Laboratorio 2 - Agrupación

Amalia Carbonell
Nicolás Arango
Mateo Rincón

Agenda

- Preparación de datos
- Algoritmos
- Análisis y conclusión

Preparación de datos

Preprocesamiento

Limpieza y escalamiento

Valores duplicados, fuera de rango, invalidos y escalar las variables para mejorar la coherencia de los clusters y evitar segmentación.

Preprocesamiento

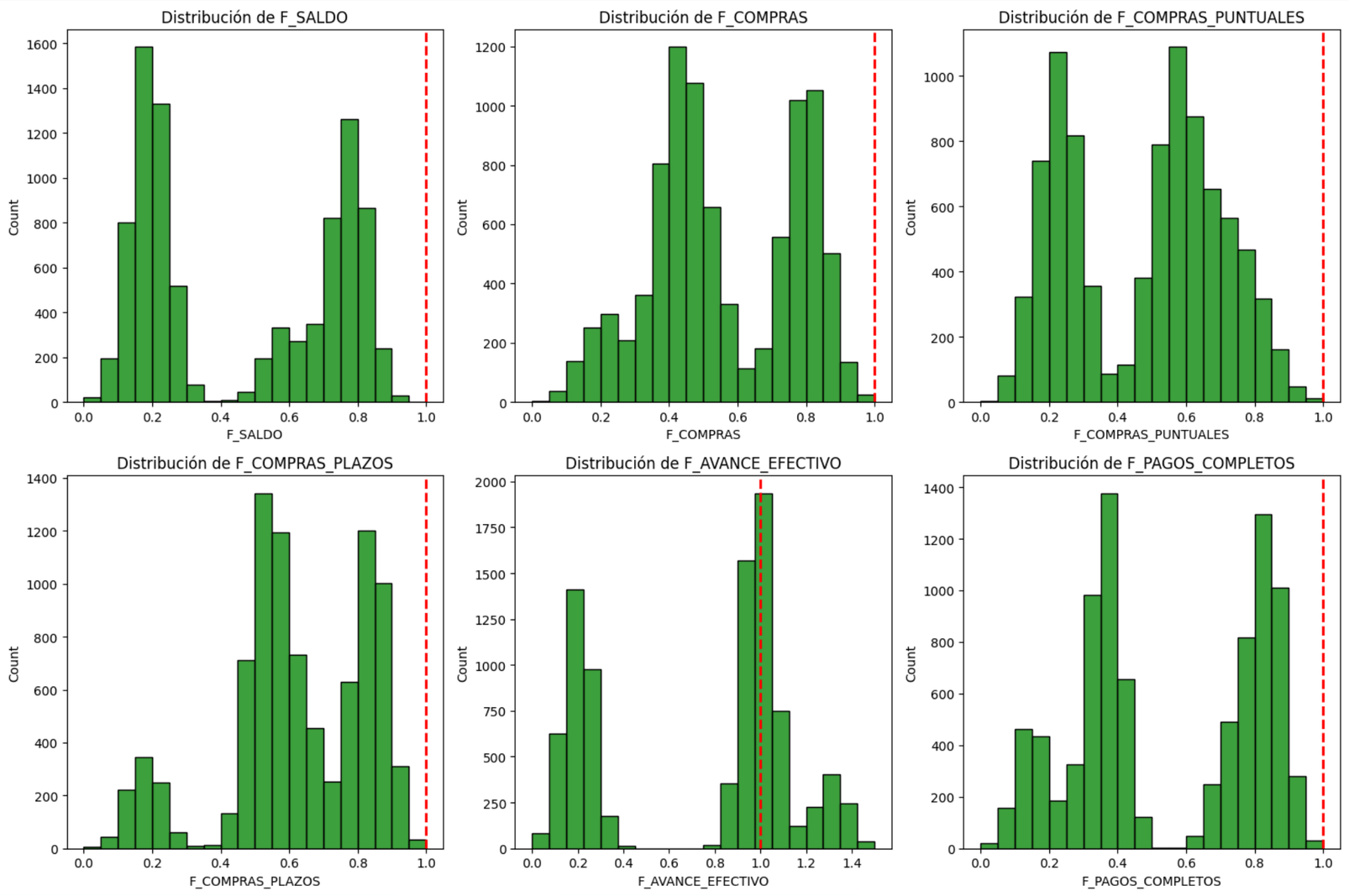
	Atributo	Completitud (%)
0	PAGOS_MINIMOS	96.502793
1	LÍMITE_CREDITO	99.988827
2	SALDO	100.000000
3	ID	100.000000
4	F_SALDO	100.000000
5	COMPRAS	100.000000
6	AVANCE_EFECTIVO	100.000000
7	F_COMPRAS	100.000000
8	COMPRAS_PUNTUALES	100.000000
9	COMPRAS_PLAZOS	100.000000
10	F_COMPRAS_PLAZOS	100.000000
11	F_COMPRAS_PUNTUALES	100.000000
12	P_AVANCE_EFECTIVO	100.000000
13	F_AVANCE_EFECTIVO	100.000000
14	P_COMPRAS	100.000000
15	PAGOS	100.000000
16	F_PAGOS_COMPLETOS	100.000000
17	MESES_CLIENTE	100.000000

Analisis de completitud:
Se reemplazaron los valores faltantes con la mediana de cada uno de los datos

Analisis de Unicidad:
No se encontraron datos duplicados

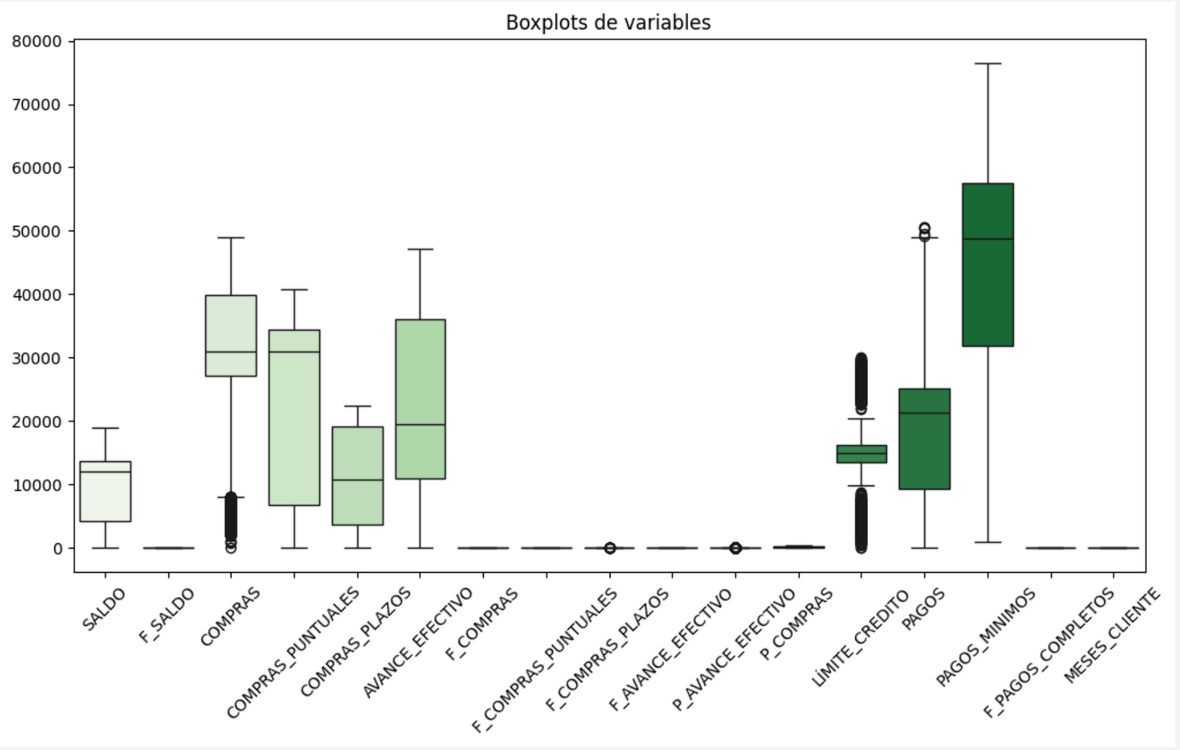
```
Cantidad de filas duplicadas: 0
Porcentaje de filas duplicadas: 0.0000%
```

