Sistema de recomendación de canasta 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

Hoy en día, en las tiendas en línea, los usuarios llegan, navegan, se registran, compran o abandonan la tienda, generando una infinidad de registros en cada interacción; “Olist Store” es una de estas tiendas en línea y quiere aprovechar su histórico de registros de órdenes de compra para incrementar sus ventas, por lo que se plantea el problema de:
¿Qué productos se pueden recomendar al cliente a partir de las compras realizadas por usuarios similares o por compras anteriores hechas por él mismo?

Para dar respuesta a esta pregunta en el presente documento, se comienza con una revisión preliminar de la literatura relacionada al problema, se describen los datos disponibles entregados por la empresa y finalmente se presenta una propuesta de metodología de aprendizaje no supervisado para la recomendación de productos como alternativa de respuesta a la pregunta planteada.

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; 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 también 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 las compras realizadas por usuarios similares o por compras anteriores hechas por él mismo?

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 filtrado colaborativo, recomendar al cliente productos adicionales para su compra.

3 Revisión preliminar de la literatura

Que productos vender, cuando y donde, se ha convertido en una pregunta reiterativa para las áreas de mercadeo y ventas de las compañías, que productos se deben ofrecer ha sido un tema que históricamente se ha desarrollado intuitivamente con conocimientos subjetivos de un experto. Hoy con el avance de la tecnología se han generado nuevas estrategias para la toma de estas decisiones.

En la universidad de Maribor, se a realizado una investigación sobre el impacto de los modelos de aprendizaje automático en las predicciones de ventas, la idea del proyecto es entender la complejidad de la dinámica empresarial que a menudo obliga a los tomadores de decisiones a tomar decisiones basadas en modelos mentales subjetivos, que reflejan su experiencia. Sin embargo, la investigación ha demostrado que las empresas se desempeñan mejor cuando aplican la toma de decisiones basada en datos. Esto crea un incentivo para introducir modelos de decisión inteligentes basados en datos, que son integrales y respaldan la evaluación interactiva de las opciones de decisión necesarias para el entorno empresarial. (Bohanec & Kljajić Borštnar, 2017, 2-10). Han propuesto una nueva metodología de explicación general, que apoya la explicación de modelos de predicción de caja negra de última generación. Presentan un uso novedoso de esta metodología dentro de un sistema inteligente en un caso real de pronóstico de ventas de empresa a empresa (B2B) por medio de un modelo de recomendación construido a partir de Support Vector Machine y un modelo de clasificación.

En otro estudio se encuentra que a lo largo del tiempo se han producido diferentes estrategias para la generación de estas recomendaciones, en una investigación de la universidad SRM en China, realizaron un proyecto donde generaban recomendaciones utilizando minería de datos y algoritmos de aprendizaje automático, este proyecto plantea que con el aumento en la demanda de los sitios web de comercio electrónico, surgen muchos problemas debido a que los usuarios enfrentan dificultades para encontrar la información relevante que coincida con sus preferencias. Por lo tanto, representaron un sistema que recomienda productos alimenticios al usuario en función de su compra. El producto alimenticio se recomienda en función de la salud del día a día y enfermedades del usuario, esta información se genera a través de un perfilamiento del usuario utilizando minería de datos y posteriormente un modelo de recomendación con Support Vector Machine. (Roy, 2017, 1-9).

Al igual que en el presente proyecto, las investigaciones mencionadas buscan crear un sistema de recomendación que ayude a las compañías a tomar decisiones basadas en datos encontrando similitudes con el presente proyecto en cuanto que utilizan sistemas de reconocimiento de patrones de compra y se diferencian en los métodos usados, ya que se enfocan principalmente en Support Vector Machine mientras que en este proyecto se propone el uso de un sistema de recomendación de filtrado colaborativo.

4 Descripción de los datos

La empresa en análisis permite el acceso a diferentes fuentes de información de las cuales las siguientes son las relevantes para responder la pregunta planteada:

- Órdenes: Relación de Orden y Cliente entre otras características.
- Órdenes-Items: Relación de Productos comprados en cada Orden entre otras características.
- Productos: Productos y sus características.
- Clientes: Clientes y sus características.

Las tablas aparentemente vienen depuradas pero se verificará su consistencia, se extraerán las columnas necesarias para resolver la pregunta y se realizarán los cruces respectivos para aplicar los algoritmos propuestos.

Tabla: Descripción tabla Órdenes

index	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date
count	99441	99441	99441	99441	99281	97658	96476	99441
unique	99441	99441	8	98875	90733	81018	95664	459
top	e48…	9ef…	delivered	2018-04-11 10:48:14	2018-02-27 04:31:10	2018-05-09 15:48:00	2018-05-08 23:38:46	2017-12-20 00:00:00
freq	1	1	96478	3	9	47	3	522

Tabla de Órdenes: Consta de 99.441 registros y las columnas que pueden ser útiles para el proyecto son **order_id** como conector de Órdenes y **customer_id** como conector de Cliente.

