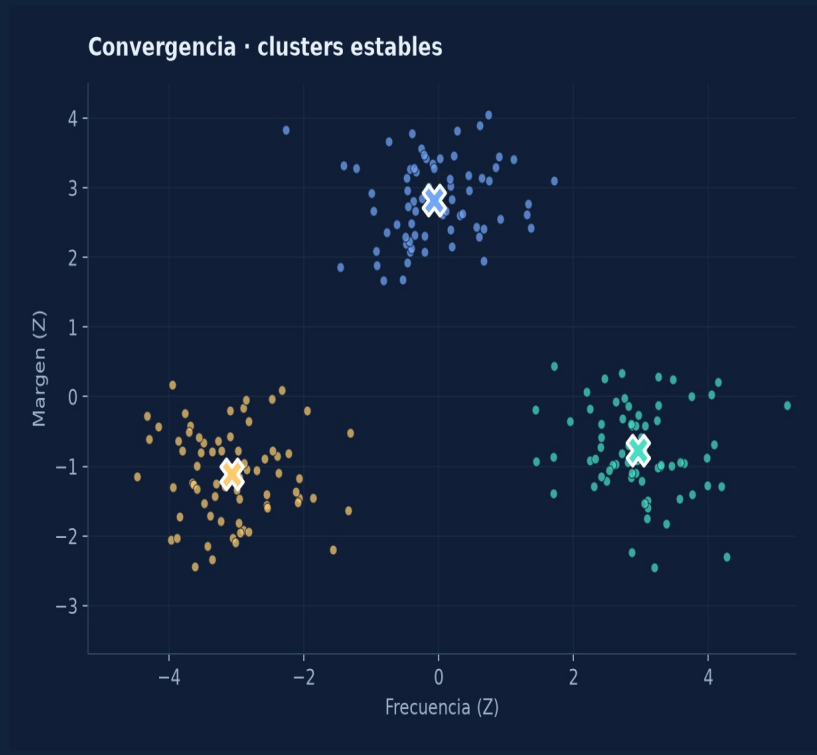
CIENCIA DE DATOS PARA DECISIONES DE NEGOCIO

MÓDULO 10

K-Means · Segmentación por centroides

Cómo se mueven los centroides, cómo elegir K y cuándo desconfiar de los grupos.

Clases conceptuales · Matemáticas · Evidencia visual · Decisión de negocio



Sergio Gevatschnaider · Material de cátedra

La pregunta correcta antes de elegir K

IDEAS CLAVE

- Segmentar para definir acciones diferenciadas, no para decorar un dashboard.
- Precisar a quién se segmenta: clientes, sucursales, productos o eventos.
- Seleccionar variables, distancias, escala y número de segmentos.
- Medir estabilidad, capacidad descriptiva y costos de implementación.

PREGUNTAS DE DECISIÓN

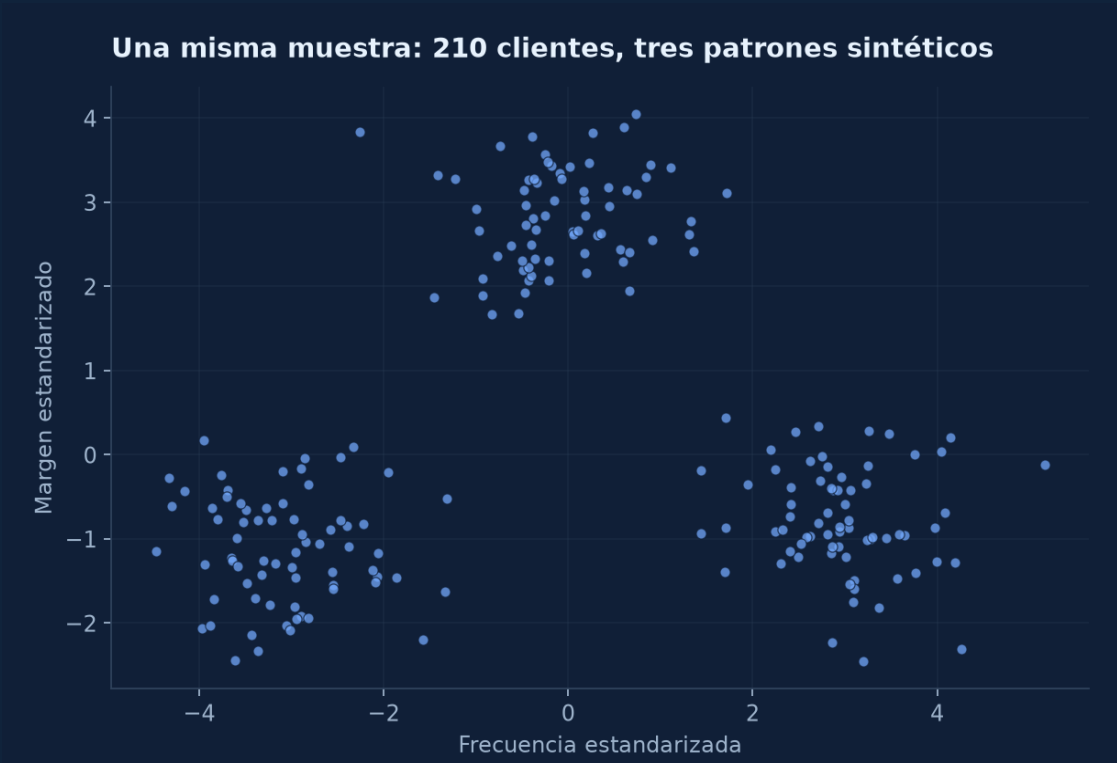
1. ¿Qué decisión cambia entre segmentos?
2. ¿Es comparable cada variable?
3. ¿Se pueden ejecutar acciones diferentes?
4. ¿Los segmentos resisten una nueva muestra?

IDEA QUE DEBE QUEDAR **Un cluster no tiene valor comercial hasta que cambia una decisión.**

Un mismo dataset sin etiquetas de segmento

IDEAS CLAVE

- 210 clientes sintéticos expresados en frecuencia y margen estandarizados.
- La nube parece exhibir tres concentraciones principales.
- El algoritmo no conoce los grupos originales que generaron los datos.
- El reto es detectar y describir una estructura sin sobreinterpretarla.

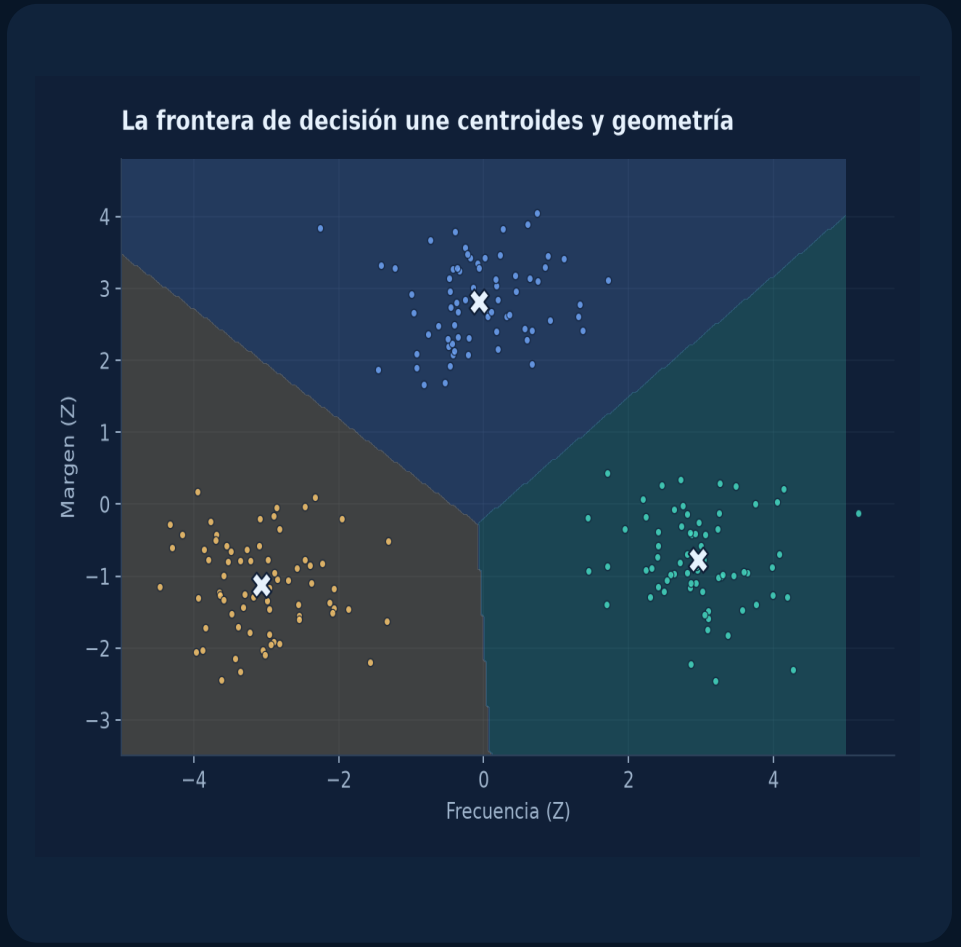


ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Los grupos descubiertos dependen de distancia, escala y supuestos.**

01 · Qué optimiza K-Means

Una partición “buena” minimiza las distancias cuadráticas de los puntos a sus centroides.



