Tema TFM

Clasificación de alimentos de origen vegetal según sus macronutrientes mediante

Análisis de Cluster

Entrega 3

Grupo 2

Miembros del grupo

Infante Ortega, Lady Viviana.

Sánchez Fabre, Alex Antonio.

Rodriguez Muñoz, Ricardo.

Sánchez Fabre, Felipe Andrés.

12 de septiembre de 2023

Contenido:

1. Desarrollo de un modelo predictivo aplicando técnicas de machine learning y/o aplicando algunos de los algoritmos vistos en los módulos 6 y 7.

Sobre la base de este requerimiento se decidió emplear el algoritmo no supervisado denominado Local Outlier Factor (LOF), éste permite la detección de productos con comportamientos atípicos a nivel multivariante. Esta técnica requiere el cálculo de la densidad local de un punto de datos obtenido con respecto a sus vecinos. Es local en el sentido de que el grado depende de qué tan aislado esté el objeto con respecto al vecino más cercano.

Para ello, se consideran como valores atípicos los elementos que tienen una densidad inferior a la de sus vecinos. Se ha configurado en la función un total de 40 vecinos para definir la comparación. Adicionalmente, el punto de corte de la distribución LOF elegido es el percentil 90. Los resultados con una densidad superior a 1 (uno) son considerados atípicos, por lo tanto, se han excluido para la conformación de clústeres.

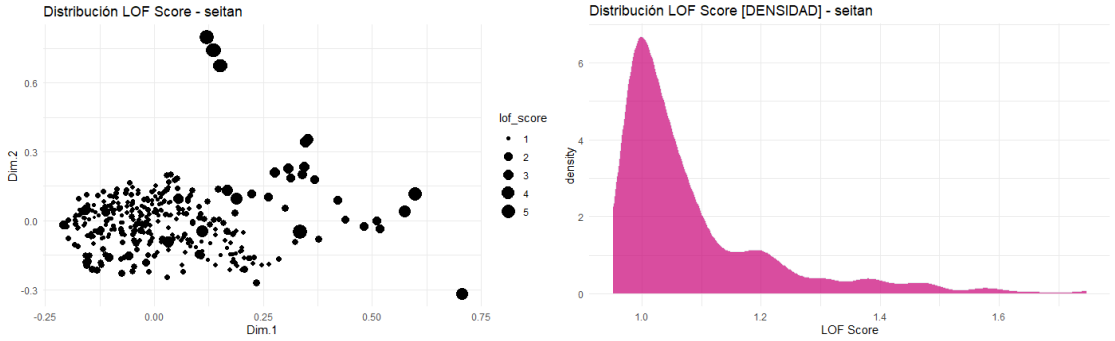
Se implementó este algoritmo para el seitán, tofu y la soja. A continuación, se presentan los resultados para el seitán:

Tabla 1

Seitán por cada 100 gramos	**Kilocalorías**	**Grasas**	**Carbohidratos**	**Proteínas**
Valores de referencia Cluster 1	*231*	*10.6*	*7.5*	*25.49*
Valores de referencia Cluster 2	*139*	*1.8*	*6.2*	*21*
Valores de referencia Cluster 3	*224*	*9.9*	*20.04*	*12.16*
Outilier 1: Seitán	81	1.3	4.9	12
Outilier 2: Seitán Curry	197	6.5	7.10	25.6
Outilier 3: Spiced Seitan Tenders	287	15.6	17.3	17.7

Está técnica ha propuesto prescindir de **50** productos de seitán, estos presentan un valor en alguna de sus variables diferente a sus vecinos más cercanos. Por ejemplo, se presentan en la tabla 1 los valores referencia promedio por cluster, así como **3** resultados identificados como atípicos por esta técnica en este producto.

Gráfico 1



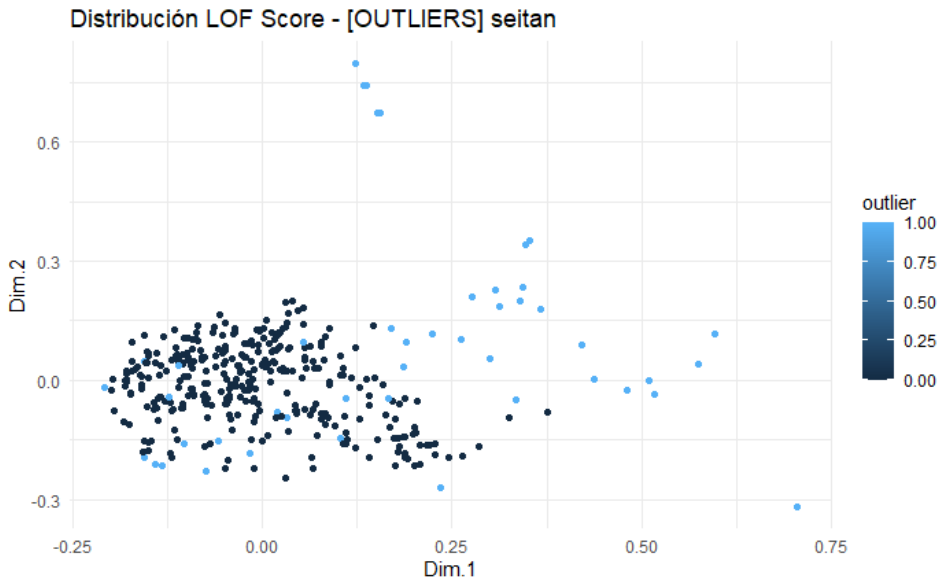
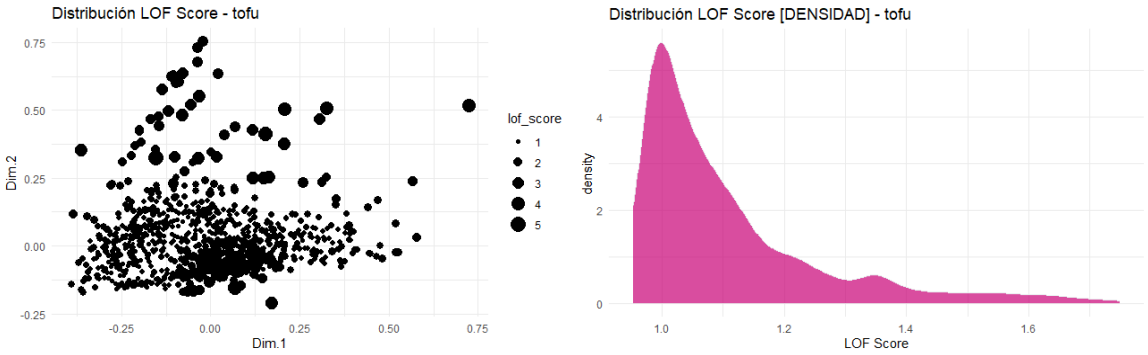
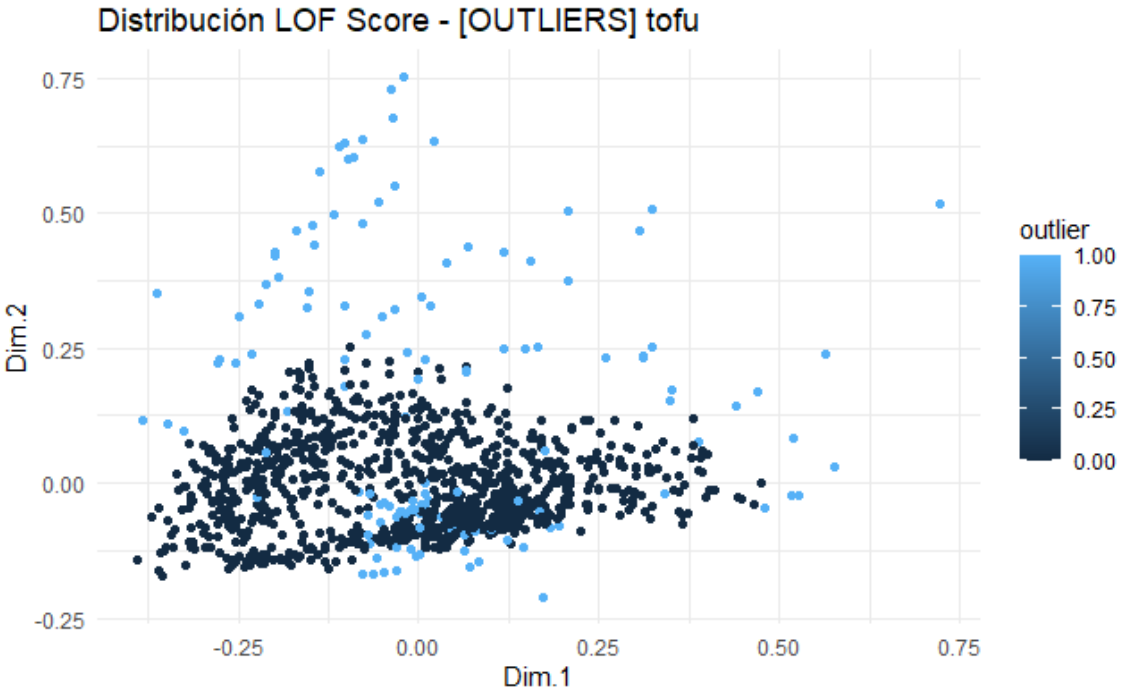


