UNIVERSIDAD PRIVADA BOLIVIANA

Tópicos Selectos en Inteligencia Artificial - Checkpoint Final - Open Food Facts

Presentado por: Alejandro Ledezma

Docente: Ariel Hugo Alba Rios

Cochabamba, 01 de noviembre de 2022

Dataset a utilizar

Para el trabajo final se usarán los datasets de Open Food Facts y Free Music Archive. El dataset de Open Food Facts recopila un montón de productos alimenticios con un montón de datos recopilados sobre cada uno. Tenemos datos como: el valor de nutrición, valor nova (cuanto tarda en echarse a perder, categoría, empresa que los produce, etc.). Por otro lado Free Music Archive nos muestra un montón de datos sobre canciones, canciones con características como: artista, género musical, duración, número de reproducciones, álbum al que pertenece, etc.

Nuestro dataset principal es el de Open Food Facts.

Objetivos (Responder a las preguntas)

Las preguntas a realizar son:

- ¿Cuál es la relación entre **Valor Nutritivo**, **Valor Nova** y la **Marca** del alimento porcentualmente? Se puede hacer con un modelo NO SUPERVISADO con CLUSTERING con los 3 features (podemos agregar más pesos (multiplicando por un valor (en este caso los valores agregados serán los datos de GloVe para tokenizar y hacer embedding de los nombres de los productos)))

Preprocesamiento del dataset

Para el preprocesamiento de los datos, primero tenemos que abrir el archivo de 6.2 GB. Esto lo haremos con ayuda de una función que selecciona datos saltando datos, así podemos ir seleccionando varios datos en “orden”, porque en realidad es solo una sección y no está verdaderamente ordenado, para así no caer en datos super sesgados (como ser todos los que empezarían con A, o todos los que están agrupados en una sola categoría o marca.). Después de esta selección lo mandaremos a un archivo .csv llamado “food.csv” que pesa 64 MB, tiene casi 25 mil filas y otro llamado “food 2.csv” que pesa 200MB y tiene 77 mil filas de datos para poder trabajar.

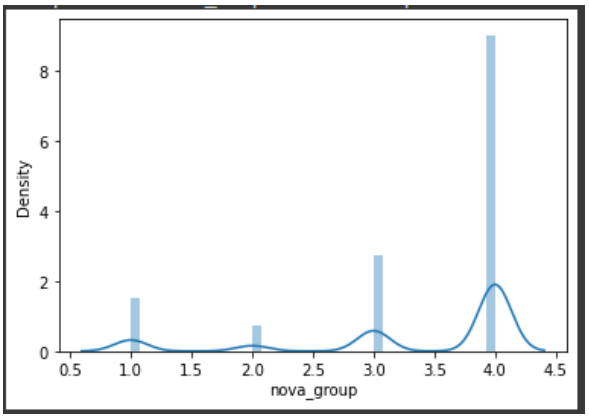
La primera parte de la limpieza la haremos eliminando las filas que tienen el nombre del producto como un valor nulo.

Crearemos un dataset con features destacables y que usaremos para calcular en los modelos, eliminando TODAS las filas que tendrían los valores de las columnas como nulos.

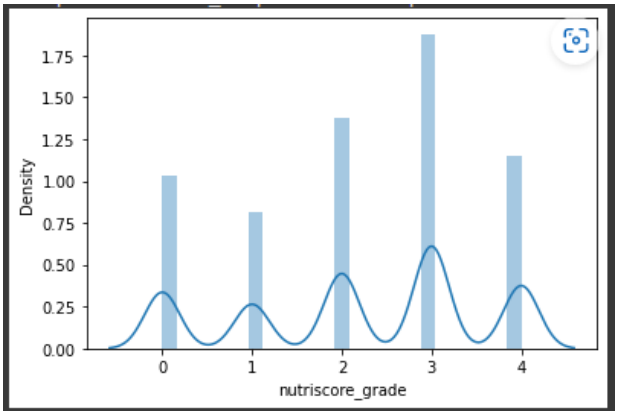
Finalmente tokenizaremos la columna de los nombres de los productos con ayuda de la librería de nltk, después haremos el embedding con los datos de GloVe, de un sample de palabras en twitter, que tiene 25 features. Agregamos estos valores para tener más datos a cada producto.

