PROYECTO: “Descubriendo los Secretos del Carrito de Compras: Un Análisis de Canasta en un Minimarket.”

Desarrollado por:

Juan Sebastián Hernández Ramírez
Jhocel Duvan Suescun Torres
Karen Rojas Giraldo
Juan Pablo Mogollón Avaunza

Limitación de uso de datos

Para el desarrollo de este proyecto se usará la data real de una cadena de mini supermercados de bajo costo. La empresa no ha autorizado el uso de esta información con fines académicos. Razón por la se ha cambiado su nombre real a “Abarrotes selectos”.

Resumen

El proyecto busca analizar y entender los patrones de compra de los clientes, enfocándose en los productos comprados en una misma transacción. El objetivo es mejorar la toma de decisiones para nuevas iniciativas de negocio, como la creación de combos para aumentar el ticket promedio y la fidelización de clientes, y la redistribución de productos en el local. Se realizará un análisis detallado de la composición de las compras por factura, sin considerar cantidades, utilizando datos de ventas detallados. Tras validar y corregir los datos, se llevará a cabo un análisis exploratorio de demanda, share de venta y asociaciones entre productos. Se transformarán los datos en una matriz binaria para aplicar los algoritmos Apriori y FP-Growth para descubrir asociaciones entre productos. También se considera la aplicación opcional de clustering y reducción de dimensionalidad para obtener y analizar resultados adicionales.

Contexto

Abarrotes Selectos es una cadena de mini supermercados de bajo costo que ofrece una variedad de productos de consumo inmediato, como alimentos preparados, productos para preparar comidas, artículos de aseo e higiene personal, y otros productos misceláneos. El gerente de la tienda quiere entender los hábitos de consumo de los clientes para diseñar estrategias que maximicen los ingresos. Para ello, ha compartido los detalles de las ventas del primer trimestre de 2024 y planea aumentar el ticket promedio y la fidelización mediante la creación de combos y la redistribución de productos en el local.

Introducción

En el competitivo mundo del retail, conocer los hábitos de compra de los clientes es fundamental para diseñar estrategias que maximicen las ventas. Nuestro minimarket, situado en un vecindario popular, ha visto un crecimiento constante, pero sospechamos que hay potencial no explotado en nuestras ventas. Para capitalizar estas oportunidades, hemos decidido llevar a cabo un análisis profundo de las transacciones mediante técnicas de aprendizaje no supervisado, enfocándonos en el análisis de canasta.

Pregunta/Problema a Resolver

El principal objetivo de nuestro proyecto es responder a las siguientes preguntas clave:

1. ¿Cuáles son las combinaciones de productos que los clientes suelen comprar juntos?
2. ¿Existen productos que, al ser promovidos juntos, podrían aumentar el valor de las compras?
3. ¿Podemos reorganizar la disposición de productos en el minimarket para incentivar la compra conjunta de ciertos artículos?

Para abordar estas preguntas, nos centraremos en tareas de clustering y asociación. El análisis de canasta se enfocará en descubrir reglas de asociación entre productos, mientras que el clustering podría ayudar a identificar segmentos de clientes con comportamientos de compra similares. Si encontramos que los datos

son demasiado complejos o numerosos, también podríamos considerar técnicas de reducción de dimensión para simplificar la interpretación.

Motivación

- Maximización de ingresos: Al identificar qué productos se compran juntos, se pueden crear promociones y combos que incentiven compras adicionales.
- Optimización del layout del minimarket: Mejorar la disposición de productos para facilitar el acceso a combinaciones de productos frecuentemente comprados juntos, lo que podría incrementar las ventas por impulso.
- Mejora de la satisfacción del cliente: Ofrecer una experiencia de compra más fluida y conveniente al anticipar sus necesidades y preferencias.

Clasificación del Problema

Este problema pertenece principalmente a la tarea de asociación (análisis de canasta). Sin embargo, podría incluir elementos de clustering para segmentar clientes y entender patrones de compra más detallados.

Revisión preliminar de antecedentes en la literatura

La identificación de patrones de compra ha sido un tema de estudio durante años con diversas técnicas. Un estudio de 2017 por Kaur y Sing aplicó K-Means para agrupar productos minoristas según el comportamiento de los clientes. Hernández y Villalobos (2021) utilizaron mapas auto-organizados (SOM) para detectar patrones y mejorar estrategias de marketing.