Tabla 2

Tofu por cada 100 gramos	Kilocalorías	Grasas	Carbohidratos	Proteínas
Valores de referencia Cluster 1	180	7.4	19.4	7.577
Valores de referencia Cluster 2	133	7.5	1.9	13
Valores de referencia Cluster 3	219	14.07	5	17
Outilier 1: Tofu De Huevo	31.00000	1.400000	4.400000	0.300000
Outilier 2: Wurstel Di Tofu Gustosi	188.00000	13.000000	1.300000	15.000000
Outilier 3: Tartinable Tofu Pistou	314.00000	29.400000	2.400000	8.900000

Gráfico 2





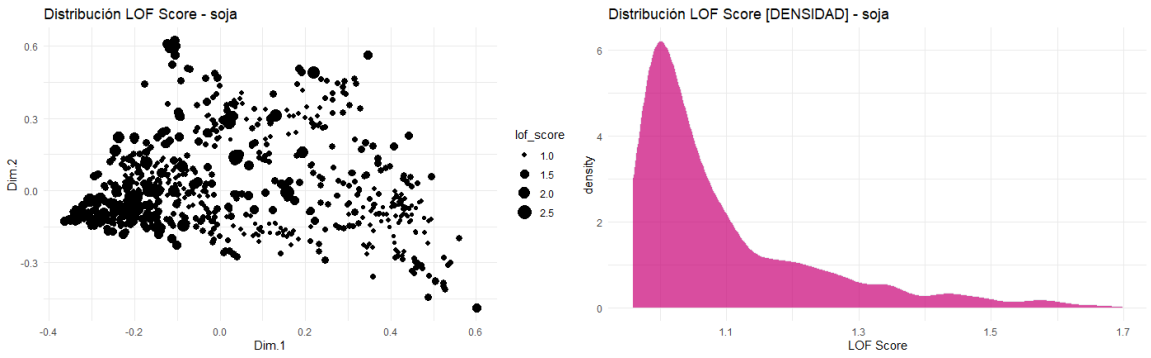
En el caso del **tofu**, está técnica ha propuesto prescindir de **144** productos, estos presentan un valor en alguna de sus variables diferente a sus vecinos más cercanos. Por ejemplo, se presentan los valores referencia por cluster, así como **3** resultados identificados como atípicos por esta técnica en este producto:

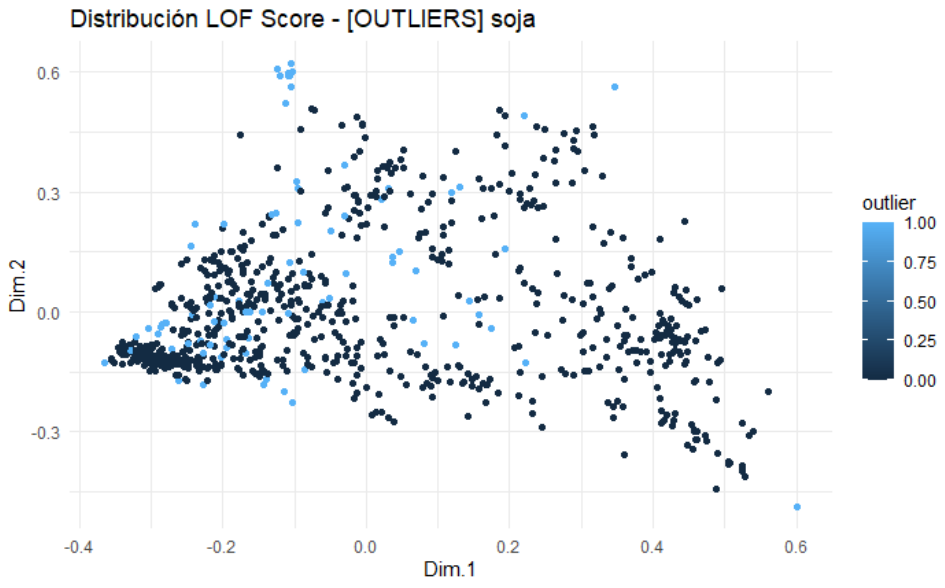
Los valores de referencia, para los productos de ejemplo como el Tofu De Huevo, Wurstel Di Tofu Gustosi y Tartinable Tofu Pistou, presenta niveles muy disímiles respecto a los valores promedio del Cluster 1, 2 y 3 o en todas o en alguna de las variables medidas.

Tabla 3

Soja por cada 100 gramos	Kilocalorías	Grasas	Carbohidratos	Proteínas
Valores de referencia Cluster 1	447.8	25.22	21.70	34.91
Valores de referencia Cluster 2	398	12.94	57.50	11.48
Valores de referencia Cluster 3	124	4	10	5.8
Outilier 1: Soia Natural Senza Zuccheri	32	1.8	0.40	3.3
Outilier 2: Bruschette Di Soia	379	2.4	72	15
Outilier 3: Tahin Blanc Eco Natursoy	588	49.7	11.6	17.7

Gráfico 3





Para el producto **soja**, está técnica ha propuesto prescindir de **91** productos, estos presentan un valor en alguna de sus variables diferente a sus vecinos más cercanos. Por ejemplo, se presentan los valores referencia por cluster, así como **3** resultados identificados como atípicos por esta técnica en este grupo:

Los valores de referencia, para los productos de ejemplo como el Soia Natural Senza Zuccheri, Bruschette Di Soia y Tahin Blanc Eco Natursoy, presenta niveles muy disímiles respecto a los valores promedio del Cluster 1, 2 y 3 o en todas o en alguna de las variables medidas.

