PROPUESTA
SENECAFÉALPES

Andrés Neira
Esteban Castelblanco
Daniel Stiven Roa Uribe

Índice de
CONTENIDOS

Sobre el Proyecto

INTRODUCCIÓN

El presente proyecto busca aplicar técnicas de aprendizaje no supervisado para segmentar granos de café de la empresa SenecaféAlpes a partir de atributos morfológicos y de procesamiento. El objetivo es identificar patrones ocultos y agrupaciones naturales en los datos que permitan apoyar la caracterización de la calidad del grano, optimizar procesos de selección y generar valor para la empresa mediante una mejor comprensión de su producto.

OBJETIVOS

01 Identificar agrupaciones naturales en los granos de café a partir de sus atributos físicos y de procesamiento.

02 Evaluar distintos algoritmos de clustering para determinar cuál ofrece la mejor segmentación de los datos.

03 Generar información útil para SenecaféAlpes que contribuya a mejorar la clasificación y el control de calidad del grano.

LIMPIEZA DE DATOS: PROCESO

01 Se identificaron variables categóricas y se transformaron mediante normalización de texto y One-Hot Encoding.

02 Se detectaron y corrigieron valores negativos en variables numéricas (ajustándolos con valor absoluto).

03 Se realizó imputación de valores faltantes utilizando medidas estadísticas (media) y fórmulas definidas en el diccionario de datos (ej. Redondez).

04 Se eliminaron filas duplicadas (≈4.9%), asegurando mayor unicidad en el dataset.

05 Se descartaron variables sin variabilidad o redundantes para evitar sesgo en los modelos.

LIMPIEZA DE DATOS: RECOMENDACIONES

01 Estandarizar la captura de datos en campo y laboratorio para reducir la presencia de errores (valores negativos o inconsistentes).

02 Implementar un protocolo de control de calidad de datos antes de consolidar bases (detección temprana de duplicados, validación de rangos esperados).

03 Unificar nomenclaturas y categorías desde el origen (ej. “Normal” vs “normal”) para evitar procesos adicionales de limpieza.

04 Establecer una estrategia de manejo de valores faltantes, definiendo criterios claros de imputación o descarte según la relevancia de la variable.

05 Crear un repositorio centralizado de datos para asegurar la trazabilidad y facilitar futuras actualizaciones o integraciones con otros sistemas analíticos.

CONSTRUCCIÓN DE MODELOS

01 K-Means

02 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

03 Gaussian Mixture Models (GMM)

K-MEANS(DESCRIPCIÓN)

K-Means es un algoritmo de agrupamiento no supervisado que organiza los datos en grupos (clusters) basados en la similitud de sus características. Su objetivo es minimizar la variabilidad dentro de cada grupo y maximizar la diferencia entre ellos, utilizando la distancia a los centroides como criterio principal.

CARACTERISTICAS

01 Divide los datos en un número k de clusters predefinido.

02 Utiliza la distancia euclidiana para asignar cada punto al centro más cercano.

03 Los centroides representan el punto medio de cada cluster.

04 Tiene como limitación la sensibilidad a outliers y a la elección inicial de k.

05 Es sensible a la escala de los datos → requiere estandarización.

06 Funciona mejor con clusters esféricos y de tamaño similar.

K-MEANS(RENDIMIENTO)

El modelo de K-Means aplicado con k=7 logró identificar grupos diferenciados de granos en función de su tamaño, forma y método de secado. Aunque la separación entre clusters no fue perfecta, el algoritmo capturó patrones útiles para el negocio y demostró ser una herramienta práctica para caracterizar la variabilidad de los granos.

