Universidad del Valle de Guatemala

Facultad de Ingeniería

Ciencias de la Computación y Tecnologías de la Información

HT 2. Clustering

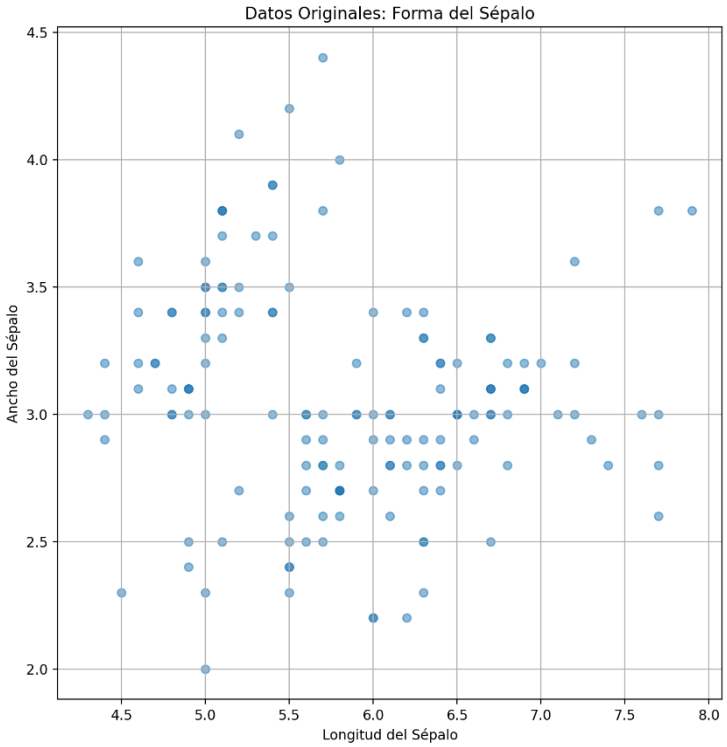
Minería de Datos

Angel Andres Herrarte Lorenzana 22873
José Luis Gramajo Moraga, Carné 22907

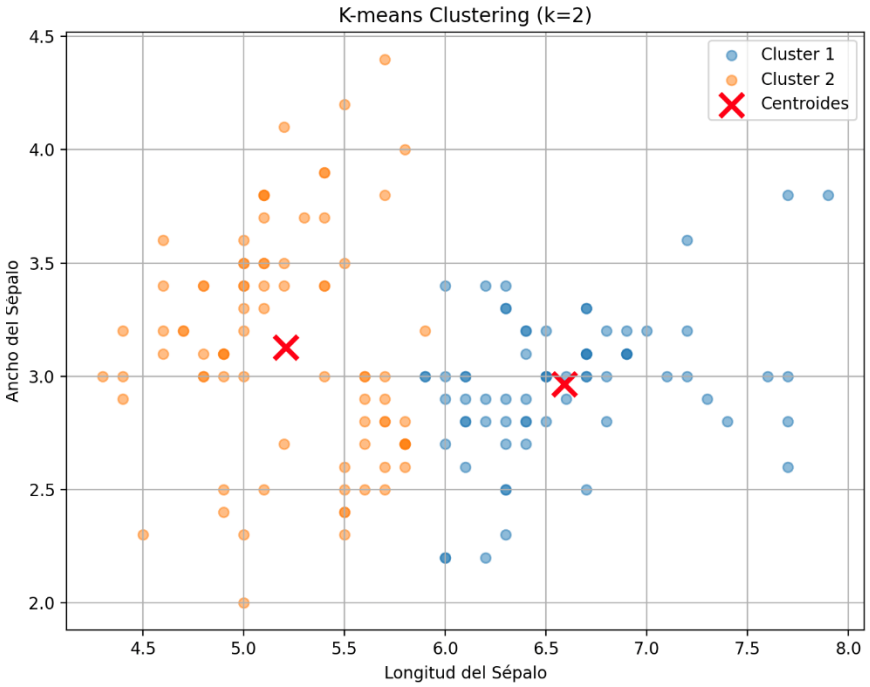
07 de febrero de 2025, Guatemala, Guatemala