Si bien es cierto, que algunos pueden corresponder a registros válidos, esta técnica establece los límites de definición de outliers en función del total de vecinos sobre los que hará la comparación de las distancias y además sobre el punto de corte de la distribución LOF, en este caso ha sido de a partir del percentil 90.

Para complementar esta interpretación se presentan los clustergramas para cada grupo de productos, esta representación gráfica ha sido propuesta para visualizar patrones de agrupación (clustering) en un conjunto de datos. El objetivo principal es mostrar cómo los productos se agrupan en función de sus similitudes/diferencias en los macronutrientes, en este sentido, se pretende identificar la clasificación de los artículos a medida que se incrementan los numeros de conglomerados.

Gráfico 4 - Seitán

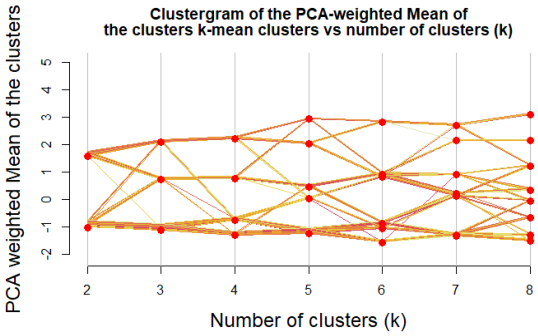


Gráfico 5 - Tofu

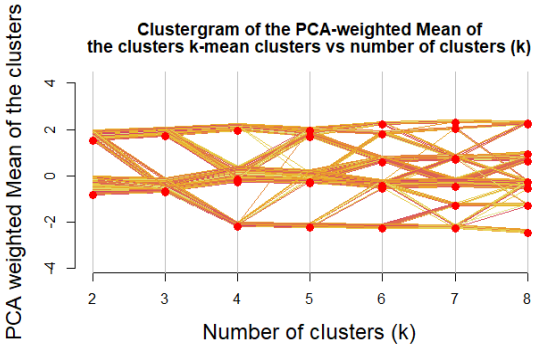
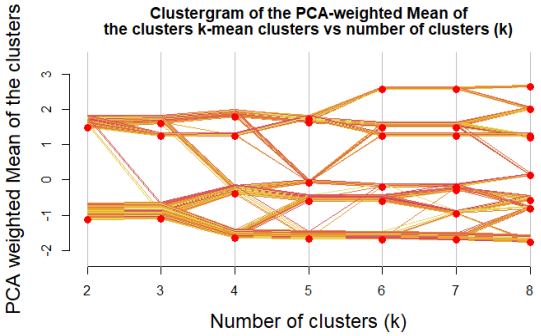


Gráfico 6 - Soja



No se visualizan resultados concluyentes en ninguno de los clustergramas, por lo que podemos concluir que los clusters tendrán cierto solapamiento entre sí en las tres categorías.

2. Comparativa, desarrollo y explicación con respecto al modelo predictivo propuesto y desarrollado en la entrega anterior.

Seitán

A continuación se presentan los resultados luego de haber realizado las exclusiones de outliers.

Estos resultados han sido calculados sobre la base de **445** productos.

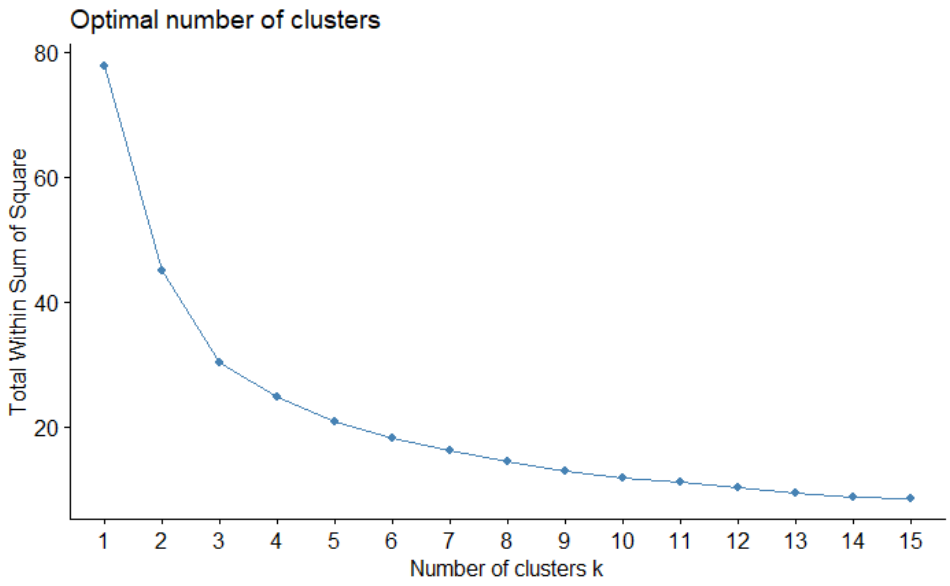
El analisis de componentes principales realizado previamente muestran como los dos primeros autovalores recogen el 77% de variabilidad total. Así mismo, se observa como las contribuciones al eje 1 estan marcadas principalmente por ENERGY, FAT y CARBOHYDRATES, mientras que PROTEINS contribuye más al eje 2.

```
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4
Variance               0.020   0.017   0.009   0.003
% of var.             41.944  34.902  17.744   5.410
Cumulative % of var.  41.944  76.846  94.590 100.000

Individuals (the 10 first)
                        Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
2                   |  0.117 | -0.106  0.111  0.817 |  0.041  0.020  0.121 | -0.029  0.020  0.061 |
5                   |  0.105 | -0.076  0.057  0.517 |  0.058  0.040  0.300 | -0.044  0.045  0.172 |
6                   |  0.085 | -0.024  0.006  0.078 |  0.060  0.043  0.503 | -0.054  0.069  0.412 |
7                   |  0.053 | -0.011  0.001  0.041 |  0.028  0.009  0.272 | -0.042  0.041  0.624 |
9                   |  0.525 |  0.510  2.567  0.943 | -0.001  0.000  0.000 |  0.126  0.368  0.057 |
11                  |  1.009 |  0.597  3.528  0.351 |  0.117  0.163  0.013 |  0.804 15.113  0.635 |
19                  |  0.136 | -0.105  0.109  0.599 | -0.058  0.040  0.184 | -0.062  0.090  0.209 |
20                  |  0.087 | -0.043  0.019  0.246 | -0.045  0.024  0.261 | -0.061  0.086  0.485 |
22                  |  0.151 | -0.150  0.224  0.997 |  0.007  0.001  0.002 |  0.003  0.000  0.000 |
23                  |  0.116 | -0.105  0.109  0.825 |  0.044  0.023  0.141 | -0.021  0.010  0.033 |

Variables
                      Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
ENERGY_KCAL_100G   |  0.088 37.874  0.722 |  0.034  6.993  0.111 |  0.022  5.659  0.046 |
FAT_100G           |  0.054 14.125  0.342 | -0.006  0.198  0.004 |  0.068 54.196  0.555 |
CARBOHYDRATES_100G |  0.099 47.695  0.676 | -0.037  8.127  0.096 | -0.055 35.431  0.212 |
PROTEINS_100G      |  0.008  0.306  0.004 |  0.120 84.683  0.951 | -0.020  4.714  0.027 |
```

Gráfico 7 - Seitán



El gráfico previo muestra el número óptimo de clusters que pueden ser elegidos para conformar los grupos. Se puede puede hacerse una clasificación de máximo 15 clusters, aunque ya se ha evidenciado que tres clusters pueden ser suficientes.

