Sistema de recomendación de productos para una tienda en línea

Tabla: Integrantes Grupo 7

Nombre	Código	Correo
Miryam Alejandra Maturna Cordoba	202121442	m.maturanac@uniandes.edu.co
Daniel Alfaro Rojas	202121104	d.alfaror@uniandes.edu.co
Daniel Andrés Londoño Galvis	200422981	dan-lond@uniandes.edu.co
Kevin Alexander Leguizamo Almanza	201924222	k.leguizamoa@unaindes.edu.co

1 Resumen

"Olist Store" es una empresa brasileña con un Ecommerce que reúne varias marcas dentro de su página y para incrementar sus ventas necesita resolver la pregunta:

¿Qué productos se pueden recomendar al cliente a partir de los históricos de compras realizadas en la tienda?

Para resolver la pregunta, a partir de los datos transaccionales históricos entregados, se exploran diferentes métodos no supervisados denominados sistemas de recomendación, tales como: reglas de asociación con algoritmo Apriori, filtrado colaborativo con media ponderada y filtrado colaborativo basado en embeddings.

Con base en los resultados obtenidos y las dificultades presentadas, se recomienda el uso del algoritmo Apriori debido a que se puede utilizar toda la base de datos y que su costo computacional es reducido en comparación a los otros métodos.

Finalmente se evidencia que es recomendable tener otros datos como las descripciones de los productos, los datos demográficos de los usuarios u otra información de contexto como los ingresos y el clima, con lo que se podría integrar más información al algoritmo y producir recomendaciones más sofisticadas y específicas para responder la pregunta.

2 Introducción

Las tiendas en línea buscan incrementar sus ventas por medio de atraer nuevos clientes y retenerlos. Hoy, los usuarios llegan, navegan, se registran, compran o abandonan la tienda; y en cada interacción se generan infinidad de registros con los que se pueden identificar patrones de comportamiento y así predecir sus necesidades de compra con la ayuda de la analítica de datos.

Esta problemática le sucede a "Olist Store", empresa brasileña con un Ecommerce que reúne varias marcas dentro de su página. Teniendo en cuenta lo anterior surge la pregunta:

¿Qué productos se pueden recomendar al cliente a partir de los históricos de compras realizadas en la tienda?

Para dar respuesta a esta pregunta se propone el uso del aprendizaje no supervisado, el cual es uno de las categorías de Machine Learning, en la que se parte de datos sin etiquetas o clases previamente definidas y su objetivo es encontrar grupos similares en el conjunto de datos, detectando relaciones y/o tendencias que puedan ayudar a una empresa a tomar decisiones. En este caso, lo que se busca es tomar todo el histórico de productos comprados por pedidos realizados a "Olist Store" y por medio de un sistema de recomendación de canasta, recomendar al cliente productos adicionales para su compra.

Dicha recomendación se dará haciendo uso del algoritmo Apriori, el cual se utiliza para encontrar grupos de itemsets que aparecen juntos con frecuencia en un conjunto de datos y para Olist Store, este puede ser el primer paso para encontrar nuevas formas de promocionar los productos. Este algoritmo fue propuesto por Agrawal en 1994 y menciona

que “todo subconjunto de un conjunto de ítems frecuentes también será un conjunto de ítems frecuentes” (Agrawal & Srikant, 1994, 487-499). Es por esta razón que el algoritmo siempre obtendrá en primer lugar todo conjunto de datos más frecuentes y después de hacer esta evaluación irá, cuantas veces se le indique al mismo, formando los demás subconjuntos que encuentre.

En el ámbito académico, se ha buscado optimizar el mismo o crear uno nuevo con base al propuesto por Agrawal. Uno de estos casos es el trabajo de Chen en 2002, en donde forma los conjuntos generados por el algoritmo pero con reglas mucho más simples sin llegar a perder información al abordarlo y sobre todo sin incurrir a un gran costo computacional (CHEN, 2002, 721-733).

