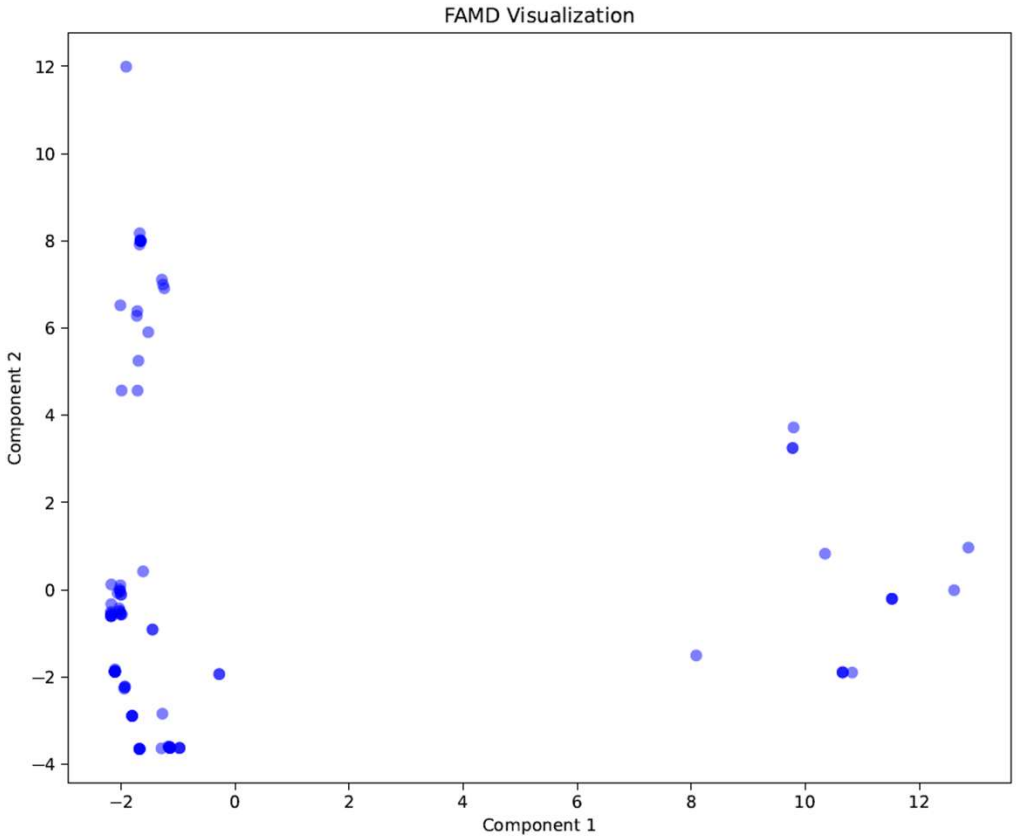




PCA es la unica técnica de reduccion de dimensiones?

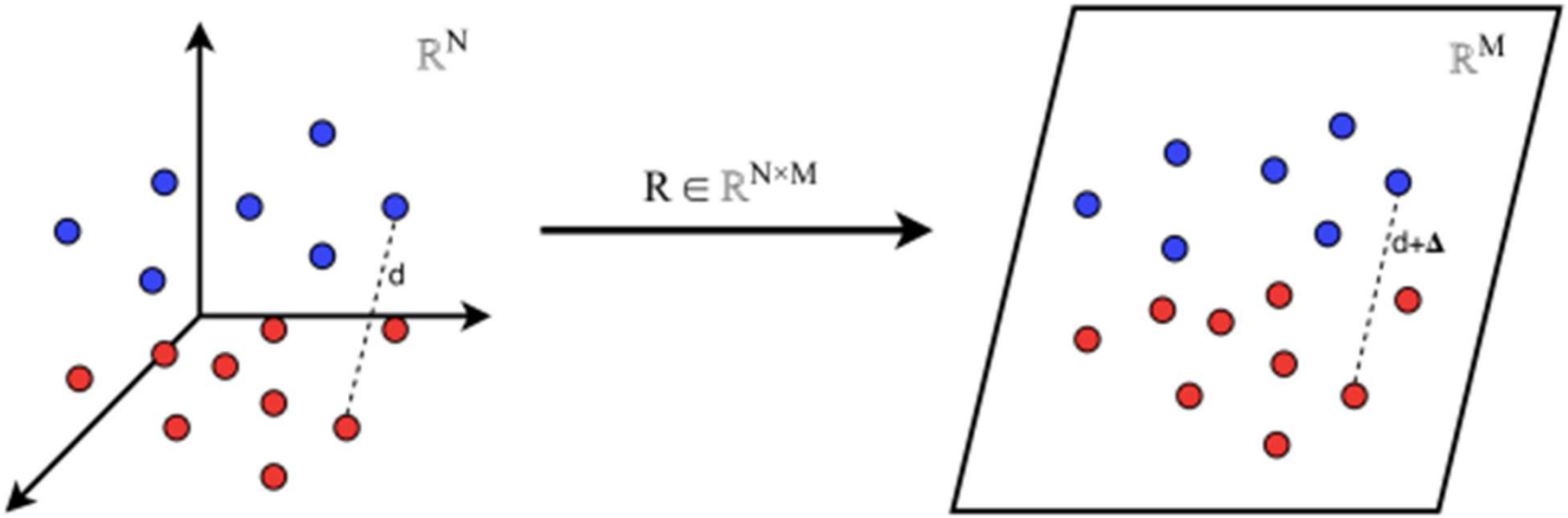
UMAP





Porqué no montar el modelos de cluster directamente en los componentes principales?

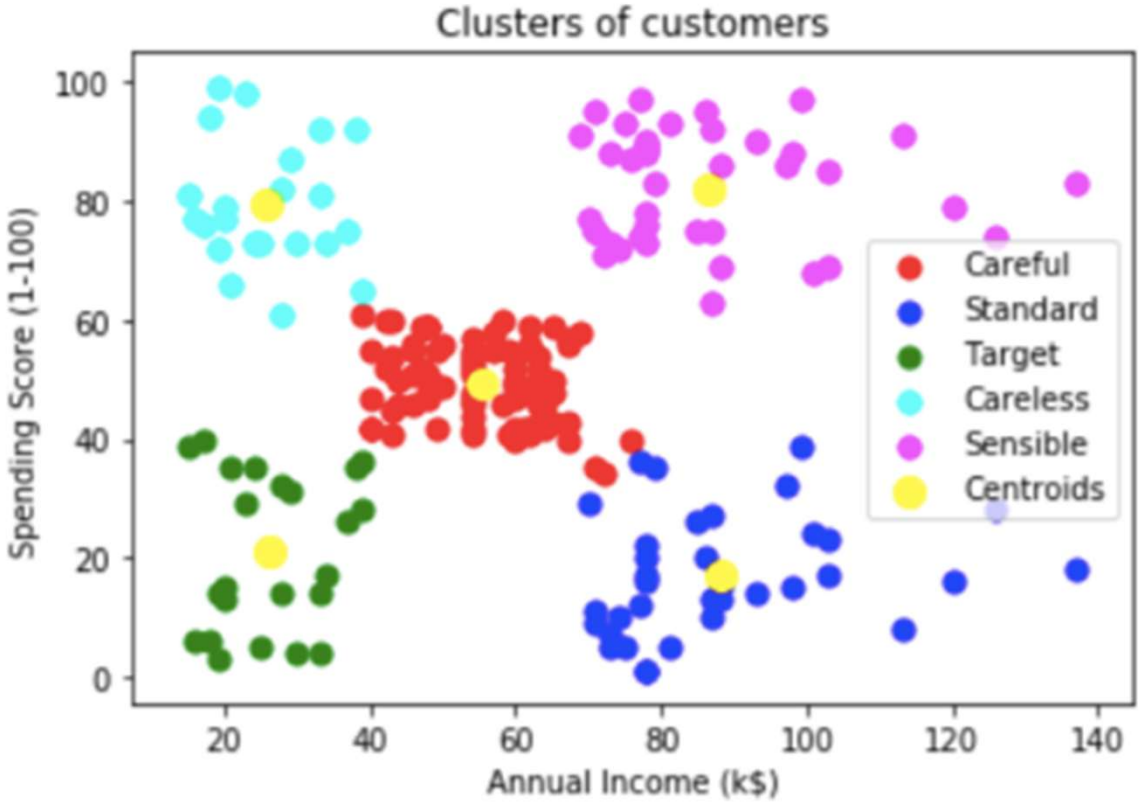
- Interpretabilidad
- Expectativa del regulador
- “Perdida de informacion”