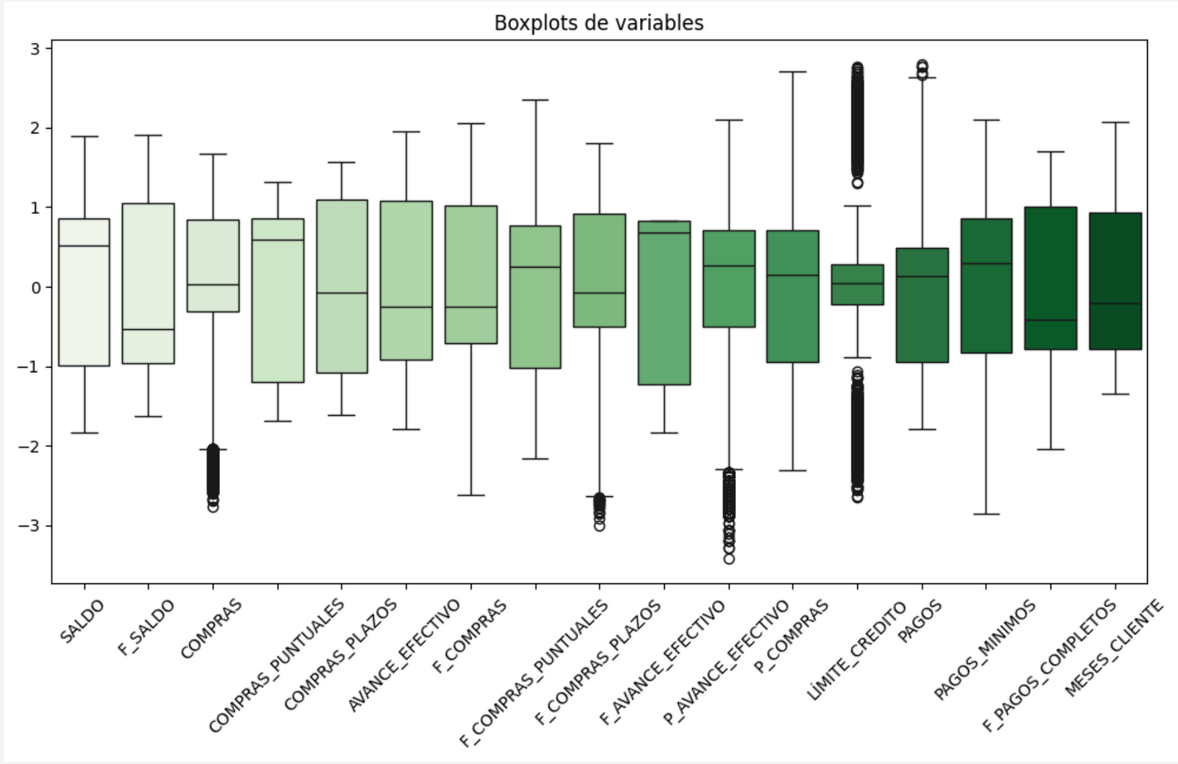
Preprocesamiento



Analisis de Validez: Se encontraron 3040 datos por fuera del rango. Se reemplazaron con el valor maximo posible en el rango.

Preprocesamiento





Escalamiento de datos:

Se escalarán los datos para asegurar que todas las variables tengan la misma influencia en el análisis, evitando que aquellas con valores más grandes dominen el clustering.

Algoritmos

-K-means
-DBSCAN
-MeanShift

K-means

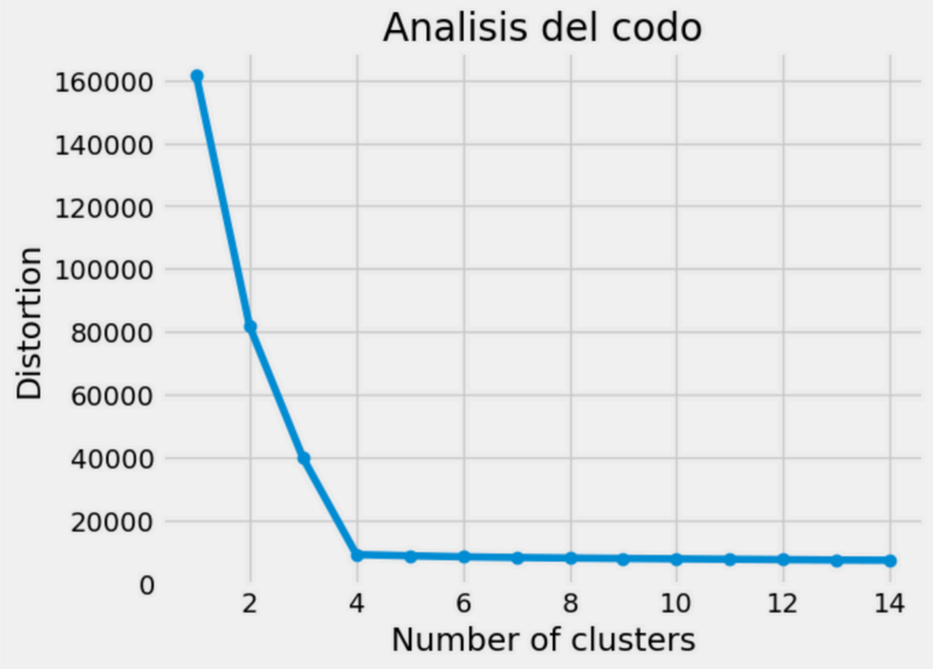
¿Por qué usar K-Means?

Simplicidad y eficiencia: K-Means es un algoritmo rápido y fácil de implementar, ideal para grandes volúmenes de datos.

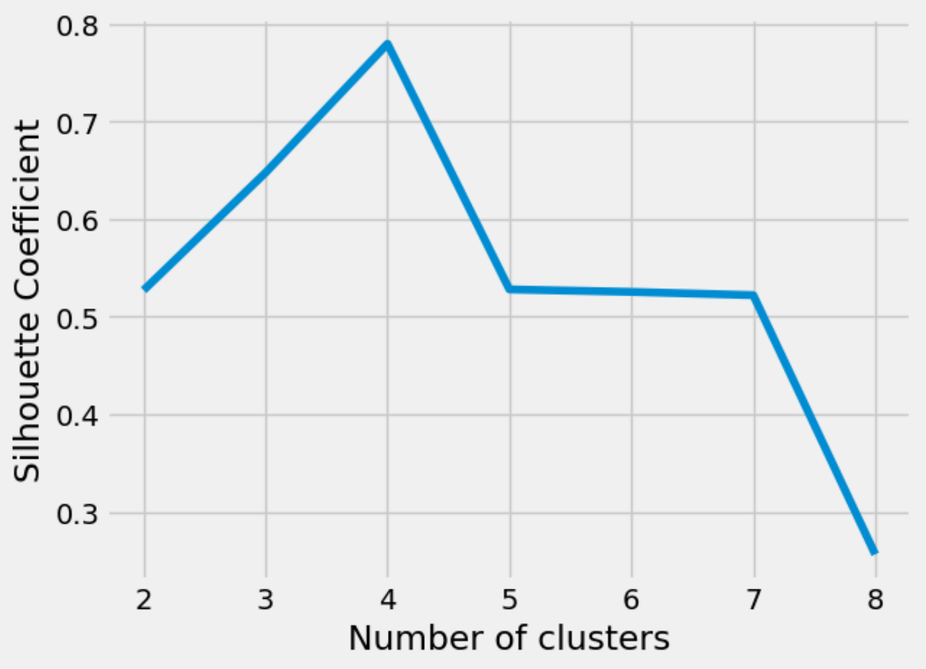
Interpretabilidad: Genera clusters bien definidos con centroides claros, lo que facilita la interpretación de los resultados.

Flexibilidad en distintas aplicaciones: Es ampliamente utilizado en segmentación de clientes, compresión de datos y reconocimiento de patrones.

