Segmentación de productos y análisis por tienda usando BigMart (Kaggle)

Big Mart Sales Prediction

El dataset Big Mart Sales Prediction contiene datos de ventas de 1,559 productos en 10 tiendas de distintas ciudades, incluyendo atributos del producto y de la tienda:

https://www.kaggle.com/datasets/gauravduttakiit/big-mart-sales-prediction

Objetivo

Diseñar un flujo de Business Intelligence apoyado en clustering que:

1. Realice un clustering de productos (Item_Identifier) según su comportamiento de ventas y características.
2. Analice, para cada tienda (Outlet_Identifier), qué mezcla de clusters de productos vende.
3. Presente los resultados en un dashboard de BI que permita a la gerencia responder:
    - ¿Qué tipos de productos existen en nuestro portafolio?
    - ¿Qué tiendas venden más de cada tipo de producto?
    - ¿Qué oportunidades de negocio se observan por tienda/segmento de producto?

Fases de trabajo

Fase A. Descarga y exploración inicial

En esta fase, se realiza la descarga del dataset de Kaggle y se realizará una exploración básica del archivo principal (Train.csv). Deberán identificar el número de registros, el significado de las variables clave (por ejemplo, Item_Identifier, Item_MRP, Item_Type, Outlet_Identifier, Item_Outlet_Sales) y describir en uno o dos párrafos el contexto de negocio.

Fase B. Construcción del dataset a nivel producto

Aquí se construye un dataset agregado donde cada fila representa un producto (Item_Identifier). A partir de los datos originales producto–tienda, se calcularán métricas como ventas totales del producto (suma de Item_Outlet_Sales), ventas promedio por tienda, número de tiendas donde se vende cada producto, así como promedios de Item_MRP, Item_Weight y Item_Visibility. Además, se codificarán variables categóricas relevantes (como Item_Type y Item_Fat_Content) para obtener un conjunto de variables numéricas adecuado para el clustering, tratando valores faltantes y escalando los datos cuando sea necesario.

Fase C. Clustering de productos

Con el dataset a nivel producto listo, se seleccionará un subconjunto de variables numéricas (ventas, precio, visibilidad, presencia en tiendas, tipo de producto codificado, etc.) y se

aplicará dos aproximaciones de algoritmos de clustering (particional y jerárquico). En este punto se elegirá un número de clusters razonable y entrenará el modelo final, asignando a cada producto un cluster_producto. Finalmente, interpretará los clusters describiendo qué caracteriza a los productos de cada grupo.

Fase D. Análisis por tienda usando los clusters de producto

En esta fase se vuelve al dataset original producto–tienda, se incorpora la columna cluster_producto y se calcula, para cada tienda (Outlet_Identifier), la mezcla de clusters que vende. Se obtendrán métricas como el porcentaje de ventas de la tienda que proviene de cada cluster, el número de productos distintos por cluster y las ventas promedio por producto dentro de cada cluster. Con estas métricas se construirá un dataset a nivel tienda que permita responder preguntas como: qué tiendas dependen más de productos de alto volumen y bajo precio, cuáles venden más productos premium y cómo se relaciona esto con el tipo de tienda o su ubicación.

Fase E. Dashboard de Business Intelligence (2 vistas)

Diseñar un dashboard de BI que incluya al menos dos vistas:

1. Clusters de Productos
2. Mezcla de Clusters por Tienda

Entregables

1. Notebook con cada una de las fases.
2. Archivo del dashboard (ej. .pbix) con:
    - Vistas de clusters de productos.
    - Vistas de mezcla de clusters por tienda.
    - Filtros y elementos de navegación básicos.
3. Resumen ejecutivo (3 páginas):
    - Objetivo de negocio del proyecto.
    - Descripción clara de los clusters de productos.
    - Hallazgos clave por tienda/tipo de tienda.
    - Recomendaciones accionables.