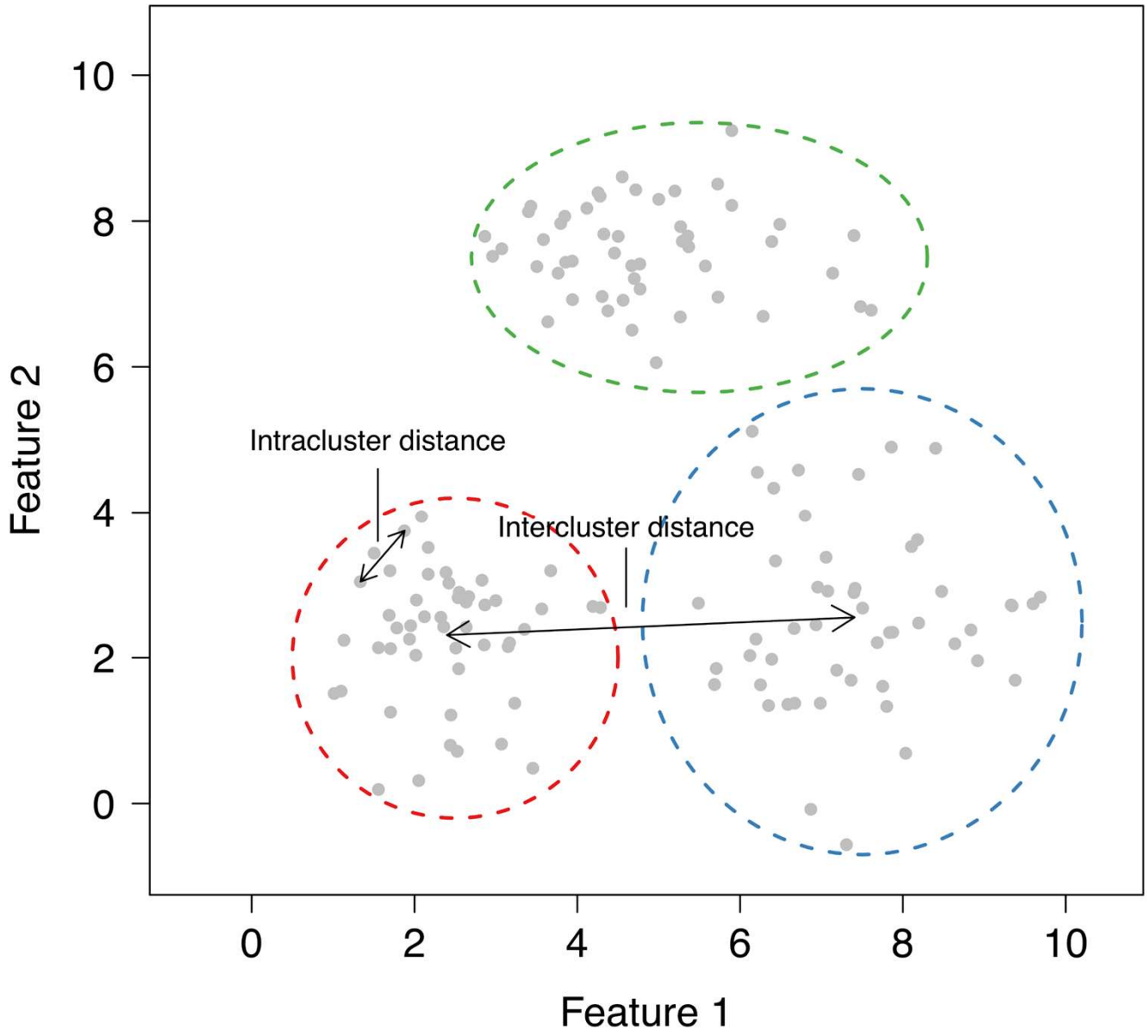


4. Selección del número óptimo de segmentos

Porque seleccionar un número de segmentos?

- El parámetro k es un *input* del modelo
- De su elección dependen aspectos críticos, p.e., definición de las senales de alerta.




Intracluster distance
Intercluster distance
Feature 2
Feature 1
0
2
4
6
8
10

4.a El criterio del codo (Elbow criterion)