En Perú, Cam Gensollen (2022) desarrolló un proyecto sobre la segmentación de clientes y un sistema de recomendación en supermercados europeos, utilizando k-medias y la librería LightFM de Python. Este estudio también destacó la importancia del análisis exploratorio de datos y el proceso TDSP. Atencio et al. (2022) propusieron una segmentación en Tottus, aplicando preprocesamiento de datos, PCA, y técnicas de clustering como K-means, K-medoids y clustering jerárquico, evaluando la idoneidad de estos métodos mediante el codo y el dendograma.

El análisis de canasta, descrito por Efrat et al. (2020), examina los hábitos de compra para descubrir relaciones entre productos. El algoritmo Apriori se usa para identificar asociaciones entre productos, ajustando valores de soporte y confianza. Rao et al. (2023) aplicaron Apriori para identificar enfermedades frecuentes en áreas específicas. Singha et al. (2024) usaron el algoritmo FP-Growth para identificar asociaciones significativas y diseñar campañas de marketing cruzado. Mustakim et al. (2018) compararon Apriori y FP-Growth, concluyendo que FP-Growth es más eficiente para grandes conjuntos de datos debido a su capacidad para reducir exploraciones y generación de candidatos.

Exploración preliminar de los datos

Se obtuvo del sistema de facturación del minimarket, la facturación es detallada de un trimestre (enero a marzo de 2024), la base de datos cuenta 152.781 registros y 11 columnas así:

Fecha: fecha de la factura, **codigo_factura:** código de la factura, **mnt_total_a_pagar:** valor total a pagar, **cant_items:** cantidad de productos, **familia:** 11 grupos al que pertenece el artículo (Aseo, alimentos sal, bebidas,etc…), **categoría:** se encuentran 31 categorías, **subcategoria:** se encuentran 186 subcategorías, **codigo_producto:** código del producto, **cantidad_vendida:** cantidad de unidades vendidas del artículo, **venta_bruta_producto:** valor bruto del artículo, **venta_neta_producto:** valor neto del artículo.

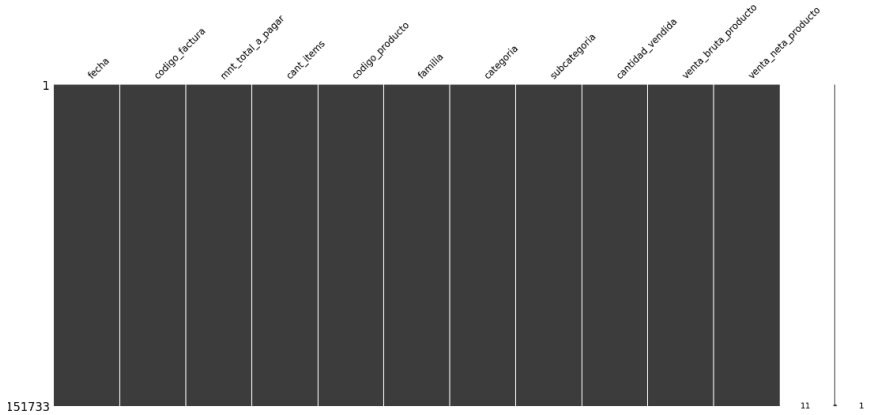
Completitud y coherencia de los datos

La revisión de la base de datos muestra que no hay datos faltantes en ninguna columna. Sin embargo, las columnas 'codigo_factura', 'codigo_producto', 'familia', 'categoria', y 'subcategoria' se cargaron como tipo objeto y deberán ser convertidas a categorías.

Las estadísticas de las columnas numéricas revelan valores negativos en 'mnt_total_a_pagar', 'cantidad_vendida', 'venta_bruta_producto', y 'venta_neta_producto', lo cual es inconsistente. Se aclaró que los valores negativos en 'cantidad_vendida', 'venta_bruta_producto', y 'venta_neta_producto' representan devoluciones, y que el valor negativo en 'mnt_total_a_pagar' se repite en todos los registros de una factura anulada.