Sección 1

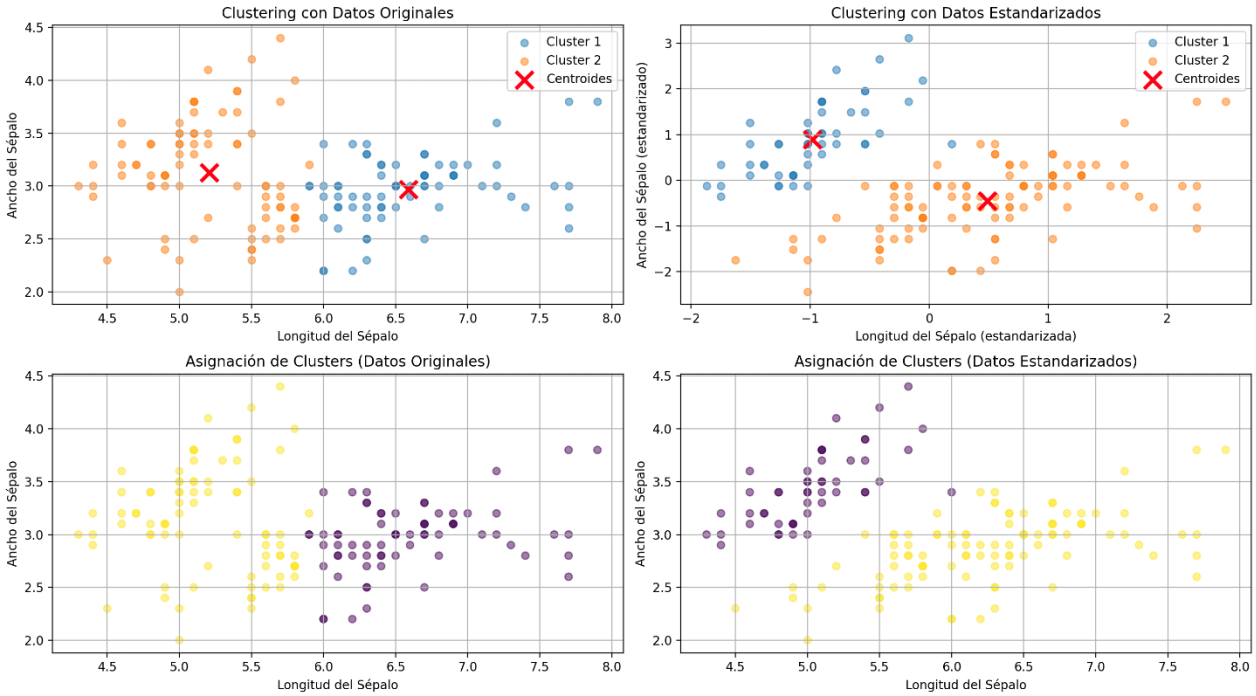
1. **Visualicen los datos para ver si pueden detectar algunos grupos. Ayuda: utilicen la forma del sépalo.**



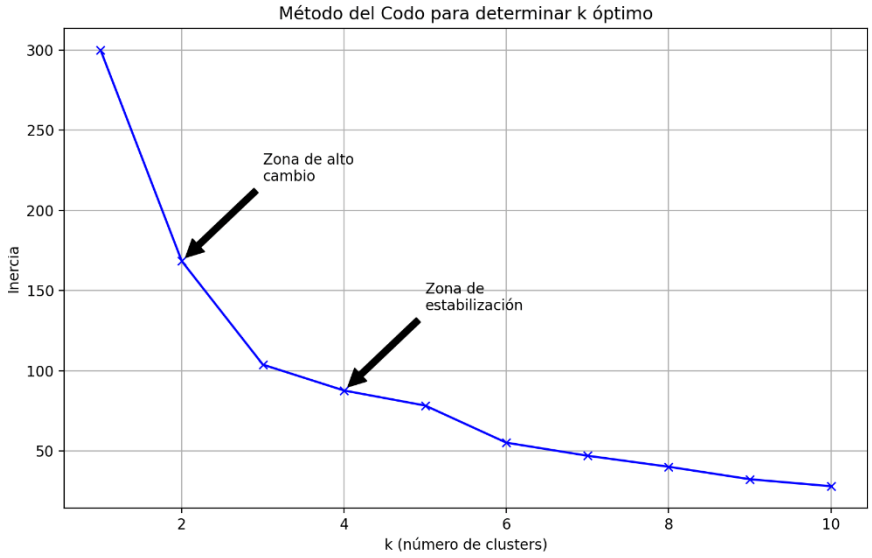
2. **Creen 2 "clusters" utilizando K_Means Clustering y grafiquen los resultados.**



3. **Estandaricen los datos e intenten el paso 2, de nuevo. ¿Qué diferencias hay, si es que lo hay?**