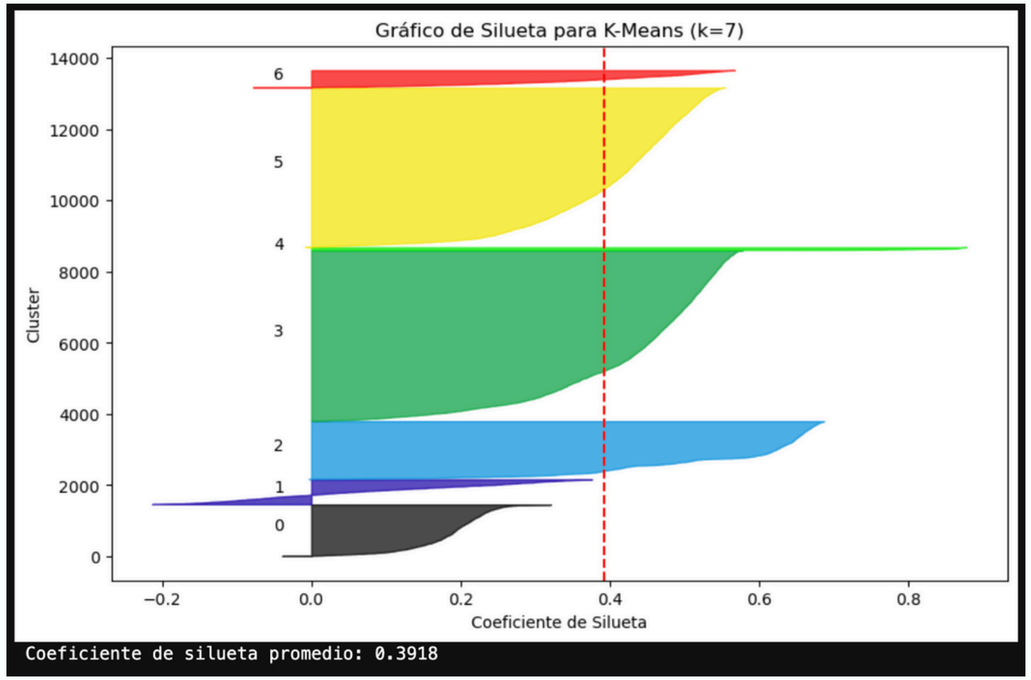
01 Coeficiente de silueta promedio ≈ 0.39, indicando una calidad de clustering moderada.

02 Separación clara de clusters según tamaño (Área y Perímetro).

03 Diferenciación de granos redondeados vs alargados mediante Excentricidad y Redondez.

04 Asociación de clusters con métodos de secado (ej. lavado vs natural).

05 Limitación: algunos clusters presentan traslape y menor definición.



DBSCAN(DESCRIPCIÓN)

Es un algoritmo de agrupamiento basado en densidad que forma clusters identificando regiones con alta concentración de puntos. A diferencia de K-Means, no requiere especificar el número de clusters y puede detectar outliers como ruido, lo que lo hace útil en datos con estructuras más complejas o no esféricas.

CARACTERISTICAS

01 No necesita definir k (número de clusters) previamente.

02 Clasifica puntos en tres tipos: núcleo, borde y ruido.

03 Requiere dos parámetros clave: eps (radio de vecindad) y min_samples (mínimo de puntos para formar un cluster).

04 Identifica outliers explícitamente, lo que mejora el análisis de datos ruidosos.

05 Capaz de detectar clusters de formas irregulares y de diferentes tamaños.

06 Limitación: la calidad depende fuertemente de la elección de eps y min_samples.

DBSCAN(RENDIMIENTO)

El modelo DBSCAN permitió identificar 2 clusters principales en los datos y clasificar un grupo reducido de observaciones como ruido. Aunque la calidad de la separación fue moderada, el algoritmo aportó valor al detectar puntos atípicos y diferenciar sutilmente los granos según algunas características morfológicas.