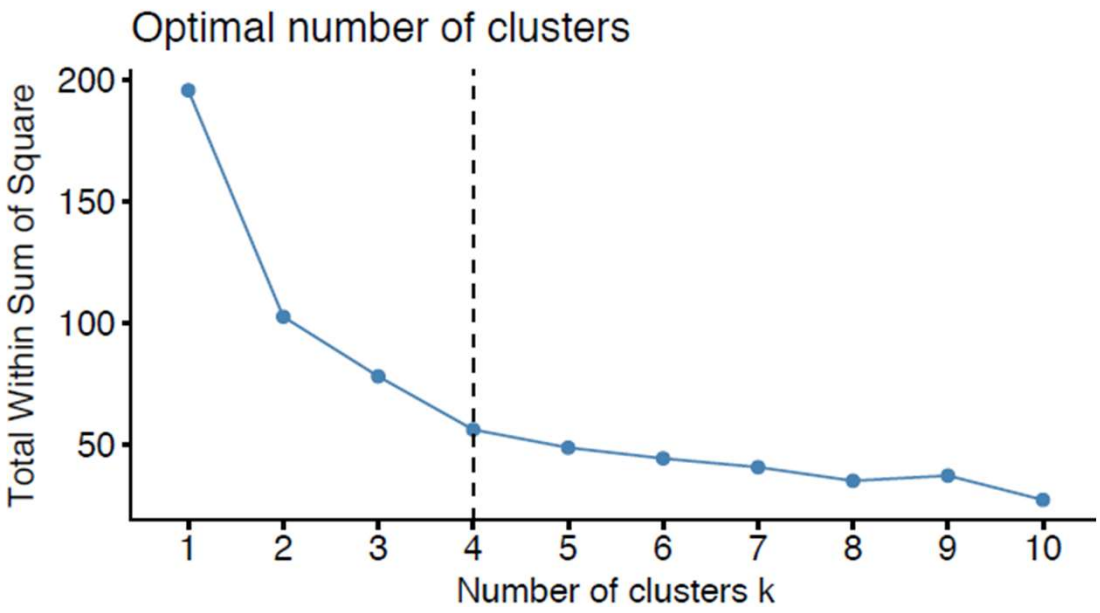
Criterio del codo

Importancia del centroide

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

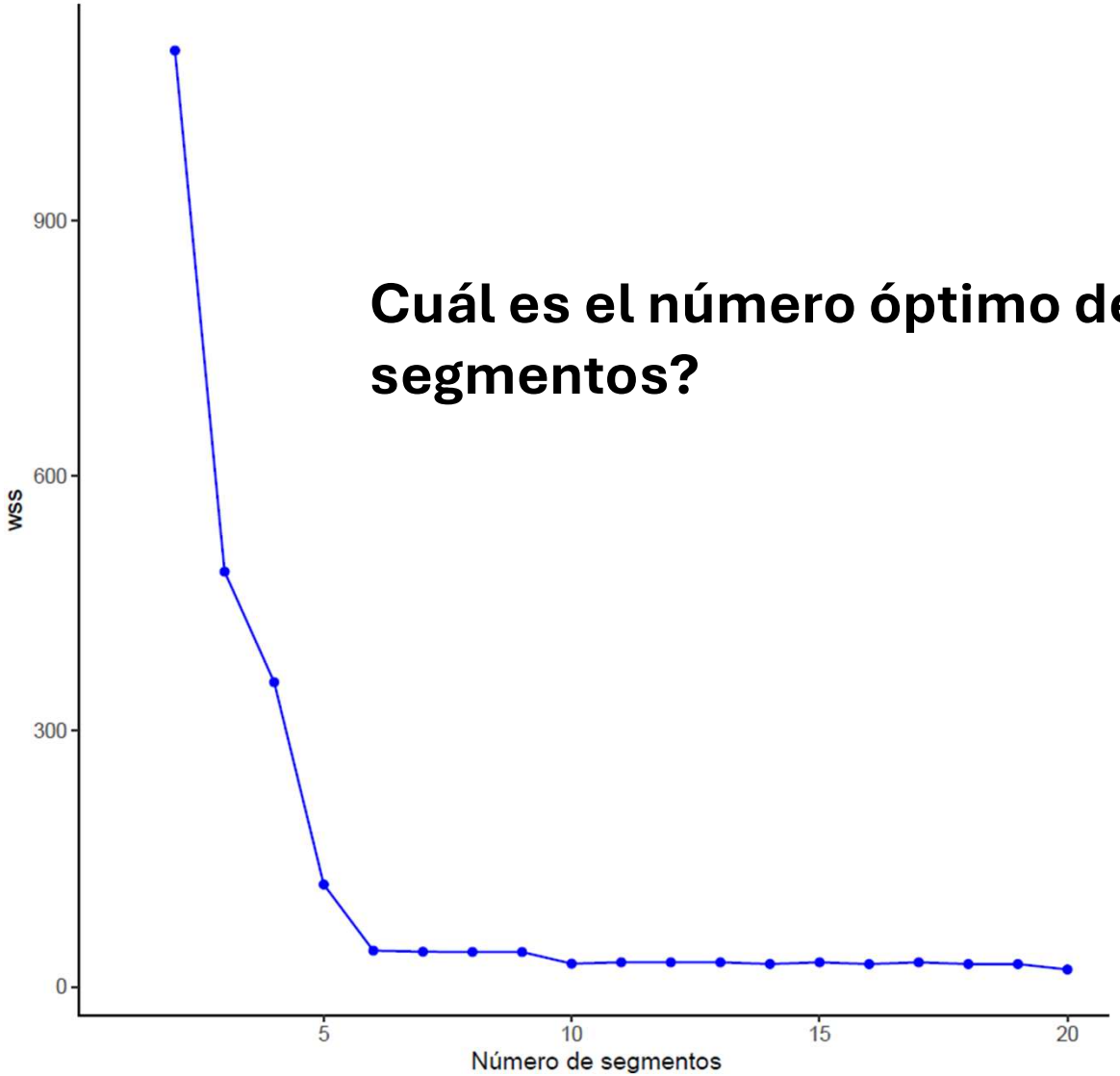
x_i design a data point belonging to the cluster C_k
μ_k is the mean value of the points assigned to the cluster C_k

$$ot.withinss = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$



Taller

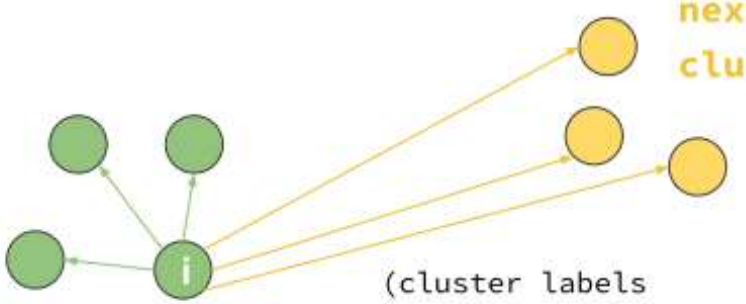
Criterio del codo



Cuál es el número óptimo de segmentos?

4.b Otros métodos

b(i): distance between i and its **next nearest cluster centroid**



(cluster labels assigned from ground-truth classes)

a(i): average distance between i and all of the other points in **its own cluster**

For a single point, i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$-1 \leq s(i) \leq 1$$