Sin embargo, como todo algoritmo, tiene sus limitaciones y Apriori no es la excepción. La mayor limitación de Apriori, son la cantidad de veces que se tienen que escanear todos los datos, en busca de las frecuencias entre ellos. Esto hace que el algoritmo no pueda aplicarse en situaciones con una cantidad de datos considerables, pero de acuerdo a la literatura, se han desarrollado diferentes adaptaciones para que esta limitación sea mínima (MARTÍNEZ MANZO, 2020)

3 Materiales y Métodos

La empresa en análisis permite el acceso a diferentes fuentes de información (https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_order_reviews_dataset.csv) de las cuales las siguientes son las tablas relevantes para responder la pregunta planteada:

Tabla: Descripción de las tablas

Dataset	Tabla	Descripción	Registros
Original **Fuente:** https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_order_reviews_dataset.csv	**Órdenes**	Relación de Orden y Cliente entre otras características	99441
	Órdenes-Items	Relación de Productos comprados en cada Orden entre otras características	112650
	Productos	Productos y sus características	32951
	Clientes	Clientes y sus características	99441
	Calificaciones	Calificaciones realizadas por los Clientes a las Órdenes	99224
Final	**Órdenes-Items-Consolidado**	Tabla resultante de cruzar de las tablas originales las columnas que aportan información para resolver la pregunta	112650

3.1 Pre procesamiento

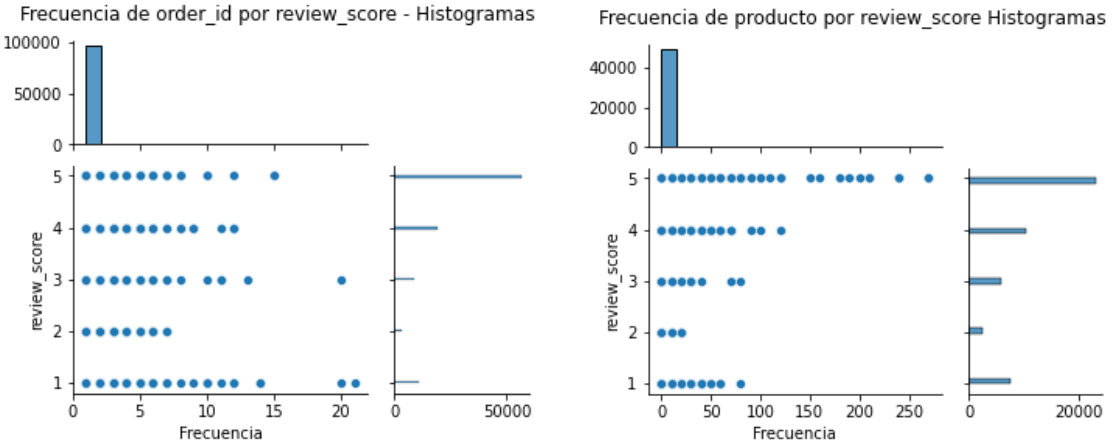
A partir de las tablas originales se crea una nueva tabla denominada Órdenes-Items-Consolidado resultante del cruce de las diferentes tablas y compuesta por las columnas que se consideran necesarias para resolver la pregunta, creando una nueva columna denominada ‘producto’, resultante de la combinación de la categoría de producto y el id del producto.

Al realizar el cruce se identifican faltantes en las columnas de Categorías y review_score, las cuales son imputadas por “no_category” y “3” respectivamente.

Tabla: Descripción de las variables de la tabla Órdenes-Items-Consolidado

Variable	Descripción	Únicos	Frecuencia	Tipo
order_id	ID de Órden	98666	21	Categórica
producto	Unión de Categoría y ID de Producto	32951	527	Categórica
customer_id	ID de Cliente	98666	21	Categórica
review_score	Calificación dada por el Cliente a la Órden	5		Numérica

Imagen: Gráficas descriptivas de review_score por order_id y producto



A partir de las gráficas se identifica que:

- El 50% de los datos tienen calificación (review_score) de cinco (5)
- El 78% de las órdenes de compra solo tenían un (1) producto
- El 16% de los productos ofertados se compró solo una (1) vez

3.2 Algoritmo utilizado

Se utilizan las “Reglas de Asociación con algoritmo Apriori”, conformado por dos etapas:

1. Identificar todos los productos que ocurren con una frecuencia por encima de un determinado límite (productos frecuentes)
2. Convertir esos productos frecuentes en reglas de asociación. Este método permite el uso de la matriz de calificaciones con el 100% de los datos, pero solo utilizando los que superan el soporte por lo que es más eficiente computacionalmente.