Determinacion numero de clusters

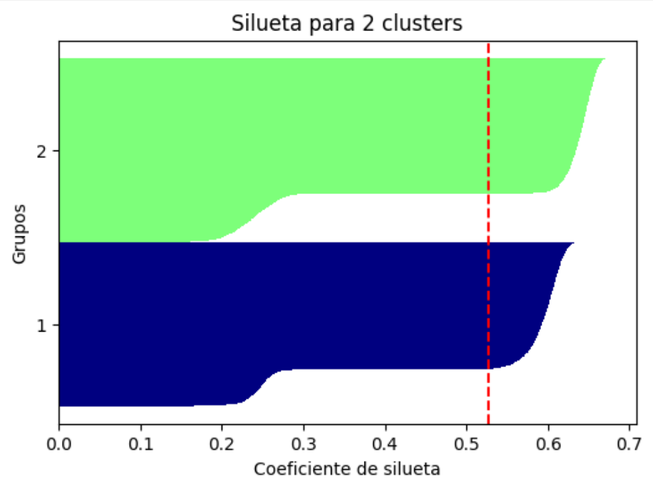


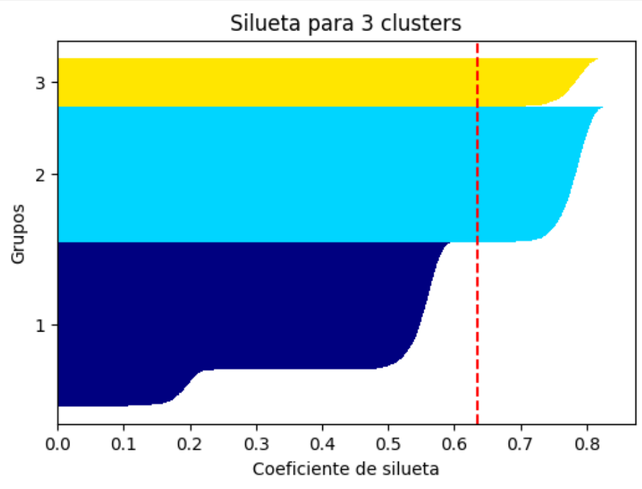
El método del codo se utilizó para identificar el número óptimo de clusters, determinando el punto donde la inercia se estabiliza, evitando una segmentación excesiva.

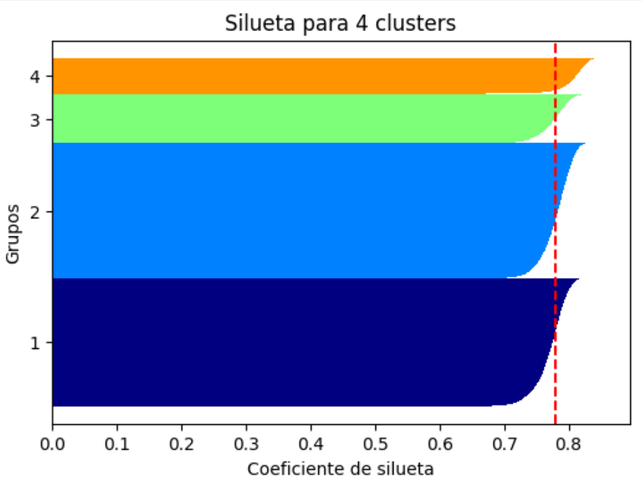


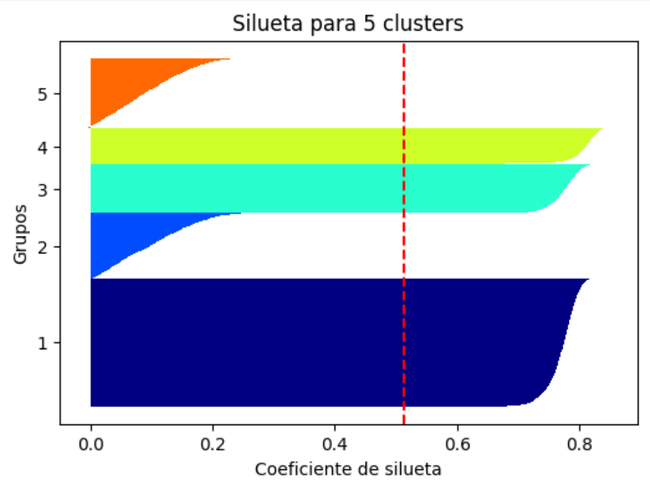
El método del coeficiente de silueta se utilizó para evaluar la calidad de la agrupación, mostrando gráficamente el coeficiente de silueta que permite medir qué tan bien separados y compactos están los clusters.