La función objetivo es la inercia (WCSS)

IDEAS CLAVE

- Para cada cliente i se asigna una etiqueta z_i entre 1 y K.
- Cada cluster j tiene un centroide μ_j, media de sus observaciones.
- Se minimiza la suma de distancias euclídeas al cuadrado.
- La solución es local y depende de la inicialización.

MODELO MATEMÁTICO

$$J = \sum_i \|x_i - \mu z_i\|^2$$

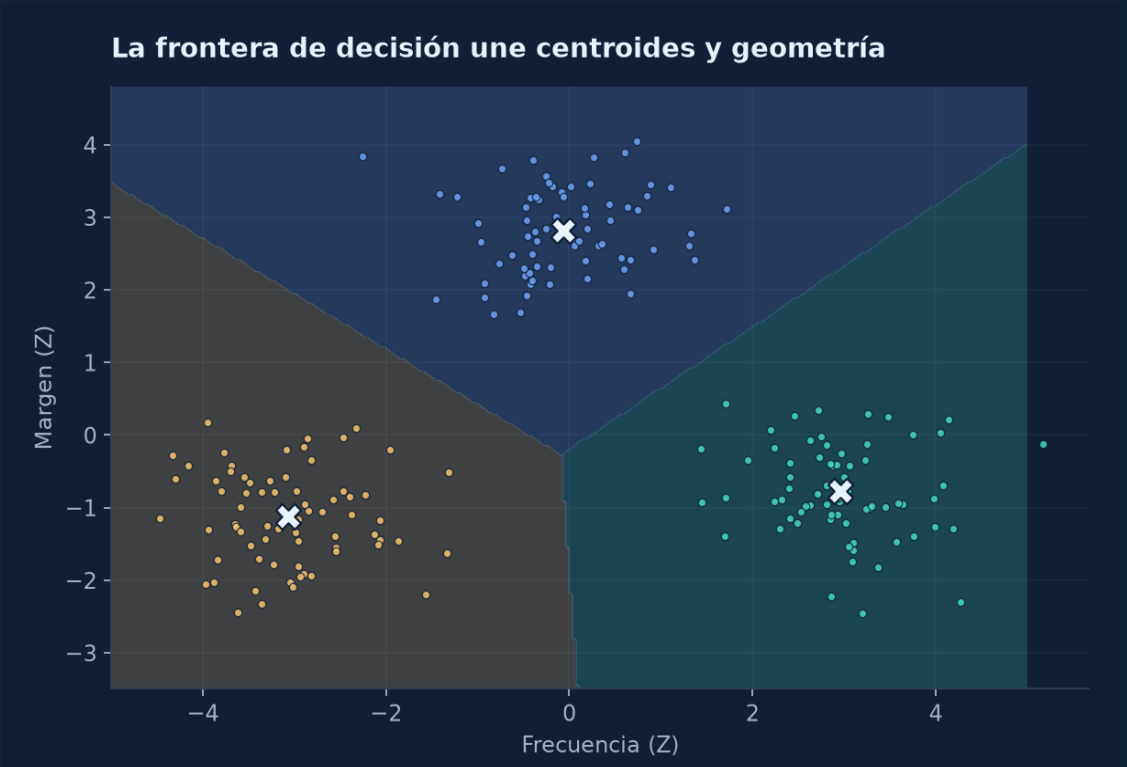
- x_i: vector de características del cliente.
- z_i: etiqueta asignada de 1 a K.
- μ_j: promedio del cluster j.
- K-Means favorece grupos compactos.

IDEA QUE DEBE QUEDAR **La inercia es interna; no mide directamente la utilidad de un segmento.**

Asignación: cada punto elige su centroide más cercano

IDEAS CLAVE

- Fijados los centros, calcular la distancia euclídea de cada x_i.
- La etiqueta se actualiza según el centro de distancia mínima.
- El plano queda separado en regiones de Voronoi.
- Un cambio en los centros altera qué cliente cae en cada zona.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **En cada iteración se reasignan clientes por proximidad.**

Actualización: la media minimiza la SSE intracluster

IDEAS CLAVE

- Para cada grupo se calcula la media de los puntos asignados.
- La media es el minimizador de la suma de distancias al cuadrado.
- Repetimos asignación y actualización hasta converger.
- Un cluster vacío requiere una estrategia explícita de reinicio.

MODELO MATEMÁTICO

$$\mu_j \leftarrow (1 / n_j) \sum_{i:z_i=j} x_i$$

- Asignar puntos al centro más cercano.
- Calcular la media de cada nuevo grupo.
- Repetir y detener ante estabilidad / tolerancia.
- Controlar iteraciones, vacíos y semilla.

IDEA QUE DEBE QUEDAR **El algoritmo de Lloyd alterna dos optimizaciones condicionadas.**

Demostración manual con seis puntos

IDEAS CLAVE

- A=(1,1), B=(2,1), C=(2,2); D=(7,7), E=(8,7), F=(8,8).
- Con K=2 y una asignación natural: ABC frente a DEF.
- Centroides: μ_1=(1,67;1,33), μ_2=(7,67;7,33).
- Cada punto debe quedar asignado al centroide más próximo.

MODELO MATEMÁTICO

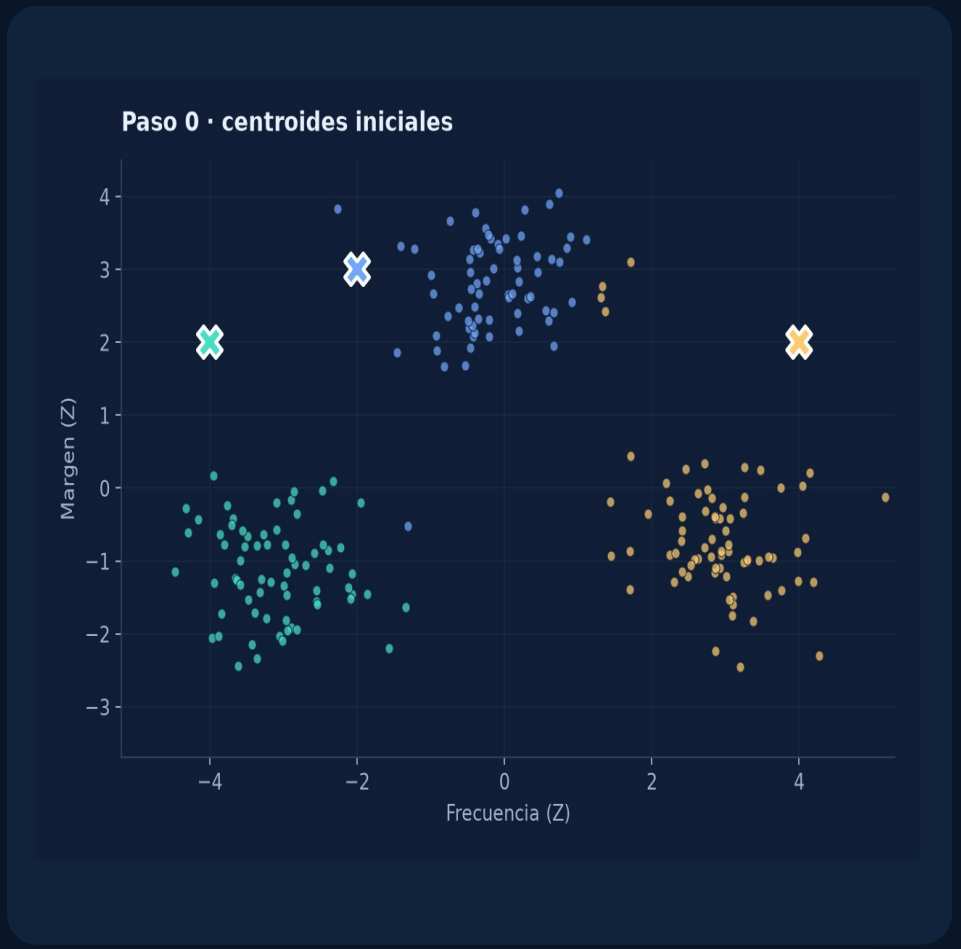
$$\mu_1 = (5/3, 4/3) \quad \cdot \quad \mu_2 = (23/3, 22/3)$$

- Pedir a la clase justificar la primera asignación.
- Calcular ambas medias coordenada a coordenada.
- Verificar distancias y si cambiaron etiquetas.
- Comparar con K=1 y K=3.