Sprawling, overlapped clusters

Tight, well-separated clusters

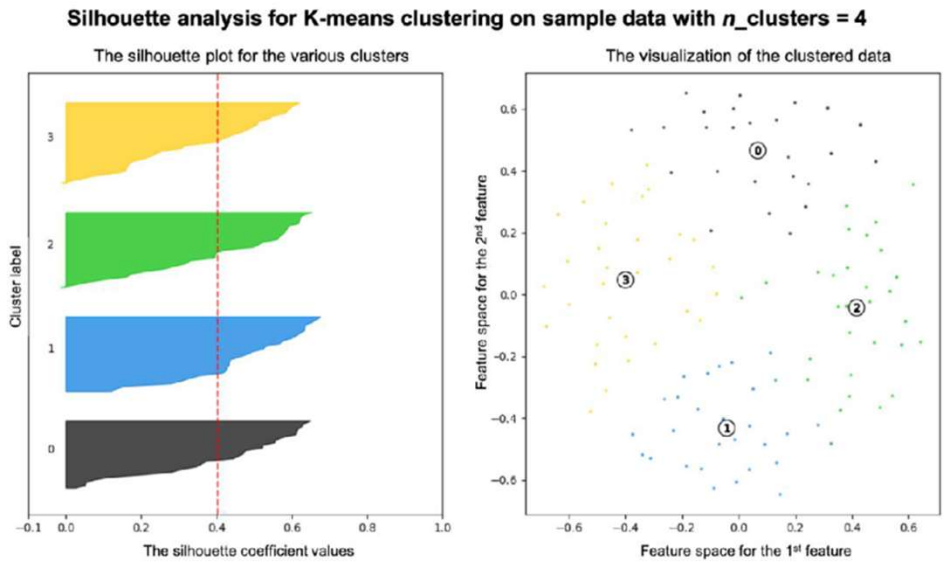
Silueta

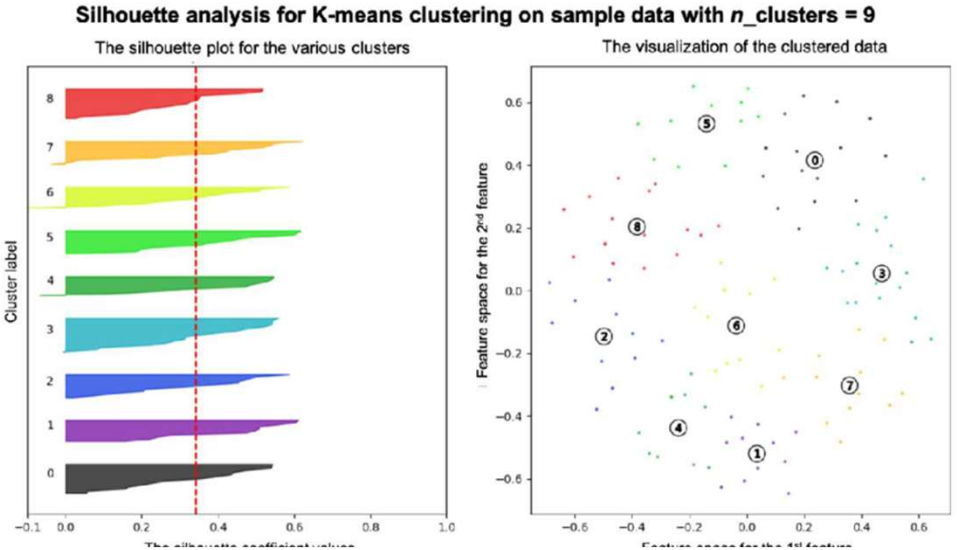
Para cada punto se calcula:

- La distancia promedio a todos los demas puntos pertenecientes al mismo cluster (cohesion)
- La distancia promedio a todos los puntos pertenecientes al segmento mas cercano (separacion)

La silueta se calcula como:

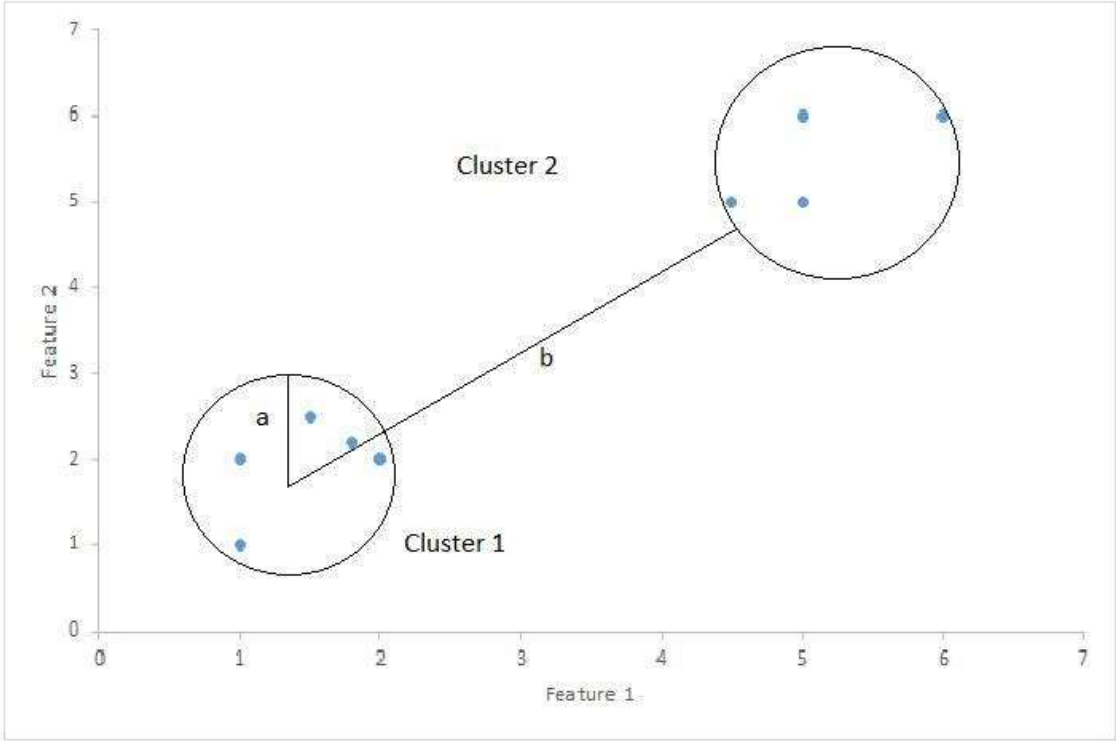
- (separacion – cohesion) / max (separacion y cohesion)





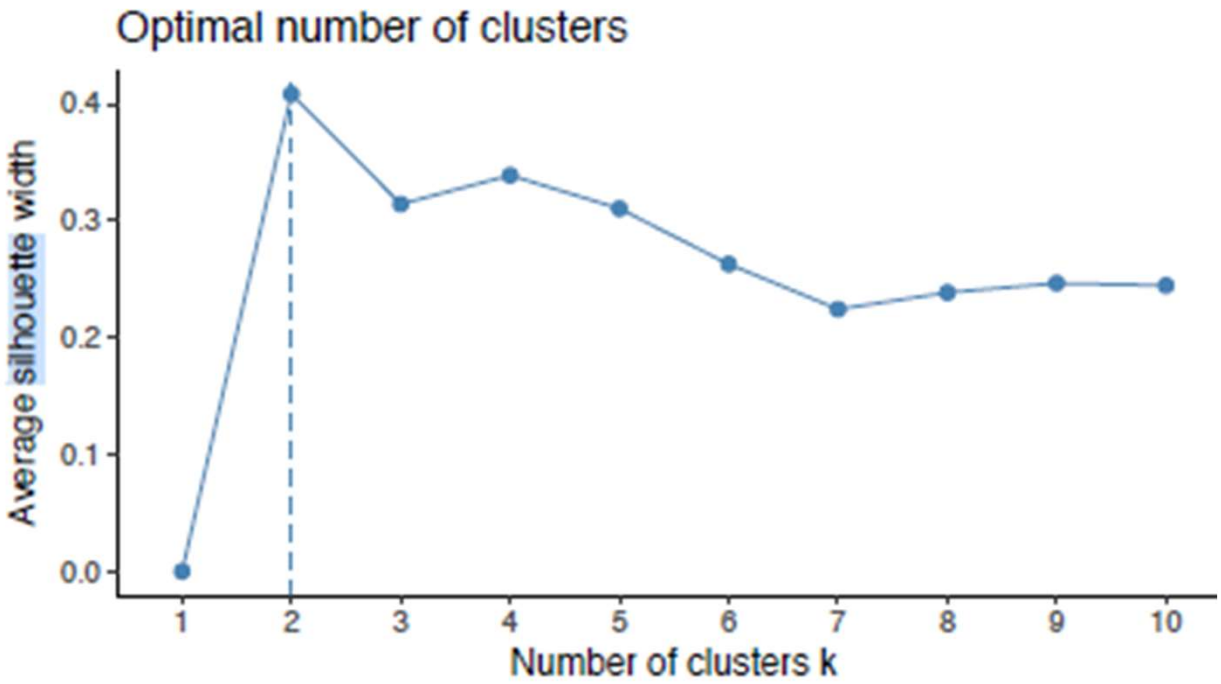
Silueta

- Valores de silueta:
- 1: Los segmentos estan separados entre si y estan claramente diferenciados
- 0: Los segmentos son indiferentes, o que la distancia es insignificante
- -1:La asignacion de segmentos ha sido errónea