Se optó por la opción de tomar varios datos del dataset original e ir quitando filas donde existían columnas con features nulos, esto por conveniencia de que tenemos varios datos y aun eliminando varias columnas nulas, aun tenemos muchos datos con los que trabajar.

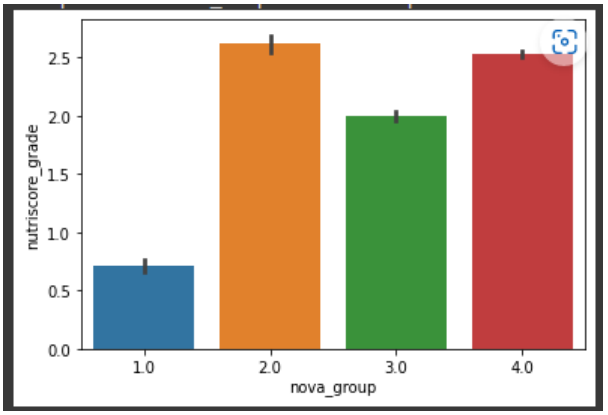
Gráficos de los Features



Densidad del Valor Nova. Este valor indica el nivel de procesamiento del alimento. 1: son los alimentos MENOS procesados, o sea, que se pudren mucho más rápido pero son más naturales y tienen menos conservantes. Por otro lado en 4: son los alimentos MÁS procesados, o sea, se pudren con menos facilidad pero tienen muchos más conservantes y normalmente tienden a ser más dañinos.



Densidad del Valor Nutritivo. Este valor indica el valor nutritivo de los productos. Fue discretizado de A, B, C, D y E, donde A=1, indica el mayor nivel nutritivo, calculado por todos los nutrientes y cosas positivas que tiene el producto. Mientras que apuntando a E=4, indica el nivel más bajo de nutrición, o sea, que tiene más grasas saturadas, pocas vitaminas, nutrientes y en general es mal alimento.



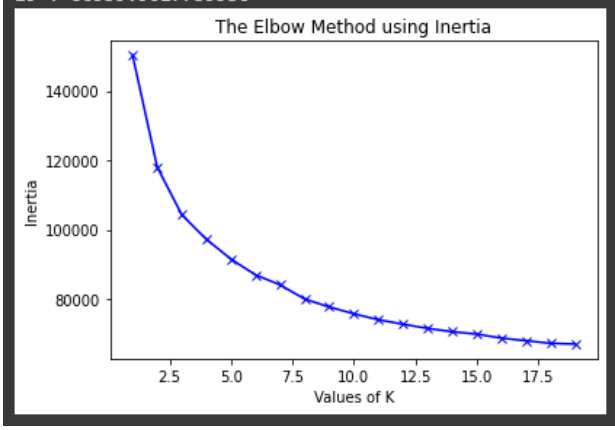
Relación de media de los valores nova y los valores nutritivos de TODOS los productos. Se nota que los productos con un nova bajo (menos procesados) tienden a ser los que tienen un valor nutritivo más alto (tienden a 1 y 0 => B y A).

Modelos a utilizar y el porqué se los escogieron

Usaremos un modelo no supervisado de clustering, en este caso k means, esto porque queremos hacer un análisis del comportamiento de los datos respecto a sus features de valor nova, valor nutritivo y nombre. Obviamente no sabemos muy bien el resultado, por eso tiene que ser NO SUPERVISADO.

Feature Engineering (Hyperparameter tuning)

```
1 : 150262.5947217715
2 : 118038.19375009804
3 : 104282.09903880373
4 : 97202.3787626742
5 : 91480.00415797473
6 : 86923.68436847605
7 : 84043.68250420313
8 : 79963.27689377051
9 : 77678.86374668528
10 : 75692.86391564536
11 : 73996.85693012152
12 : 72690.4837485576
13 : 71494.87777858517
14 : 70526.15727810262
15 : 69840.47154165502
16 : 68692.92601123128
17 : 67938.01154669959
18 : 67193.22183377751
19 : 66959.9017759936
```

