Análisis de Clustering con K-Means

Alan Miranda, Marlon Caviedes, christian orozco

Abstract

En este artículo se muestra un estudio práctico de agrupación (clustering) aplicado a datos de COVID-19 utilizando el algoritmo K-Means implementado de forma manual. Se explica el proceso de lectura, limpieza y conversión de los datos, la aplicación del algoritmo utilizando la distancia Euclidiana y la interpretación de los resultados obtenidos.

1 Introducción

El clustering es una técnica básica en el análisis de datos que se utiliza para agrupar elementos similares. En este trabajo se emplea el algoritmo K-Means para dividir un conjunto de datos relacionados con COVID-19 en distintos grupos. Aunque es un método sencillo, permite identificar patrones y tendencias en la información.

2 Objetivos del Análisis

Los objetivos de este estudio son:

- Aplicar el algoritmo K-Means a un conjunto real de datos.
- Dividir los datos de COVID-19 en grupos homogéneos.
- Interpretar los resultados obtenidos.

3 Descripción del Análisis

El estudio se realizó en los siguientes pasos:

1. **Lectura y limpieza de datos:** Se cargó el archivo CSV ubicado en `/content/archivo_organizado.cs` y se realizaron ajustes para corregir valores faltantes.
2. **Transformación de datos:** Se seleccionaron las variables relevantes y se convirtieron a datos numéricos adecuados para la aplicación del algoritmo.
3. **Aplicación de K-Means:** Se utilizó la distancia Euclidiana para asignar cada dato a su centroide más cercano, y se recalcularon los centroides hasta que dejaron de variar significativamente.

4 Interpretación y Conclusiones

El uso del algoritmo K-Means permitió descubrir subgrupos en el conjunto de datos de COVID-19. La implementación manual, basada en la distancia Euclidiana, fue suficiente para dividir los datos en 3 grupos distintos. Los resultados indican que:

1. Existe una tendencia natural en los datos que puede ayudar en estudios epidemiológicos.
2. La técnica aplicada es adecuada para identificar grupos en conjuntos de datos complejos.
3. La interpretación de los centroides, calculados como el promedio de las características de cada grupo, permite entender la estructura interna de los datos.

En resumen, este trabajo demuestra la efectividad del algoritmo K-Means para la segmentación de datos reales y sienta las bases para investigaciones futuras en salud pública.