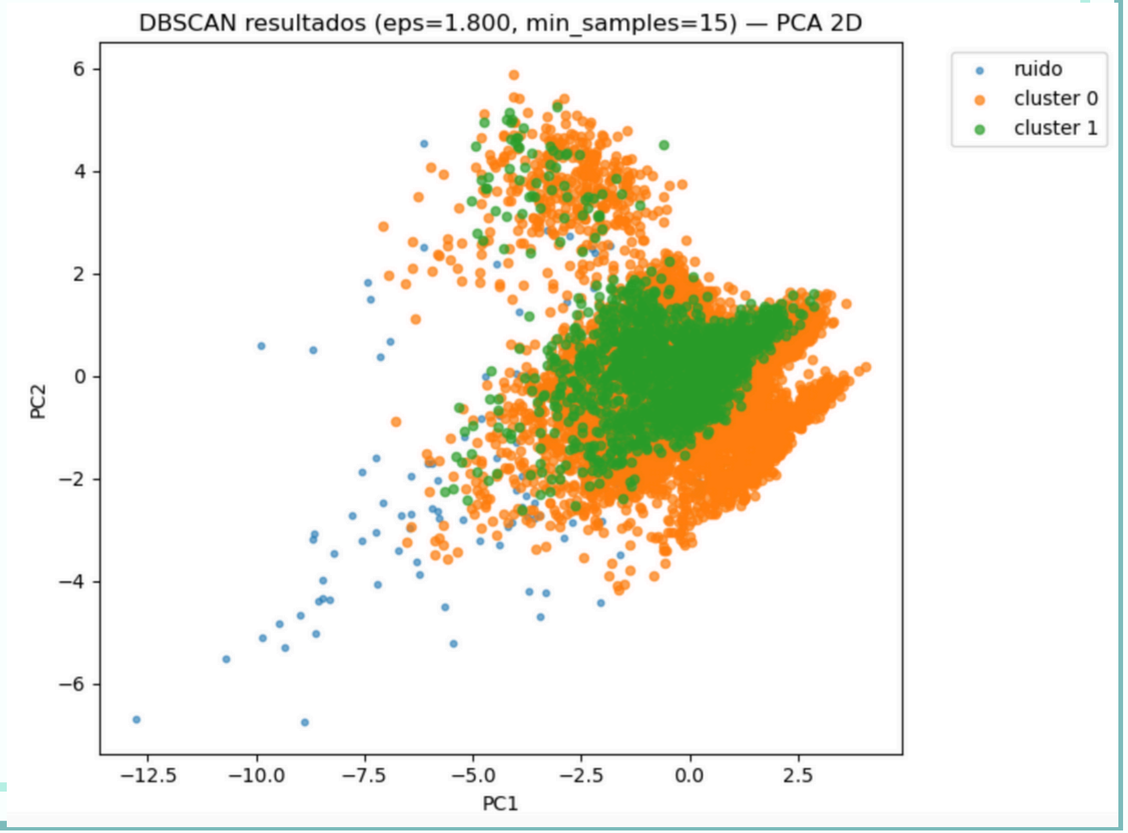
01 Configuración óptima: eps=1.8, min_samples=15.

02 Clusters detectados: 2 principales, con 101 observaciones clasificadas como ruido.

03 Silhouette Score ≈ 0.30, indicando separación moderada y cierto solapamiento entre clusters.

04 Diferencias clave en Perímetro y Medida, mientras que el resto de variables resultaron muy similares.

05 Los clusters obtenidos fueron pocos y con diferencias reducidas, lo que limita la profundidad del análisis.



GMM(DESCRIPCIÓN)

Enfoque probabilístico de clustering que asume que los datos provienen de una combinación de varias distribuciones normales (gaussianas). A diferencia de K-Means, GMM permite que los clusters tengan formas elípticas, tamaños distintos y que cada punto pertenezca a un cluster con cierta probabilidad, en lugar de una asignación rígida.

CARACTERISTICAS

01 Modelo de soft clustering: asigna a cada punto una probabilidad de pertenencia a cada cluster.

02 Utiliza el algoritmo Expectation-Maximization (EM) para estimar los parámetros.

03 Permite clusters de formas elípticas y no solo esféricas.

04 Más flexible que K-Means, pero también más costoso computacionalmente.

05 Requiere definir el número de componentes (clusters) a ajustar.

06 Puede capturar estructuras más complejas en los datos, siempre que las gaussianas sean una buena aproximación.

GMM(RENDIMIENTO)

El modelo Gaussian Mixture Models (GMM) permitió identificar 10 clusters en los datos, ofreciendo una segmentación probabilística que asigna a cada observación una probabilidad de pertenencia a los grupos. Los resultados muestran que las asignaciones fueron consistentes, con un promedio de certeza de 0.98 y una desviación estándar baja (0.072). Sin embargo, el Silhouette Score obtenido (≈ 0.096) indica que la separación entre clusters fue débil, lo que sugiere un importante solapamiento entre grupos a pesar de la alta confianza en las clasificaciones.