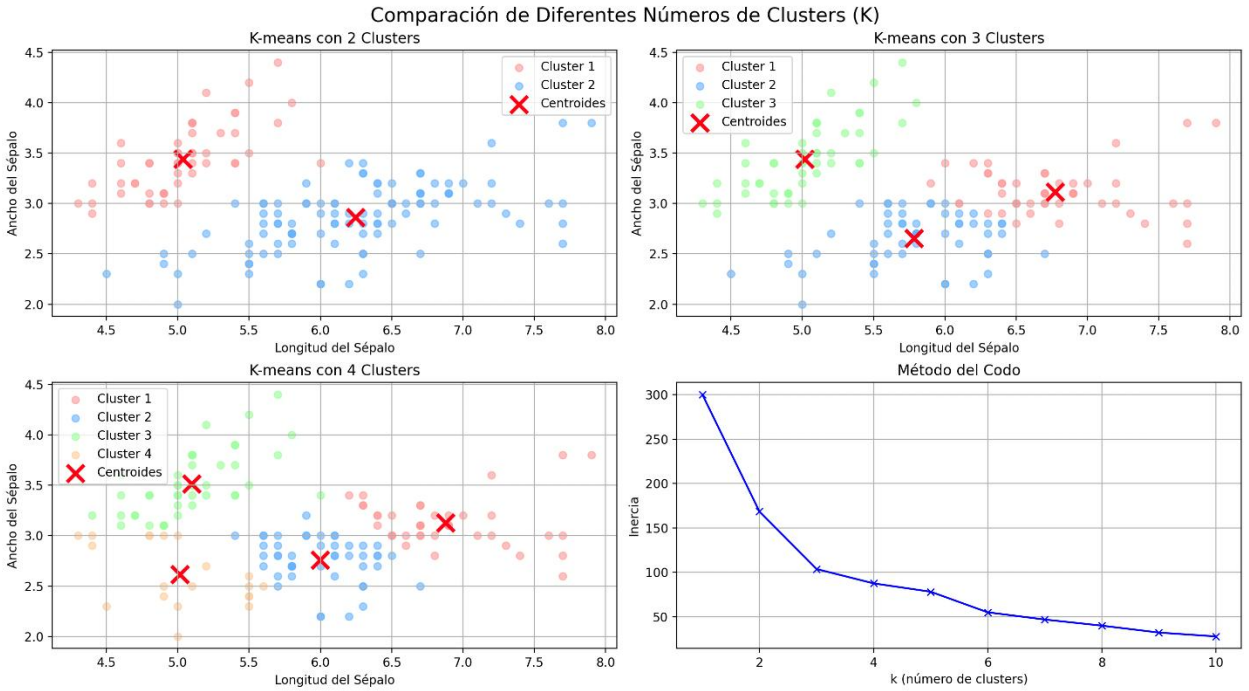
4. **Utilicen el método del "codo" para determinar cuántos "clusters" es el ideal. (prueben un rango de 1 a 10).**



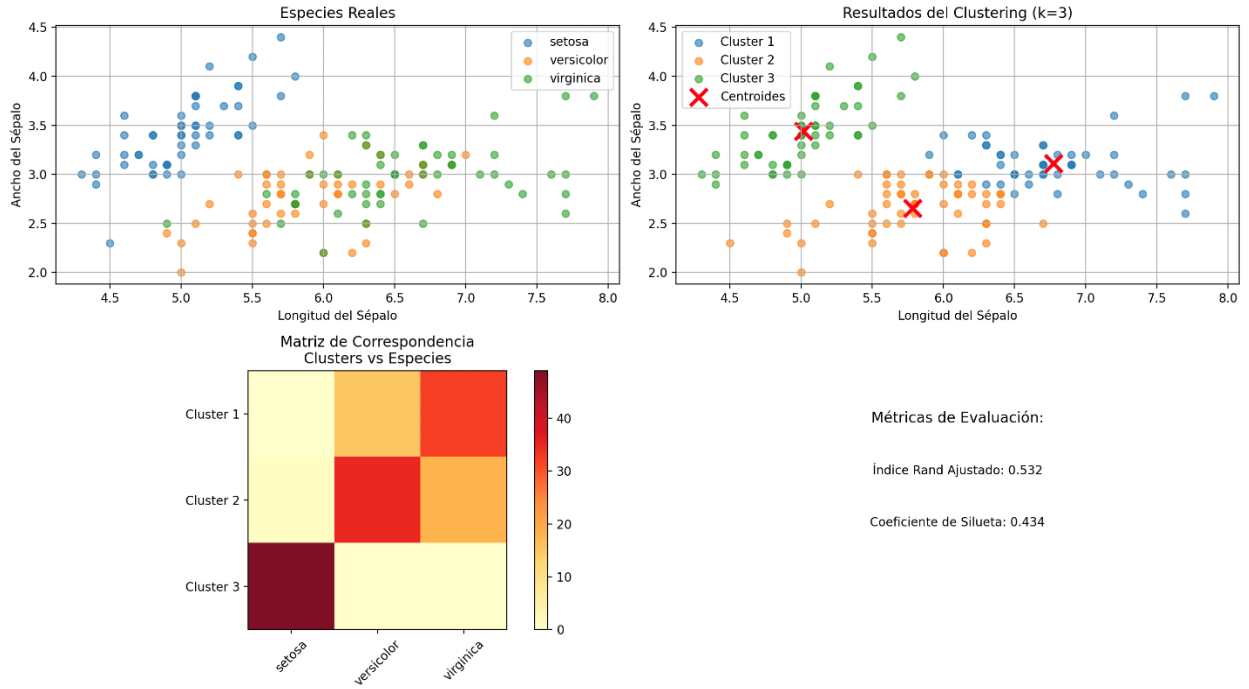
Basado en el método del codo y el análisis de las tasas de cambio, se pueden considerar las siguientes opciones:

