



Clustering para caracterizar café – SenecaféAlpes

LAB 2 - BI, Sección 2 Grupo 27

Contenidos

Introducción

Modelo 1 - K-Means

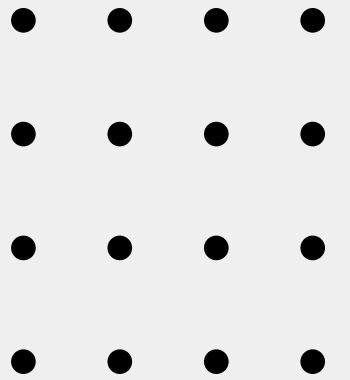
Modelo 2 - DBSCAN

Modelo 3 - Agglomerative Clustering

Modelo escogido

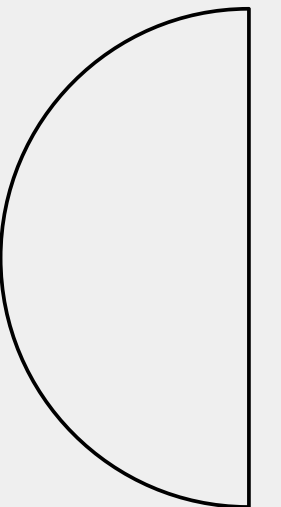
Recomendaciones


USO IA



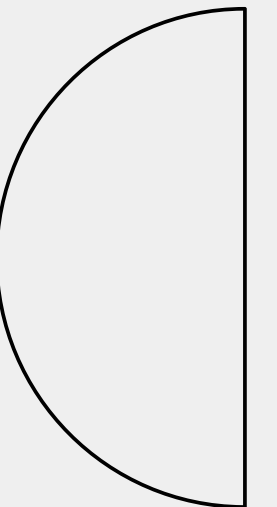
Introducción

- El objetivo de este análisis es explorar patrones en las características morfológicas de los granos de café para apoyar la clasificación y el control de calidad.
- Se aplicaron 3 algoritmos de clustering: K-Means, DBSCAN y Agglomerative Clustering
- Se prepararon los datos y se utilizaron variables físicas de los granos (área, perímetro, solidez, redondez, factores de forma, tipo de secado, relación de aspecto).





Nota sobre el coeficiente de silueta: Es un número que mide qué tan bien quedan separados los grupos (clústeres) entre sí y qué tan compactos son internamente. Sus valores van de -1 a 1: valores cercanos a 1 indican que los grupos están bien definidos; valores cercanos a 0 significan que los grupos se mezclan entre sí; y valores negativos indican que varios datos están mal asignados al grupo. En general, valores superiores a 0.25–0.30 son aceptables y valores mayores a 0.5 se consideran buenos.