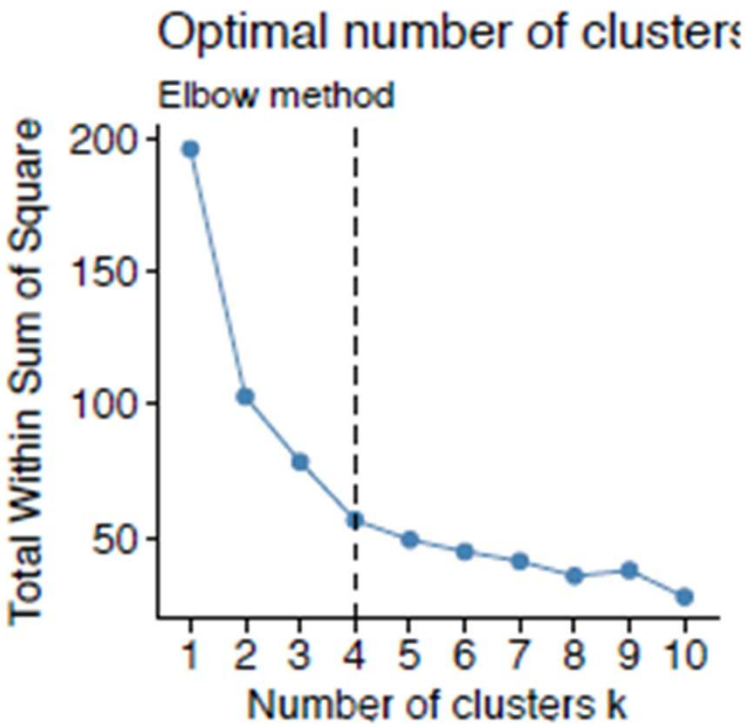


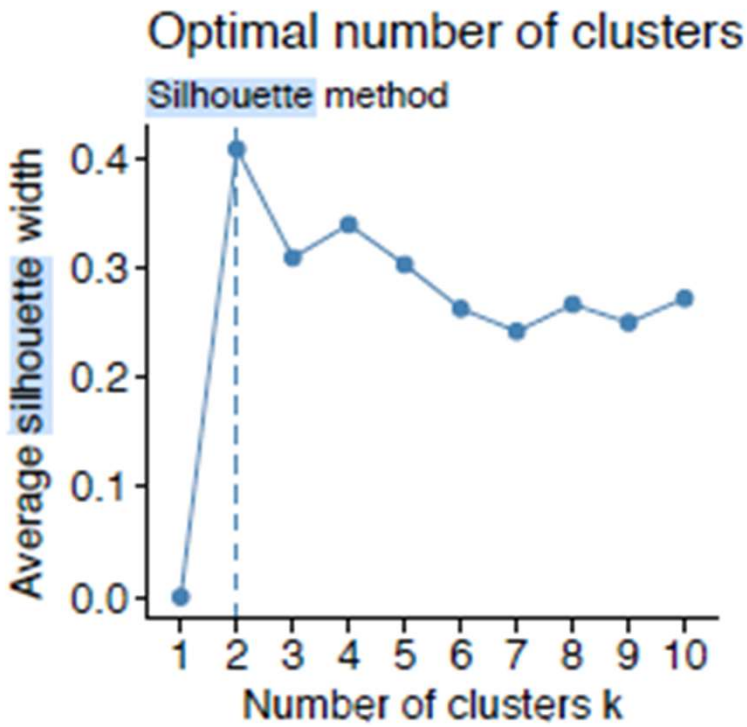
Cual valor de K seleccionar en function de la Silueta?

Silueta



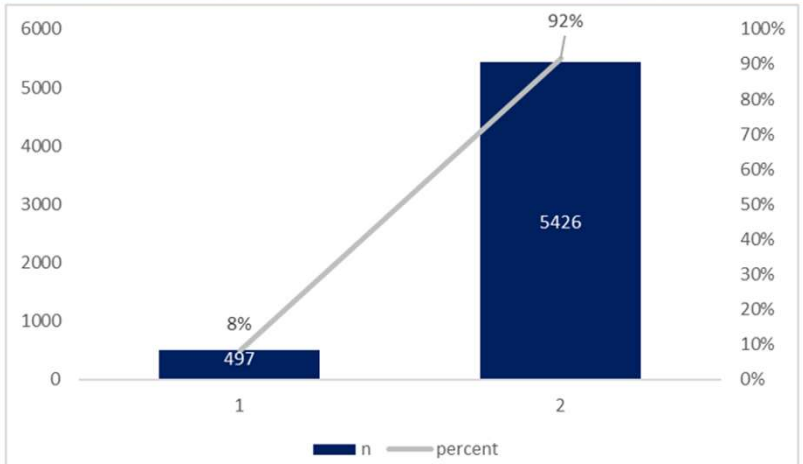
Putting all together

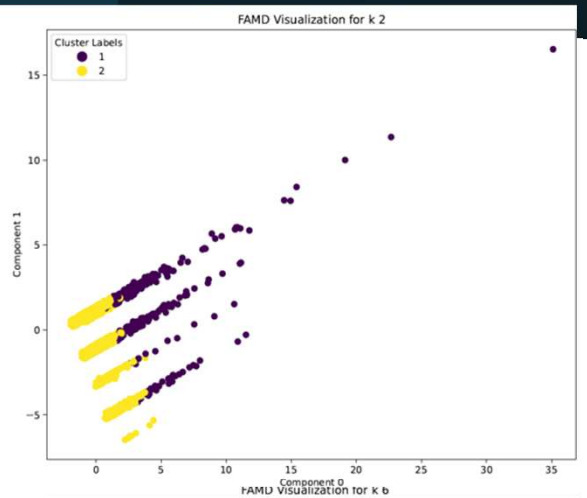




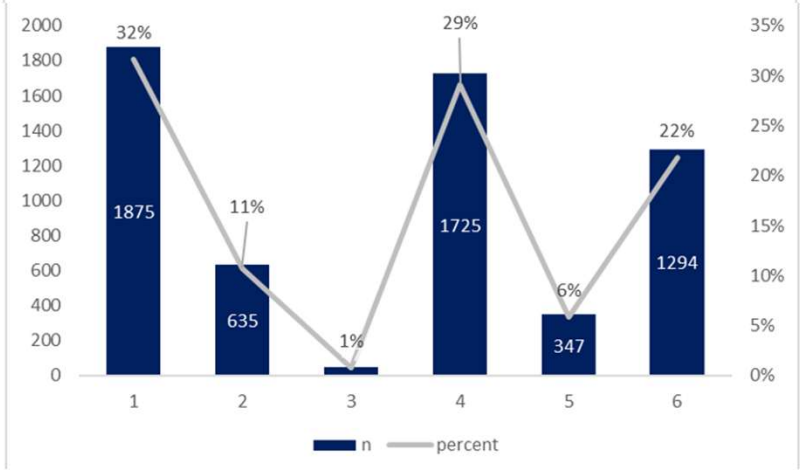
Otros aspectos a evaluar: Concentración de segmentos