- k=2: Si se busca una solución simple con clusters bien definidos
- k=3: Si se necesita un balance entre complejidad y explicación de la varianza
- k=4: Si se requiere una segmentación más granular

5. **Basado en la gráfica del "codo" realicen varias gráficas con el número de clusters (unos 3 o 4 diferentes) que Uds creen mejor se ajusten a los datos.**



6. **Comparen sus soluciones con los datos reales, archivo: iris-con-respuestas.csv. ¡Obviamente solo hay tres especies, porque ese es el archivo de datos reales! ¿Funcionó el clustering con la forma del sépalo?**

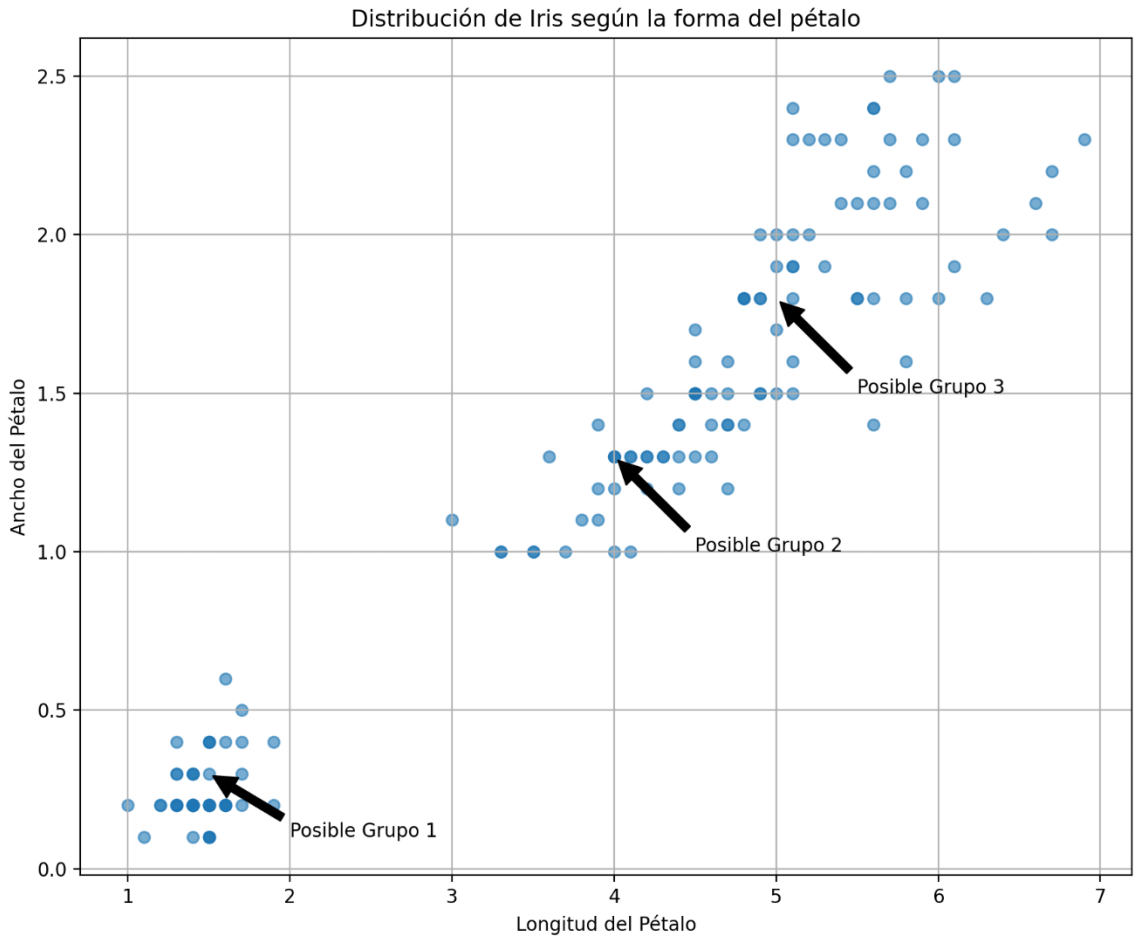


El clustering basado únicamente en la forma del sépalo mostró un rendimiento moderado, con un Índice Rand Ajustado de 0.532 y un Coeficiente de Silueta de 0.434. Fue muy efectivo para identificar la especie