4 Resultados y Discusión

Reglas de Asociación con algoritmo Apriori:

Se busca identificar la frecuencia de productos que existen entre cada una de las órdenes realizadas en Olist Store (order_id y producto) con un análisis de canasta de compra.

Se seleccionan los siguientes valores para los parámetros del mismo:

- min_support=0.00003, este parámetro ayuda a seleccionar los elementos con valores de soporte superiores al que se le indique. El soporte es el número de transacciones que contienen un producto dividido por el número total de transacciones.
- min_confidence=0.1, la función de este parámetro es filtrar todas las reglas que tengan mínimo este valor, es decir, es un umbral para filtrar reglas. La confianza es la probabilidad de que también se compre el artículo Y si se compra el artículo X. Se calcula como el número de transacciones que contienen X e Y dividido por el número de transacciones que contienen X.
- min_lift=1, este parámetro filtra las reglas generadas por medio del umbral que se le ingrese. Para poder examinar esa asociación de los productos evaluados se da el valor de 1. Se calcula como el cociente de la confianza y el soporte.
- min_length=2, este parámetro indica la cantidad mínima de elementos que se desean en las reglas que se generen. Se decide dejarlo en 2 teniendo en cuenta que se espera que la cantidad mínima de elementos en las reglas sea 2 productos.

Como resultado de estos parámetros se obtuvieron 51 reglas de asociación.

Tabla: Resultados algoritmo Apriori

Producto Recomendado	Producto	Soporte	Confianza	Lift
automotor_060cb193	automotor_98d61056	0.000061	0.240000	623.153684
automotor_f4f67cca	automotor_4fcb3d9a	0.000172	0.191011	336.541332

automotor_a50acd33	automotor_dfb97c88	0.000051	0.147059	174.815734
belleza_salud_0a4093a4	belleza_salud_e0cf7976	0.000041	0.100000	75.317557
belleza_salud_9bb8ca33	belleza_salud_b8a0d73b	0.000030	0.100000	308.331250

Como se observa, primero se obtiene la regla de asociación de la combinación de productos X -> Y que se venden juntos, con un ***Soporte*** de menos del 1%, una ***Confianza*** que oscila entre el 10% y 25% con algunas excepciones que son superiores al 50%, esta es la probabilidad de que al comprarse el producto Y también se compre el producto X. Finalmente tenemos el ***Lift*** mayor a 70, el cual nos indica de que al comprar el producto Y es tantas veces más probable que se compre también el producto X.

Filtrado colaborativo con media ponderada:

Para aplicar el algoritmo se trató de realizar la matriz pivote de clientes contra productos y como valor se estableció el review de la compra, pero por el alto coste computacional de esta operación no se pudo realizar el pivote ya que se estaban cruzando más de 99 mil clientes y 32 mil productos.

Teniendo en cuenta esta limitación se excluyeron los productos que tuvieron menos de dos calificaciones y sus respectivos usuarios, obteniendo como resultado una matriz pivote de 9803 usuarios contra 8157 productos y como valores sus reviews, la cual ya se puede procesar pero tiene el inconveniente de pérdida de información generada por la exclusión realizada, aunque sirve para identificar los productos con una alta rotación de ventas que se pueden recomendar a los usuarios que todavía no los han comprado.

Sobre la matriz se identificó la relación de los productos teniendo en cuenta la similaridad de los usuarios por medio de la similitud de coseno, obteniendo 5 recomendaciones para cada usuario.

Tabla: Resultados Filtrado colaborativo con media ponderada

Producto Recomendado	Promedio
herramientas_jardin_368c6c73	3.4
agro_industria_y_comercio_026f43af	0.0
muebles_decoracion_586bb93b	0.0
muebles_decoracion_61b82c5d	0.0
muebles_decoracion_60cfed7b	0.0

Filtrado colaborativo basado en embeddings:

Con el método de descomposición en valores singulares se busca una representación más compacta de la matriz de calificaciones en donde se encuentra una estructura latente de los datos en un subespacio de baja dimensión conocido como embedding.

El primer paso es filtrar para dejar en la matriz usuarios que tengan al menos 3 calificaciones.