Modelo 1 - K-Means

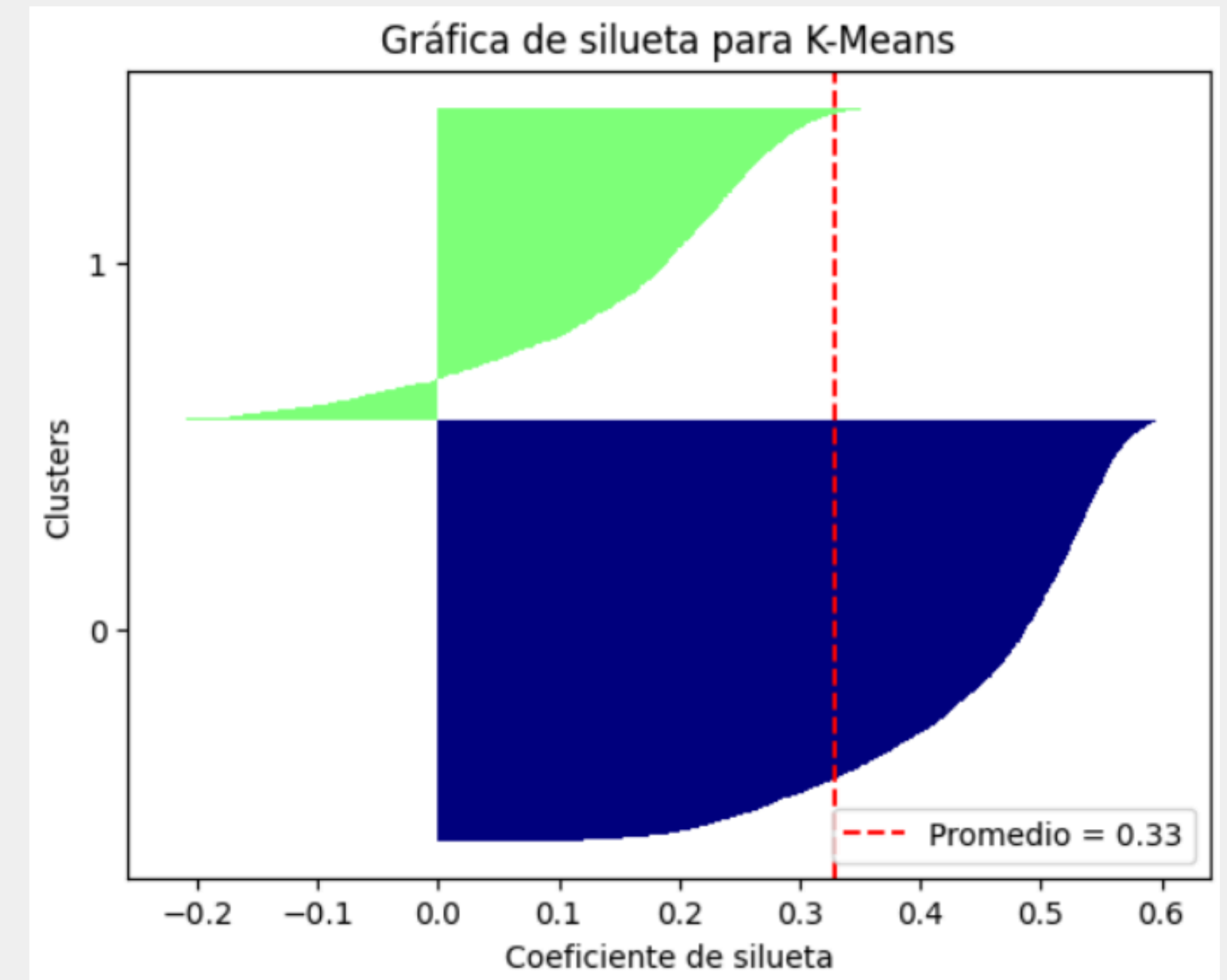
Agrupar los datos en k grupos o clusters según su similitud. Funciona buscando los centros (centroides).

Distribución de registros:

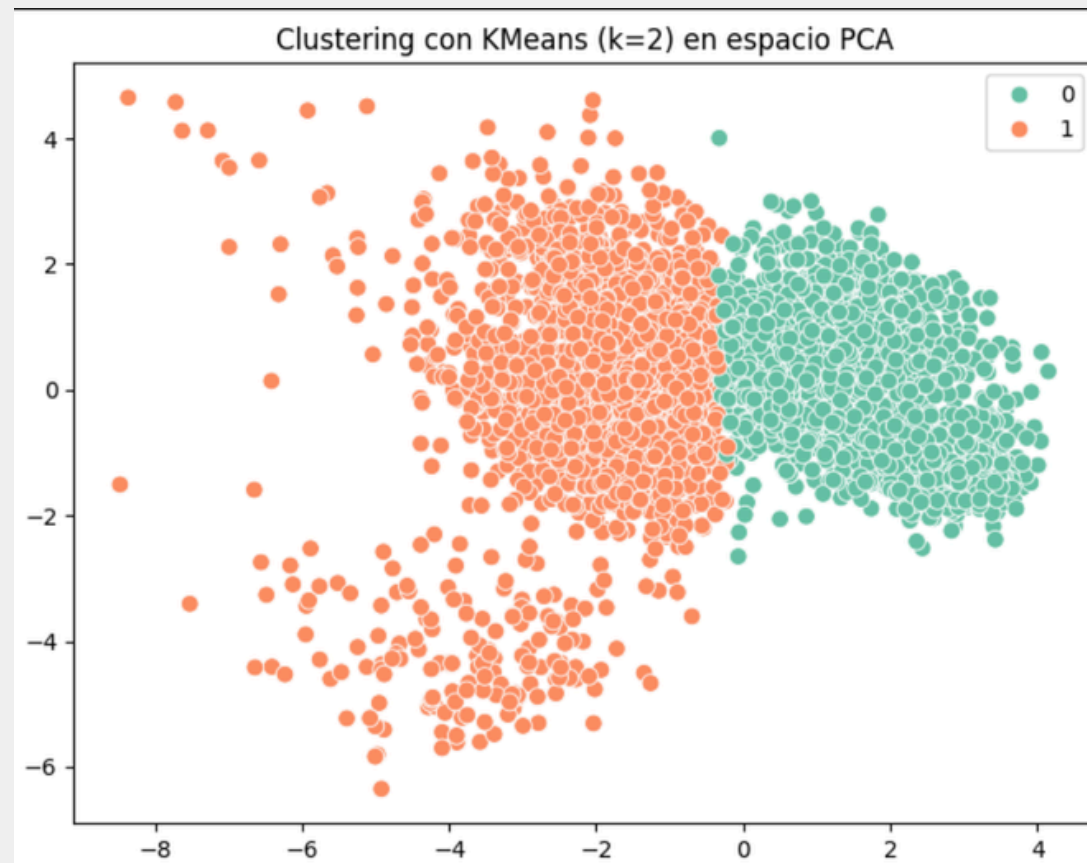
- **Cluster 0:** 2715 granos ($\approx 58\%$ de los datos).
- **Cluster 1:** 2004 granos ($\approx 42\%$).