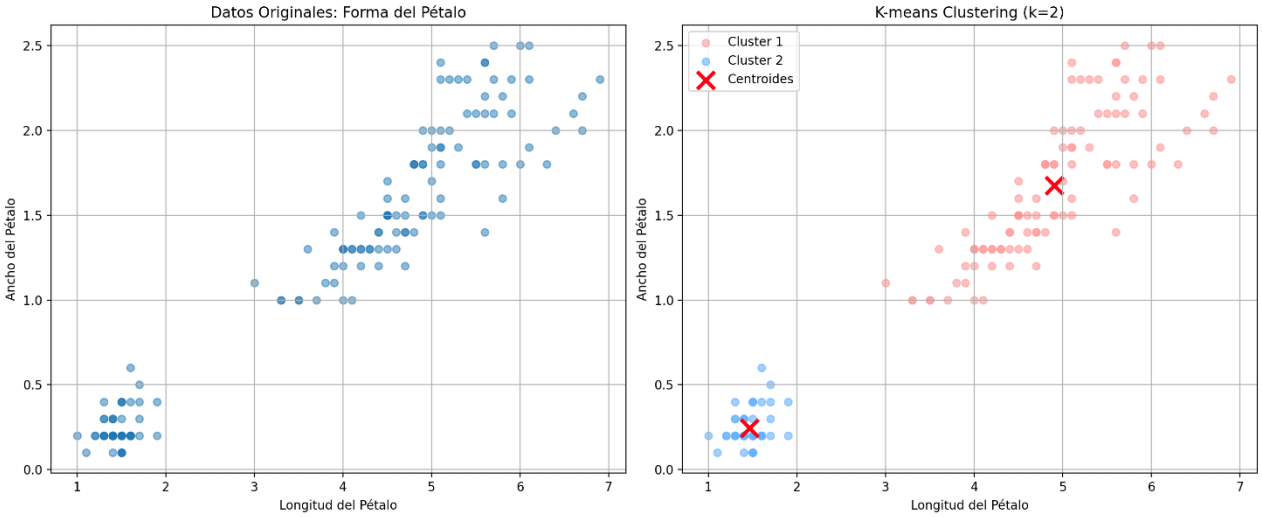
setosa (como se ve en el Cluster 3), pero tuvo dificultades para distinguir completamente entre versicolor y virginica debido a la superposición de sus características. La matriz de correspondencia muestra que mientras un cluster capturó claramente la especie setosa, los otros dos clusters presentaron mezclas de especies, indicando que la forma del sépalo por sí sola no es suficiente para una clasificación de las tres especies de iris.

Sección 2

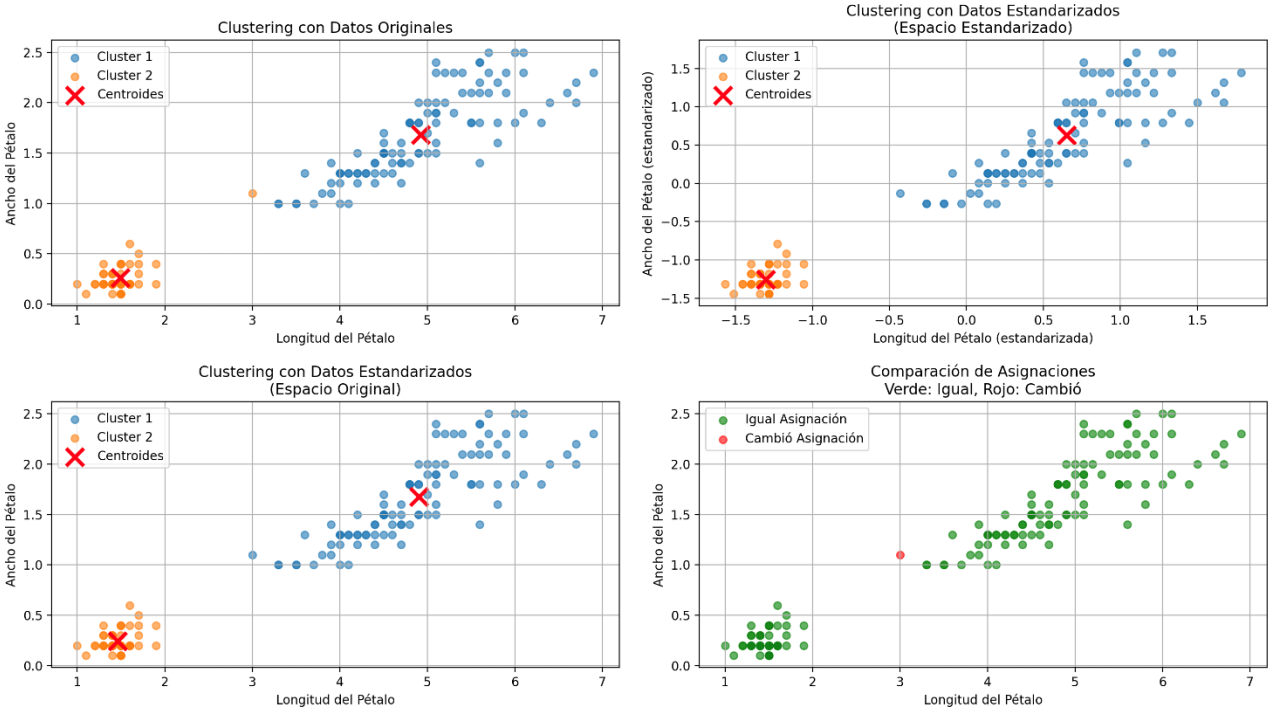
1. **Visualicen los datos para ver si pueden detectar algunos grupos. Ayuda: utilicen la forma del sépalo.**