IDEA QUE DEBE QUEDAR **La media es vectorial: cada coordenada se promedia dentro de su grupo.**

02 · La dinámica visible

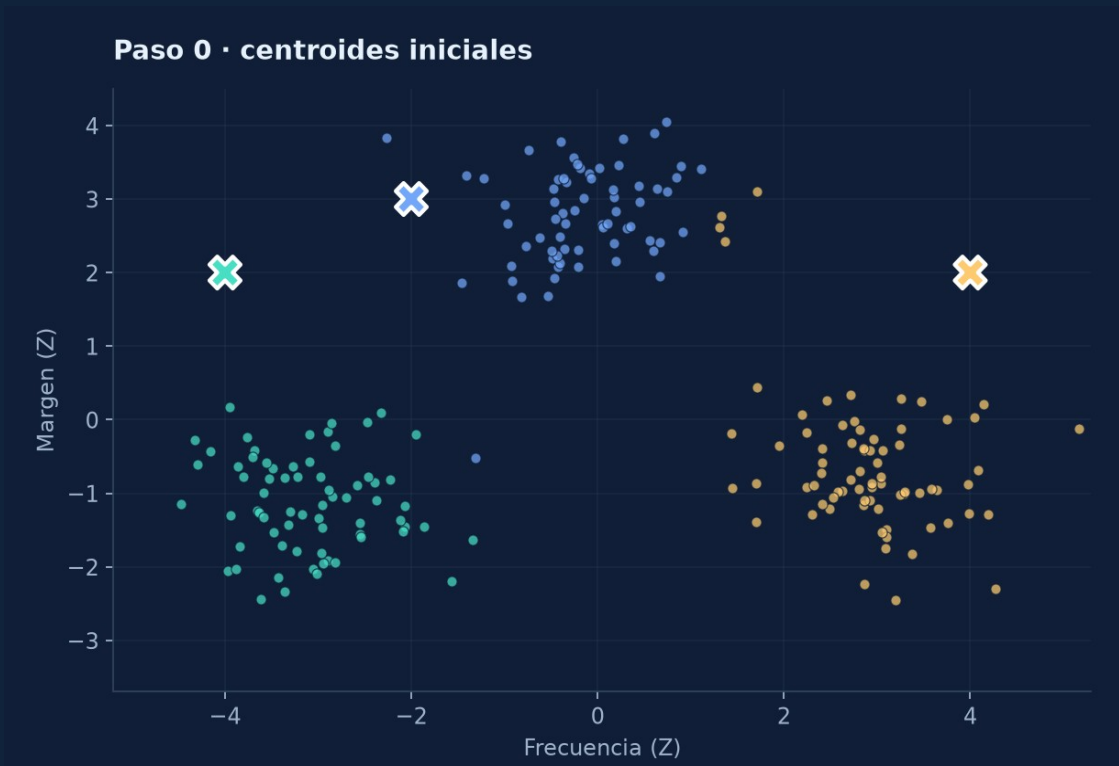
Acompañar cada desplazamiento del centroide explica mucho más que mostrar el cluster final.



Inicialización: una elección con consecuencias

IDEAS CLAVE

- Elegimos K=3 y tres centroides iniciales, potencialmente alejados.
- Se colorean los clientes según el centro más cercano.
- Una elección desfavorable puede inducir un mínimo local de peor calidad.
- El objetivo todavía puede disminuir de forma importante.



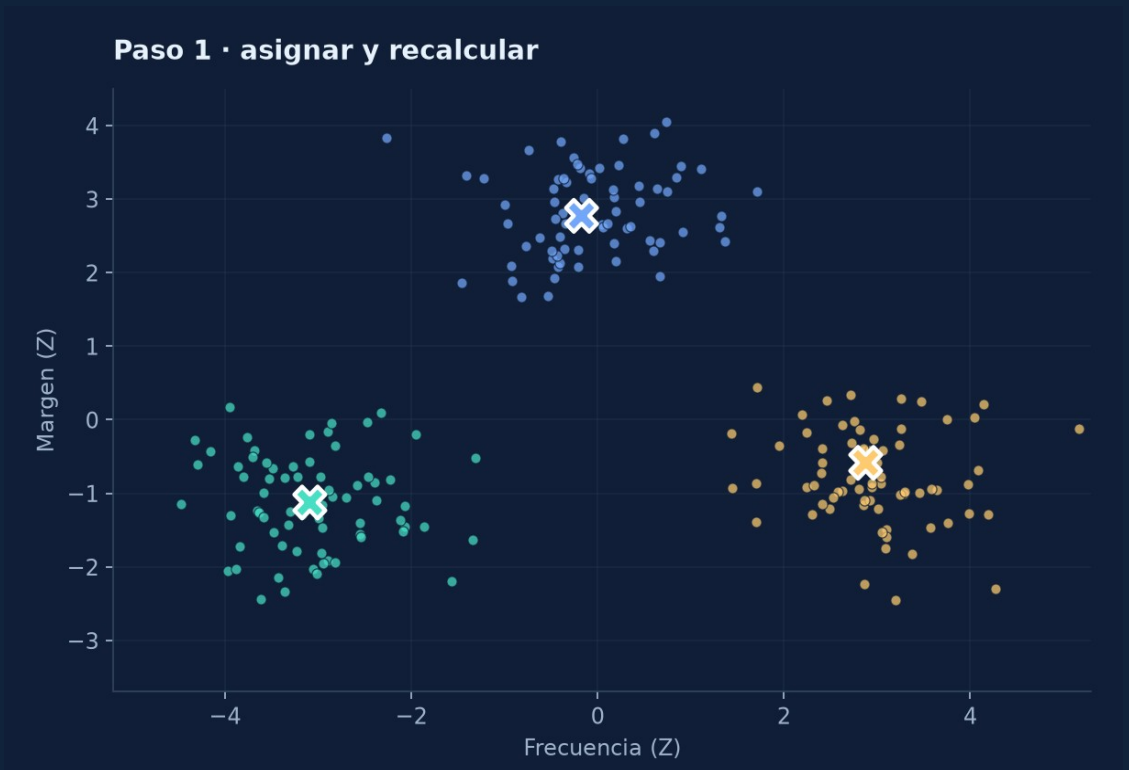
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **S05 del futuro simulador permitirá mover centroides a mano.**

Primer ciclo: nuevas medias, nuevas asignaciones

IDEAS CLAVE

- Con las etiquetas actuales, recalculamos tres promedios.
- Los centroides migran hacia las regiones con más puntos asignados.
- Luego cambia la frontera de decisión.
- La inercia no aumenta cuando se ejecutan correctamente ambos pasos.