Evaluación:

- **Coeficiente de silueta promedio:** 0.328, separación moderada entre los grupos. Hay granos que probablemente estén mal agrupados.



Modelo 2 - K-Means



La estructura de los clústeres es más definida que la obtenida con otros modelos, lo que sugiere que el método de partición logra capturar patrones relevantes en los datos.

Cluster 0 – Granos Compactos y Uniformes:

- Áreas más pequeñas (37,805) y perímetros menores (637,101), además de tener
- Valores ligeramente mayores de solidez y redondez, lo que indica granos más compactos y regulares.
- Más granos secados por el método lavado (64.8%) y una proporción más alta de granos con forma redondeada (21%)

Cluster 1 – Granos Grandes e Irregulares:

- Mayor tamaño (área promedio $\approx 73,185$) y perímetros considerablemente más grandes ($\approx 946,668$).
- Valores ligeramente menores en los factores de solidez y redondez. Lo que indica formas irregulares.
- Proporción casi nula de formas redondeadas ($\approx 0.3\%$), sugiriendo que agrupa granos de mayor tamaño pero menos uniformes.

Modelo 2 - DBSCAN

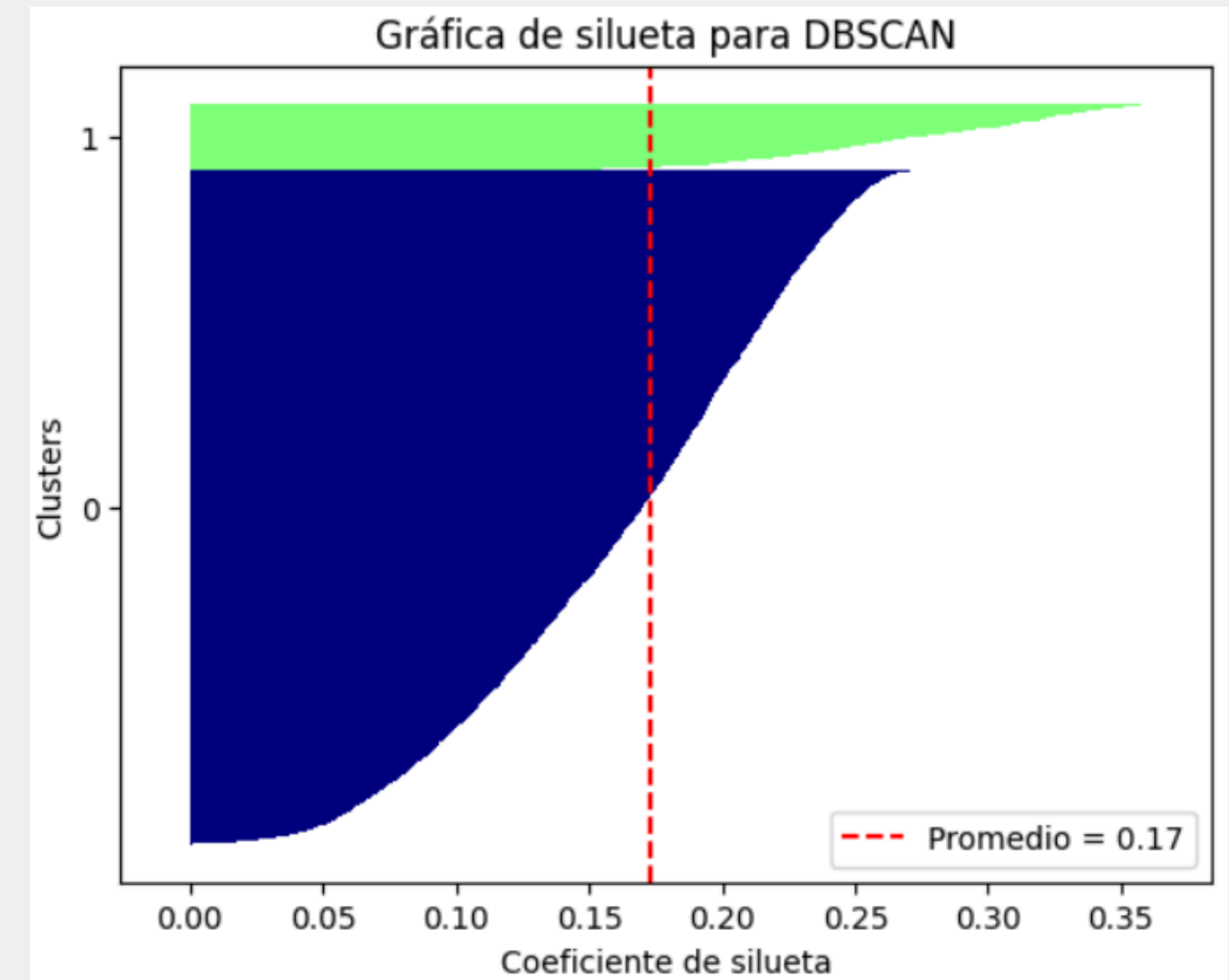
Agrupar datos según densidad y permite detectar muestras atípicas.

Distribución de registros:

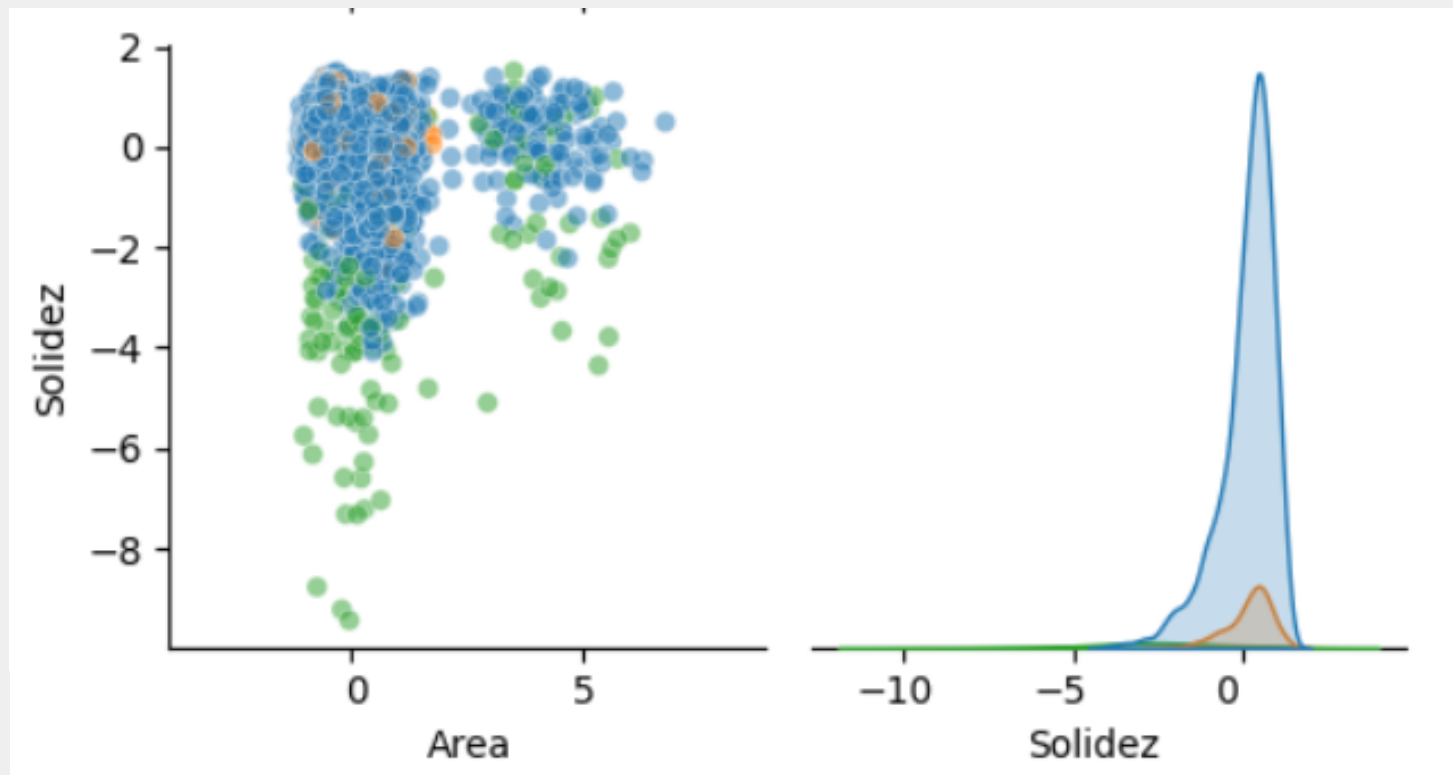
- **Cluster 0:** 4169 granos ($\approx 86\%$ de los datos).
- **Cluster 1:** 412 granos ($\approx 9\%$).
- **Ruido (-1):** 138 granos ($\approx 3\%$), ruido.