Esta reducción de dimensionalidad, permitiría construir el analisis de cluster a partir de los resultados obtenidos con las componentes principales que mayor variabilidad aportan al modelo, dando paso a la aplicación del análisis de cluster mediante método de K-means y ACP, este algoritmo de clasificación no supervisada agrupa objetos en k grupos basándose en la mínima suma de distancias entre cada objeto y el centroide de su grupo o cluster con las nuevas variables creadas en el ACP..

A continuación se presentan los hallazgos obtenidos mediante esta técnica.

Gráfico 8 - Seitán

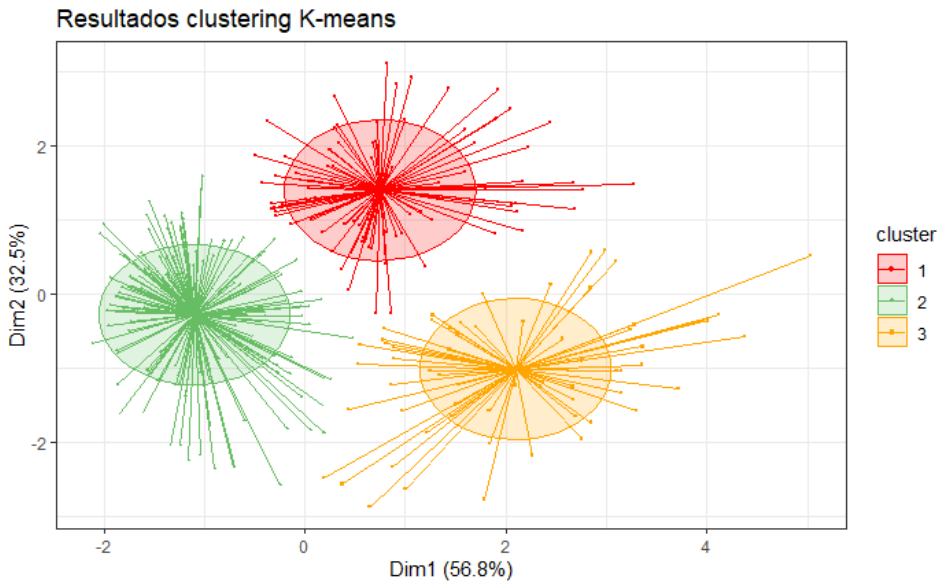
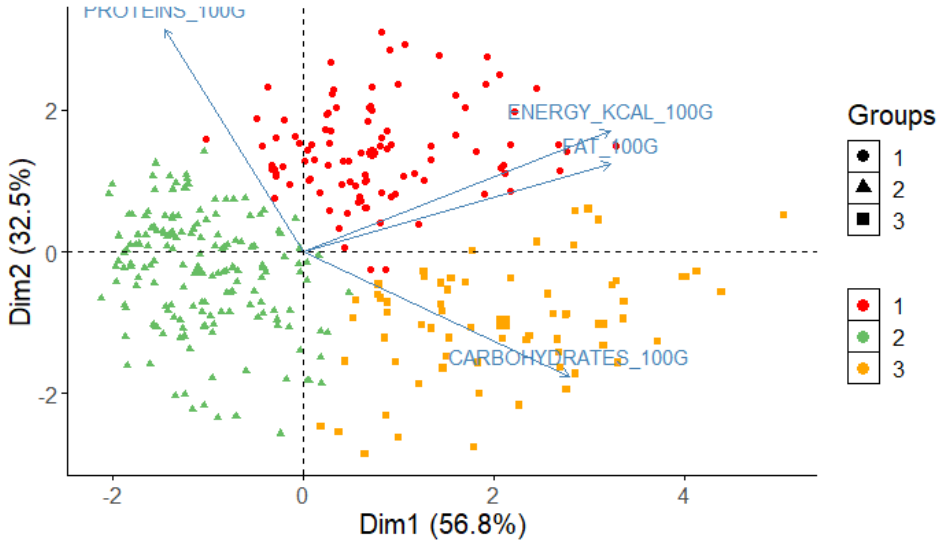


Gráfico 9 - Seitán



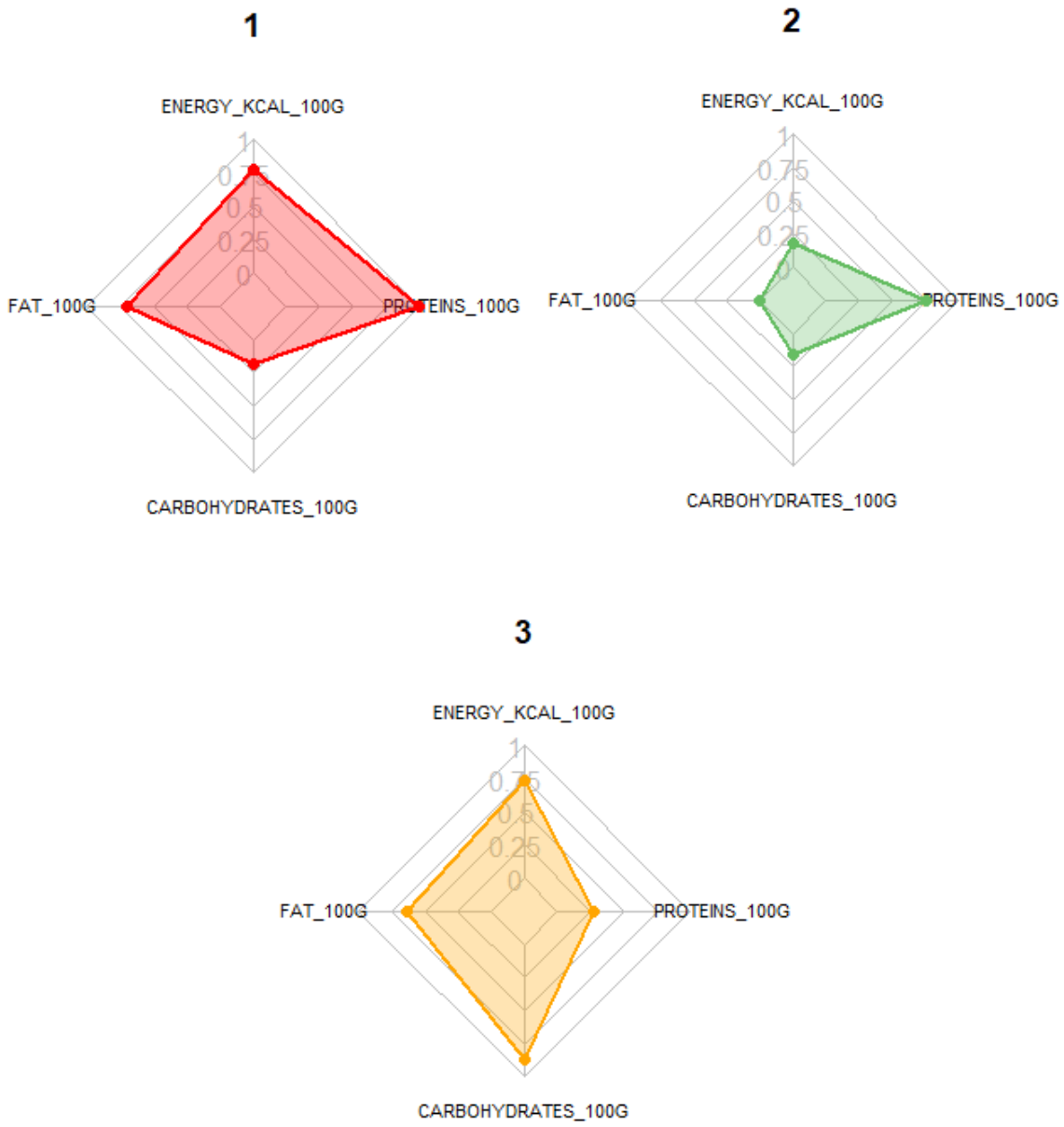
El gráfico previo ilustra los productos en el plano, se observa primero como las variables ENERGY y FAT se encuentran altamente correlacionadas con el eje 1, por otra parte, PROTEINS esta correlacionada con el eje 2, finalmente CARBOHYDRATES esta explicada parcialmente por el eje 1 y 2. Productos cercanos a la dirección de los vectores indican valores altos en esas variables, productos opuestos a la dirección de los vectores indican bajos valores en esas variables. En este sentido, se puede identificar lo siguiente:

Grupo 1. Productos de seitán con valores bajos en energía, grasas y proteínas pero con ligero grado de carbohidratos.

Grupo 2. Productos con valores altos en grasa, proteínas y energía, pero bajos en carbohidratos.

Grupo 3. Productos con valores altos en carbohidratos grasas y energías, pero con valores bajos en proteínas.

Gráfico 10 - Seitán



El siguiente gráfico se ha construido a partir del cálculo de dos medidas de análisis denominadas densidad calórica y densidad nutricional.

La **densidad calórica** en un alimento se refiere a la cantidad de calorías (energía) que contiene por unidad de peso o volumen. En general, los alimentos con una alta densidad calórica tienden a ser más ricos en calorías por unidad de peso o volumen, lo que significa que proporcionan más energía sin necesariamente proporcionar una mayor cantidad de nutrientes esenciales. Esto puede ser problemático si se consumen en exceso, ya que puede contribuir al aumento de peso.

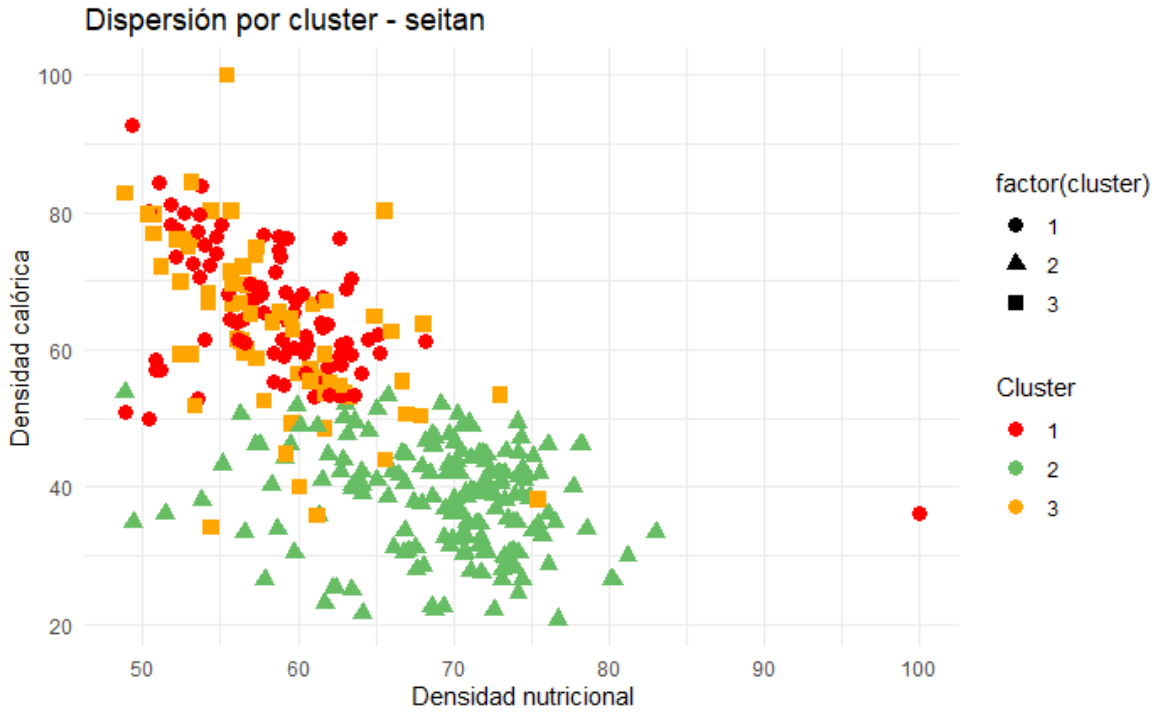
$$Desidad\ Calórica\ = \frac{Valor\ de\ Kilo\ Calorías}{Peso\ (gr)}$$

Por otra parte, **la densidad nutricional** de un alimento se refiere a la cantidad de nutrientes, como vitaminas, minerales, proteínas, fibra y otros compuestos beneficiosos para la salud, que contiene en relación con su contenido calórico o peso. En otras palabras, se trata de la cantidad de nutrientes esenciales que un alimento proporciona en comparación con la energía que aporta. Para cálculo de este indicador, se ha decidido tomar los macronutrientes como medida de densidad nutricional a nivel de macronutrientes.

Los alimentos con alta densidad nutricional suelen contener una gran cantidad de nutrientes por cada caloría que aportan, lo que los hace más saludables y beneficiosos para la dieta. Los alimentos con baja densidad nutricional son aquellos que contienen muchas calorías en comparación con su contenido de nutrientes esenciales. Esto incluye alimentos ricos en azúcares añadidos, grasas saturadas y calorías vacías, como bebidas azucaradas, dulces, pasteles y comida rápida.

$$Desidad\ Nutricional\ = \frac{(Proteínas\ +\ Grasas\ +\ Carbohidratos)\ gr}{Valor\ de\ Kilo\ Calorías}$$

Gráfico 11 - Seitán



Se puede observar la relación inversamente proporcional que existe entre densidad calórica y densidad nutricional, es decir, a medida que un alimento es denso nutricionalmente disminuye su densidad calórica. Según estos resultados se aprecia que los productos del cluster 2 presentan mayor densidad nutricional y menor densidad calórica, lo cual nos indica que son los productos mas eficientes energéticamente.

Entre los productos más representativos del **seitán** en cada clúster se encuentran:

Cluster 1 - Seitán:

Nombre del Producto	Energía(kcal)	Grasas	Carbohidratos	Proteínas
SEITAN AUFSCHNITT SPICY BEANS	206.0000	6.8000	9.3000	25.2000
SEITAN WIENER	202.0000	7.7000	6.9000	25.0000
SEITAN VEGE-JAMBON FUME	211.7647	8.2353	12.9412	22.3529

Cluster 2 - Seitán:

Nombre del Producto	Energía(kcal)	Grasas	Carbohidratos	Proteínas
SEITAN AL NATURAL	132	1.6	7.2	22.3
SEITAN NATURALE	129	1.4	8.8	19.0
SEITAN TRADIZIINALE A FETTE	129	4.0	8.8	19.0

Cluster 3 - Seitán:

Nombre del Producto	Energía(kcal)	Grasas	Carbohidratos	Proteínas
FILETES DE BIO SEITAN REBOZADOS	267	12.68	16.09	22.1
BISTECCA DI SEITAN VEGAN	265	15.00	11.00	21.0
FILETES DE SEITAN	269	15.00	11.00	21.0

Tofu

Para el tofu, los resultados han sido calculados sobre la base de **1293** productos sometidos a revisión.