Figura 1. Información de la base de datos y matriz de datos faltantes

```
<class 'pandas.core.frame.DataFrame'>
Index: 151733 entries, 0 to 52491
Data columns (total 11 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   fecha                151733 non-null  datetime64[ns]
 1   codigo_factura       151733 non-null  category
 2   mnt_total_a_pagar    151733 non-null  float64
 3   cant_items           151733 non-null  int64
 4   codigo_producto      151733 non-null  category
 5   familia              151733 non-null  category
 6   categoria            151733 non-null  category
 7   subcategoria         151733 non-null  category
 8   cantidad_vendida     151733 non-null  float64
 9   venta_bruta_producto 151733 non-null  float64
 10  venta_neta_producto  151733 non-null  float64
dtypes: category(5), datetime64[ns](1), float64(4), int64(1)
memory usage: 10.9 MB
```



Fuente: Elaboración Propia

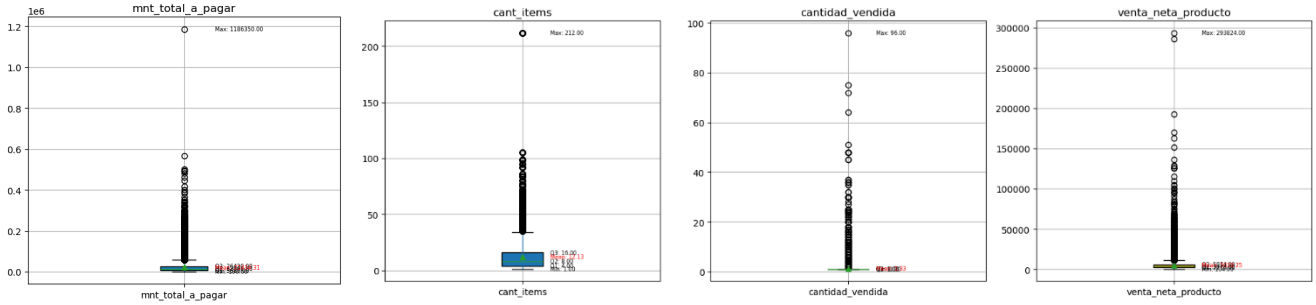
Se realizó una transformación en los datos agrupando por 'fecha', 'codigo_factura', 'codigo_producto', 'familia', 'categoria', y 'subcategoria', y sumando 'cantidad_vendida', 'venta_bruta_producto', y 'venta_neta_producto'. Se excluyeron las facturas con 'mnt_total_a_pagar' igual o menor a cero.

La base de datos resultante, con 151,733 registros (reducción de 152,781), contiene cinco variables numéricas, cinco categóricas y una de tipo fecha. Las variables numéricas son: 'mnt_total_a_pagar', 'cant_items', 'cantidad_vendida', 'venta_bruta_producto', y 'venta_neta_producto'; las categóricas son: 'codigo_factura', 'codigo_producto', 'familia', 'categoria', y 'subcategoria'.

Distribución de variables:

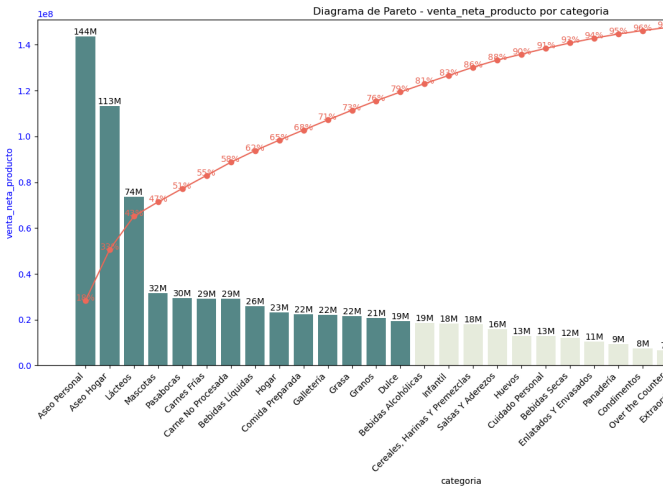
- **mnt_total_a_pagar:** Promedio de 22,431 con un rango extenso, mayormente concentrado en valores bajos y algunos valores atípicos altos.
- **cant_items:** Mediana baja (8), indicando que la mayoría de las compras incluyen pocos ítems, con algunos pedidos con un número inusualmente alto de ítems.
- **cantidad_vendida:** La mayoría de los valores están en el extremo inferior del rango, con varios valores atípicos en el extremo superior.
- **venta_neta_producto:** Alta concentración en valores bajos y un amplio rango de valores atípicos positivos.