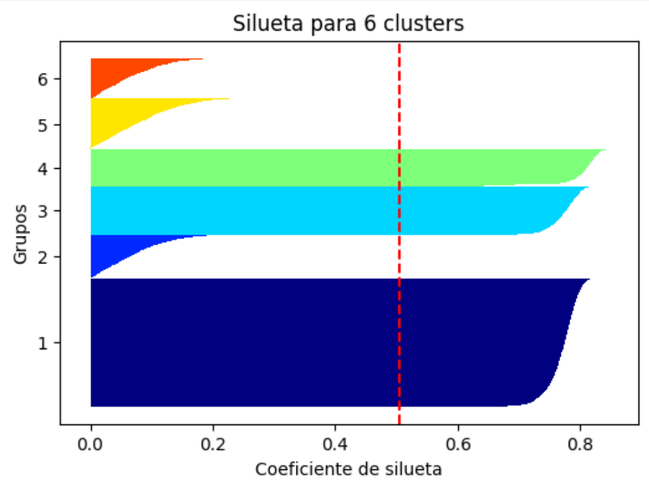
Graficas de silueta





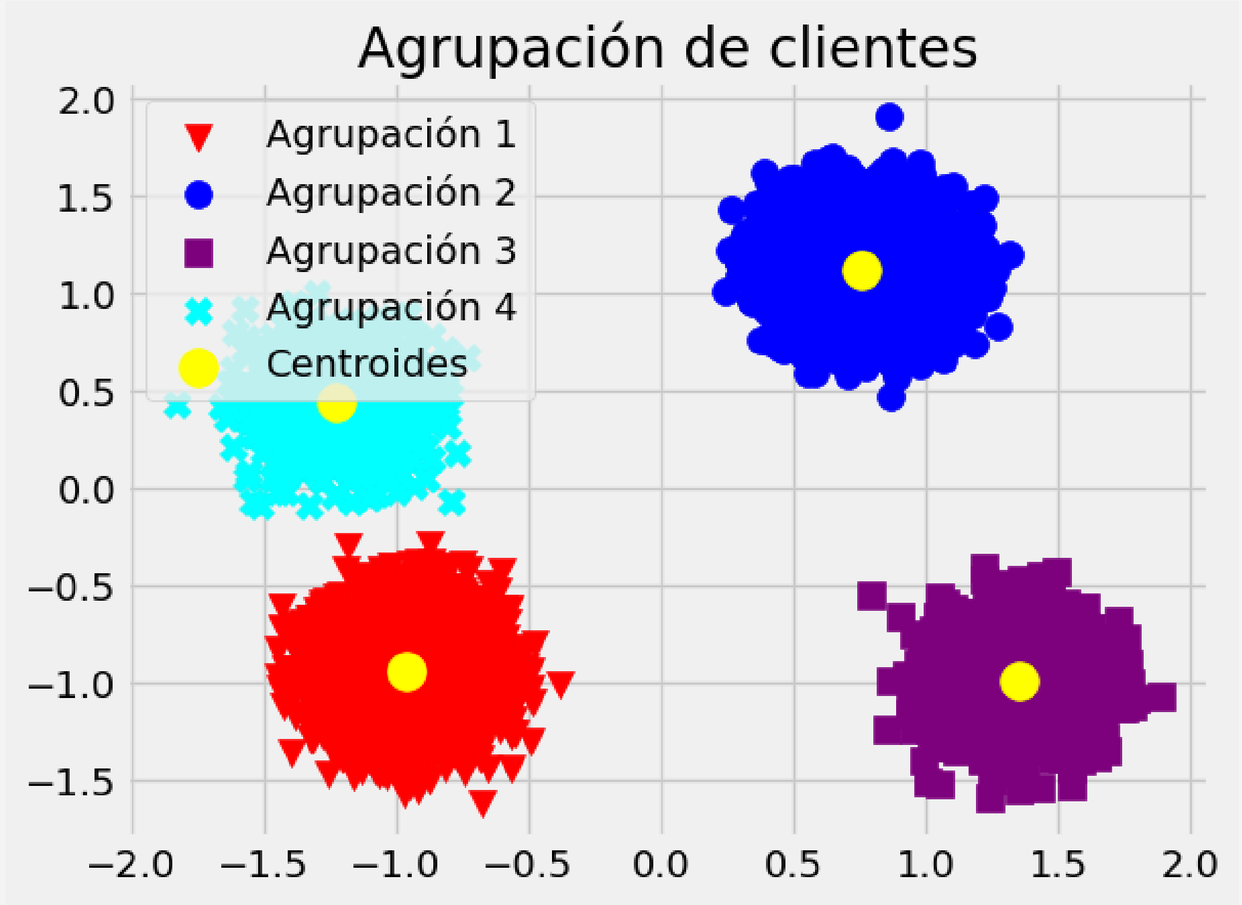




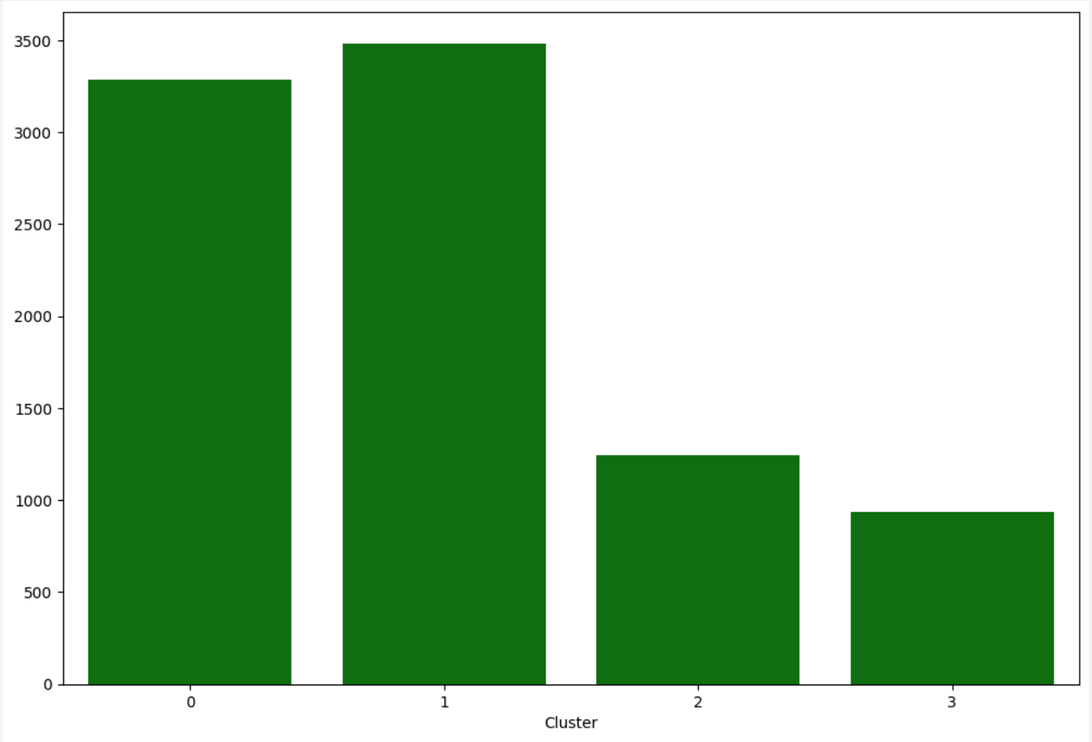


La gráfica de siluetas se utilizó para evaluar el número óptimo de clusters, analizando la cohesión y separación de cada grupo a través de la distribución del coeficiente de silueta.

Visualizácion de los datos



4 clusters armados



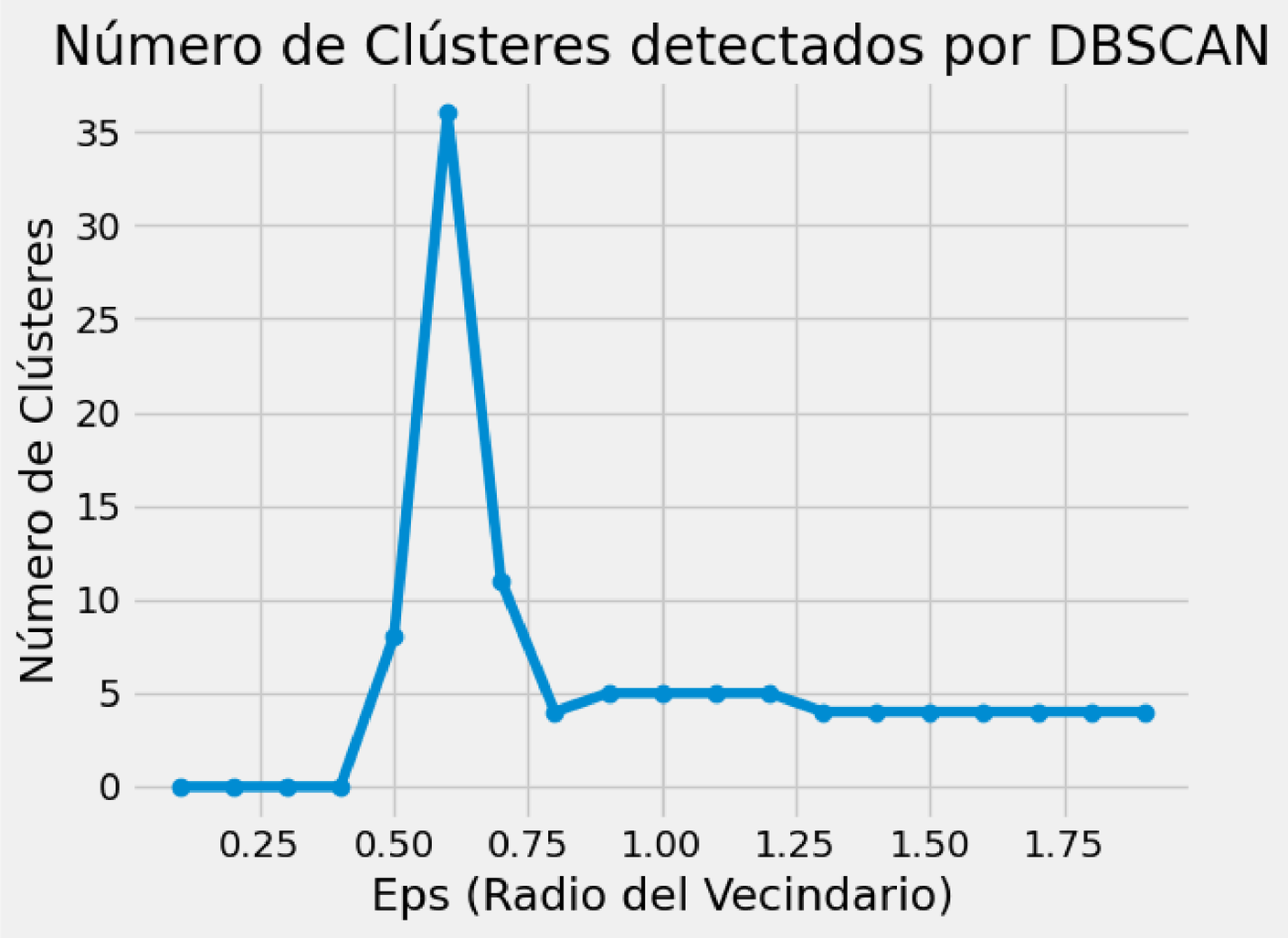
Conteo de datos por cluster

DBSCAN

¿Por qué usar DBSCAN

- Manejo de ruido y outliers: DBSCAN clasifica automáticamente los puntos aislados como ruido, lo que ayuda a reducir el impacto de valores atípicos en el análisis.
- Detección de clusters automática: DBSCAN no recibe por parámetro el número de grupos esperados cómo si lo hace K-means
- Escalabilidad: DBSCAN es un algoritmo oportuno para agrupar muchos datos en un numero mediano de grupos.