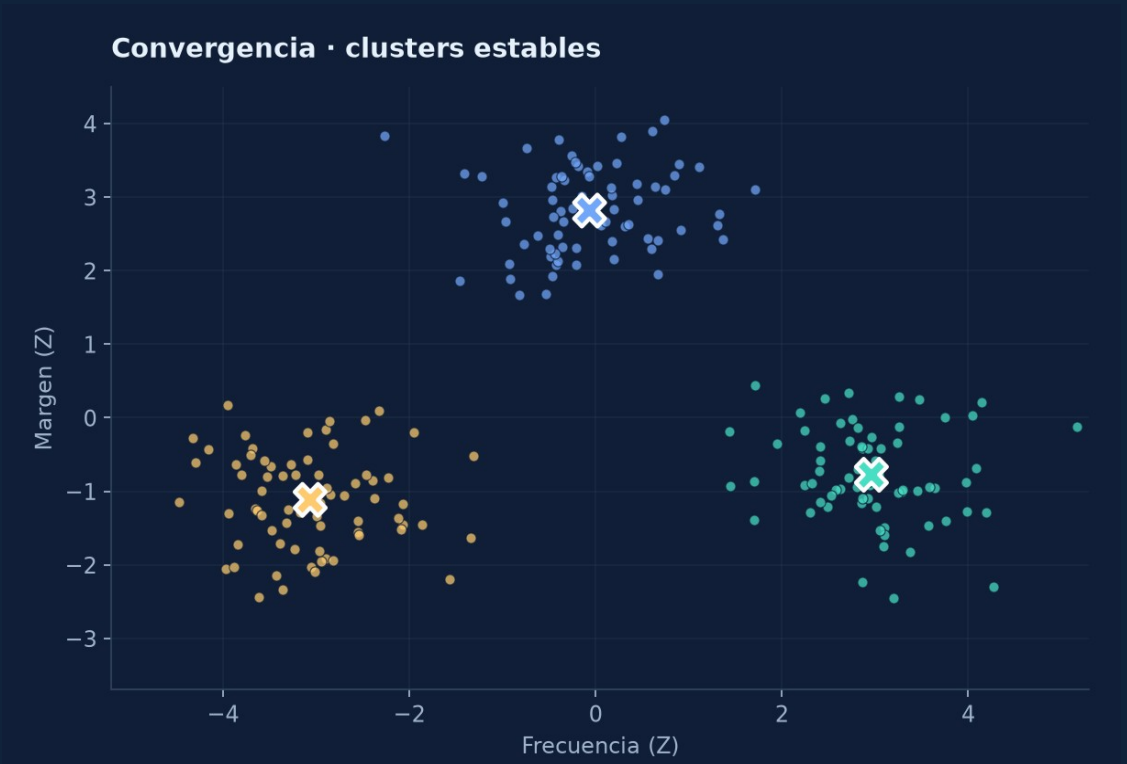
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Cada iteración reduce o mantiene el objetivo; no garantiza óptimo global.**

Convergencia: se estabilizan los grupos

IDEAS CLAVE

- El algoritmo termina cuando no hay mejora significativa o se alcanza un límite.
- Los tres segmentos sintéticos pasan a ser territorialmente compactos.
- La posición final de cada centro resume promedios, no clientes reales.
- El orden numérico de las etiquetas 0, 1, 2 es arbitrario.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **La convergencia técnica no certifica que el modelo sea estable o útil.**

Por qué conviene K-Means++ y varios reinicios

IDEAS CLAVE

- K-Means++ tiende a separar centroides iniciales mediante distancias.
- Disminuye el riesgo de soluciones iniciales claramente pobres.
- Usar varios reinicios y escoger la menor inercia obtenida.
- Guardar random_state, n_init, tolerancia y versión de librería.

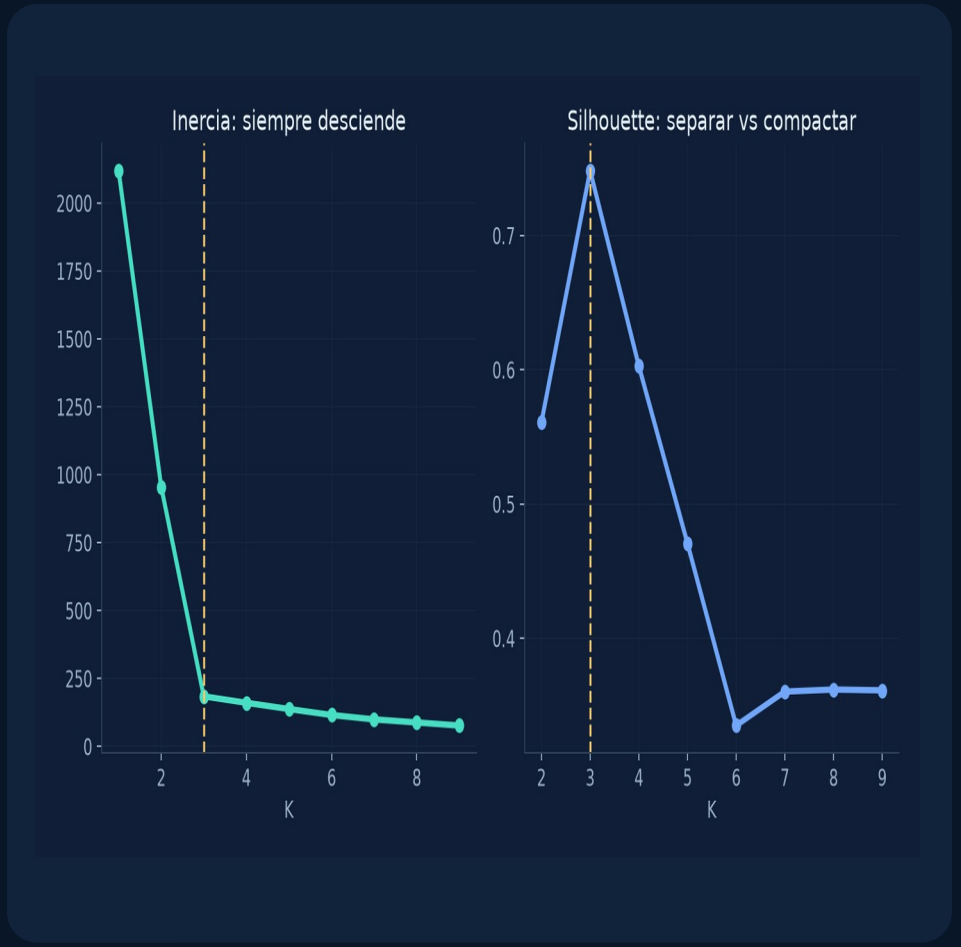
DISEÑO REPRODUCIBLE

1. Elegir K y variables
2. Sembrar centros con K-Means++
3. Ejecutar varios reinicios
4. Comparar inercia y estabilidad
5. **Guardar parámetros / perfil resultante**

IDEA QUE DEBE QUEDAR **Más reinicios reducen dependencia del azar; no eliminan los supuestos del método.**

03 · Elegir K sin autoengaño

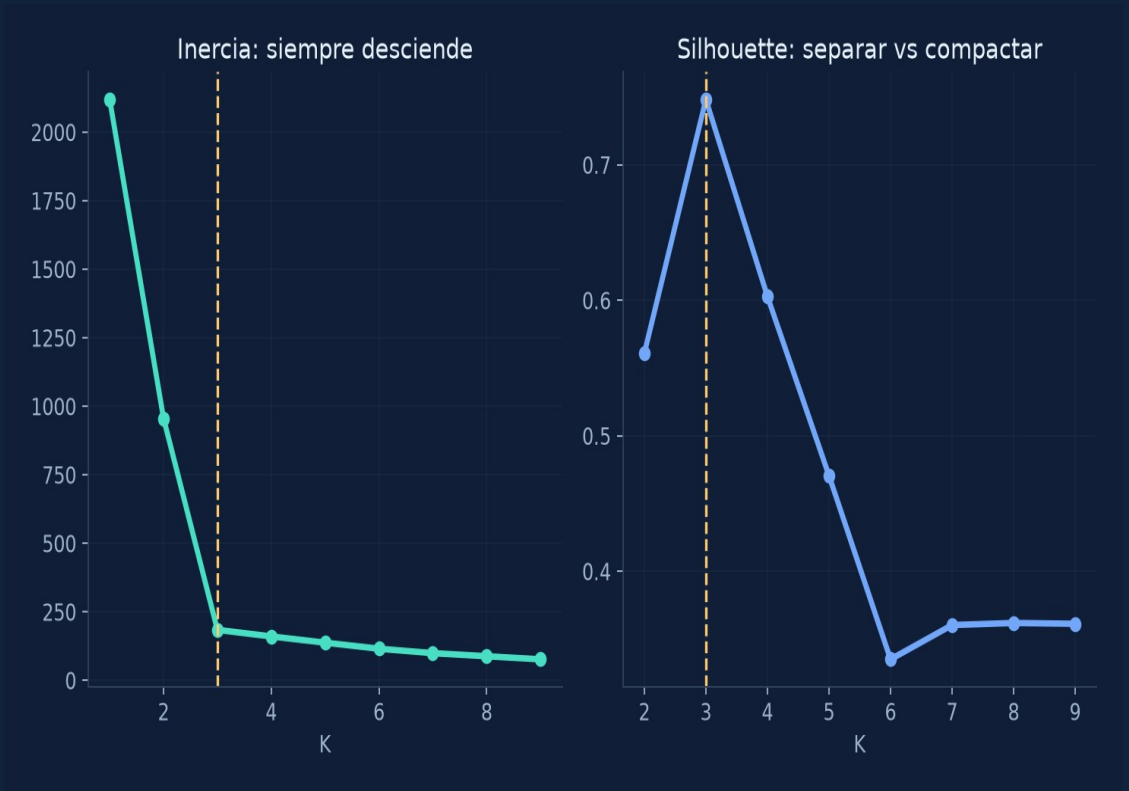
Una curva bonita no sustituye un criterio de decisión ni la validación de estabilidad.