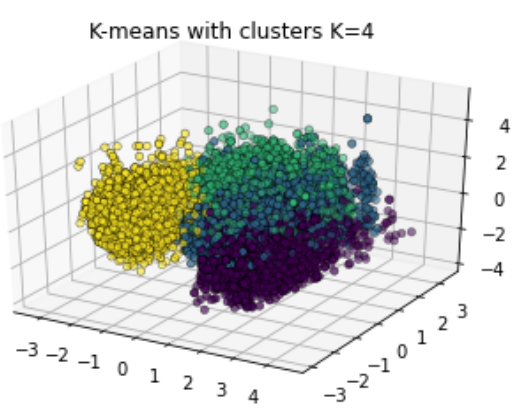


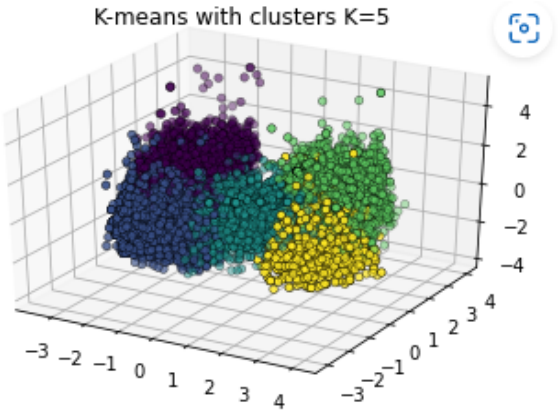
Se hicieron 20 iteraciones con diferentes valores K (número de clusters) para hacer diferentes modelos con k means. Observando el gráfico y los errores, podemos ver que el valor de K (número de clusters) óptimo ronda entre 3 y 5. Usaremos K=4 para mostrar los datos.

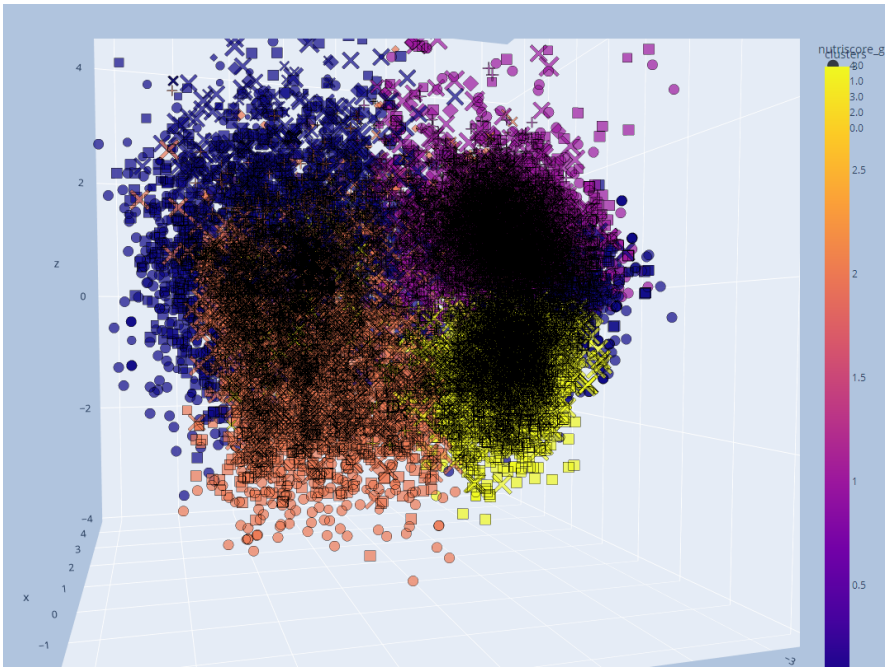
Reducción de dimensionalidad (si es aplicado, en caso contrario colocar las razones por qué no se lo utilizó)

Si se usó reducción de dimensionalidad, para esto usamos una librería de PCA. En nuestro dataset ya limpio teníamos 30 features para trabajar, todos estos features no podemos meterlos a graficar, así que primero entrenamos el k means con los 30 features y finalmente redujimos la dimensionalidad a 3 componentes: “pca1”, ”pca2”, ”pca3”. Estos 3 datos (los más relevantes) ya podemos meterlos a graficar.