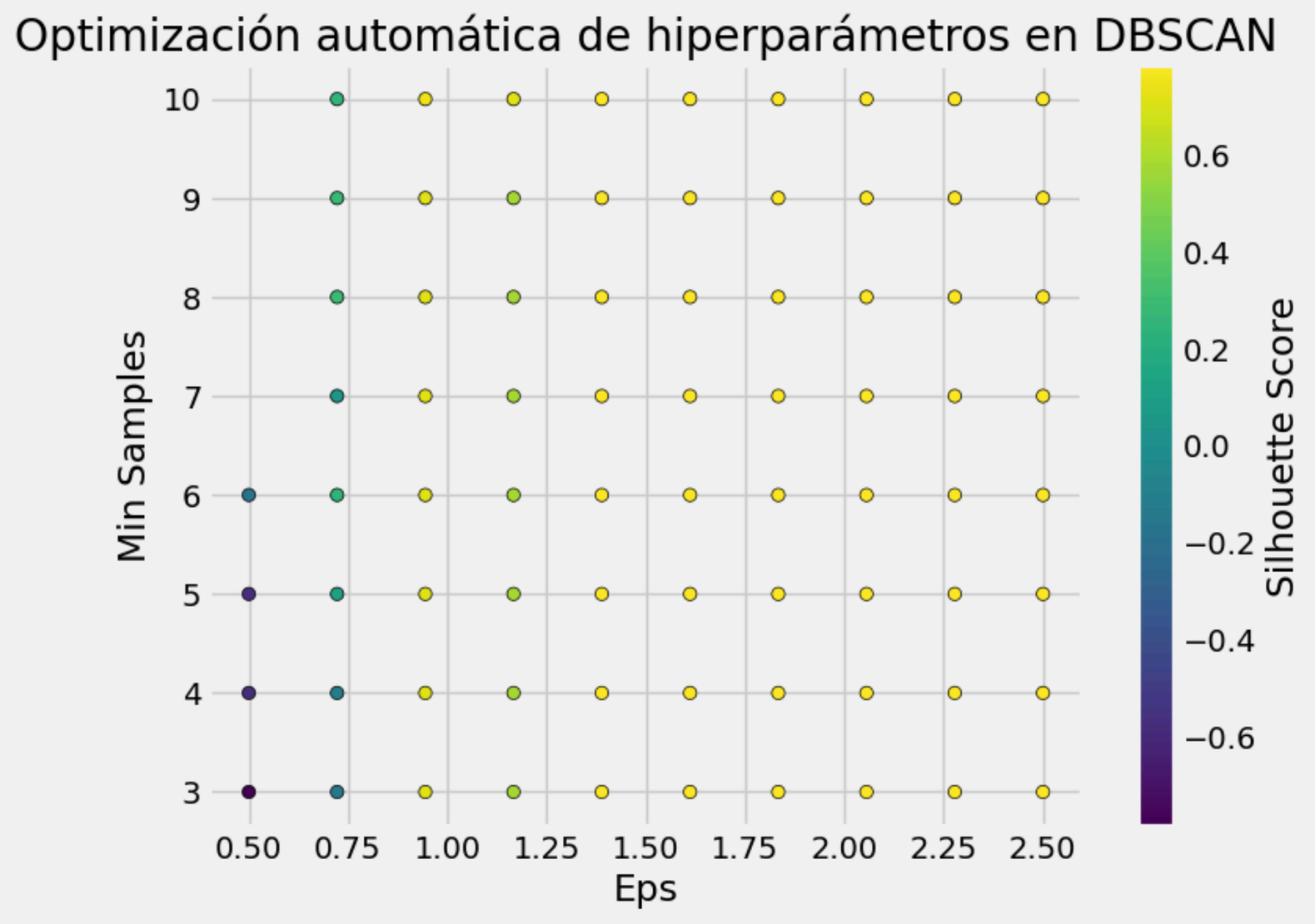
Punto de codo



- Para valores de Eps menores a 0.5 se encuentran 0 grupos ya que el radio es muy pequeño

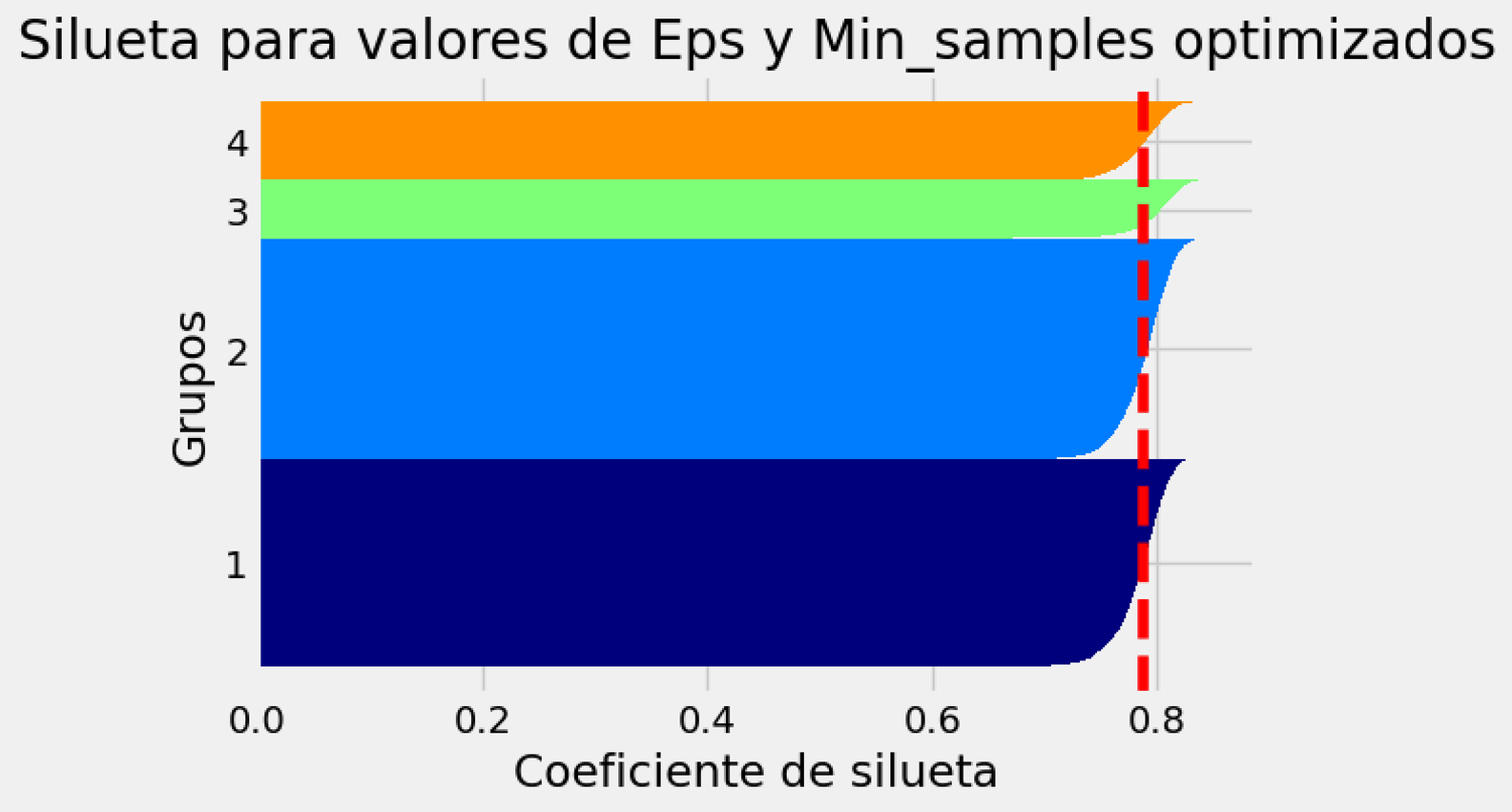
- El pico ocurre cuando Eps=35, a partir de este punto el número de grupos empieza a disminuir y se estabiliza cuando EPS=0.75

Modelo óptimo



- Se usa Silhouette Score como métrica para encontrar el mejor modelo
- Hiperparámetros:
Eps: tamaño del radio
Min samples: mínimo de puntos en una vecindad
- Hiperparámetros óptimos:
Eps: 1.3888888888888888
Min samples: 3

