Machine learning aplicado a la categorización de productos en e-commerce.

Savian, Pablo Bruno; Tornillo, Julián E.; Pascal, Guadalupe
Becarios EVC-CIN 2023/2024: Velázquez Ramos, Alejo; Berta, Adrián.
Instituto de Investigaciones en Ingeniería Industrial (I4) ; Cátedra de Investigación Operativa, Facultad de Ingeniería.

Resumen y problemática

El contexto epidemiológico a causa del COVID-19 ha sido el gran disparador del e-commerce en América Latina y gran parte del mundo. Diversas investigaciones de mercado demuestran que los consumidores se vuelven cada vez más exigentes en cuanto a los tiempos de entrega, donde la logística juega un papel fundamental. En contraparte, el creciente mix de productos comercializados en línea aumenta la complejidad de las operaciones logísticas que buscan optimizar estos tiempos de entrega, siendo necesario adoptar un riguroso sistema de categorización de productos que permita mejorar la toma de decisiones.

Marketplaces Plataformas E-Commerce
Ecosistema E-Commerce

Última milla a consumidor final

Cualquier tipo de seller y producto

Servicios logísticos

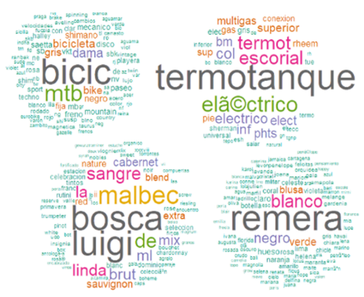
Problemática

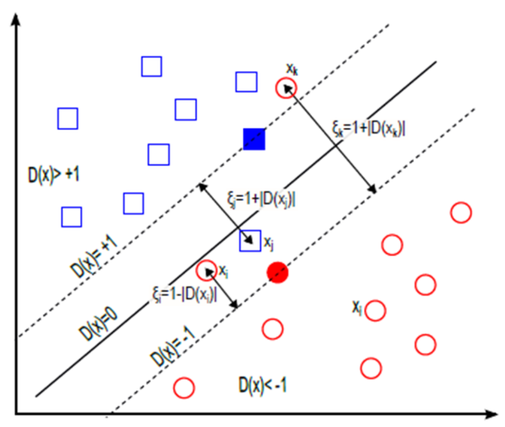
Modelo de negocio de la empresa proveedora de los datos.

Metodología

1. **Muestreo:** se seleccionó una muestra que comprende más de 50.000 órdenes, formando 327 categorías de producto
2. **Tokenización:** se convirtió la descripción de producto en tokens identificatorios.
3. **Limpieza de datos:** se eliminaron las palabras y caracteres que pueden producir sesgo (stopwords, caracteres especiales, etc.).
4. **TF-IDF:** se identificaron tokens con una alta frecuencia en ciertas categorías y baja en otras.
5. **SVM (Support Vector Machine):** se aplicó un algoritmo de clasificación de productos en categorías a partir del parámetro TF-IDF.
6. **Optimización de hiperparámetros:** se optimizó el hiperparámetro para la clasificación mediante validación cruzada.

Resultados





$$(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(t,d) : t \in d\}} \qquad \mathrm{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

$$\text{minimize} \quad \frac{1}{2}\mathrm{w}^T\mathrm{w} + C\sum_{i=1}^{l}\xi_i\,,$$

subject to

$$y_i[\mathrm{w}^T\mathrm{x}_i + b] \geq 1 - \xi_i,\ i = 1,\ l,\ \xi_i \geq 0\,,$$

i.e., subject to

$$\mathrm{w}^T\mathrm{x}_i + b \geq +1 - \xi_i,\ \text{for } y_i = +1,\ \xi_i \geq 0,$$
$$\mathrm{w}^T\mathrm{x}_i + b \leq -1 + \xi_i,\ \text{for } y_i = -1, \xi_i \geq 0\,.$$

$$y_i[\mathrm{w}^T\mathrm{x}_i + b] \geq 1 - \xi_i,\ i = 1,\ l,\ \xi_i \geq 0\,,$$

description	subcategory	weight	vol	prediction	check
aire acondicionado split inverter frão calor samsung	Aires Acondicionados	24.0	0.1320	Aires Acondicionados	FALSE
aire acondicionado split inverter frão calor samsung ar bsh	Aires Acondicionados	42.0	0.2027	Aires Acondicionados	FALSE
ano nuevo juego de cartas	Juegos de Mesa y Cartas	0.8	0.0017	Juegos de Mesa y Cartas	FALSE
apple airpods rd gen white	Auriculares	0.8	0.0017	Auriculares	FALSE
aspirad liliana la aquaterra plus	Aspiradoras	7.0	0.0520	Aspiradoras	FALSE
aspiradora black decker vcbd ar bolsa lavable	Aspiradoras	5.0	0.0270	Aspiradoras	FALSE
camara deportiva noblex acn action cam sensor sony lcd	Camaras y Accesorios	0.6	0.0068	Pequeños Electrodomesticos	TRUE
cava de vinos vondom linea negra botellas	Cavas Conservadoras	23.0	0.2527	Bebidas Alcoholicas	TRUE
almohadon caracola rosa	Almohadas	0.3	0.0024	Manicuria y Pedicuria	TRUE
almohadon maggie rosa	Almohadas	0.5	0.0480	Manicuria y Pedicuria	TRUE

Discusiones y líneas futuras

El algoritmo propuesto logra clasificar los productos en categorías con un accuracy del 95,92%.Entre las líneas futuras nos proponemos:

- Generar un modelo predictivo que permita clasificar nuevos productos correspondientes a nuevas categorías.
- Elaborar un algoritmo de predicción de ventas por categorías.