Universidad de Manizales
Facultad de Ciencias e Ingeniería
Ingeniería en Analítica de Datos

Actividad: Aplicación de K-Means en Datos Públicos

Objetivo:

Los estudiantes buscarán un conjunto de datos público, lo analizarán y aplicarán **K-Means** para identificar patrones y realizar agrupamientos significativos. Aprenderán a identificar qué datos son adecuados para la clusterización, preprocesarlos y usar K-Means para extraer información útil.

Instrucciones:

Búsqueda de datos: Los estudiantes deberán buscar y seleccionar un conjunto de datos público para aplicar el algoritmo **K-Means**. Estos datos deben cumplir con las siguientes características:

- Tener al menos **tres** variables numéricas que permitan identificar patrones (como edad, ingresos, puntaje de crédito, etc.).
- El tamaño del conjunto de datos debe ser razonable (mínimo 100 filas y máximo 10,000 filas).

Algunos recursos donde los estudiantes pueden buscar conjuntos de datos:

- Kaggle Datasets
- UCI Machine Learning Repository
- Google Dataset Search
- Open Data Portals (gobiernos y organizaciones públicas)

Recomendación para los estudiantes: Deben elegir un conjunto de datos relevante y justificar su selección. Deberían considerar si los datos son adecuados para la clusterización con K-Means.

Pasos a seguir:

1. Exploración de los datos:

- Una vez que encuentren el conjunto de datos, deben cargarlo y realizar un análisis exploratorio para entender sus características.

- Realizar visualizaciones iniciales de las variables, como histogramas, gráficos de dispersión, o gráficos de correlación entre variables numéricas.
- Justificar si el conjunto de datos es adecuado para K-Means (es decir, si existen patrones claros o relaciones entre las variables que podrían llevar a buenos agrupamientos).

2. Preprocesamiento de los datos:

- Escalar las variables numéricas usando técnicas como **StandardScaler** o **MinMaxScaler**, asegurándose de que todas las características estén en la misma escala.
- Manejar valores faltantes, si existen, mediante imputación o eliminación, y realizar cualquier otra limpieza necesaria (como manejo de valores atípicos).

3. Aplicación de K-Means:

- Aplicar el algoritmo de **K-Means** a los datos preprocesados.
- Utilizar el **método del codo** para seleccionar el número óptimo de clusters. Esto implica ejecutar K-Means con diferentes números de clusters (por ejemplo, entre 2 y 10) y graficar la **inercia** o el **error cuadrático medio** para identificar el punto donde se estabiliza la mejora.
- Opcional: Pueden utilizar el **coeficiente de silueta** para complementar su análisis.

4. Interpretación de los clusters:

- Una vez determinado el número óptimo de clusters, deben aplicar K-Means y asignar cada instancia a su respectivo cluster.
- Analizar los **centroides** de cada cluster para interpretar qué características describen a los grupos formados.
- Realizar visualizaciones que muestren cómo se distribuyen los clusters en el espacio de las variables más representativas.

5. Presentación de resultados:

- Describir los diferentes grupos (clusters) formados, interpretando lo que cada cluster representa. Por ejemplo, si los datos son sobre clientes, ¿qué tipo de clientes están en cada cluster? ¿Clientes jóvenes con altos ingresos? ¿Clientes de edad avanzada con baja puntuación de crédito?
- Proponer posibles aplicaciones prácticas de los clusters. Si es un conjunto de datos de clientes, ¿qué tipo de estrategias de marketing podrían dirigirse a cada grupo?

2. **Producto final a entregar:**
    - Un **notebook** de Jupyter con el código desarrollado para realizar la clusterización. Esto debe incluir:
        - El análisis exploratorio de los datos.
        - Preprocesamiento de las variables.
        - Aplicación de K-Means con justificación del número de clusters seleccionado.
        - Visualizaciones claras que muestren los clusters formados.
    - Un **informe en PDF** que incluya:
        - Descripción del conjunto de datos seleccionado y justificación de su elección.
        - Explicación del proceso seguido para preprocesar los datos y aplicar K-Means.
        - Interpretación de los clusters obtenidos.
        - Conclusiones y posibles aplicaciones de los resultados.

Criterios de evaluación:

- **Elección del dataset**: Los estudiantes deben justificar por qué seleccionaron ese conjunto de datos y si es adecuado para aplicar K-Means.
- **Preprocesamiento**: Deben mostrar un análisis adecuado de los datos y realizar las transformaciones necesarias para que sean aptos para el modelo.
- **Uso del método del codo**: Deben usar este método para seleccionar el número de clusters y justificar la elección.

- **Visualizaciones y análisis**: Deben incluir visualizaciones claras que muestren los clusters y su interpretación.
- **Informe final**: La calidad del informe escrito será importante, evaluando si describen bien los resultados y proponen conclusiones o aplicaciones prácticas de los clusters.