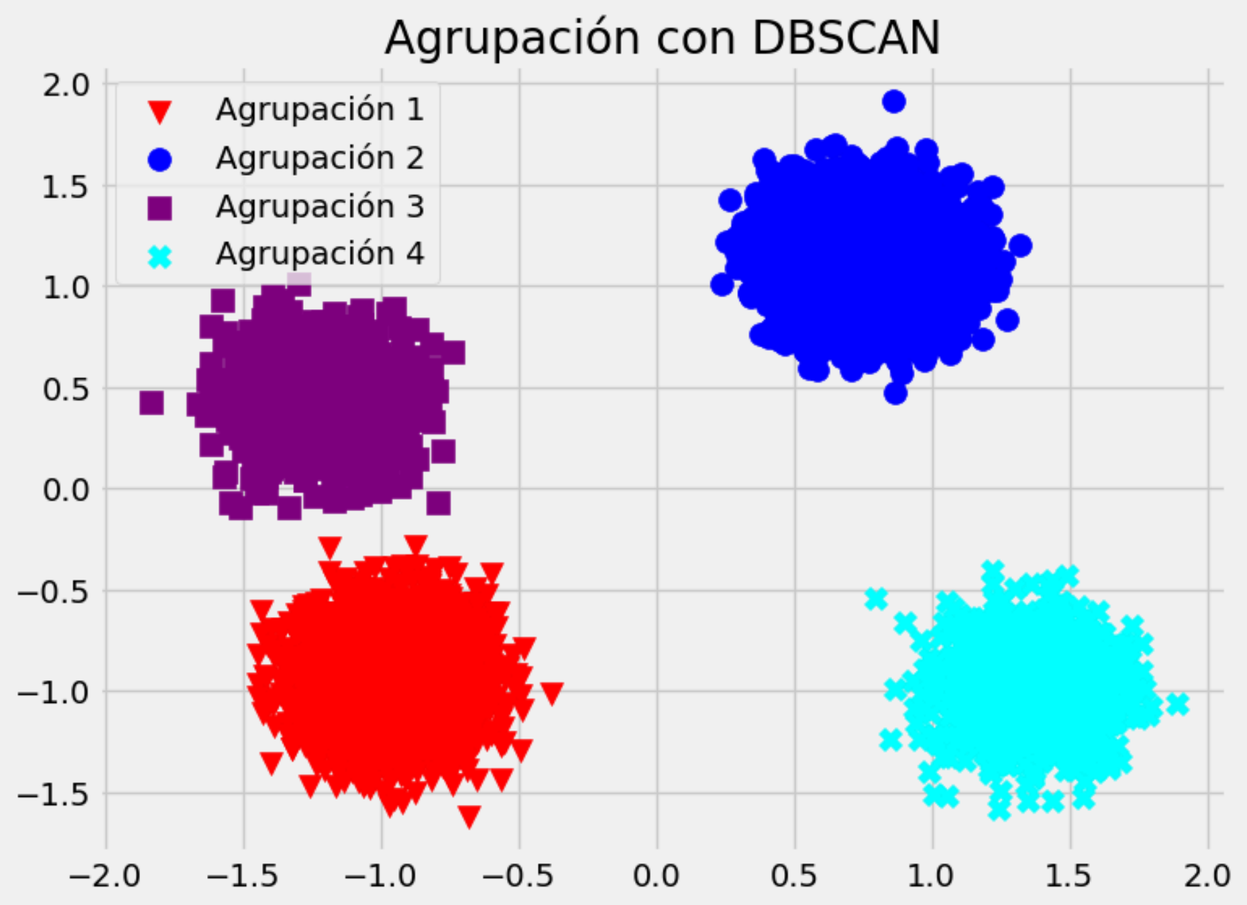
Silueta para el modelo optimizado



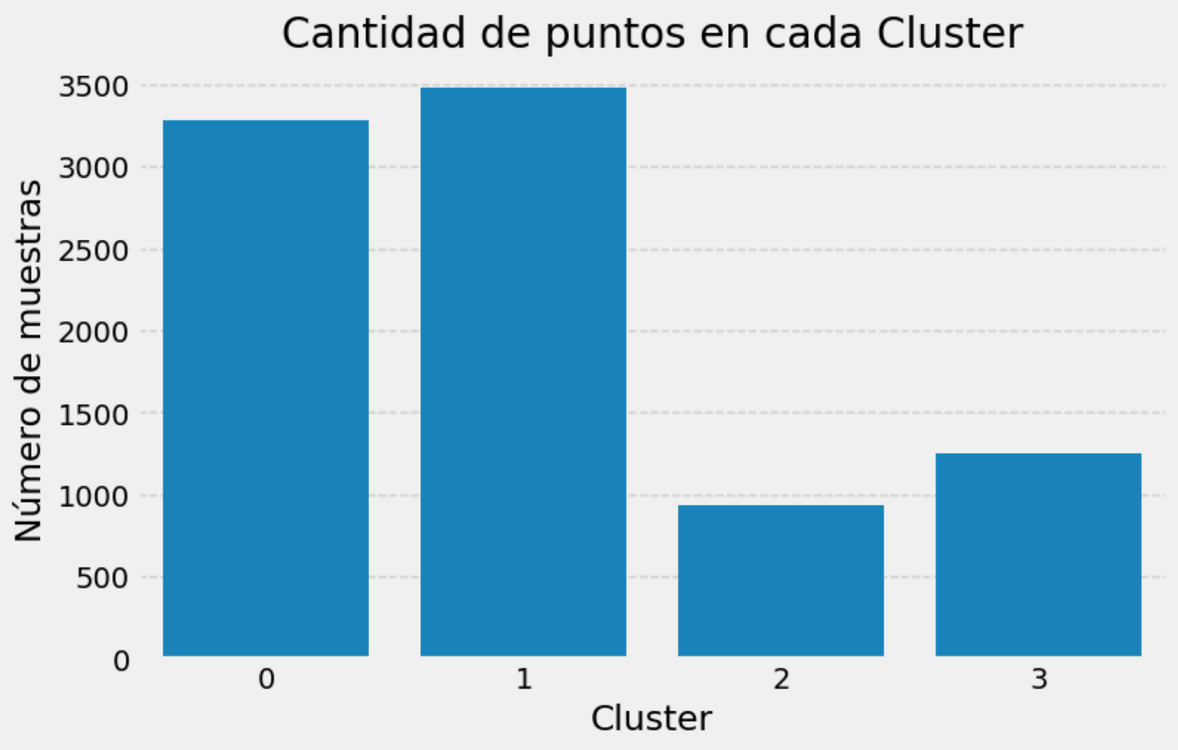
- Se usa Silhouette Score como métrica para encontrar el mejor modelo
- 4 grupos encontrados
- Silhouette Score: 0.788

Visualización de los datos

- 4 grupos encontrados
- *DBSCAN no usa centroides



- Conteo de datos por grupo



MeanShift

Volver a la página de agenda

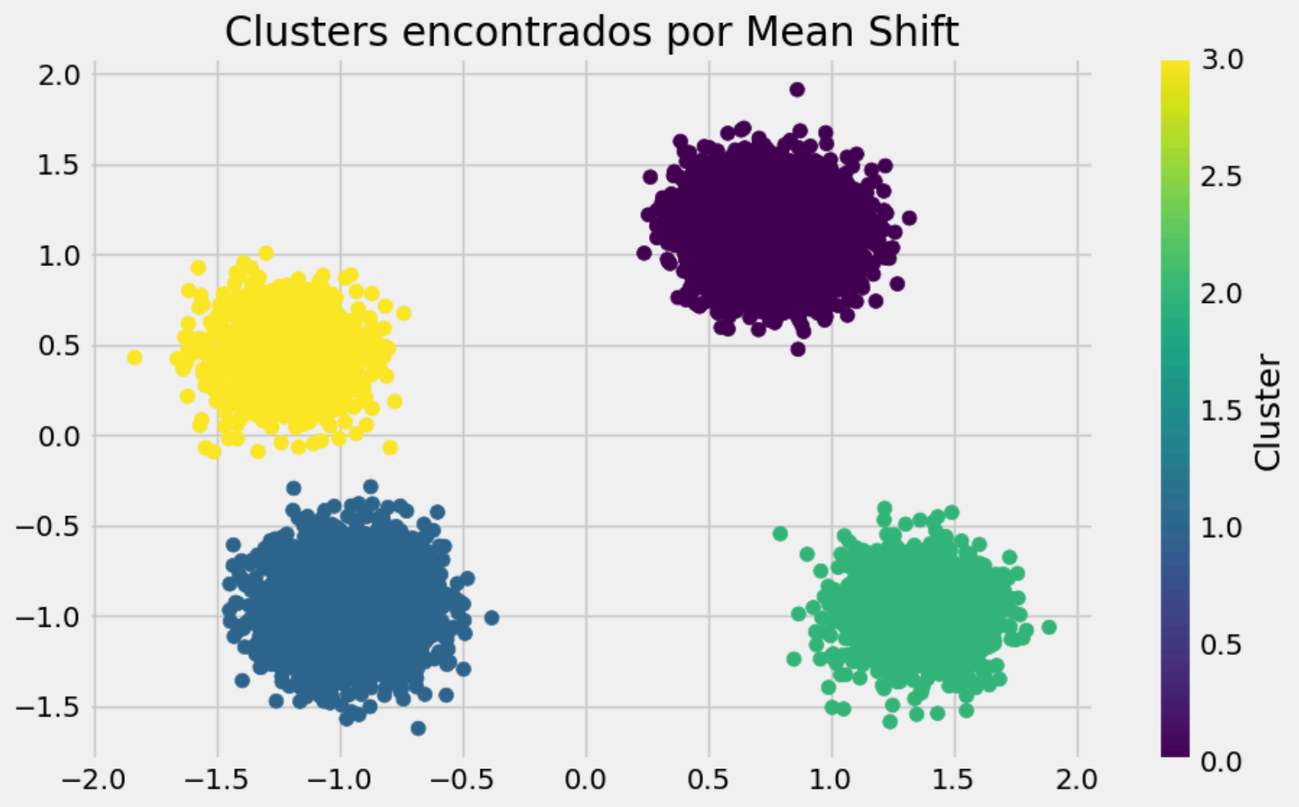
Porque usar MeanShift?