Elbow: buscar rendimientos marginales decrecientes

IDEAS CLAVE

- El mínimo teórico de inercia no crece con K; un ajuste local puede diferir.
- Se observa si la mejora comienza a ser marginal al añadir clusters.
- Algunas muestras no muestran un codo claro.
- Nunca maximizar K exclusivamente para minimizar la inercia.



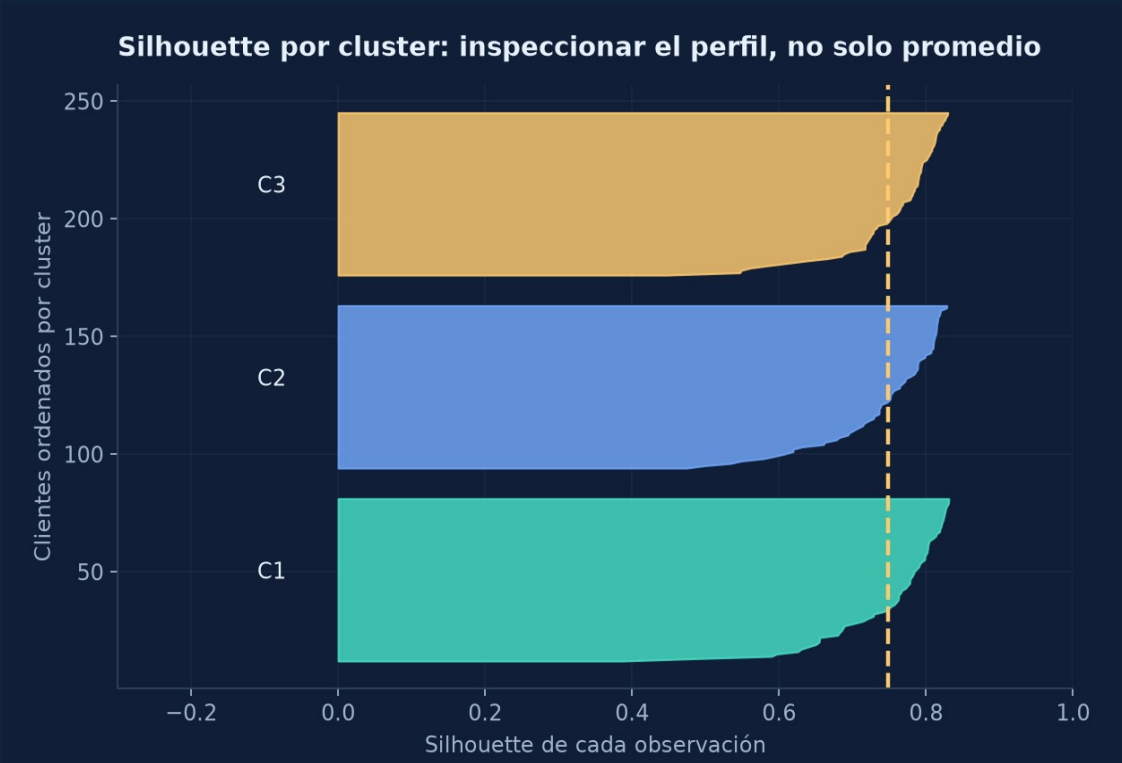
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **El “codo” es una heurística visual, no una prueba estadística universal.**

Silhouette: equilibrio de cohesión y separación

IDEAS CLAVE

- a(i): distancia media del punto a su propio cluster (excluyéndose a sí mismo).
- b(i): menor distancia media a otro cluster.
- s(i) = [b(i) − a(i)] / max{a(i), b(i)}.
- Cercano a 1 indica mayor separación; negativo sugiere asignación discutible.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Usar tanto la media como el perfil por cluster; requiere grupos de tamaño adecuado.**

Triangular K con evidencias diferentes

IDEAS CLAVE

- Calcular inertia, silhouette, tamaño de grupos y estabilidad entre muestras.
- Revisar si los perfiles originales son distintos e interpretables.
- Medir utilidad: retención, conversión, uplift o ahorro operativo.
- Preferir K interpretable cuando diferencias de métricas son marginales.

REGLAS DE DECISIÓN

K

No siempre hay óptimo único

SSE

Compacidad interna

S(i)

Cohesión / separación

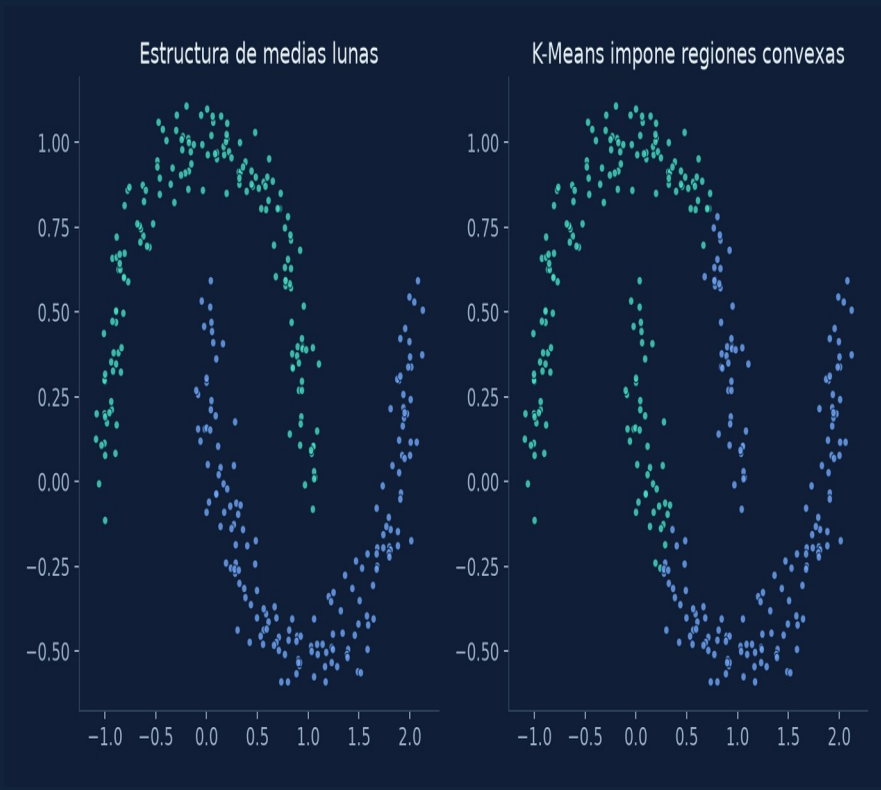
€

Acción / valor esperado

IDEA QUE DEBE QUEDAR **La selección de K es multicriterio y debe declararse antes de actuar.**

04 · Límites estructurales

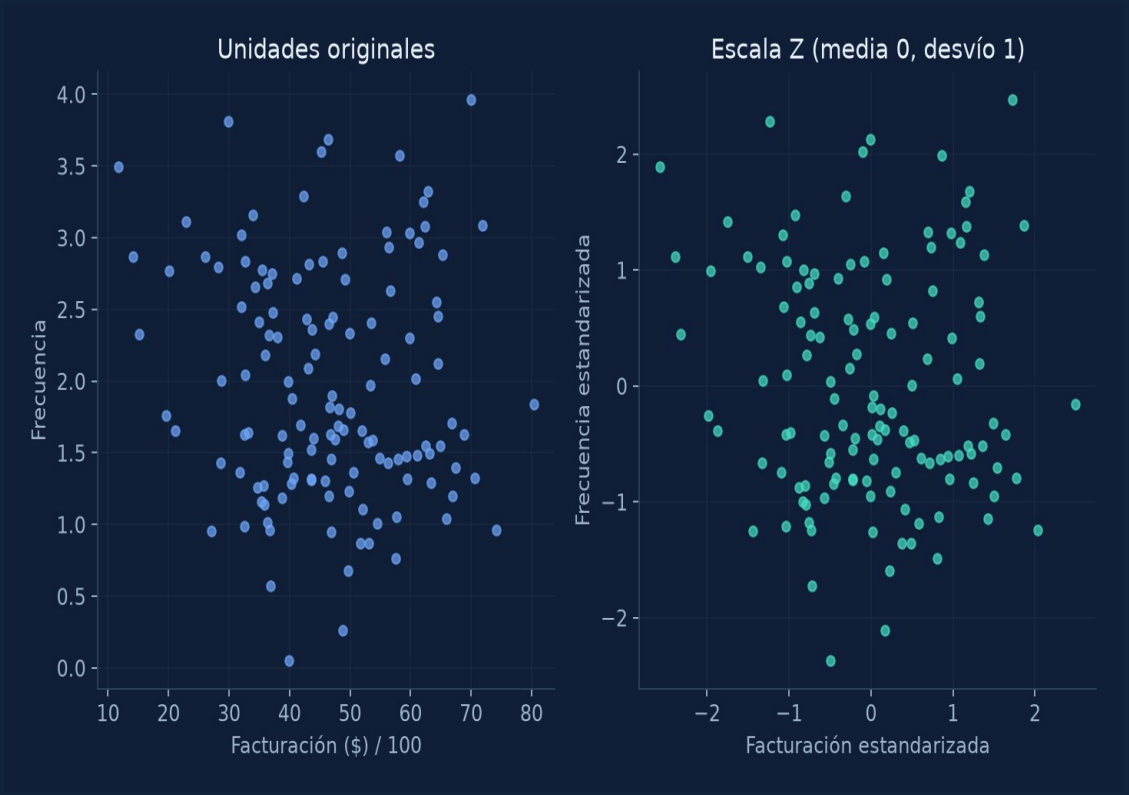
La forma de los datos, los outliers y las decisiones de escala determinan qué segmentos son posibles.