Evaluación:

- **Coefficiente de silueta promedio:** 0.173 (excluyendo ruido), lo que sugiere una separación débil entre clústeres.



Modelo 2 - DBSCAN



Se muestra un importante solapamiento entre clusters, indicando que las variables actuales no logran separar completamente los grupos.

Cluster 0 – Granos predominantes:

- Valores intermedios de área y perímetro (área promedio ≈ 52.000 píxeles y perímetro ≈ 847.00)
- Mayor solidez (≈ 0.75) y redondez moderada.
- Predominio del método de secado lavado ($\approx 59.5\%$)
- Bajo porcentaje de granos con forma redondeada ($\approx 12\%$).

Cluster 1 – Granos diferenciados:

- Área y perímetro algo más bajos (≈ 48.000 y 822).
- Alta solidez y redondez (similares a Cluster 0).
- Más granos lavados ($\approx 60.7\%$).
- Leve incremento en la relación de aspecto redondeado ($\approx 13.3\%$).

Modelo 3 - Agglomerative Clustering

Une de forma iterativa las observaciones más cercanas hasta formar el número deseado de clústeres.

Distribución de registros:

- **Cluster 0:** 1673 granos ($\approx 35\%$ de los datos).
- **Cluster 1:** 3046 granos ($\approx 65\%$).

Evaluación:

- **Coeficiente de silueta promedio:** 0.320, lo que sugiere una separación débil entre clústeres.