Detección automática de clusters: No requiere especificar el número de clusters previamente

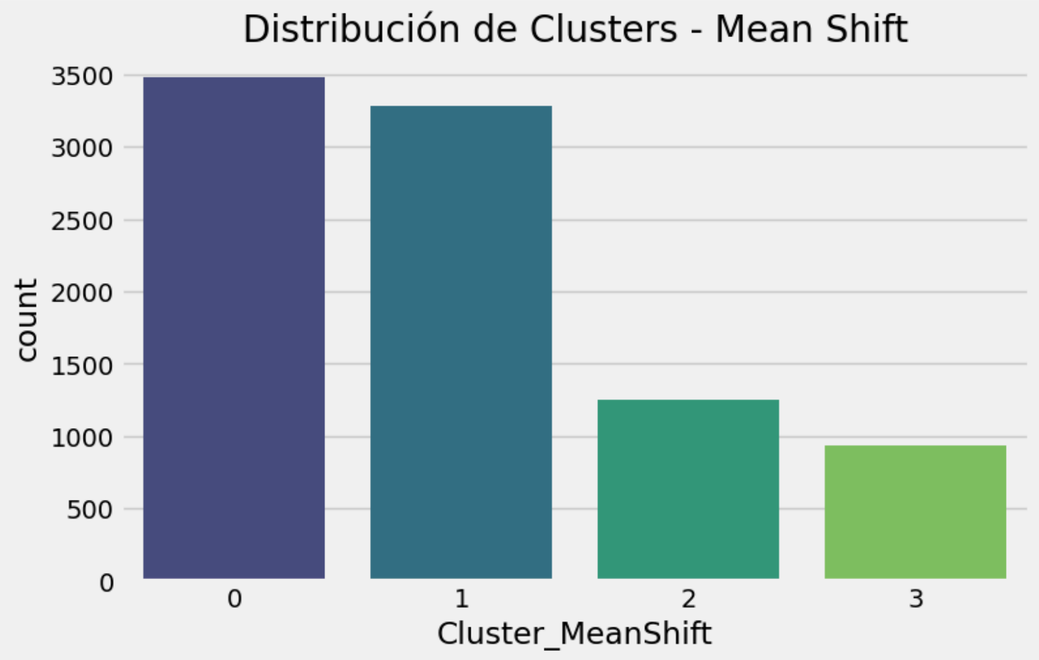
Adaptabilidad a formas complejas: Puede identificar clusters de diferentes formas y tamaños sin asumir una estructura esférica.

No depende de la inicialización: No es sensible a la selección inicial de centroides, lo que mejora la estabilidad de los resultados.

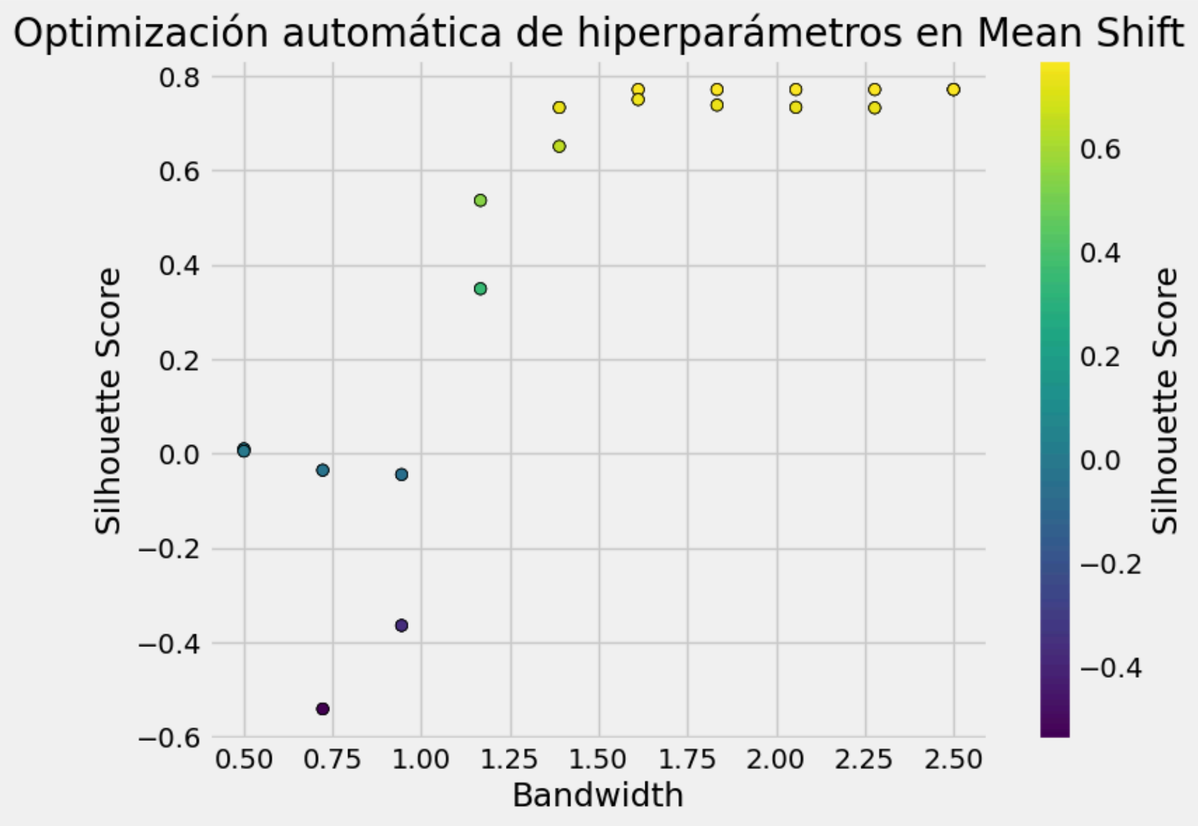
Visualización de los datos antes de optimizar