La escala cambia la noción de parecido

IDEAS CLAVE

- Euclidean² suma diferencias al cuadrado por variable.
- Si una característica tiene varianza enorme puede dirigir la partición.
- StandardScaler, transformaciones log y ponderaciones cambian los grupos.
- Evaluar cambios de perfiles y estabilidad frente a transformaciones.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Una distancia solo tiene sentido luego de justificar la escala y las variables.**

K-Means no es “clustering con cualquier distancia”

IDEAS CLAVE

- La actualización con medias deriva de distancias euclídeas al cuadrado.
- Con distancias arbitrarias el centroide puede dejar de ser el minimizador.
- K-Medoids usa observaciones representativas, útil con métricas más generales.
- Para datos mixtos o categóricos convienen modelos de agrupamiento adecuados.

COMPARACIÓN

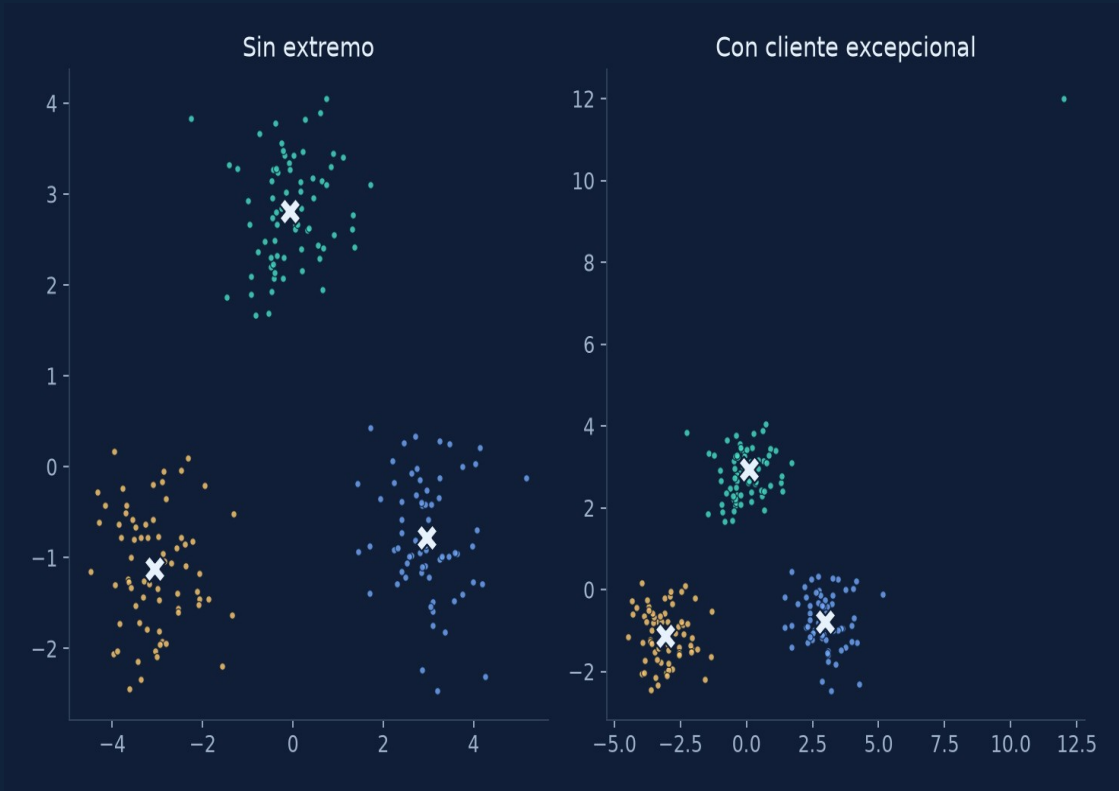
K-MEANS	Centroide medio / Euclidean².
K-MEDOIDS	Medoide real / distancias elegidas.
K-MODES	Moda para variables categóricas.
GOWER	Una posible distancia para datos mixtos.

IDEA QUE DEBE QUEDAR **No cambiar la función de distancia sin reconsiderar el algoritmo completo.**

Outliers: un punto puede arrastrar centroides

IDEAS CLAVE

- La pérdida cuadrática penaliza fuertemente las desviaciones extremas.
- Un punto aislado puede desplazar uno o más centroides.
- Comparar ajustes con y sin casos extremos documentados.
- K-Medoids, clustering robusto o reglas específicas pueden ser alternativas.



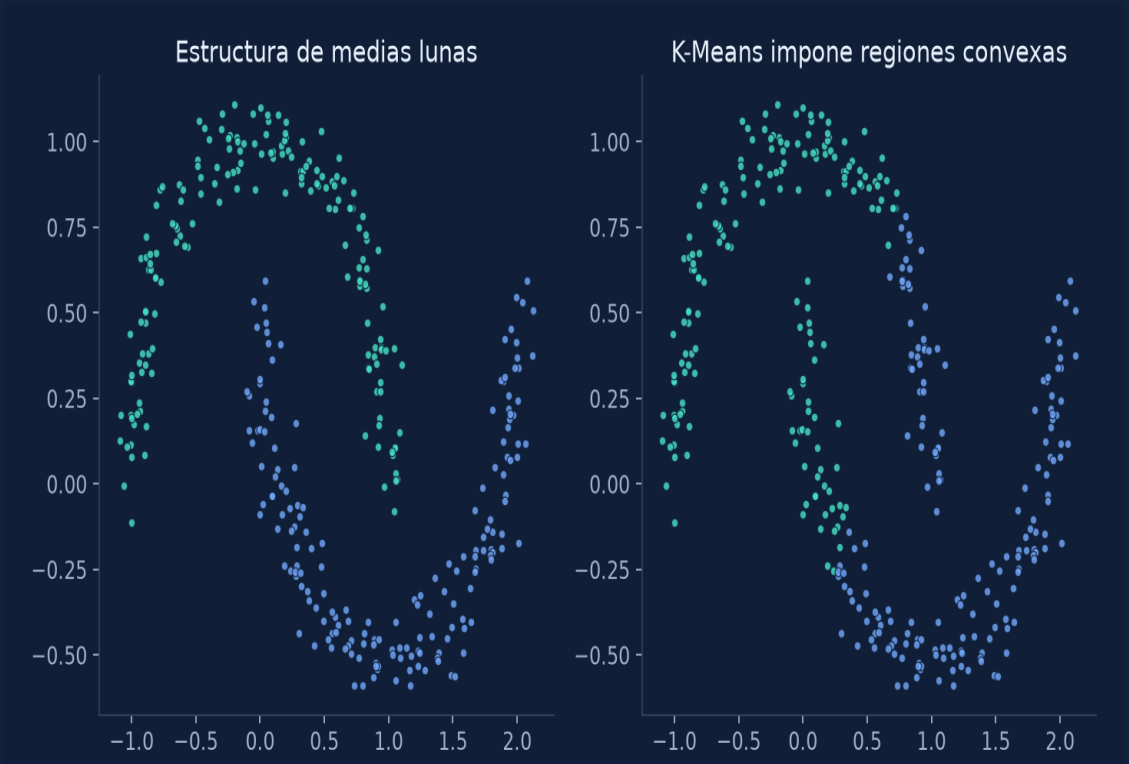
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **No decidir automáticamente que los clientes excepcionales son “errores”.**

Clases no convexas: K-Means puede cortar la estructura

IDEAS CLAVE

- Con medias lunas, el agrupamiento real sigue una trayectoria curva.
- K-Means crea celdas convexas y no sigue automáticamente la curva.
- El criterio de inercia puede verse razonable pese a una partición equivocada.
- Examinar alternativas como métodos de densidad o espectrales.