Resultados gráficos de las optimizaciones







El gráfico final es un clustering con K=4 (el más óptimo según la regla del codo) graficados después de redimensionar los datos. En el gráfico:

Los símbolos (forma de los puntos):

Símbolo	Valor Nutritivo
símbolo de suma (+)	0 (A)
rombo (◊)	1 (B)
equis (X)	2 (C)
cuadrado (□)	3 (D)
círculo (O)	4 (E)

Los colores (color de los puntos):

Color	Cluster
azul	0
magenta	1
naranja/salmón	2
amarillo	3

Los tamaños (tamaño de los puntos):

Tamaño	Valor Nova
1 (mas pequeño)	1 (menos procesado)
2	2

3	3
4 (más grande)	4 (mas procesado)

Conclusiones

- **Cluster amarillo:** La mayoría de los productos con **valor nutritivo** A (símbolo suma) rondan entre los clusters y un **valor nova** rondando 1 y 3, siendo 1 el más dominante. También pertenecen a **marcas** como: Carnes, Cereales, Lácteos y algunas Frutas. Con todo esto podemos concluir que: Los productos más nutritivos están compuestos por Frutas, Carnes, Lácteos y algunas Frutas que también son productos con pocos o ningún conservante. Asi que el cluster mas nutritivo y saludable es el **amarillo**, mas que todo SUS EXTREMOS, donde comparte algunos de estos productos nutritivos con el **cluster magenta.**
- **Cluster azul:** La mayoría de los productos tienen **valores nutritivos** que rondan D y C, algunos que se mezclan con el **cluster magenta** son C. En el **valor nova** predominan 4 y 3, más que todo el 4, o sea, son muy procesados. Dentro de las **marcas** hay muchos: Snacks Azucarados y Salados, Dulces y tirando para el cluster **magenta** algunos Lácteos y Quesos procesados. Definitivamente el **cluster azul** es el extremo opuesto al **amarillo**, hay demasiados productos con valor nutritivo muy bajo y procesados, es el menos nutritivo y el que aporta menos nutrientes.
- **Cluster magenta:** La mayoría de los productos tienen **valores nutritivos** que rondan E y D. En el **valor nova** predominan 4 y 3 por igual, o sea, son muy procesados, aparte hay una gran población de productos con **valor nutritivo** A y **nova** 4, productos saludables…pero muy procesados. Revisando las **marcas** tenemos: Grasas, Bebidas azucaradas, Salsas, y productos Lácteos como Yogurt, además de algunas Comidas Congeladas y Cereales. Por el extremo donde hay más productos con **valor nutritivo** A tenemos más comidas Congeladas, Cereales, Frutas y algunos snacks. Este cluster está muy interesante, ya que tiene productos procesados, pero nutritivos en sus extremos que rozan al **cluster amarillo**, podriamos decir que aqui van productos fitness y procesados, pero pesandos precisamente para entrenamientos y ser saludables. Este es un cluster con productos empaquetados y procesados que no caen dentro de la categoría de snack que no aporta nada nutritivo.
- **Cluster naranja:** La mayoría de los productos tienen **valores nutritivos** que rondan C y B, además de varios A en sus extremos, se mezcla con el **cluster amarillo** en C. Respecto a los **valores nova**, hay una gran variedad entre todos los valores, pero predominan 3 y 4. Tenemos las **marcas** de: Lácteos, Pescado, Huevos, Brebajes sin azúcar, Carnes. En los extremos con valor nutritivo A y nova 4 tenemos bastante fruta que deben ser enlatados en su mayoría. Este cluster representa productos más pesados que los del **cluster amarillo**, en su mayoria son Lacteos, Carnes y Pescados que estan congelados o procesados para ser comidas que calientas y te sirves una sola vez. Este cluster es bastante neutral, tiene productos procesados, pero tampoco son lo peor del mundo, son las comidas pre hechas del supermercado.
- **Final**: El **cluster amarillo** es el mejor, seguido del **magenta**. Recomiendo basar nuestras dietas en base a estos productos, consumir los del **cluster naranja** cuando no tenemos mucho tiempo para preparar nuestra comida y tratar de evitar los del **azul**, más que todo tratarlos como premios ya que en su mayoría son dulces, pero tienen muchos conservantes y aportan muy poco valor nutritivo.