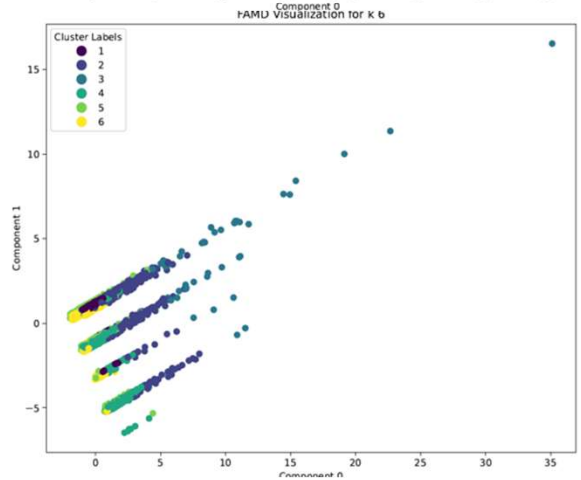
k = 2





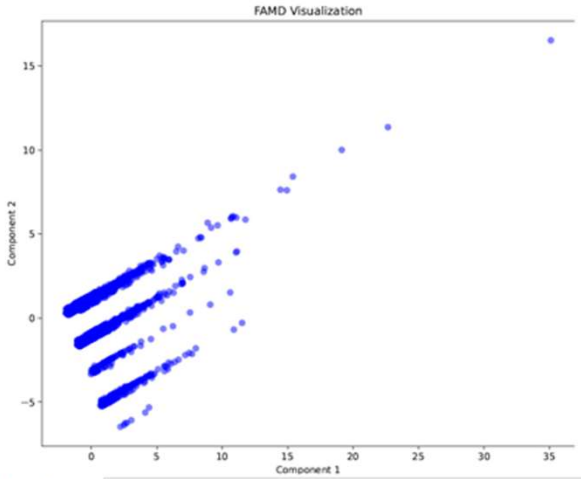
k = 6



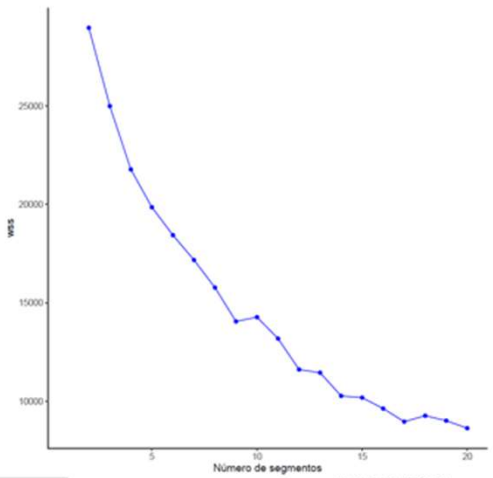


Ejercicio 2: Cuál sería su recomendación?