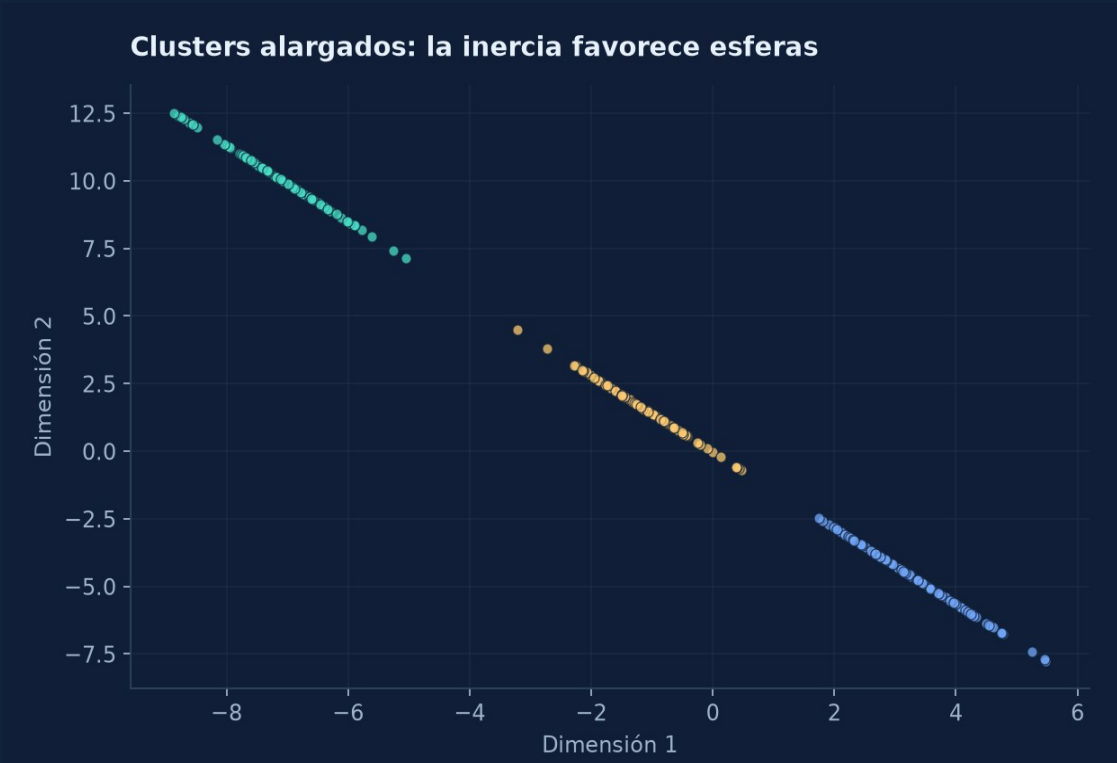
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Una silhouette elevada no corrige supuestos geométricos inadecuados.**

Clusters alargados, tamaños distintos y densidades desiguales

IDEAS CLAVE

- Los centros esféricos sufren cuando las covarianzas son anisotrópicas.
- Clusters grandes pueden absorber pequeños pero importantes.
- La diferencia de densidad deteriora la frontera por cercanía.
- Gaussian Mixture o modelos de densidad pueden resultar más adecuados.



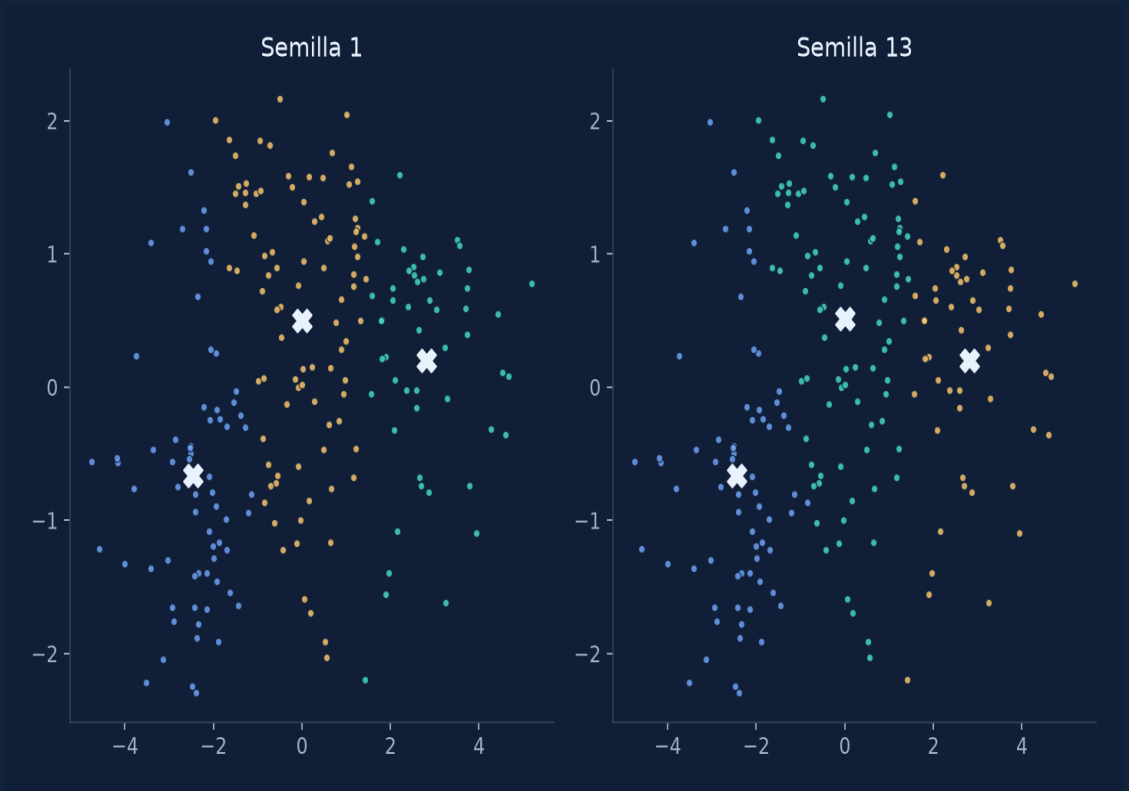
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **El negocio puede preferir un segmento pequeño que K-Means diluye.**

Estabilidad frente a semillas y remuestreo

IDEAS CLAVE

- Correr diferentes semillas y revisar inercia, silhouette y tamaños.
- Repetir el análisis sobre bootstrap o cortes temporales comparables.
- Evaluar acuerdo entre particiones mediante ARI u otras métricas.
- No comparar etiquetas numéricas sin alinear las particiones.

