Data set: Online Retail

Las variables contenidas en este conjunto de datos son:

InvoiceNo: Número de factura. Un número entero de 6 dígitos asignado de forma única a cada transacción. Si este código comienza con la letra 'c', indica una cancelación.

StockCode: Código de producto. Un número entero de 5 dígitos asignado de forma única a cada producto distinto.

Description: Nombre del producto.

Quantity: Cantidades de cada producto por transacción.

InvoiceDate: Fecha y hora de la factura. El día y la hora en que se generó cada transacción.

UnitPrice: Precio por unidad. Precio del producto por unidad en libras esterlinas.

CustomerID: Número de cliente. Un número entero de 5 dígitos asignado de forma única a cada cliente.

Country: Nombre del país. El nombre del país donde reside cada cliente.

Objetivo: Comprender el funcionamiento del algoritmo de K-Means aplicándolo en un conjunto de datos real de comercio electrónico con el fin de identificar perfiles de clientes y optimizar las estrategias de marketing para incrementar las ganancias anuales. Recuerde que el ciclo de vida de un proyecto no es lineal, requiere ir y volver.

1. Comenten el conjunto de datos y discuta las razones para utilizar el algoritmo de K-Means y cómo puede ayudar a la empresa.

2. Cargue el conjunto de datos de transacciones y realizar un análisis exploratorio de los datos para entender su estructura y las variables disponibles. Explique la estructura y utilidad del dataset.

3. Limpie y transforme los datos según sea necesario. Esto puede implicar la eliminación de valores faltantes, la conversión de variables categóricas en numéricas, la creación de nuevas variables (por ejemplo, total de ventas por cliente), y la normalización o estandarización de las variables.

4. Visualice los datos y las relaciones entre las variables. Esto puede incluir histogramas, gráficos de dispersión, y gráficos de caja. Justifique sus decisiones.

5. Aplique el algoritmo al conjunto de datos procesado. Experimente con diferentes números de clústeres (k) y observe cómo cambian los resultados.

6. Interprete los clústeres encontrados en el contexto del problema de negocio. ¿Cómo se diferencian los clústeres entre sí? ¿Cómo se pueden utilizar estos clústeres para informar las decisiones de marketing?

7. Visualice los clústeres en relación con las variables clave utilizando gráficos de dispersión y gráficos de barras.

8. Elabore un párrafo que explique las principales decisiones, sobre todo del paso 5, y los pasos seguidos. Investigue que utilidad tendría utilizar PCA y que aporte podría significar para el negocio.