
# Informe tarea final

Asignatura:
Introducción a la Ciencia de Datos

Grupo 3

Julio 2024, Montevideo, Uruguay

*Integrantes*
Dominguez, Leandro
Romani, Tiziana

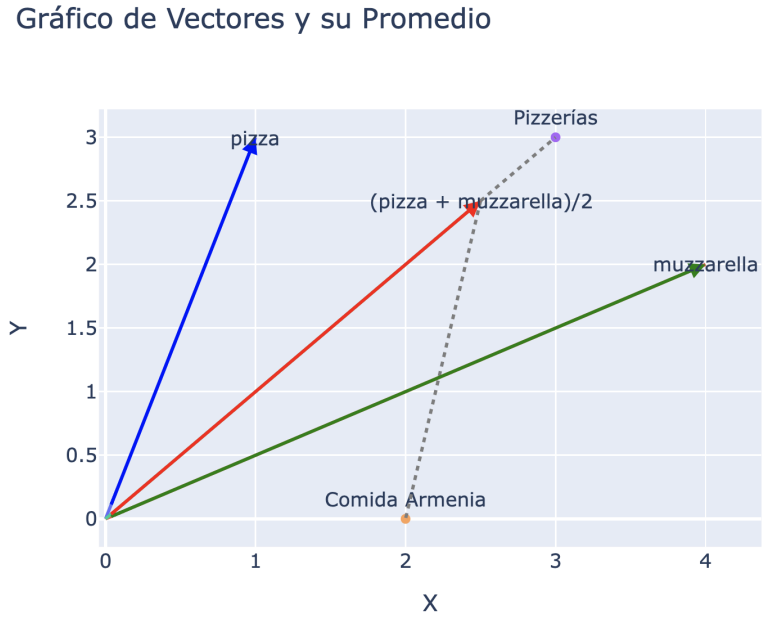
# índice

# Introducción

PedidosYa es una empresa de delivery de comida que se encarga de mostrar el menú de restaurantes, procesar pagos y hacer el delivery de los productos comprados.
La categorización de productos en determinadas categorías es una parte fundamental del proceso para que las personas usuarias encuentren lo que buscan así como para realizar recomendaciones personalizadas.
Además la categorización de productos se utiliza para filtrar tipos de comida y para clasificar los restaurantes según la proporción de productos que tenga en su menú y la cantidad de órdenes que tenga cada uno.

## Modelo actual

El modelo actual tiene varios años (se desarrolló en 2017) y básicamente se basa en extraer palabras “relevantes” de las descripciones de los productos dada de una lista de 5 mil palabras clave. Los productos se representan con un subconjunto de palabras claves contenidas en su descripción y nombre.
Por ejemplo, el producto “Deliciosa pizza muzzarella crocante al horno” se representará por las palabras claves [“pizza”, “muzzarella”].
Luego se utiliza un modelo de embeddings Word2Vec para representar cada palabra y el producto se representará como el promedio de los embeddings de sus palabras claves.
A su vez, cada categoría se representa como el embedding del promedio de los embeddings de las palabras que la componen (si tiene más de una) y para asignar productos a las categorías se toma la categoría más cercana en el espacio utilizando Distancia Coseno.
Una idea gráfica puede verse a continuación, en este caso “pizza muzzarella” quedaría clasificada como Pizzerías por ser la categoría más cercana:



## Limitaciones del modelo actual

Si bien el modelo actual tiene una performance aceptable, tiene muchas falencias y limitaciones:

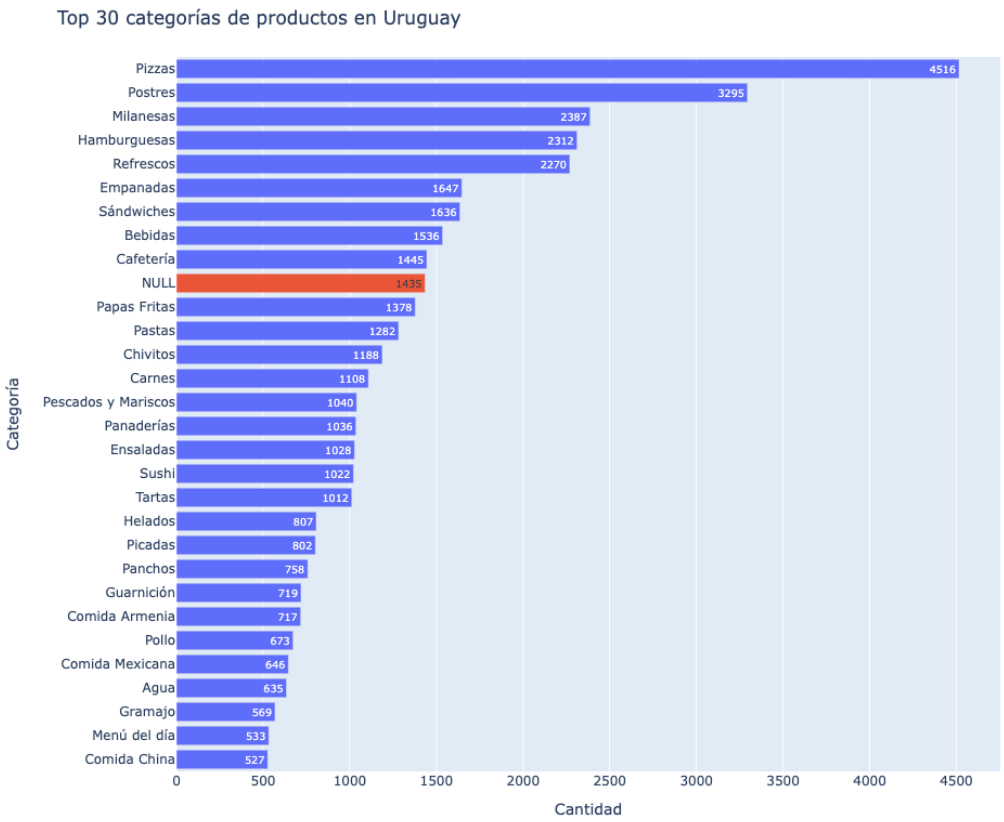
- La lista de palabras clave no es mantenida con regularidad debido al trabajo que conlleva, lo que hace que no sean tenidos en cuenta muchos de los productos nuevos.
- Los promedios de vectores no siempre son una buena representación del producto, terminando cerca de conceptos que no tienen mucho que ver con las palabras originales.
- El modelo es aceptable para Argentina y Uruguay pero tiene grandes sesgos de la región, performando pobremente en mercados del norte. Esto se debe a que fue desarrollado por personas uruguayas y tanto la lista de etiquetas así como las categorías están pensadas para estas latitudes. Un ejemplo concreto es que todo lo que contenga las palabras “taco” o “tortilla” con probabilidad muy alta será clasificado como Comida Mexicana, incluso en México.
- Si no se encuentra exactamente una palabra clave dentro del texto del producto, el mismo quedará sin clasificar.
- Word2Vec es un modelo bastante viejo que no considera el orden de las palabras ni maneja mecanismos de atención para darle más importancia a ciertas palabras dependiendo del contexto.

# Objetivo

Proponer un nuevo algoritmo de clasificación de productos del catálogo de comida de PedidosYa mediante la utilización de embeddings obtenidos por Large Language Models para mejorar la precisión del actual modelo.

# EDA del dataset

Es el conjunto de productos de la plataforma PedidosYa en Uruguay junto con su clasificación actual (que tiene gran posibilidad de mejora).
La tabla contiene las tuplas con el siguiente esquema:
*(product_id, product_name, product_description, product_category)*
A continuación se hace un histograma de categorías para entender cómo se conforma el catálogo para Uruguay:

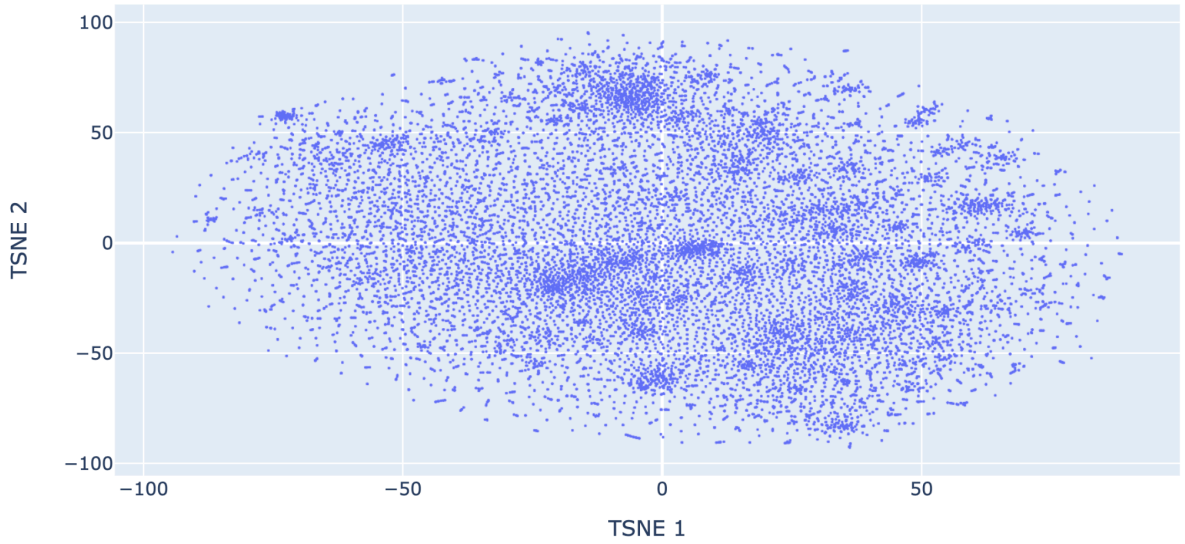



Se aprecia que en lugar 10, con más de un 3% del total de productos del catálogo, se encuentra la categoría nula (sin asignar). Esto representa un número bastante considerable teniendo en cuenta que hay varias categorías que tienen menos porcentaje que este.
Por otro lado se sabe por experiencia que la calidad de muchas de las categorías no son buenas y presenta errores:

- Rolls de canela etiquetados como “Sushi” (por contener la palabra Roll)
- Tartas saladas etiquetadas como “Postres” por tener la palabra “Tarta”
- Casi cualquier producto que contenga la palabra “Carne” (aunque sea una empanada) será categorizado como “Carnes” con alta probabilidad.

## Visualización del dataset

Se utilizan embeddings de ChatGPT para representar los textos de los productos y se utiliza TSNE para visualizar una muestra de 10 mil productos:




A simple vista pueden encontrarse algunas agrupaciones semánticas en nubes que posiblemente sean productos que están relacionados. Es usando esta noción de similitud en el espacio generado por el modelo de embeddings que queremos construir un clasificador.

# Problemas a resolver usando conocimientos del curso

## Problema supervisado

El problema a resolver puede atacarse claramente como un problema de clasificación, donde las clases ***y*** son las posibles 30 categorías y las observaciones ***X*** son los embeddings de los productos. Se probarán distintos modelos de clasificación, haciendo tuneo de hiperparámetros con las técnicas vistas en el curso (por ejemplo grid search). Algunos pueden ser KNN, Support Vector Machines, entre otros. Además, para maximizar la certeza de las métricas de cada modelo, así como para usar todos los datos para entrenar y testear cada modelo.

## Desbalance del dataset

El dataset original se encuentra claramente desbalanceado, predominando categorías como **Pizzas** y **Postres** sobre clases subrepresentadas como **Comida China** y **Gramajo**, entre otros. Para solucionar este potencial problema a la hora de entrenar un clasificador, se podrían usar técnicas de oversampling y/o undersampling para disminuir las clases con mayor peso y/o aumentar el tamaño de las clases con menores ejemplos.

## Construcción de dataset validado

Es necesario construir un dataset etiquetado con ejemplos de cada una de las 30 categorías para entrenar cualquier clasificador. Podría usarse el actual modelo para un pre-filtrado de las categorías corrigiendo las muestras correspondientes.

## Pruebas con modelos no supervisados

Podría probarse algún algoritmo de clustering como K-Means con 30 clases para entender si allí caen las deseadas categorías.

## Métricas a utilizar

La principal métrica que tiene sentido es la de Precisión. Esto se debe a que interesa mostrar carruseles de productos de determinadas categorías minimizando los productos que no pertenecen a la categoría inicial.
Se construirán matrices de confusión para entender qué casos son los que el modelo tiende a confundir y clasificar para otras clases.

## Preguntas a responder

1) ¿Cuáles son las categorías con más órdenes?
   Cruzando la información de productos con la información de la tabla de órdenes se podría responder esta pregunta.
2) ¿Cuáles son las categorías con mayor oferta de productos? ¿Son las mismas con el modelo nuevo y el viejo?
   Con un breve análisis que cuente la categoría de cada producto se podría resolver.
3) ¿Qué modelo es el más adecuado para nuestros datos?
   Midiendo performances y dejando un conjunto de validación se podría responder esta pregunta.
4) Con el nuevo algoritmo, ¿baja la cantidad de productos sin clasificar?
   Alcanza con medir esta categoría antes y después, aunque la nueva solución forzará a todos los productos a tener una categoría.
5) ¿Existen categorías que se podrían eliminar o agrupar?
   Existen categorías con pocos productos como la Comida Mexicana, Comida China y Comida Armenia. ¿Tiene sentido crear una nueva categoría llamada “Comida Internacional”?