Figura 2. Diagrama de cajas y bigotes, y estadísticas básicas para las variables numéricas



	cantidad_vendida	venta_bruta_producto	venta_neta_producto	cant_items	mnt_total_a_pagar
count	151733.00	151733.00	151733.00	151733.00	39123.00
mean	1.33	5784.49	5039.25	12.13	22434.31
std	1.50	5940.98	5217.42	13.63	30961.94
min	1.00	190.00	104.00	1.00	190.00
25%	1.00	2650.00	2276.00	4.00	5680.00
50%	1.00	4300.00	3773.00	8.00	12990.00
75%	1.00	6790.00	5874.00	16.00	26420.00
max	96.00	349650.00	293824.00	212.00	1186350.00

Fuente: elaboración propia

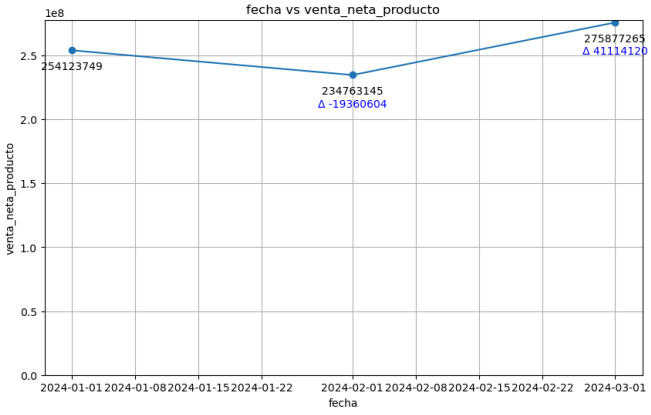


Distribución de las ventas. Pareto de las ventas por categoría.

Se puede observar que la categoría aseo personal es la más dominante, con un total de 144 millones en ventas netas, lo que representa el 18% del total acumulado, la categoría de aseo hogar sigue con 113 millones, que representa el 33% del acumulado con aseo personal y otras categorías como lácteos y mascotas también contribuyen significativamente con 74 y 43 millones respectivamente, alcanzando juntos un 47% del total acumulado (ver figura 2).

Evolución de las ventas netas a lo largo del trimestre

Se puede observar un decrecimiento de las ventas en el mes de febrero de casi 19 millones entre enero y febrero, por su parte para marzo se presenta un aumento de 41 millones, con este gráfico se puede observar una estacionalidad en las ventas, esto nos permite investigar qué productos específicos contribuyeron a este aumento y si hay combinaciones de productos que se vendieron conjuntamente. Esto podría ser útil para identificar oportunidades de promociones.



Se crearon dos tablas de contingencia para analizar la frecuencia de aparición de productos en diferentes categorías y familias, buscando asociaciones cruzadas. Se aplicó un umbral de frecuencia para filtrar los datos: 3000 para la tabla Familia-Familia y 2000 para la tabla Categoría-Categoría. Se eliminaron registros que no cumplían con estos umbrales y filas y columnas con solo valores NaN.

Resultados:

- Familias: Se encontraron asociaciones significativas entre la familia Aseo y otras familias como Alimentos Dulces, Alimentos Sal, Bebidas, Empaques y Bazar, y Refrigerados. La familia Aseo es la que más frecuentemente se compra junto a otras familias y tiene el mayor share de venta, sugiriendo que los productos de esta categoría son los mejores candidatos para combos.

- Categorías: Las categorías Aseo Personal y Aseo Hogar tienen una fuerte correlación con las categorías Dulce, Hogar, Lácteos y Otros.
- Estos hallazgos indican posibles asociaciones que deberían ser consideradas en el proyecto para diseñar estrategias de marketing y combos.

Tabla 1. Tabla contingencia Familia-Familia.

familia familia	Alimentos Dulces	Alimentos Sal	Aseo	Bebidas	Empaques y Bazar	Refri
Alimentos Dulces	NaN	4699.0	5709.0	4527.0	NaN	
Alimentos Sal	4699.0	NaN	5898.0	4841.0	NaN	
Aseo	5709.0	5898.0	NaN	5340.0	3987.0	
Bebidas	4527.0	4841.0	5340.0	NaN	NaN	
Empaques y Bazar	NaN	NaN	3987.0	NaN	NaN	
Refrigerados	3495.0	3798.0	4073.0	3217.0	NaN	

Fuente: elaboración propia

Tabla 2. Tabla contingencia Categoría - Categoría.