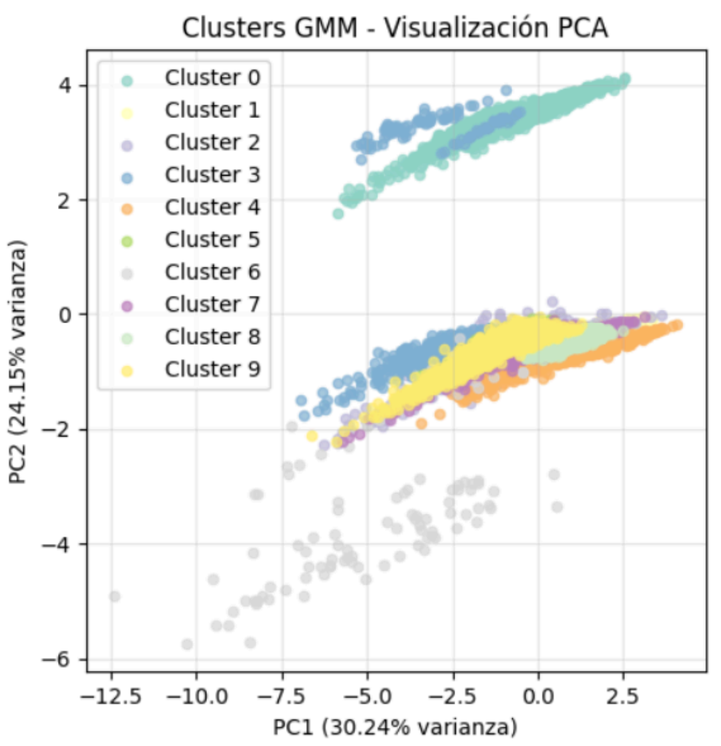
01 Se formaron 10 clusters principales, con centroides interpretados en las variables originales.

02 El modelo mostró alta certeza en las asignaciones (≈ 98%), lo que significa que cada punto fue clasificado con gran seguridad.

03 La varianza explicada por los primeros dos componentes PCA fue del 54.38%, garantizando que la visualización conserve más de la mitad de la información de los datos.