Tabla: Descripción tabla Órdenes-Items

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value
count	112650	112650.000000	112650	112650	112650	112650.000000	112650.000000
unique	98666	NaN	32951	3095	93318	NaN	NaN
top	827…	NaN	aca…	656…	2017-07-21 18:25:23	NaN	NaN
freq	21	NaN	527	2033	21	NaN	NaN
mean	NaN	1.197834	NaN	NaN	NaN	120.653739	19.990320
std	NaN	0.705124	NaN	NaN	NaN	183.633928	15.806405
min	NaN	1.000000	NaN	NaN	NaN	0.850000	0.000000
25%	NaN	1.000000	NaN	NaN	NaN	39.900000	13.080000
50%	NaN	1.000000	NaN	NaN	NaN	74.990000	16.260000
75%	NaN	1.000000	NaN	NaN	NaN	134.900000	21.150000
max	NaN	21.000000	NaN	NaN	NaN	6735.000000	409.680000

Gráfica: Frecuencia de Productos

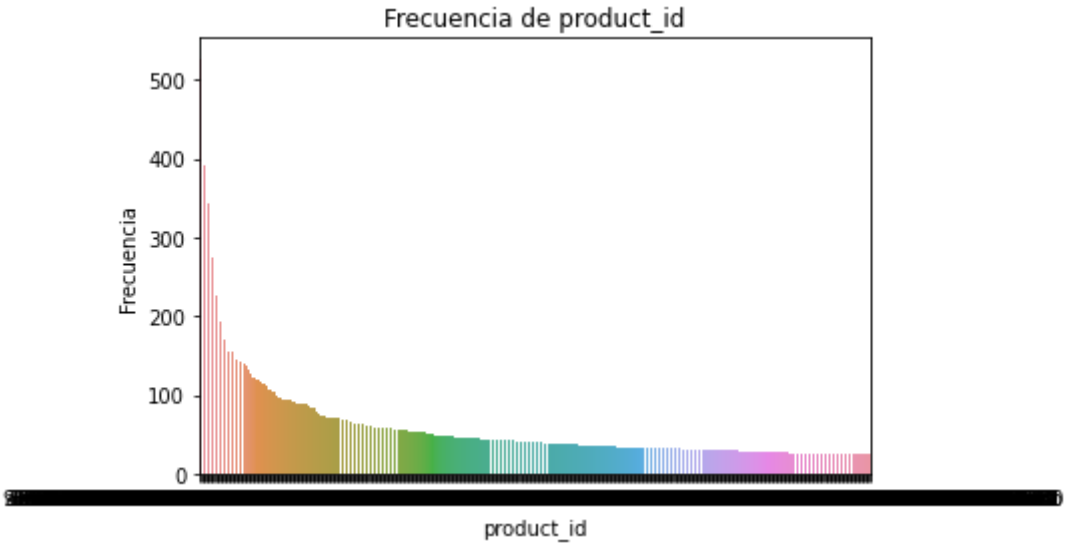

Tabla de Órdenes-Items: Es la tabla más importante para el proyecto, ya que en ella se encuentran los insumos de datos de Órdenes y Productos necesarios para el desarrollo de los algoritmos de recomendación mediante filtrado colaborativo, consta de 112.650 registros y las columnas más útiles para el proyecto son **order_id** como conector de Órdenes y **product_id** como conector de Producto.

Tabla: Descripción tabla Productos

	product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm
count	32951	32341	32341	32341	32341	32949	32949	32949	32949

unique	32951	73	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	1e9...	cama_mesa_banho	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	1	3029	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	48.4	771.4	2.1	2276.4	30.8	16.9	23.1
std	NaN	NaN	10.2	635.1	1.7	4282.0	16.9	13.6	12.0
min	NaN	NaN	5	4	1	0	7	2	6
25%	NaN	NaN	42	339	1	300	18	8	15
50%	NaN	NaN	51	595	1	700	25	13	20
75%	NaN	NaN	57	972	3	1900	38	21	30
max	NaN	NaN	76	3992	20	40425	105	105	118

Tabla de Productos: Consta de 32.341 registros y las columnas que pueden ser útiles para el proyecto son **product_id** como conector de Producto y **product_category_name** como clase de Producto.

