Hoja de Trabajo 2,
Segmentación de especies utilizando ”Cluster_Analysis”

INSTRUCCIONES:

El conjunto de datos de la flor **Iris** es uno de los más populares para el Aprendizaje de Máquina (ML). Si no lo conocen, pueden leer sobre él en:

https://en.wikipedia.org/wiki/Iris_flower_data_set

Este laboratorio debe realizarse en **PAREJAS.** Para que se pueda calificar su laboratorio debe estar inscrito en algún grupo de canvas creado para este propósito.

DESCRIPCIÓN DEL DATASET

El conjunto de datos *iris.csv* tiene cuatro variables:

- **sepal length** (longitud del sépalo),
- **sepal width** (ancho del sépalo),
- **petal length** (longitud del pétalo),
- **petal width** (ancho del pétalo).

EJERCICIOS

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 cuantos "clusters" es el ideal. (prueben un rango de 1 a 10)

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?

SECCIÓN 2:

Repitan el proceso pero ahora utilizando la forma del pétalo. Respondan a las mismas preguntas

SECCIÓN 3:

Utilicen la librería "kneed" y vean si el resultado coincide con el método del "codo" que hicieron manualmente. ¿A que podría deberse la diferencia, si la hay? ¿Les dió el número correcto de clusters, comparado a los datos reales?

Basado en los resultado que tuvieron, ¿A qué conclusiones llegaron?

EVALUACIÓN

NOTA: La evaluación de cada integrante del grupo será de acuerdo con sus contribuciones al trabajo grupal

1. Análisis con la Forma del Sépalo (30 puntos)

- Visualización inicial exploratoria de datos (5 pts)
- Implementación correcta de K-Means con 2 clusters (5 pts)
- Estandarización de datos y comparación de resultados (5 pts)
- Implementación y análisis del método del codo (5 pts)
- Visualizaciones con diferentes números de clusters (5 pts)
- Comparación con datos reales y análisis de efectividad (5 pts)

2. Análisis con la Forma del Pétalo (30 puntos)

- Repetición completa del proceso con variables del pétalo (15 pts)
- Análisis comparativo de resultados (10 pts)
- Conclusiones sobre la efectividad del clustering con pétalos (5 pts)

3. Análisis con Librería Kneed (20 puntos)

- Implementación correcta de la librería kneed (5 pts)
- Comparación con método del codo manual (5 pts)

- Análisis de diferencias encontradas (5 pts)
- Evaluación de la precisión respecto a datos reales (5 pts)

4. Documentación y Conclusiones (20 puntos)

- Código bien documentado y organizado (5 pts)
- Explicaciones claras de los procedimientos (5 pts)
- Análisis comparativo global de los métodos (5 pts)
- Conclusiones fundamentadas sobre la efectividad de diferentes enfoques (5 pts)

Aspectos a Evaluar en cada Sección:

- Correcta implementación técnica
- Calidad de las visualizaciones
- Profundidad del análisis
- Interpretación de resultados
- Comparación con datos reales

Criterios de Penalización:

- Código no funcional (-10 puntos)
- Falta de análisis comparativo entre secciones (-5 puntos)
- Visualizaciones inadecuadas o poco claras (-5 puntos)
- Conclusiones sin fundamento (-5 puntos)

Criterios de Bonificación:

- Análisis adicional más allá de lo solicitado (+5 puntos)
- Visualizaciones especialmente informativas (+3 puntos)
- Insights únicos y bien fundamentados (+2 puntos)

- Archivo .pdf con el informe del trabajo realizado.Link de Google drive donde trabajó el grupo.
- Script de R (.r o .rmd) o de Python que utilizó para responder las preguntas con el código utilizado o archivo de flujo de trabajo de Knime
- Link de github o el versionador que se utilizó.

NOTA: Si utilizan un Jupyter Notebook, pueden incluir ambos la codificación y el informe de los resultados en el mismo.