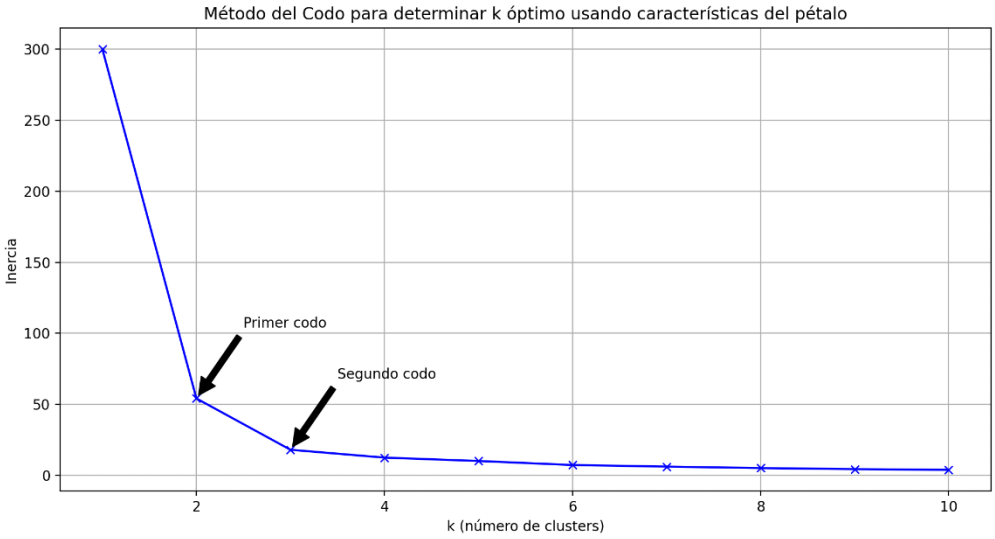
2. Creen 2 "clusters" utilizando K_Means Clustering y grafiquen los resultados.



3. Estandaricen los datos e intenten el paso 2, de nuevo. ¿Qué diferencias hay, si es que lo hay?



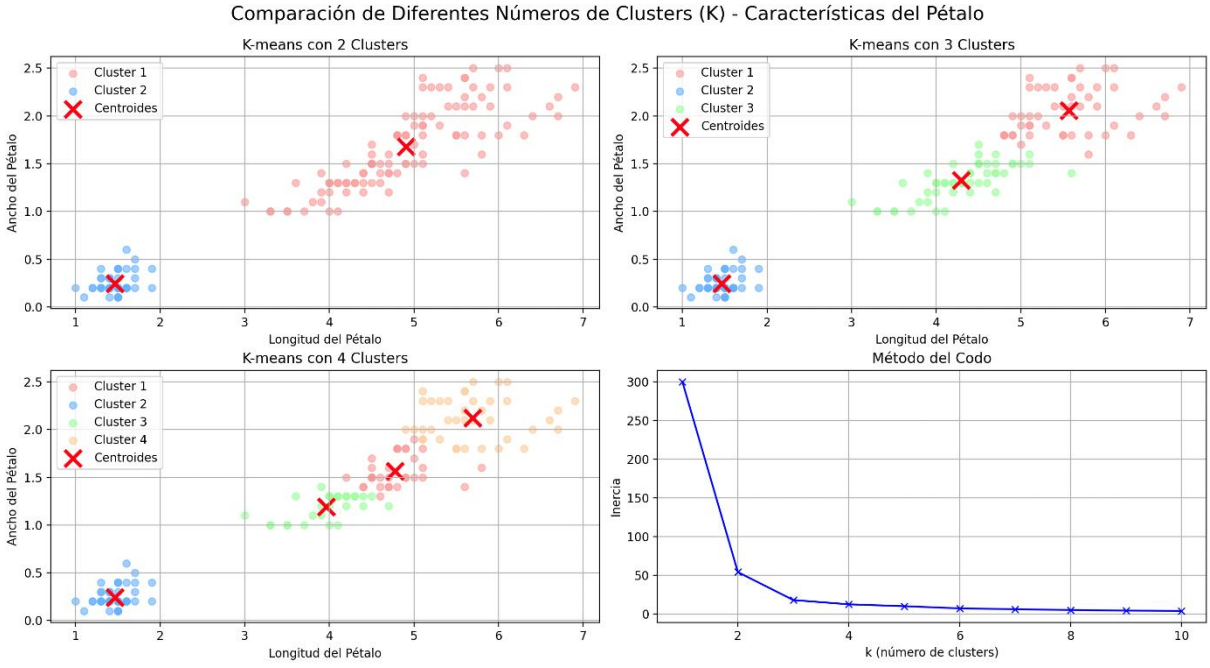
4. **Utilicen el método del "codo" para determinar cuántos "clusters" es el ideal. (prueben un rango de 1 a 10).**