La idea es descomponer la matriz de calificaciones A en la mejor representación de menor dimensión de esta matriz así: $A=U\Sigma V_T$. Aquí U es la matriz de embeddings de usuarios que también se pueden entender como la representación de cuánto le "gusta" a los usuarios cada característica, V es la matriz de embeddings de productos y representa cuán relevante es cada característica al producto, y Σ es la matriz diagonal singular que contiene los valores singulares.

Aplicando el método se obtienen 5 recomendaciones para cada usuario.

Tabla: Resultados Filtrado colaborativo con base en Embeddings

Producto Recomendado	Promedio
herramientas_jardin_368c6c73	0.999937
muebles_decoracion_be0a8cb3	0.003781

muebles_decoracion_a5341e3f	0.003781
utilidades_domesticas_8273821f	0.003781
muebles_decoracion_ea413ac4	0.003781

El algoritmo escogido fue Apriori ya que permite el uso de toda la base de datos sin reducirla ni filtrarla, manteniendo toda la información y arrojando una cantidad considerable de resultados con los cuales se puede hacer una recomendación al cliente de acuerdo a los artículos comprados frecuentemente y además su costo computacional es reducido en comparación con los otros algoritmos.

La principal desventaja de Apriori es que se requieren datos previos sobre compras y usuarios, para hacer inferencias, los cuales no están disponibles para todos los productos. Lo anterior, hace que no sea posible utilizar este tipo de sistemas en la etapa inicial de la vida de un negocio o con productos nuevos y sin compras, situación conocida como problema de "arranque frío" o "cold start". Otra desventaja es que el algoritmo requiere que se definan una cantidad de parámetros iniciales y esta definición inicial tiene un impacto en las reglas que se obtienen como resultado por lo que se requiere ser muy cuidadoso en la elección y con los objetivos claros sobre qué constituye una regla aceptable.

Este método reforzaría la compra de los productos comprados frecuentemente y no ayudaría a los productos que no se han comprado antes. Este algoritmo al trabajar por frecuencias y probabilidades funciona mucho mejor entre más datos se tengan disponibles.

Los sistemas recomendación por filtrado colaborativo y por embedding generan recomendaciones con base en información diferente, pues las reglas de asociación cuentan las frecuencias de cada producto mientras que filtrado colaborativo por usuarios hace una medida de la similaridad de usuarios con base en sus calificaciones y filtrado por embeddings hace una unión entre usuarios y productos para crear la matriz de calificación por lo cual cada uno tienen una aplicación distinta.

Finalmente se evidencia que es recomendable tener otros datos como las descripciones de los productos, los datos demográficos de los usuarios u otra información de contexto como los ingresos y el clima, con lo que se podría integrar más información al algoritmo y producir recomendaciones más sofisticadas y específicas para responder la pregunta.

5 Conclusión

La empresa brasileña “Olist Store” desea resolver la pregunta **¿Qué productos se pueden recomendar al cliente a partir de los históricos de compras realizadas en la tienda?** y para resolverla, se exploraron diferentes métodos no supervisados denominados sistemas de recomendación tales como: reglas de asociación con algoritmo Apriori, filtrado colaborativo con media ponderada y filtrado colaborativo basado en embeddings.

Teniendo en cuenta los resultados obtenidos y las dificultades presentadas se recomienda el uso del algoritmo Apriori debido a que se puede utilizar toda la base de datos y a su reducido costo computacional en comparación a los otros algoritmos.

La definición de los parámetros del algoritmo Apriori es relevante para el resultado de las reglas de asociación por lo que debe hacerse entendiendo muy bien el contexto de los datos y la definición previa de qué es una regla aceptable para el mismo.

Es recomendable tener las descripciones de los productos, los datos demográficos de los usuarios u otra información de contexto como los ingresos y el clima, con lo que se podría integrar más información al algoritmo y producir recomendaciones más sofisticadas y específicas para cada usuario.

6 Bibliografía

Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer International Publishing.