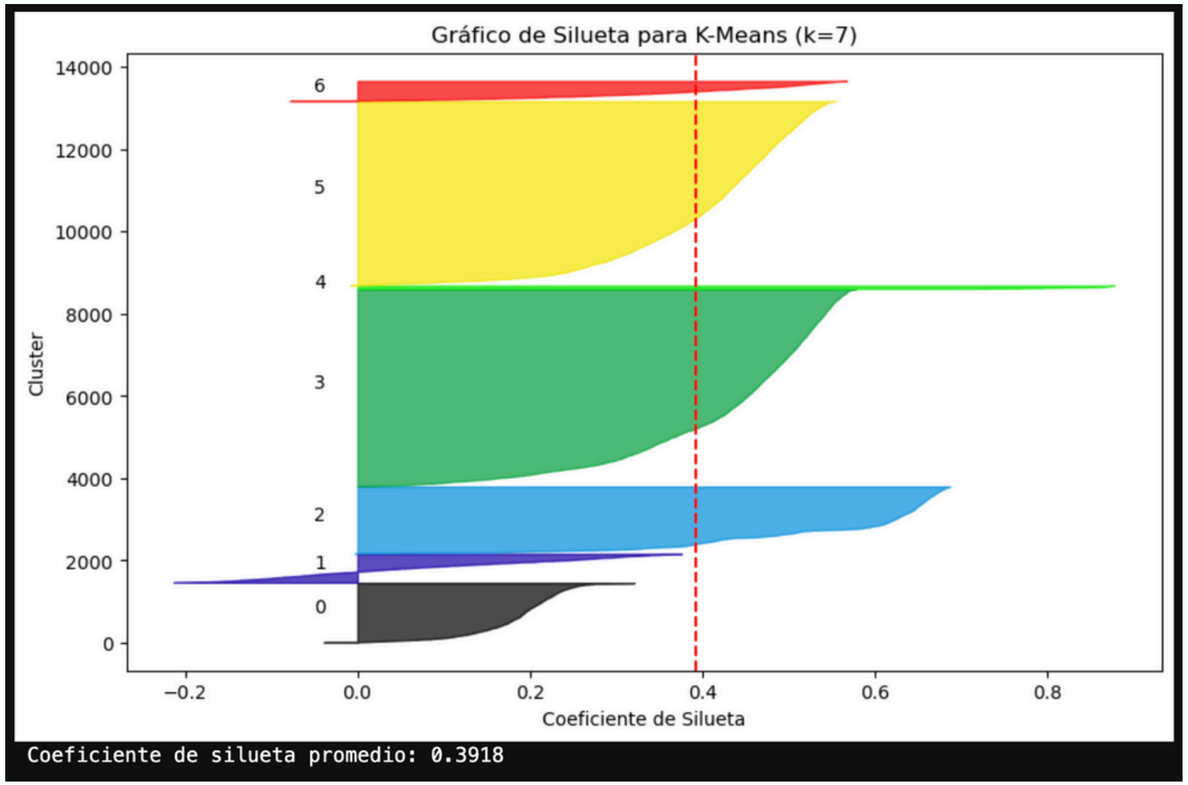
04 El Silhouette Score bajo (≈ 0.096) indica que los clusters no están claramente separados y existe un solapamiento considerable entre ellos.

05 Aunque GMM aporta flexibilidad probabilística frente a métodos como K-Means, en este caso no mejoró la calidad de la separación de los grupos.



SELECCIÓN DE MEJOR MODEL: K-MEANS

Tras comparar los modelos, K-Means resultó ser el más adecuado, con un Silhouette Score de ≈0.39 y una segmentación clara de los datos. DBSCAN aportó la ventaja de detectar ruido, pero con menor separación entre grupos (Silhouette ≈0.30), mientras que GMM presentó un solapamiento considerable (Silhouette ≈0.096). En conclusión, K-Means ofreció el mejor equilibrio entre simplicidad e interpretabilidad.



RECOMENDACIONES Y CONCLUSIONES

Recomendamos

01 Implementar K-Means como modelo base para segmentar los datos, dado que mostró el mejor desempeño en términos de calidad y claridad de clusters.

02 Usar los clusters identificados para diseñar estrategias diferenciadas de negocio, enfocándose en las características específicas de cada grupo.

03 Complementar el uso de K-Means con DBSCAN en análisis exploratorios, ya que este modelo detecta posibles outliers o comportamientos atípicos relevantes.

C1

04 Mejorar la calidad y consistencia de los datos en futuras recolecciones, con el fin de obtener clusters más definidos y robustos.

05 Explorar la combinación de modelos en futuras fases (ej. K-Means + PCA) para optimizar el rendimiento y la interpretabilidad.

06 Fomentar el uso de IA generativa y analítica avanzada para la documentación, análisis de modelos y apoyo en la toma de decisiones estratégicas.

USO DE IA GENERATIVA

En el desarrollo de este laboratorio se hizo uso de IA generativa como apoyo en diversas etapas del proyecto. En particular, se utilizó para la redacción y estructuración de textos en el informe y la presentación, así como para la investigación y comprensión de funciones, métodos y modelos de clustering. Este apoyo permitió agilizar el trabajo, asegurar una mejor claridad en la documentación y complementar el análisis técnico con explicaciones precisas y bien fundamentadas.

MUCHAS

GRACIAS