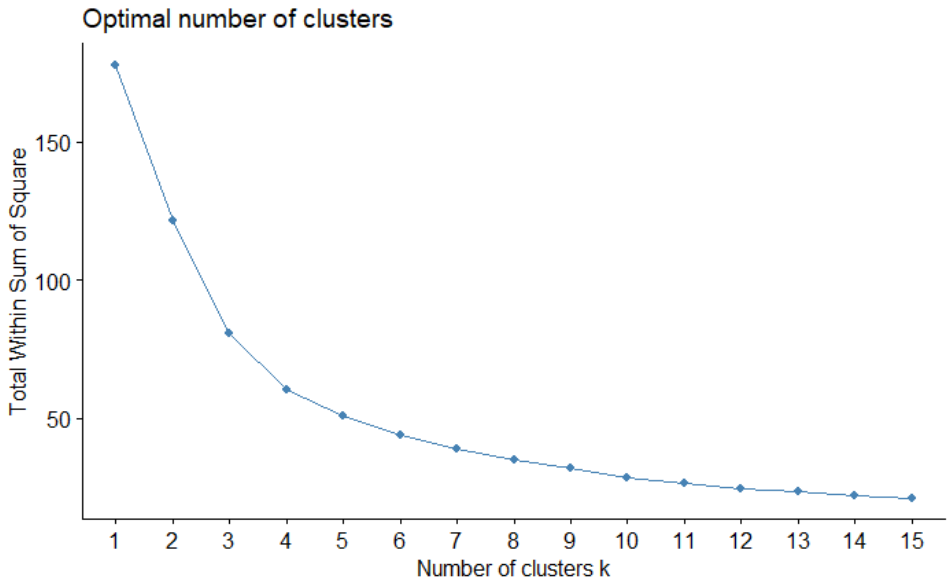
El analisis de componentes principales realizado previamente muestran como los dos primeros autovalores estarían explicado el 84% de variabilidad total con las dos primeras componentes. Así mismo, se observa como las contribuciones al eje 1 estan marcadas principalmente por PROTEINS, mientras que ENERGY, FAT y CARBOHYDRATES contribuye más al eje 2.

```
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4
Variance               0.029   0.014   0.007   0.002
% of var.             56.224  27.745  12.851   3.180
Cumulative % of var.  56.224  83.969  96.820 100.000

Individuals (the 10 first)
                        Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1                   |  0.837 | -0.021  0.001  0.001 |  0.754  2.773  0.812 | -0.358  1.346  0.183 |
7                   |  0.299 | -0.264  0.167  0.775 |  0.047  0.011  0.025 |  0.134  0.188  0.199 |
9                   |  0.145 |  0.137  0.045  0.902 | -0.020  0.002  0.020 |  0.040  0.017  0.078 |
10                  |  0.077 |  0.042  0.004  0.290 | -0.062  0.019  0.645 |  0.019  0.004  0.064 |
11                  |  0.226 | -0.219  0.115  0.939 | -0.045  0.010  0.040 |  0.032  0.011  0.021 |
12                  |  0.213 | -0.203  0.099  0.910 | -0.056  0.015  0.069 |  0.031  0.010  0.021 |
15                  |  0.119 | -0.067  0.011  0.316 | -0.098  0.047  0.675 | -0.011  0.001  0.008 |
20                  |  0.076 |  0.040  0.004  0.272 | -0.051  0.013  0.449 | -0.040  0.017  0.278 |
21                  |  0.166 | -0.008  0.000  0.002 |  0.136  0.090  0.675 | -0.093  0.091  0.316 |
22                  |  0.283 | -0.271  0.176  0.916 |  0.021  0.002  0.006 |  0.079  0.066  0.078 |

Variables
                      Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
ENERGY_KCAL_100G  |  0.029  2.957  0.096 |  0.086 51.445  0.820 | -0.004  0.199  0.001 |
FAT_100G          |  0.021  1.468  0.059 |  0.065 29.851  0.596 | -0.042 26.845  0.248 |
CARBOHYDRATES_100G| -0.029  2.892  0.105 |  0.050 17.175  0.307 |  0.067 68.159  0.564 |
PROTEINS_100G     |  0.164 92.683  0.980 | -0.015  1.529  0.008 |  0.018  4.797  0.012 |
```

Gráfico 12 - Tofu



Se ha elegido el número **3** como como valor optimo para la clasificación de productos en términos de la distancia euclídea

Gráfico 13 - Tofu

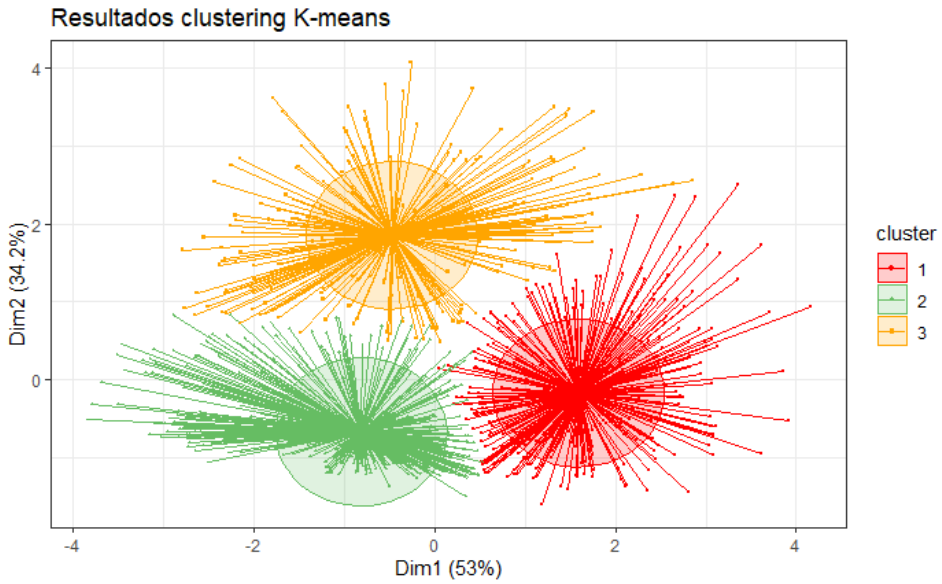
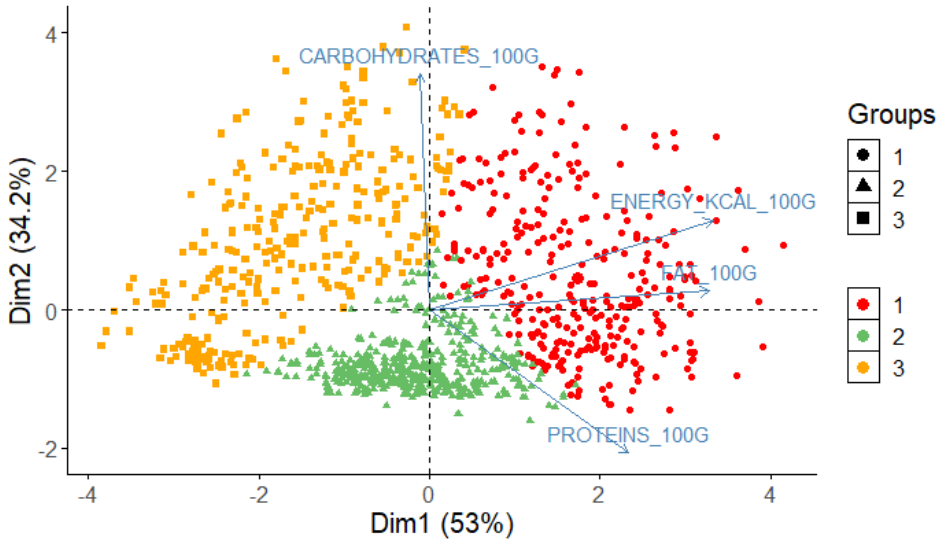