Tabla: Descripción tabla Clientes

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
count	99441	99441	99441.000000	99441	99441
unique	99441	96096	NaN	4119	27
top	06b...	8d5...	NaN	sao paulo	SP
freq	1	17	NaN	15540	41746
mean	NaN	NaN	35137.474583	NaN	NaN
std	NaN	NaN	29797.938996	NaN	NaN
min	NaN	NaN	1003.000000	NaN	NaN
25%	NaN	NaN	11347.000000	NaN	NaN
50%	NaN	NaN	24416.000000	NaN	NaN
75%	NaN	NaN	58900.000000	NaN	NaN
max	NaN	NaN	99990.000000	NaN	NaN

Tabla de Clientes: Consta de 99.441 registros y las columnas que pueden ser útiles para el proyecto son **customer_id** como conector de Cliente y **customer_city** como ciudad de residencia del Cliente.

Teniendo en cuenta que los datos disponibles corresponden únicamente a la base de datos transaccional y que no hay otros datos que aporten mayor contexto, como los detalles demográficos de los usuarios, entonces el problema a resolver se limita a filtrado colaborativo ya que los otros algoritmos de sistemas de recomendación requieren datos con un mayor contexto.

5 Propuesta metodológica

Para resolver la pregunta planteada en el proyecto se propone el uso de aprendizaje no supervisado por medio de sistemas de recomendación entre los cuales se encuentran:

- Métodos de filtrado colaborativo
- Métodos basados en el contenido.
- Métodos basados en el conocimiento.
- Métodos basados en conocimiento externo o de contexto

Se plantea el uso de modelos de filtrado colaborativo que utilizan el poder colaborativo de las calificaciones proporcionadas por múltiples usuarios para hacer recomendaciones. La idea básica es que las calificaciones no especificadas o a predecir pueden ser imputadas a partir de las calificaciones observadas ya que a menudo están altamente correlacionadas entre varios usuarios y productos.

Teniendo en cuenta que los datos disponibles para el ejercicio, son los pedidos comprados por orden en lugar de un rating específico de un usuario a un producto, se utilizan los filtros colaborativos con base en la matriz de calificaciones conocida como matriz uniaria de retroalimentación implícita. En este caso, las preferencias del cliente se derivan de sus

actividades en lugar de sus calificaciones, y corresponde específicamente a la compra o no de un producto por parte de un cliente.

Los principales tipos de algorimos a realizar teniendo en cuenta los datos disponibles según (Aggarwal, 2016, 12-34) son:

1. Filtrado colaborativo basado en usuarios: En este caso, las calificaciones proporcionadas por usuarios similares a un usuario objetivo A se utilizan para hacer recomendaciones para A. Aquí, las calificaciones objetivo o faltantes se predicen usando las calificaciones de los usuarios vecinos que corresponden a las filas de la matriz de calificaciones unarias. El supuesto es que usuarios similares tienen calificaciones similares en el mismo elemento.

2. Filtrado colaborativo basado en elementos: para hacer recomendaciones para el elemento objetivo B, el primer paso es determinar un conjunto S de elementos, que son los más similares al elemento B. Luego, para predecir la calificación de cualquier usuario A en particular para el elemento B se toman las calificaciones en el conjunto S y se utiliza el promedio ponderado de estas calificaciones para calcular la calificación prevista del usuario A para el elemento B. En este caso, se utilizan las calificaciones del mismo usuario sobre elementos vecinos que corresponden a las columnas de la matriz unaria.

3. Reglas de asociación. Aprovechando la relación natural que hay entre las reglas de asociación y el filtrado colaborativo y la utilidad que tiene para una tienda en línea, se propone un tercer análisis con el propósito de descubrir relaciones entre datos. La clave en la minería de reglas de asociación es determinar conjuntos de elementos que están estrechamente correlacionados en la base de datos de transacciones y que podrían ofrecerse en conjunto o como complemento. Estas reglas son la probabilidad condicional de que una transacción en T contenga Y, dado que también contiene X con un intervalo de confianza.

Algunas consideraciones relevantes a tener en cuenta en los algoritmos propuestos y los datos disponibles donde debe haber especial atención en dos situaciones que se presentan.

Primero, la matriz unaria es una matriz muy dispersa “sparse” que corresponde a todos los ítems ofrecidos en las columnas y todas las transacciones en las filas con una dimensiones de 32.341 columnas x 99.441 filas, por lo que se requiere un método de reducción de dimensionalidad apropiado.

Segundo, como se puede observar en la tabla de frecuencias de producto, algunos ítems son muy populares y aparecen más repetidamente empeorando la calidad de las recomendaciones y hay que darle un manejo apropiado.

Finalmente, se encuentra que otros sistemas de recomendación basados en algoritmos de contenido o de contexto, en dónde los atributos descriptivos de los elementos o de los usuarios se utilizan para hacer las recomendaciones, no pueden ser utilizados en este caso principalmente por que no se tienen estos datos disponibles.

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