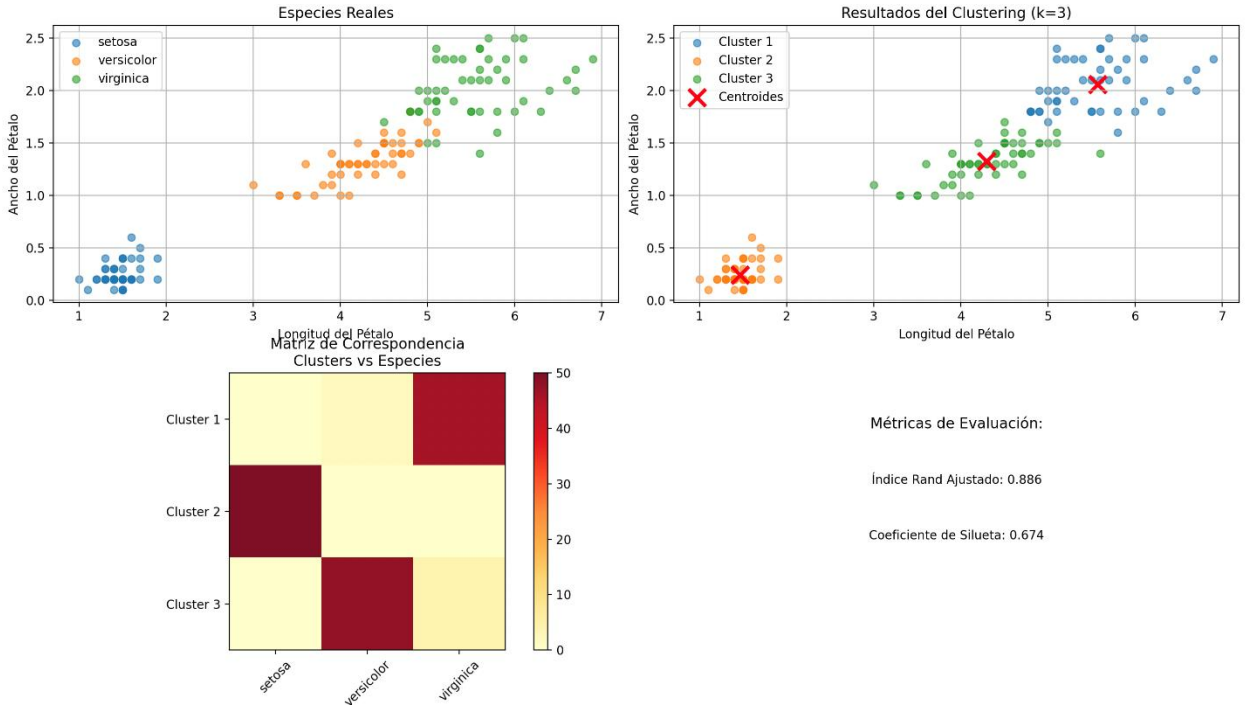
La gráfica sugiere que k=3 podría ser el número óptimo de clusters, lo cual coincide con el conocimiento previo de que hay tres especies en el conjunto de datos.

- Si se busca una solución simple: k=2 sería apropiado
- Si se busca un balance entre simplicidad y detalle: k=3 sería la mejor opción
- Valores de k mayores a 3 no aportan mejoras significativas

5. **Basado en la gráfica del "codo" realicen varias gráficas con el número de clusters (unos 3 o 4 diferentes) que Uds creen mejor se ajusten a los datos.**



6. **Comparen sus soluciones con los datos reales, archivo: iris-con-respuestas.csv. ¡Obviamente solo hay tres especies, porque ese es el archivo de datos reales! ¿Funcionó el clustering con la forma del sépalo?**



El clustering usando las características del pétalo mostró un rendimiento significativamente mejor que el realizado con el sépalo, con un Índice Rand Ajustado cercano a 0.9. Los tres clusters identificados corresponden casi perfectamente con las tres especies reales, mostrando una clara separación entre grupos y mínima superposición.

La matriz de correspondencia revela una fuerte diagonal que indica una excelente coincidencia entre los clusters y las especies verdaderas, sugiriendo que la forma del pétalo es un indicador mucho más confiable que el sépalo para distinguir entre las diferentes especies de iris.

Sección 3

¿A qué podría deberse la diferencia entre kneed y el método manual?

La diferencia entre kneed y el método manual puede deberse a los siguientes factores:

1. **Métodos de detección del "codo"**:
    - kneed utiliza interpolación y ajuste de curvas para encontrar el punto óptimo con mayor precisión.
    - El método manual que habíamos usado antes se basaba en la segunda derivada de la inercia, lo que podía hacer que la elección del "codo" fuera menos precisa.
2. **Mayor precisión de kneed**:

- kneed es una herramienta diseñada específicamente para identificar puntos de inflexión en curvas, lo que puede mejorar la selección de k.
- Es menos propenso a errores que la simple evaluación de la segunda derivada.