categoria categoria	Aseo Hogar	Aseo Personal	Dulce	Hogar	Lácteos	Otros
Aseo Hogar	NaN	6609.0	NaN	2852.0	2911.0	2575.0
Aseo Personal	6609.0	NaN	2055.0	2975.0	3315.0	2767.0
Dulce	NaN	2055.0	NaN	NaN	NaN	NaN
Hogar	2852.0	2975.0	NaN	NaN	NaN	NaN
Lácteos	2911.0	3315.0	NaN	NaN	NaN	NaN
Otros	2575.0	2767.0	NaN	NaN	NaN	NaN

Fuente: elaboración propia

Propuesta metodológica

Recolección de Datos:

- Recopilaremos todas las facturas de los últimos seis meses, incluyendo fecha, productos comprados y monto total.

Limpieza de Datos:

- Eliminación de Duplicados: Eliminaremos registros duplicados.
- Corrección de Errores Tipográficos: Verificaremos y unificaremos nombres de productos.
- Gestión de Datos Faltantes: Eliminaremos facturas con datos críticos faltantes y recuperaremos o eliminaremos registros incompletos.

Preparación de Datos para el Análisis:

- Transformación en Matriz de Transacciones: Convertiremos las facturas en una matriz binaria para representar la compra de productos.
- Codificación: Aplicaremos técnicas de codificación si es necesario.

Análisis Exploratorio:

- Realizaremos un análisis para entender la distribución de ventas y observar patrones preliminares.
- Utilizaremos gráficos para visualizar transacciones y productos.

Aplicación de Algoritmos de Aprendizaje No Supervisado:

- Asociación (Análisis de Canasta): Aplicaremos Apriori y FP-Growth para descubrir reglas de asociación entre productos.
- Clustering (Opcional): Utilizaremos k-means para segmentar clientes según patrones de compra, si se decide segmentar.
- Reducción de Dimensión (Opcional): Consideraremos PCA para reducir la complejidad si los datos son altamente dimensionales.

Interpretación de Resultados:

- Analizaremos las reglas de asociación y, si se aplicó clustering, interpretaremos los segmentos de clientes para entender patrones de compra.

Referencias

Cam Gensollen, C. R. (2022). Big data en el mundo del retail: segmentación de clientes y sistema de recomendación en una cadena de supermercados de Europa. Ingeniería Industrial, 189-216. https://doi.org/10.26439/ing.ind2022.n.5808

Atencio Manyari, S. A., De la Rosa Flores, H., Hilario Maravi, S., Navarro Huarcaya, M., & Rosas Vivanco, D. M. (2022). Propuesta de segmentación de clientes aplicando técnicas de Machine Learning para mejorar la experiencia de compra mediante un sistema de recomendación de productos de Tottus.

Efrat, A. R., Gernowo, R., & Farikhin. (2020). Consumer purchase patterns based on market basket analysis using apriori algorithms. Journal of Physics: Conference Series, 1524(1), 012109. https://doi.org/10.1088/1742-6596/1524/1/012109

Hernández, J., & Villalobos, M. (2021). Análisis de Canasta de mercado en supermercados mediante mapas auto-organizados. arXiv. https://arxiv.org/pdf/2107.10647

Kaur, H., & Singh, S. P. (2017). Clustering retail products based on customer behavior. Applied Soft Computing, 60, 752-761. https://doi.org/10.1016/j.asoc.2017.02.048

Mustakim, M., Herianda, D. M., Ilham, A., Daeng GS, A., Laumal, F. E., Kurniasih, N., Iskandar, A., Manulangga, G., Iswara, I. B. A. I., & Rahim, R. (2018). Market Basket Analysis Using Apriori and FP-Growth for Analysis Consumer Expenditure Patterns at Berkah Mart in Pekanbaru Riau. IOP Conference Series: Journal of Physics: Conference Series, 1114(1), 012131. https://doi.org/10.1088/1742-6596/1114/1/012131

Rao, A. B., Kiran, J. S., & G., P. (2023). Application of market–basket analysis on healthcare. International Journal of System Assurance Engineering and Management, 14(S924-S929). https://doi.org/10.1007/s13198-021-01298-2

Singha, K., Parthanadee, P., Kessuvan, A., & Buddhakulsomsiri, J. (2024). Market Basket Analysis of a Health Food Store in Thailand: A Case Study. International Journal of Knowledge and Systems Science, 15(1). https://doi.org/10.4018/IJKSS.333617