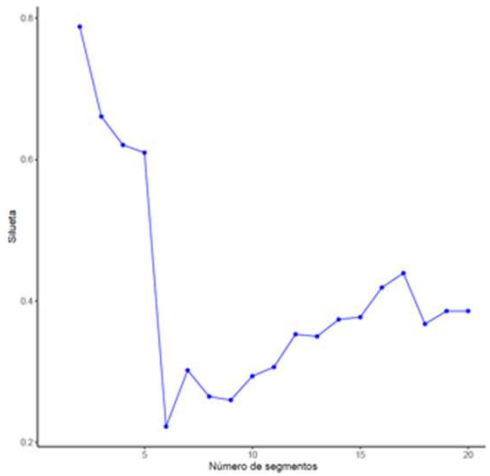
FAMD



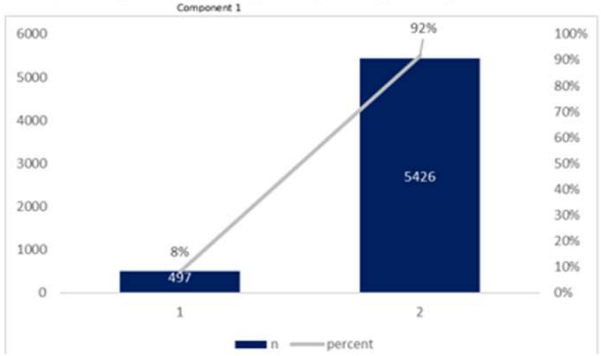
Criterio del codo

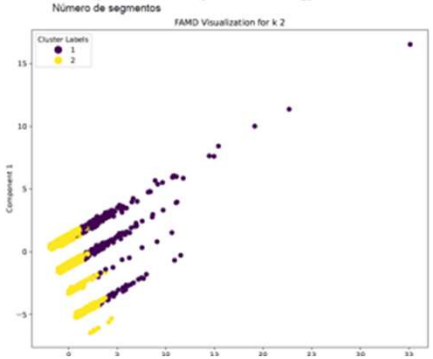


Silueta



k = 2





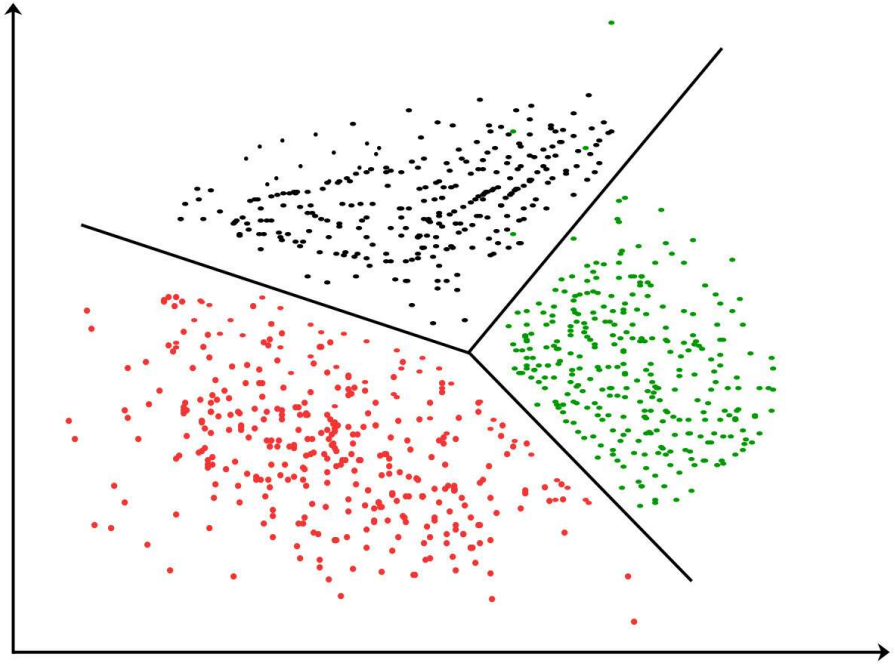
Un aspecto de crucial importancia

La prueba racional

5. Monitoreo

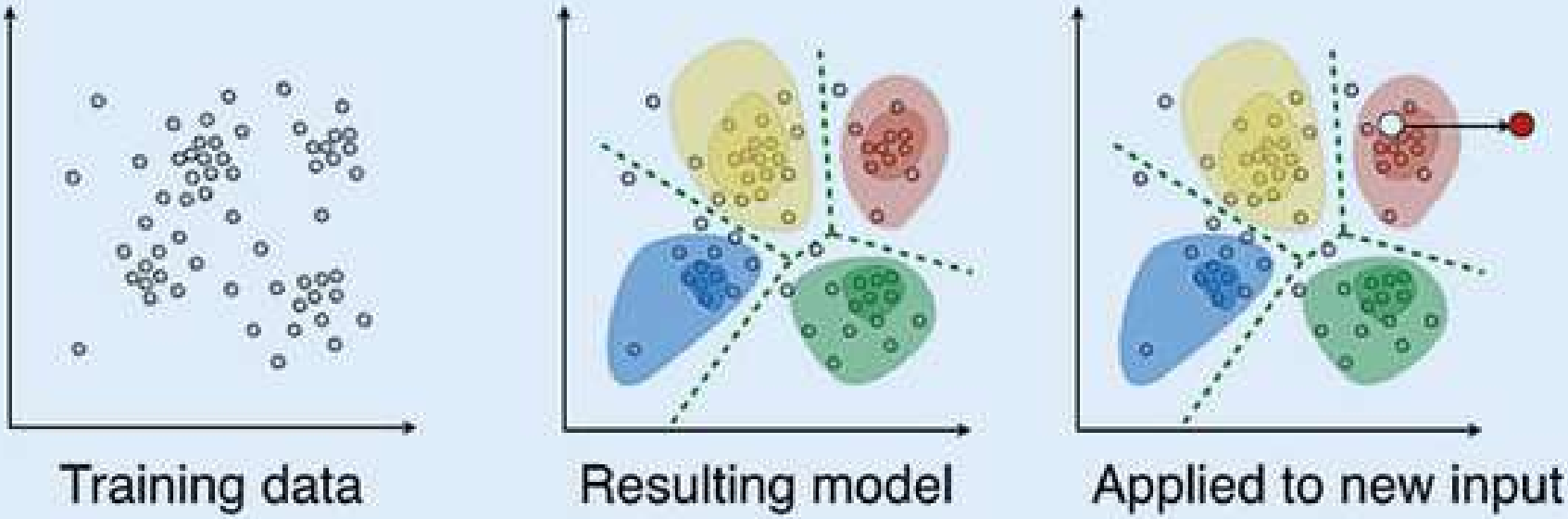
Una vez definido el modelo, como hacemos monitoreo?

- Resegmentar



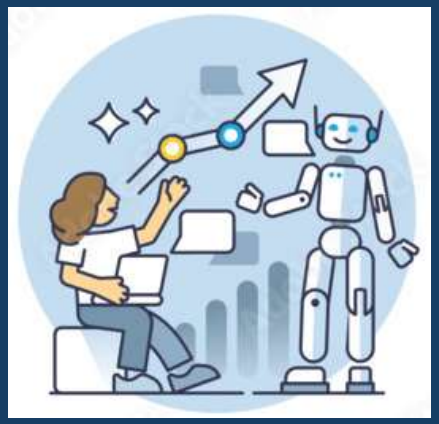
- Modelo de pronostico

Modelo de pronostico



Bajo que grupo de técnicas se podría diseñar un modelo de pronostico de segmentos