Ajitsaria, A. (n.d.). *Build a Recommendation Engine With Collaborative Filtering – Real Python*. Real Python. Retrieved September 1, 2022, from https://realpython.com/build-recommendation-engine-collaborative-filtering/

Ajitsaria, A. (n.d.). *Build a Recommendation Engine With Collaborative Filtering – Real Python*. Real Python. Retrieved September 2, 2022, from https://realpython.com/build-recommendation-engine-collaborative-filtering/

Asobancaria. (2019, December 2). *E-Commerce, crecimiento y ecosistema digital en Colombia*. Asobancaria. Retrieved September 2, 2022, from https://www.asobancaria.com/wp-content/uploads/1213.pdf

Bohanec, M., & Kljajić Borštnar, M. (2017, 4 1). Explaining machine learning models in sales predictions. *ScienceDirect*, *71*(2017), 428. https://www.sciencedirect.com/science/article/abs/pii/S0957417416306327

CEPAL. (2020, November 26). *El e-commerce en tiempos de COVID-19*. Cepal. Retrieved September 2, 2022, from https://www.cepal.org/sites/default/files/presentations/redlas_e-commerce_astarloa_0.pdf

Escalera, S., Seguí, S., Pujol, O., Dantí, F., Garrido, L., Radeva, P., Igual, L., & Puertas, E. (2017). *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*. Springer International Publishing.

Felfernig, A., Friedrich, G., Zanker, M., & Jannach, D. (2011). *Recommender Systems: An Introduction*. Cambridge University Press.

Galton, F. (n.d.). *1 Introducción | Machine Learning: Teoría y Práctica*. Bookdown. Retrieved September 3, 2022, from https://bookdown.org/victor_morales/TecnicasML/introducci%C3%B3n.html

Gorakala, S. K. (2016). *Building Recommendation Engines*. Packt Publishing.

Hall, M. (n.d.). *Capítulo 11 Aprendizaje No supervisado | Data Science con R*. Bookdown. Retrieved September 2, 2022, from https://bookdown.org/dparedesi/data-science-con-r/aprendizaje-no-supervisado.html

Jeong, Y. (2021, April 20). *Item-Based Collaborative Filtering in Python | by Yohan Jeong*. Towards Data Science. Retrieved September 2, 2022, from https://towardsdatascience.com/item-based-collaborative-filtering-in-python-91f747200fab

ND. (n.d.). *EL MACHINE LEARNING A TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD DENNIYE HINESTROZA RAMÍREZ JUAN MANUEL CÁRDENAS*. EL MACHINE LEARNING A TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD DENNIYE HINESTROZA RAMÍREZ JUAN MANUEL CÁRDENAS. Retrieved September 2, 2022, from https://repository.unilibre.edu.co/bitstream/handle/10901/17289/EL%20MACHINE%20LEARNING.pdf?sequence=1&isAllowed=y

ND. (n.d.). *Lección 7: Aprendizaje no supervisado: Agrupamiento y reglas de asociación Dentro del aprendizaje automático, además del ap*. UniMOOC. Retrieved September 2, 2022, from https://data.unimooc.com/materiales-cursos/machine-learning/Machine-Learning-7.pdf

ND. (n.d.). *PRODUCT RECOMMENDATIONS USING DATA MINING AND MACHINE LEARNING ALGORITHMS*. ARPN Journals. Retrieved September 2, 2022, from http://www.arpnjournals.org/jeas/research_papers/rp_2017/jeas_1017_6376.pdf

ND. (2001, December 12). *APLICACIÓN DE APRENDIZAJE NO SUPERVISADO AL ESTUDIO DE VIGILANCIA TECNOLÓGICA SOBRE LOS CENTROS DE INVESTIGACIÓN Y DESARROLLO*. Repositorio Universidad Autónoma de Bucaramanga. Retrieved September 2, 2022, from https://repository.unab.edu.co/bitstream/handle/20.500.12749/16263/2022_Tesis_Liceth_johama_Alvarado.pdf?sequence=1

Rogel-Salazar, J. (2017). *Data Science and Analytics with Python*. Taylor & Francis.

Roy, K. (2017, 10 1). PRODUCT RECOMMENDATIONS USING DATA MINING AND MACHINE LEARNING ALGORITHMS. *ARPN Journal of Engineering and Applied Sciences*, *12*(19), 9. http://www.arpnjournals.org/jeas/research_papers/rp_2017/jeas_1017_6376.pdf

Roy, K., Choudhary, A., & Jayapradha, J. (2017, 10 19). PRODUCT RECOMMENDATIONS USING DATA MINING AND MACHINE LEARNING ALGORITHMS. *ARPN Journal of Engineering and Applied Sciences*, *12*(19), 9. http://www.arpnjournals.org/jeas/research_papers/rp_2017/jeas_1017_6376.pdf