- 4 grupos

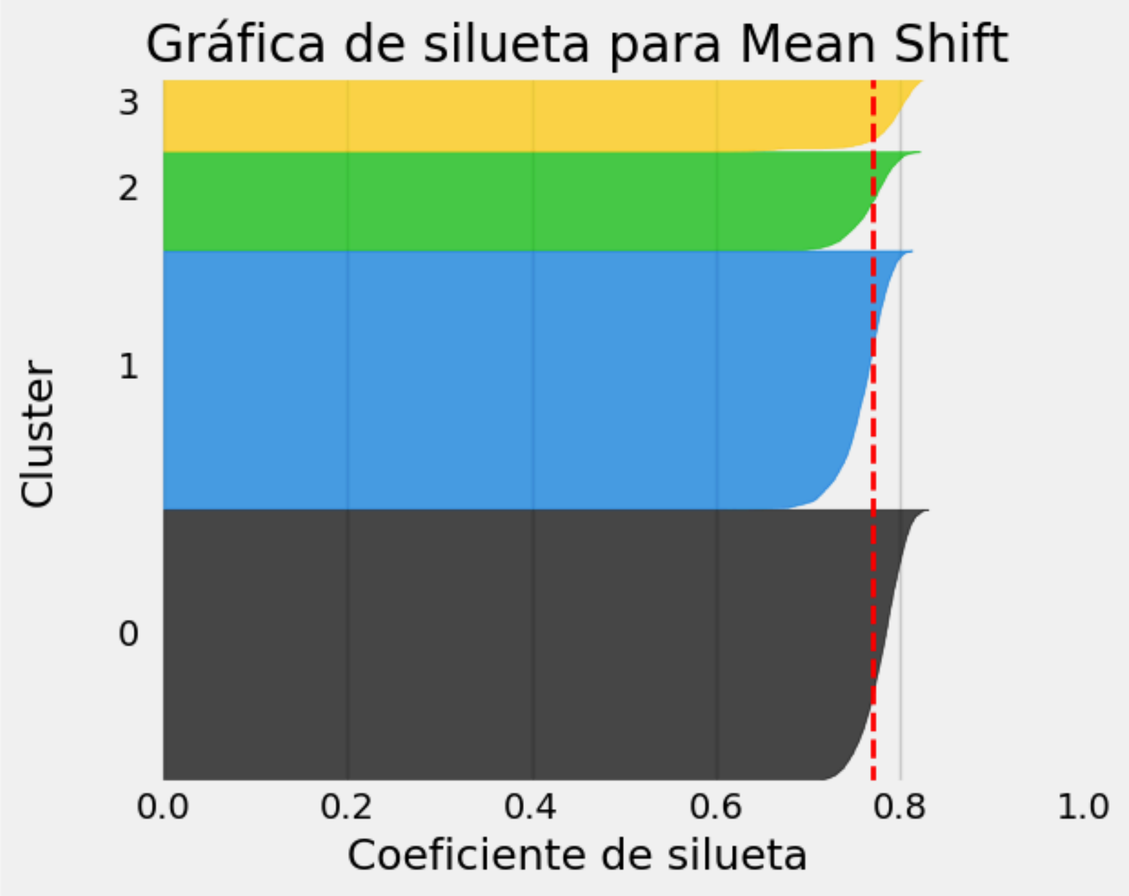


Modelo Óptimo



- Radio de Búsqueda (Bandwidth):1.611
- Número mínimo de puntos que debe tener un "bin": 1
- Todos los puntos se deben asignar a un clúster

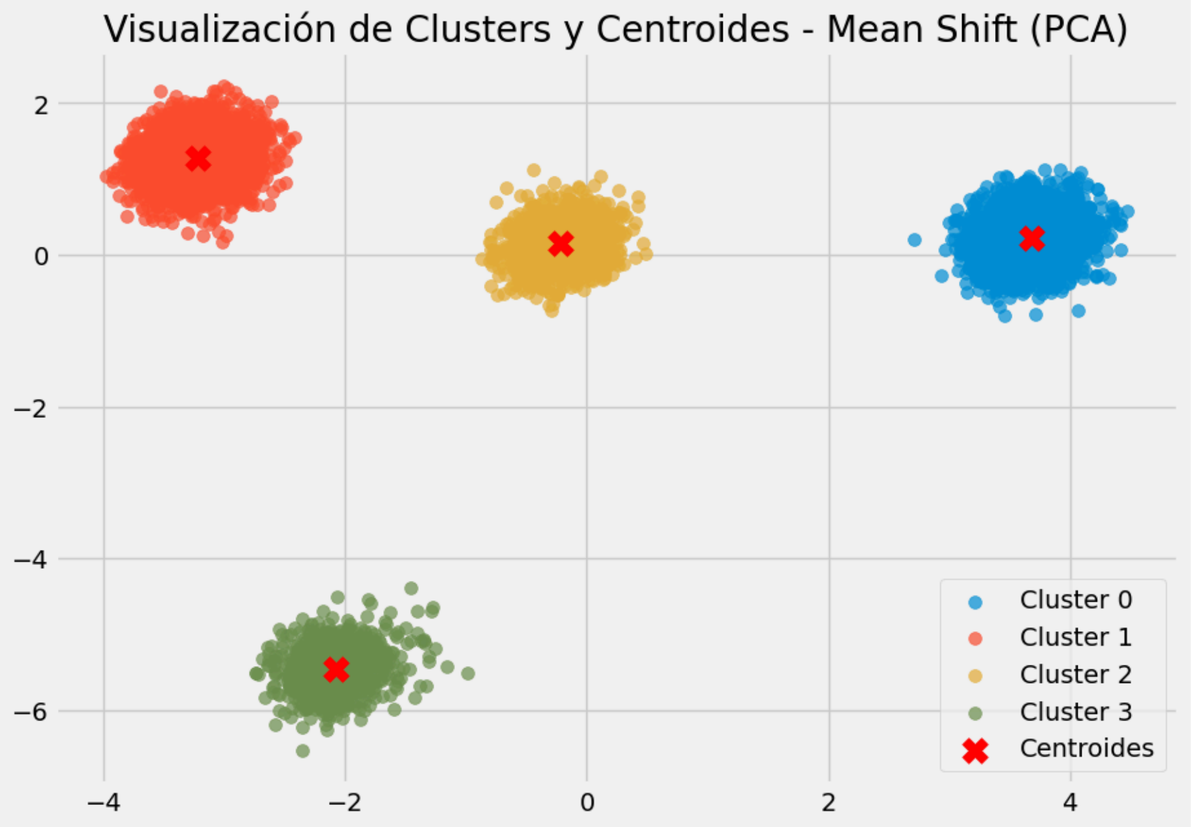
Silueta para el modelo optimizado



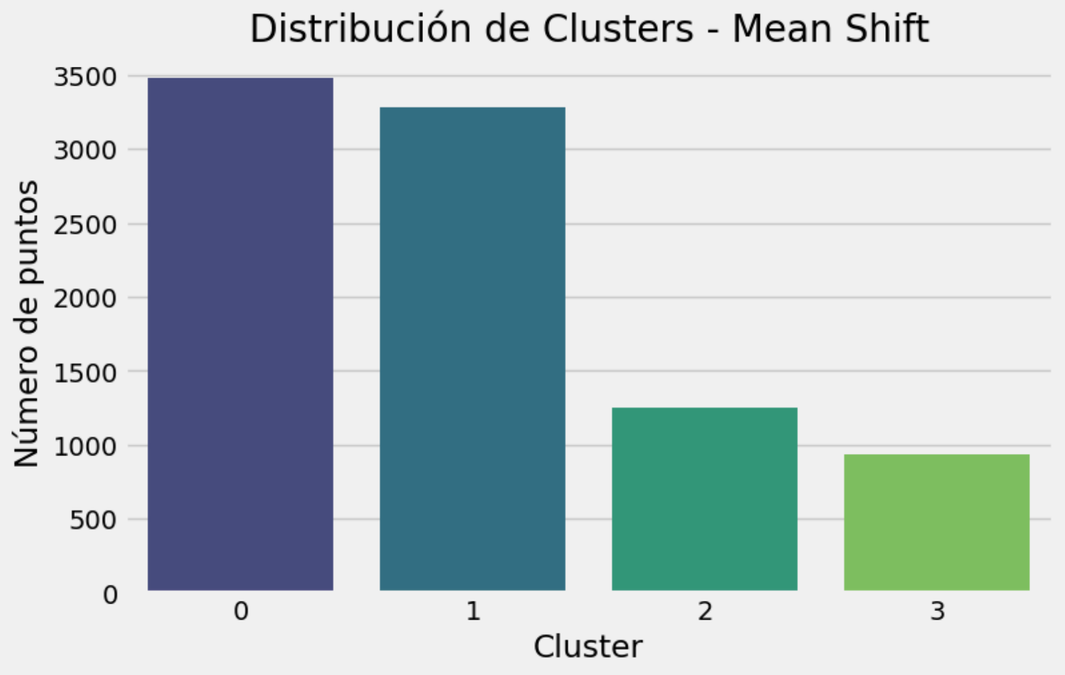
- Se usa Silhouette Score como métrica para encontrar el mejor modelo
- 4 grupos encontrados
- Silhouette Score: 0.771

Visualización de los datos

- Es muy sensible a los hiperparámetros
- MeanShift calcula los cntroides cómo la media de los puntos en el clúster



- No cambió el conteo de datos por grupo



Conclusiones

Tabla comparativa

	Numero de clusters	Silhouete Score
K Means	4	0,780
DBSCAN	4	0,788
Mean Shift	4	0,771

Algoritmo escogido

Escogimos **DBSCAN** porque permite identificar patrones sin necesidad de definir un número fijo de clusters, maneja bien los outliers y obtuvo un coeficiente de silueta de **0.788**, lo que indica una segmentación clara. Esto ayudará a FinanzasAlpes a personalizar sus estrategias de marketing y mejorar la experiencia del cliente.

El algoritmo encontró **4 clusters**, lo que indica la presencia de distintos perfiles de clientes con comportamientos de compra diferenciados. Esto permitirá a FinanzasAlpes diseñar estrategias de marketing más específicas y mejorar la personalización de sus servicios.

¡Gracias!