Aprendizaje Supervisado

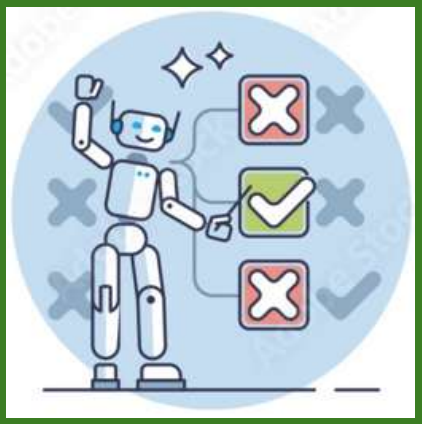


Data “marcada”
Task Drive
Predecir el siguiente valor

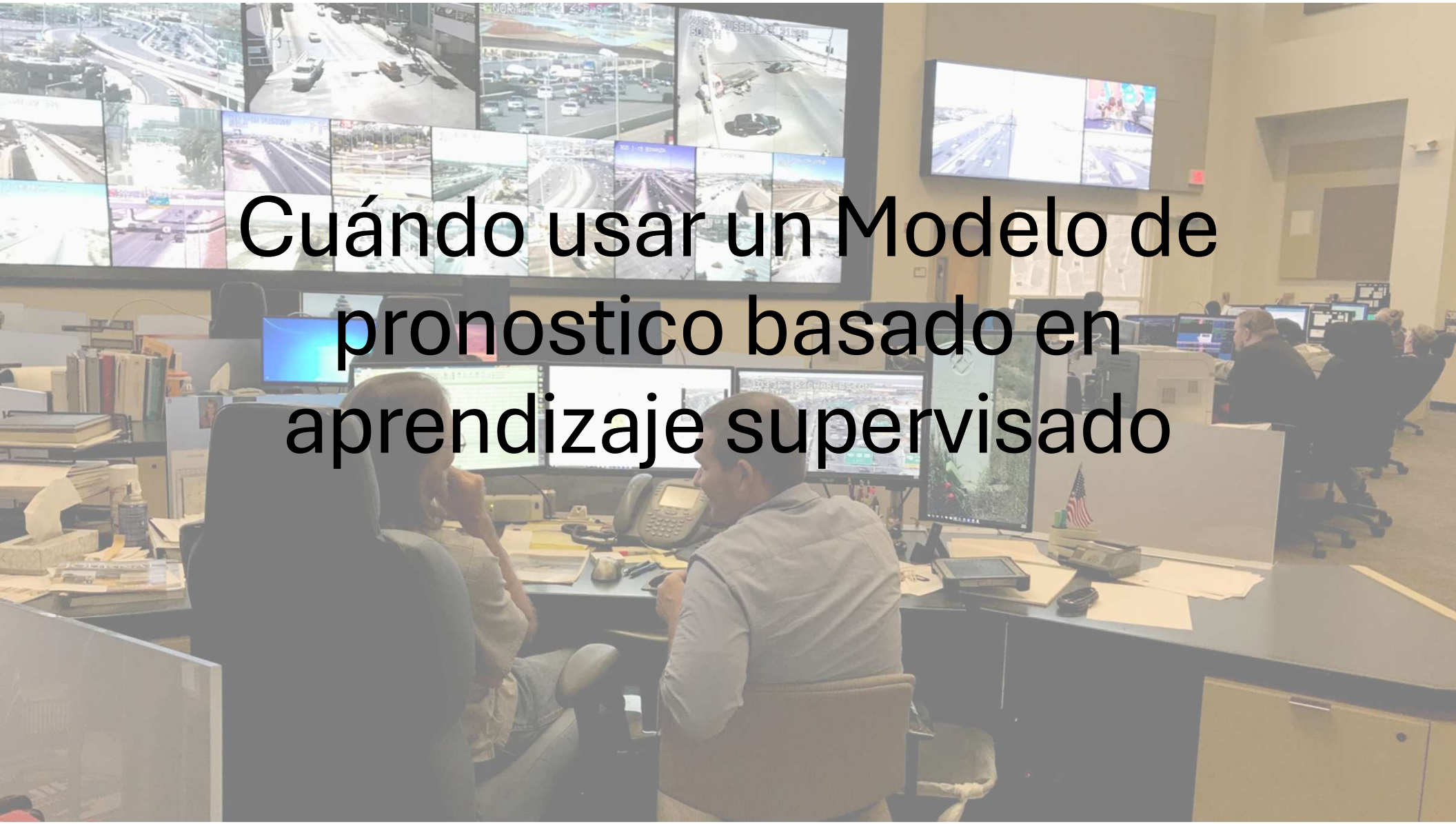
Aprendizaje No Supervisado

Data “no marcada”
Data Driven
Identificar estructura en los datos

Aprendizaje por refuerzo



Aprender de errores



Cuándo usar un Modelo de pronostico basado en aprendizaje supervisado

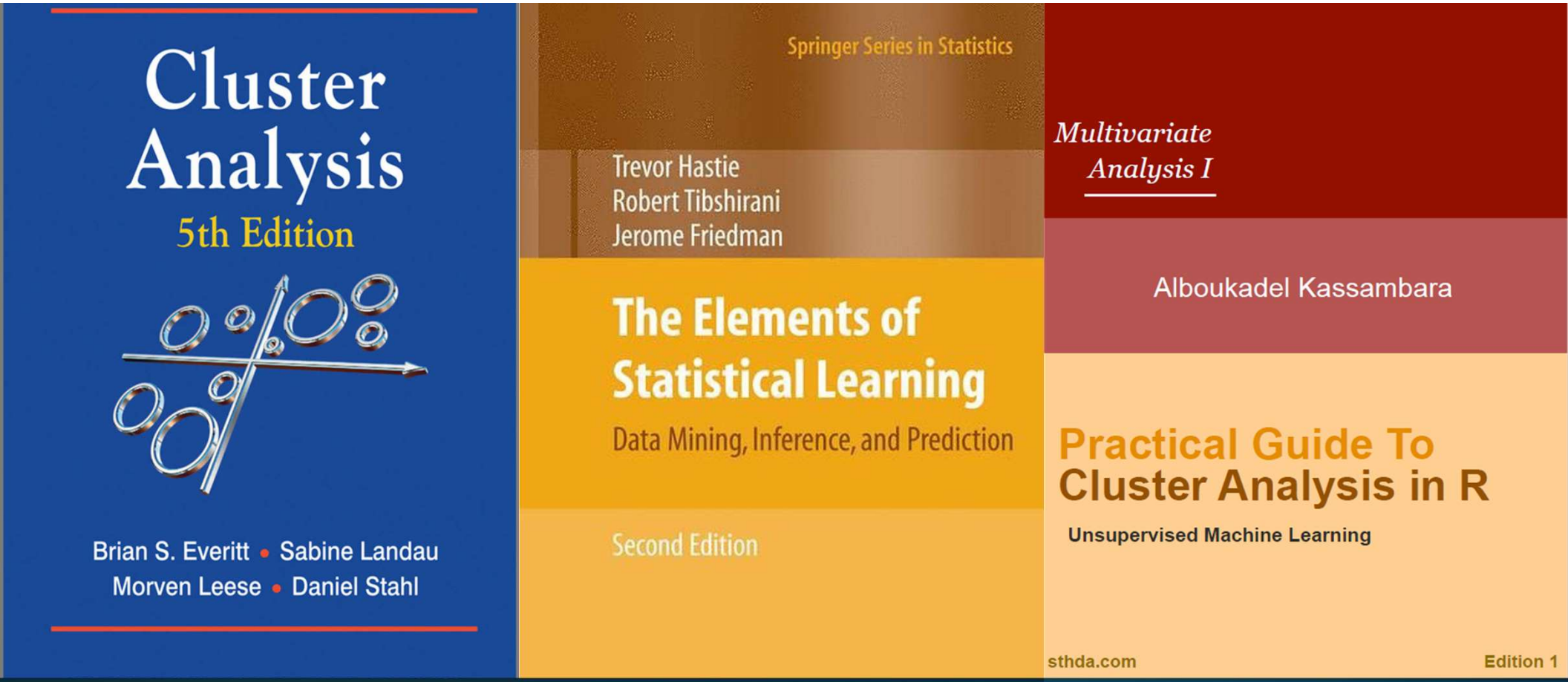
Interquartile Range & Outliers

$\leftarrow$ IQR $\rightarrow$

Outliers $\rightarrow$ $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$

Ejercicio: Diseñemos umbrales de alertamiento basados en los resultados de la segmentación

A	B	C	D	E	F	G	H	I	J
IdCliente	NRActividad	NR	Nacional	PEP	QTxsCASHIN	QTxsCASHOUT	suma_montoCASHIN	suma_montoCASHOUT	Cluster
1	Medio	Bajo	1	0	5	3	61,38175	59,5	6
2	Alto	Bajo	1	0	42	26	11079,2485	7875,72075	2
3	Alto	Alto	1	0	3	2	3318,05375	3400	4
4	Medio	Bajo	1	0	2	4	4976,6665	2741,0055	6
5	Alto	Bajo	1	0	27	22	293,23925	292,685	1
6	Alto	Bajo	1	0	1	0	0,05	0	1
7	Alto	Bajo	1	0	3	4	95,6075	95,8605	1
8	Alto	Alto	1	0	4	4	2036,7225	3195,677	4
9	Alto	Alto	1	0	5	5	455,12425	458,6805	4
10	Alto	Alto	1	0	3	2	60,55775	115,192	4
11	Alto	Bajo	1	0	2	3	156,14475	155,5	1
12	Bajo	Medio	1	0	6	2	215,96025	210	6
13	Alto	Bajo	1	0	81	0	255,22325	0	1
14	Alto	Bajo	1	0	2	1	64,58475	64	1
15	Alto	Medio	1	0	1	1	2,6705	2,5	1
16	Alto	Alto	0	0	1	0	1,11275	0	4
17	Alto	Bajo	1	0	15	20	2370,29725	2358,24425	1
18	Alto	Alto	1	0	3	9	414,1965	407,4345	4



Bibliografia