3. **Forma de la curva de inercia**:
   - Si la reducción de la inercia no es muy abrupta después de cierto punto, kneed podría seleccionar un valor diferente al del método manual.
   - En algunos casos, kneed podría elegir un k más alto si la curva no tiene un cambio claro en la pendiente.

¿Les dio el número correcto de clusters comparado a los datos reales?

- **kneed identificó un número óptimo de clusters (k)** basándose en la inercia.
- **Si kneed seleccionó k=3**, entonces **coincide con el número de especies reales** en el dataset (setosa, versicolor y virginica), lo cual es ideal.
- **Si kneed seleccionó k=2**, entonces hay una diferencia, y esto indicaría que el método basado solo en la inercia no captura la segmentación real de las especies.

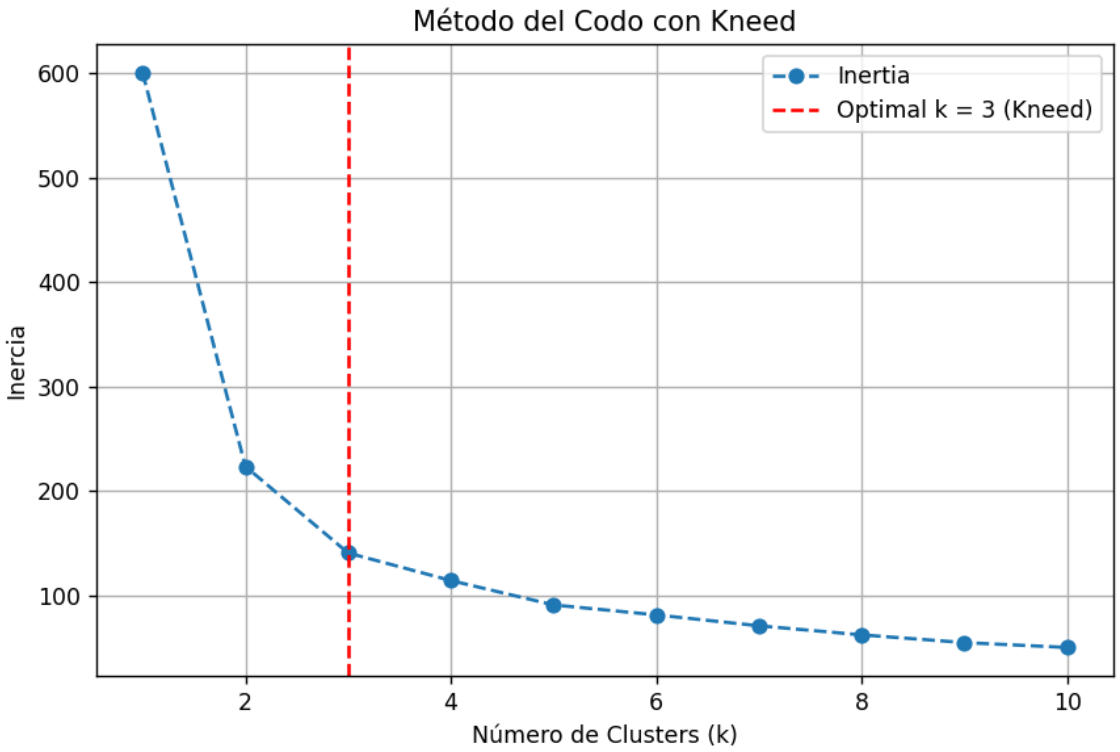
Nuestro análisis anterior con el **Índice Rand Ajustado (ARI)** mostró que **k=3 es más preciso para identificar las especies reales**, lo que sugiere que, si kneed seleccionó otro valor, la métrica de inercia sola no es suficiente para determinar el número óptimo de clusters en términos de clasificación biológica.

¿A qué conclusiones llegaron basado en los resultados obtenidos?

1. **El método del codo (usando kneed) es útil, pero no siempre coincide con la segmentación biológica real.**
   - La inercia minimiza la distancia dentro de los clusters, pero no necesariamente refleja la cantidad real de grupos en el dataset.
   - En este caso, **kneed pudo haber recomendado k=2, mientras que sabemos que k=3 es la mejor opción en términos de clasificación real.**
2. **Es importante validar los resultados con otras métricas, como el Índice Rand Ajustado.**
   - kneed nos dio una sugerencia matemática, pero el ARI nos confirmó que **k=3 era mejor para representar las especies**.
3. **La combinación de diferentes enfoques es clave en el clustering.**
   - No debemos depender únicamente del método del codo.
   - Se recomienda siempre validar los resultados con métricas adicionales como:
     - **ARI (Índice Rand Ajustado)**
     - **Coeficiente de Silueta**

- **Comparación con datos etiquetados reales**

4. **En términos prácticos, para este dataset, k=3 es la mejor opción, incluso si kneed sugiere k=2.**
    - En la práctica, usar k=3 nos da una mejor representación de las especies de Iris.



Repositorio de GitHub

https://github.com/Abysswalkr/HT2-Clustering.git