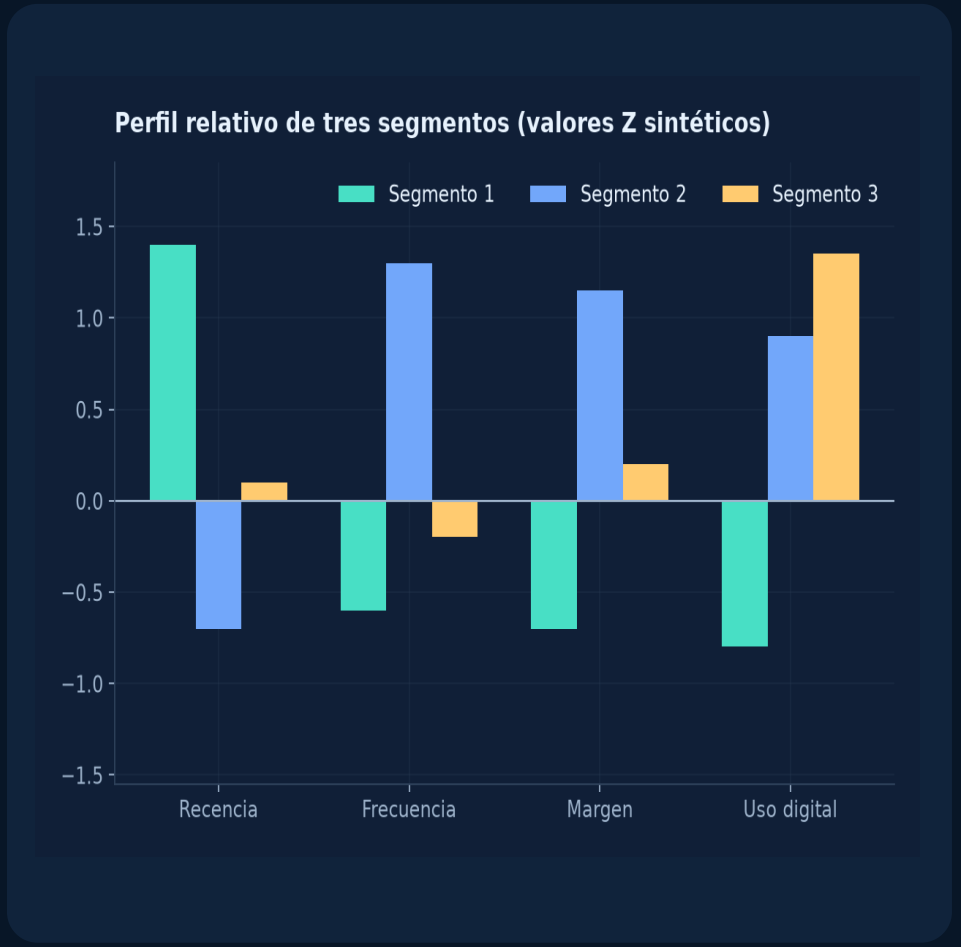


ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Un segmento que desaparece con cambios menores necesita cautela antes de activarlo.**

05 · De grupos a acciones

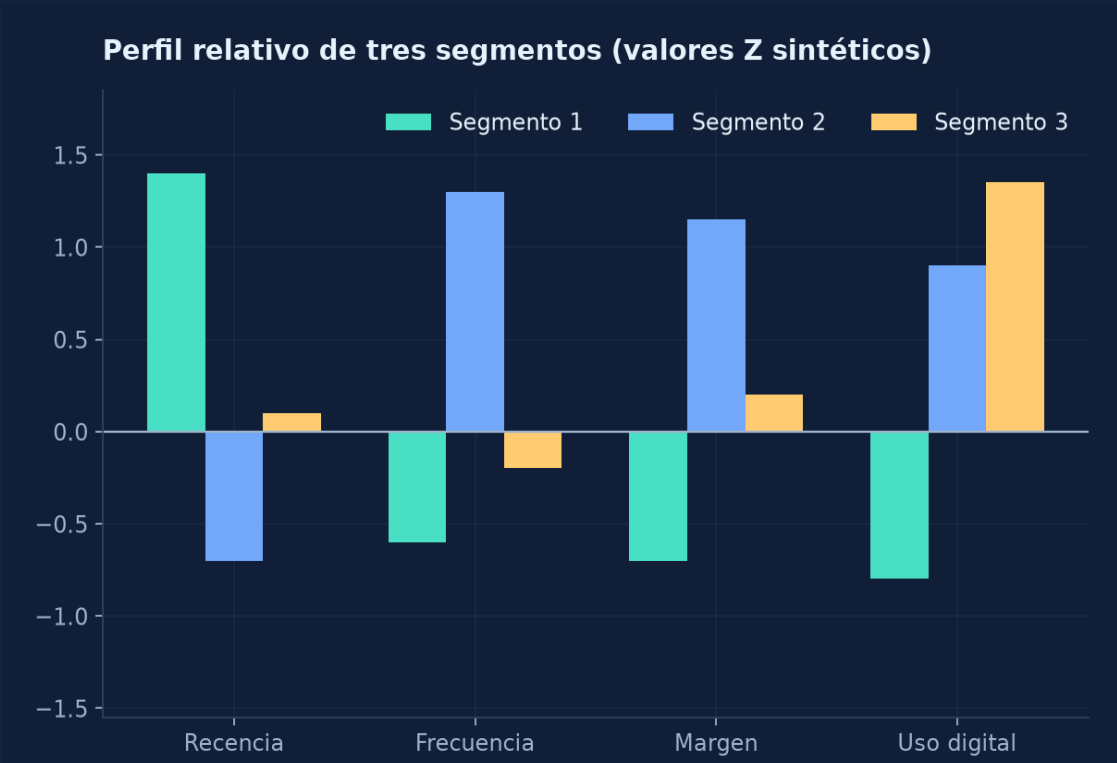
La segmentación se vuelve valiosa cuando cada perfil permite diseñar y evaluar una intervención.



Perfilado: volver de los números Z al cliente real

IDEAS CLAVE

- Calcular distribución de frecuencia, margen, recencia y canal por segmento.
- Comparar tamaño, ingreso, dispersión y riesgo de cada grupo.
- Traducir patrones a hipótesis, evitando estereotipos como “fiel” sin evidencia.
- Asignar nombres operativos solo después de validar el perfil.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **El perfil debe explicarse con unidades originales y dispersión interna.**

Diseñar acciones y testearlas

IDEAS CLAVE

- Ejemplo S1: reconexión con clientes de recencia alta y margen bajo.
- Ejemplo S2: fidelización y beneficios para frecuencia/margen altos.
- Ejemplo S3: activación digital sin otorgar descuentos indiscriminados.
- Probar campañas con grupos de control; segmentar no implica causalidad.

COMPARACIÓN

S1 · PROTEGER	Recuperación controlada / costo por contacto.
S2 · FIDELIZAR	Beneficios basados en margen incremental.
S3 · ACTIVAR	Onboarding digital y adopción de canal.
MEDIR	Lift incremental y ROI neto por segmento.

IDEA QUE DEBE QUEDAR **Una campaña segmentada debe demostrar valor incremental contra una alternativa.**

Asignación de nuevos clientes y monitoreo

IDEAS CLAVE

- Guardar escalador, variables, centros, semillas y fecha de entrenamiento.
- Calcular distancias a centros sin recalcular todo el modelo en cada predicción.
- Vigilar drift de variables, tamaño y distribución de grupos.
- Definir una política de refresco y de intervención ante anomalías.

SECUENCIA DE ANÁLISIS

1. Datos nuevos con contrato de calidad
2. Transformación fija / mismo orden
3. Asignación al centro más cercano
4. Regla de acción elegible
5. **Experimento / monitoreo / reentrenamiento**

IDEA QUE DEBE QUEDAR **El cluster se mantiene como instrumento operativo, no como etiqueta permanente.**

Simulaciones K-Means y pruebas de hipótesis

IDEAS CLAVE

- S05: arrastrar centros y observar la asignación en vivo.
- S06: ejecutar Lloyd paso a paso y ver curva de inercia.
- S07: comparar k-means++, reinicios y semillas.
- S08: comparar K mediante codo, silhouette y estabilidad.
- S09: editar escala, forma, outliers y ver fallas geométricas.

SECUENCIA DE ANÁLISIS

S05 · Asignar / actualizar

↓

S06 · Iteración / convergencia

↓

S07 · Inicio / mínimos locales

↓

S08 · K / calidad / estabilidad

↓

S09 · Fallas / perturbaciones

IDEA QUE DEBE QUEDAR **En cada simulación el alumno debe describir qué cambió y por qué.**

Prueba de transferencia: segmentar y defender

IDEAS CLAVE

- El codo sugiere K=4, pero silhouette prefiere K=3.
- Un segmento estratégico con 2 % de clientes aparece solo con K=5.
- Al cambiar una semilla, 25 % de clientes cambia de grupo.
- Argumentar un plan de selección y validación antes de activar campañas.

PREGUNTAS DE DECISIÓN

1. ¿Qué métricas compararías para K?
2. ¿Cómo proteger el segmento minoritario?
3. ¿Qué harías con la inestabilidad?
4. ¿Qué experimento prueba valor económico?

IDEA QUE DEBE QUEDAR **El mejor K equilibra estructura, estabilidad, interpretación y decisión.**

Errores frecuentes y bibliografía

IDEAS CLAVE

- Confundir K-NN (supervisado) con K-Means (no supervisado).
- Leer etiquetas de cluster como ranking o como diagnóstico causal.
- MacQueen (1967); Arthur & Vassilvitskii (2007), k-means++.
- Hastie, Tibshirani & Friedman (2009); scikit-learn User Guide: Clustering.

COMPARACIÓN

SUPUESTO	La distancia euclídea al cuadrado importa.
CONTROL	Semilla / escala / forma / outliers.
EVIDENCIA	SSE + silhouette + estabilidad.
NEGOCIO	Grupos con intervención evaluable.

IDEA QUE DEBE QUEDAR **La segmentación no es una explicación causal del comportamiento del cliente.**