Modelo 3 - Agglomerative Clustering

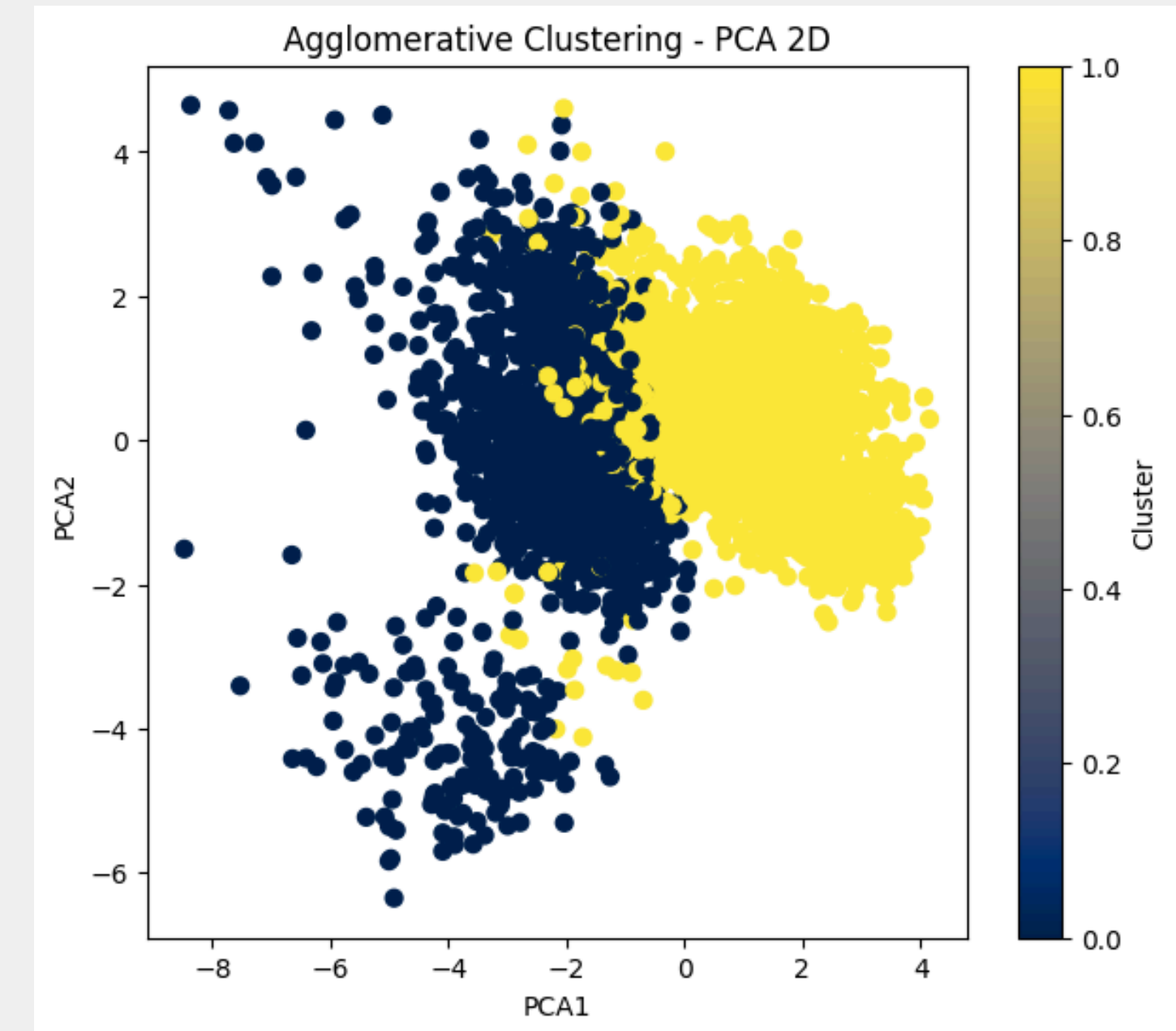
Los clústeres son mayormente diferenciados por los factores de forma, que son las variables donde existe menor solapamiento.

Cluster 0:

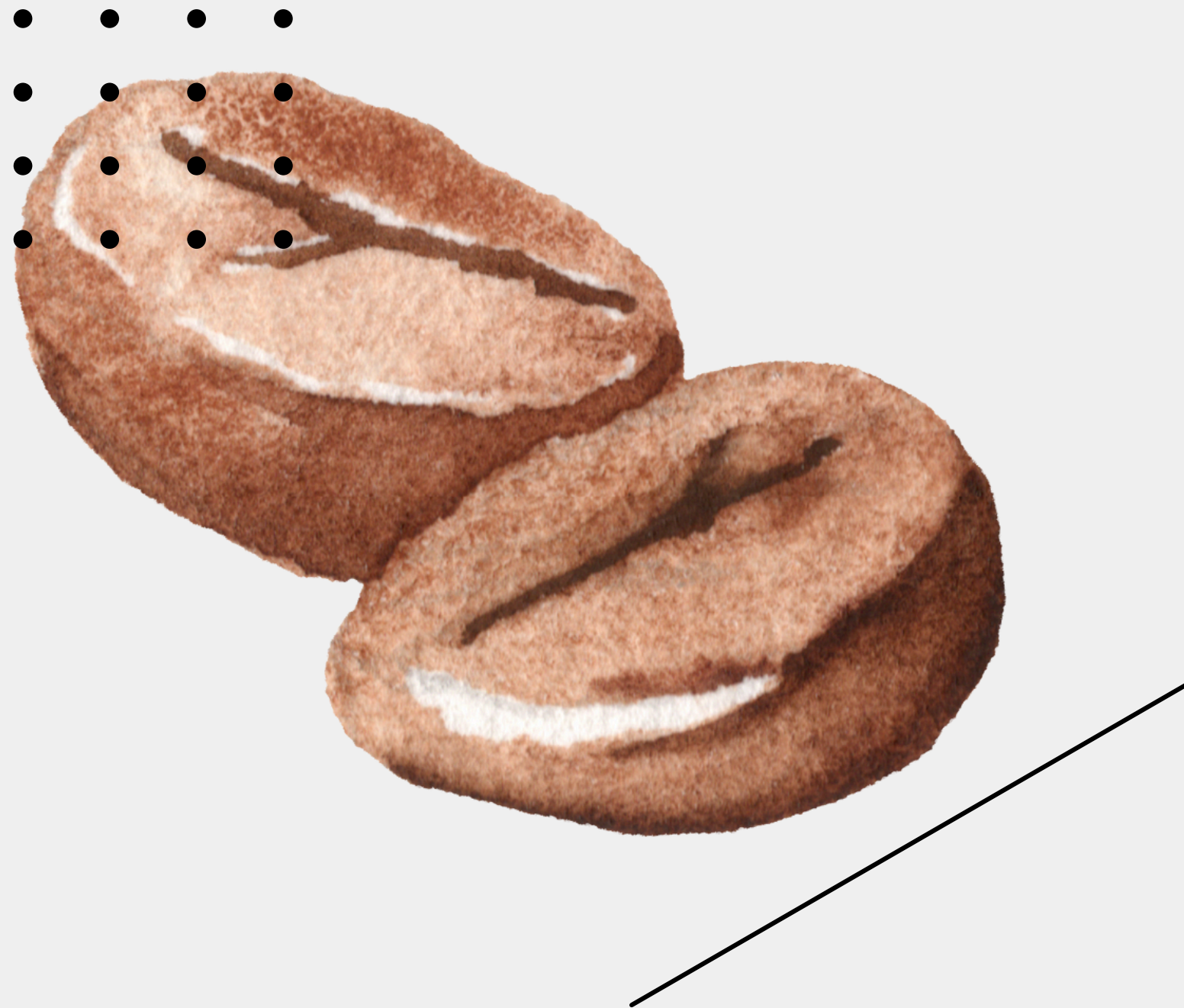
- Valores menores de Área, indicando granos más pequeños.
- En FactorForma1 se concentran en valores más bajos, lo que sugiere geometrías alargadas o menos regulares.
- En la proyección PCA se ubica más hacia la izquierda, representando la parte del conjunto con menor variabilidad en tamaño pero más diversidad en forma.

Cluster 1:

- Presenta valores de Área mayores, reflejando granos de mayor tamaño promedio.
- En FactorForma1 tienden a concentrarse en valores positivos, vinculados con granos más redondeados y consistentes.
- En la proyección PCA aparecen desplazados hacia la derecha, representando el grupo de granos más uniformes y regulares.



Modelo escogido: Kmeans



- 1** Logró el mejor coeficiente de silueta (0.328), indicando clústeres más compactos y mejor separados que los obtenidos con Agglomerative (0.320) y DBSCAN (0.173).
- 2** Se identificaron dos grupos: "Granos Compactos y Uniformes" y "Granos Grandes e Irregulares".


Recomendaciones

- 1 Adoptar K-Means como punto de partida:** Utilizar el modelo de K-Means para clasificar los granos en dos segmentos principales, apoyando la toma de decisiones en control de calidad y optimización de procesos.
- 2 Integrar más variables relevantes:** Incluir atributos adicionales como humedad, origen geográfico y características químicas de los granos para mejorar la precisión y relevancia de los clústeres para fines productivos y comerciales.
- 3 Validación práctica y continua:** Probar la clasificación en entornos reales de producción para verificar su utilidad.



Uso IA

Para esta laboratorio se hizo uso de chatbots como ChatGPT para solución de errores y debug, asistencia en graficación y sugerencias en intepretación de datos



Gracias