Gráfico 14 - Tofu



El gráfico previo ilustra los productos en el plano, se observa primero como las variables ENERGY y FAT se encuentran altamente correlacionadas con el eje 1, por otra parte, CARBOHYDRATES esta correlacionada con el eje 2, finalmente PROTEINS esta explicada parcialmente por el eje 1 y 2. Productos cercanos a la dirección de los vectores indican valores altos en esas variables, productos opuestos a la dirección de los vectores indican bajos valores en esas variables. En este sentido, se puede identificar lo siguiente:

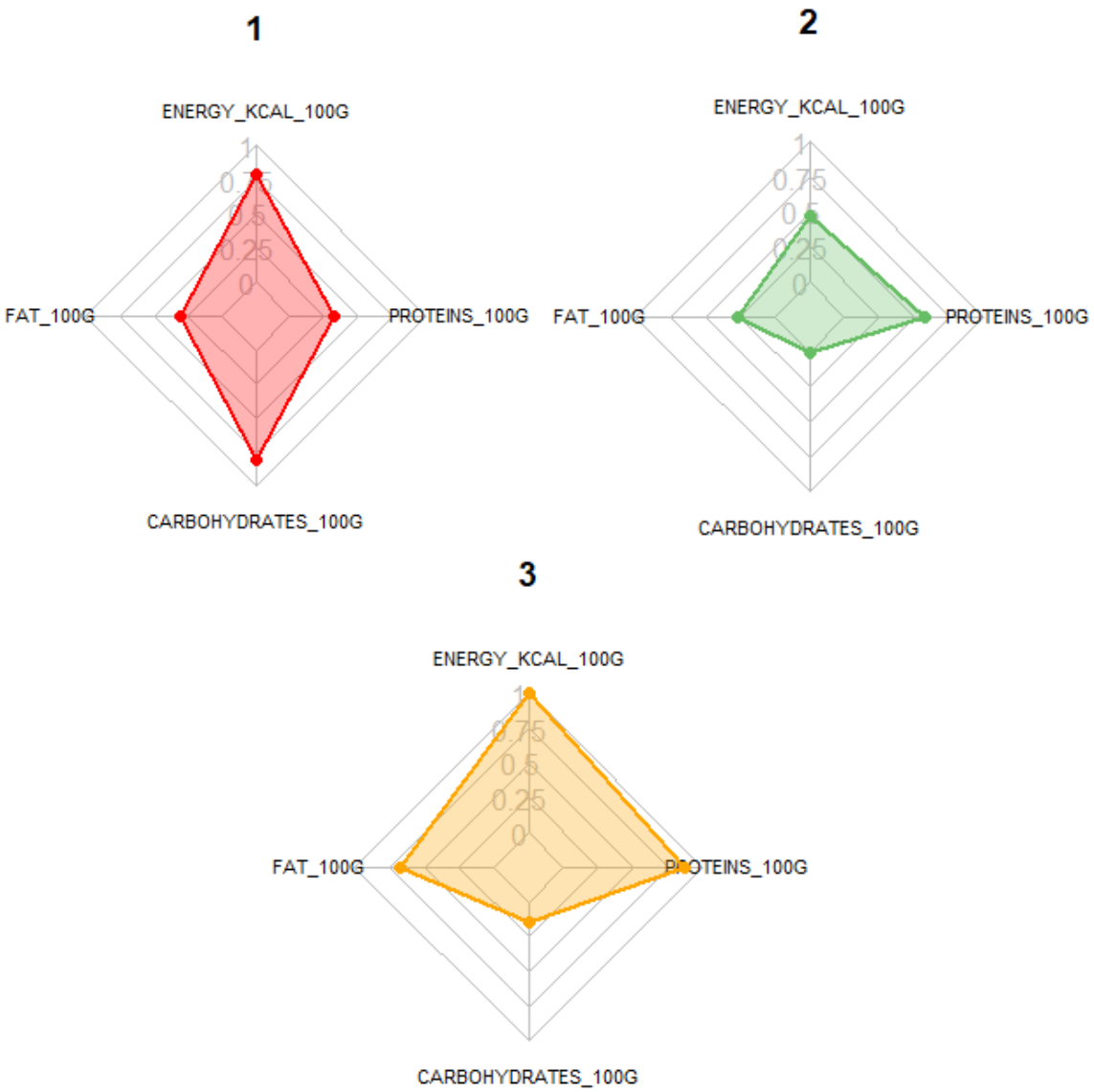
Grupo 1. Productos de tofu con valores altos en energía y carbohidratos, pero con ligero grado de grasas y proteínas.

Grupo 2. Productos con valores altos en proteínas, pero bajos en el resto de las variables.

Grupo 3. Productos con valores altos en energía, grasas y proteínas, pero con valores bajos en carbohidratos.

A continuación, se muestran estos resultados de otra manera mediante gráficos de radar:

Gráfico 15 - Tofu



Cluster 1 - tofu:

Nombre del Producto	Energía(kcal)	Grasas	Carbohidratos	Proteínas
TOFU AHUMADO	160	8.3	5.5	15
TOFU AHUMADO	160	8.3	5.5	15
BIO TOFU AHUMADO	160	8.3	5.5	15

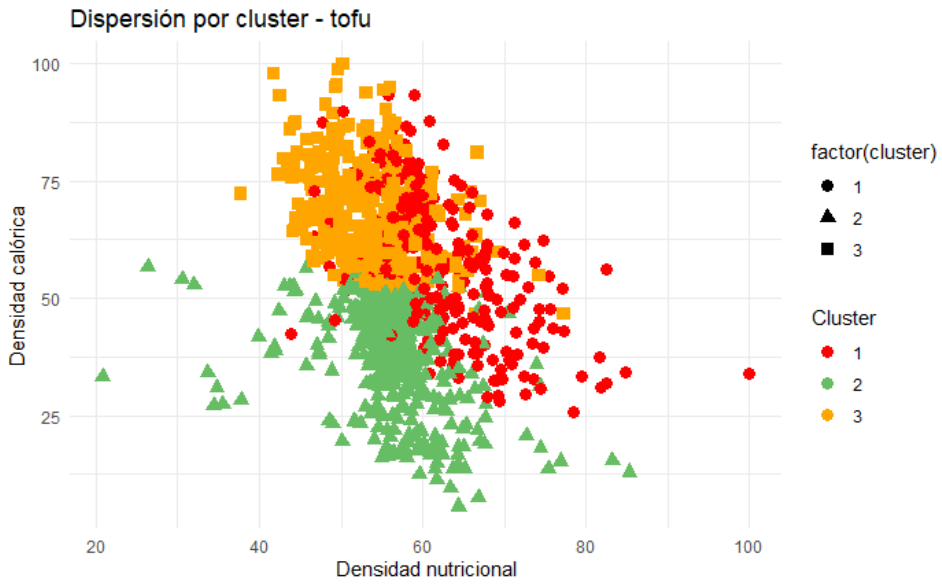
Cluster 2- tofu:

Nombre del Producto	Energía(kcal)	Grasas	Carbohidratos	Proteínas
ORGANIC TOFU CUMBERLAND SAUSAGES	235	13	12.2	14.5
CRISPY TOFU	238	15	12.0	13.0
BASTONCINI DI TOFU	238	15	13.0	12.0

Cluster 3 - tofu:

Nombre del Producto	Energía(kcal)	Grasas	Carbohidratos	Proteínas
MACRO MEDIUM CLASSIC TOFU	102	5.6	3.0	8.2
BIOLOGISCHE NATUREL TOFU	101	5.0	2.0	10.0
TOFU FINAS HIERBAS	101	5.5	1.9	11.0

Gráfico 16 - Tofu



El gráfico previo de dispersión por clúster muestra al clúster 2 como el que tiene menor densidad calórica en términos generales, mientras que el clúster 1, presenta varios productos con mayor densidad nutricional, pero mayor densidad calórica, finalmente el clúster 3 posee mayor densidad calórica y menor densidad nutricional, por lo que parece ser el menos recomendable.

Soja

Los resultados han sido calculados sobre la base de **818** productos.

El análisis de componentes principales realizado previamente muestra como los dos primeros autovalores son mayores que uno, eso nos indica que al seleccionar los dos primeros ejes, el 76% de variabilidad total estaría explicado por las dos primeras componentes resumen. Así mismo se observa como las contribuciones al eje 1 están marcadas principalmente por ENERGY Y FAT, mientras que PROTEINS Y CARBOHYDRATES contribuyen más al eje 2.

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4
Variance	1.956	1.067	0.879	0.095
% of var.	48.943	26.688	21.984	2.385
Cumulative % of var.	48.943	75.631	97.615	100.000

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
1	2.037	0.341	0.004	0.028	-1.263	0.103	0.384	1.556	0.189	0.583
2	1.599	-1.490	0.078	0.868	-0.577	0.021	0.130	0.046	0.000	0.001
3	2.185	-0.087	0.000	0.002	-0.005	0.000	0.000	-0.331	0.009	0.023
4	1.092	-0.988	0.034	0.818	-0.389	0.010	0.127	-0.256	0.005	0.055
5	2.278	2.157	0.163	0.897	0.700	0.032	0.095	-0.203	0.003	0.008
6	2.176	1.322	0.061	0.369	-0.679	0.030	0.097	1.589	0.197	0.533
8	1.820	1.658	0.097	0.830	0.744	0.036	0.167	0.102	0.001	0.003
9	1.118	-0.451	0.007	0.163	-0.503	0.016	0.203	0.889	0.062	0.632
10	2.102	-2.070	0.151	0.970	-0.034	0.000	0.000	-0.363	0.010	0.030
11	2.086	-2.049	0.148	0.965	-0.034	0.000	0.000	-0.391	0.012	0.035

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
ENERGY_100G	0.974	48.474	0.949	0.031	0.090	0.001	0.034	0.128	0.001
FAT_100G	0.733	27.453	0.537	-0.020	0.039	0.000	-0.664	50.125	0.441
CARBOHYDRATES_100G	0.553	15.655	0.306	-0.628	36.977	0.395	0.533	32.348	0.284
PROTEINS_100G	0.406	8.419	0.165	0.819	62.895	0.671	0.391	17.398	0.153

Se ha decidido construir los conglomerados con 3 grupos, al visualizar el gráfico del codo.

Gráfico 17 - Soja

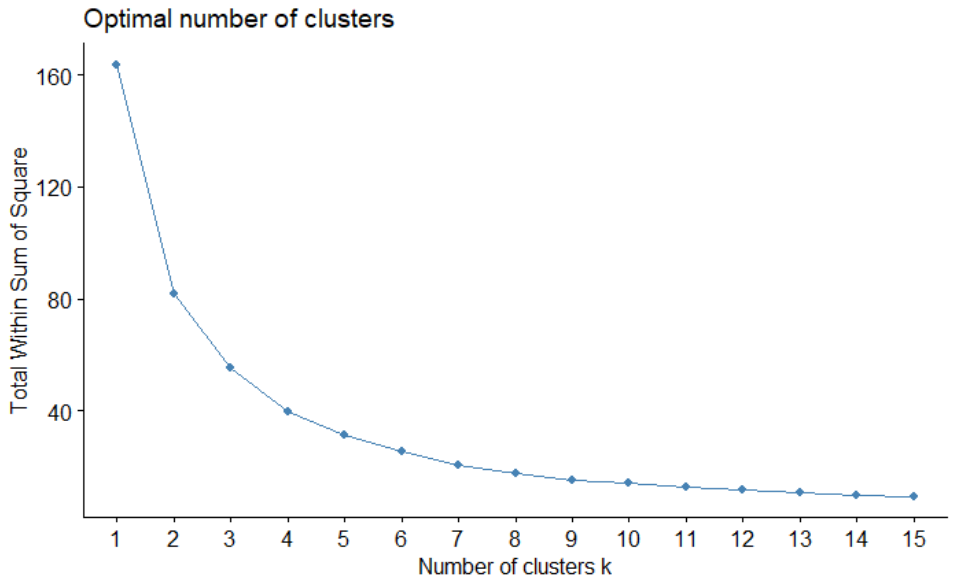


Gráfico 18 - Soja

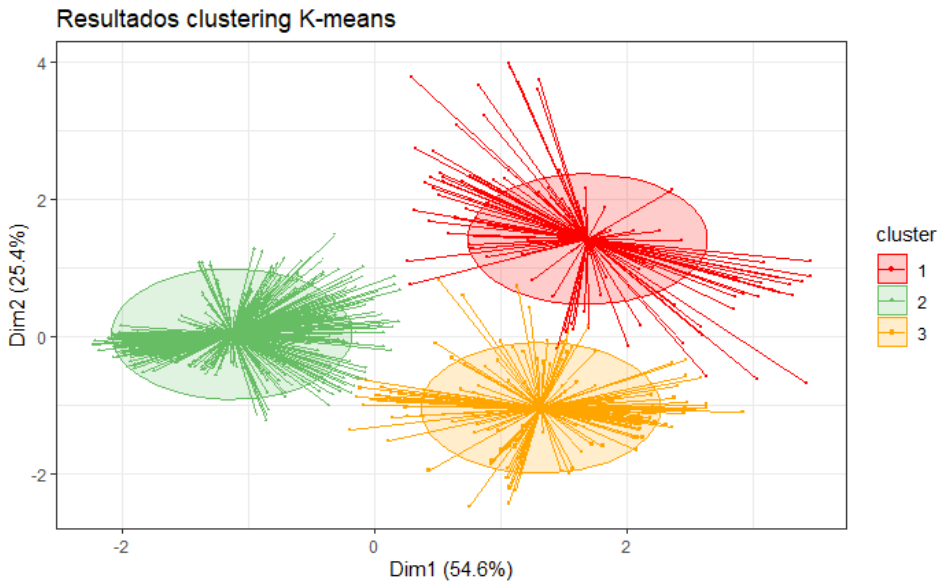
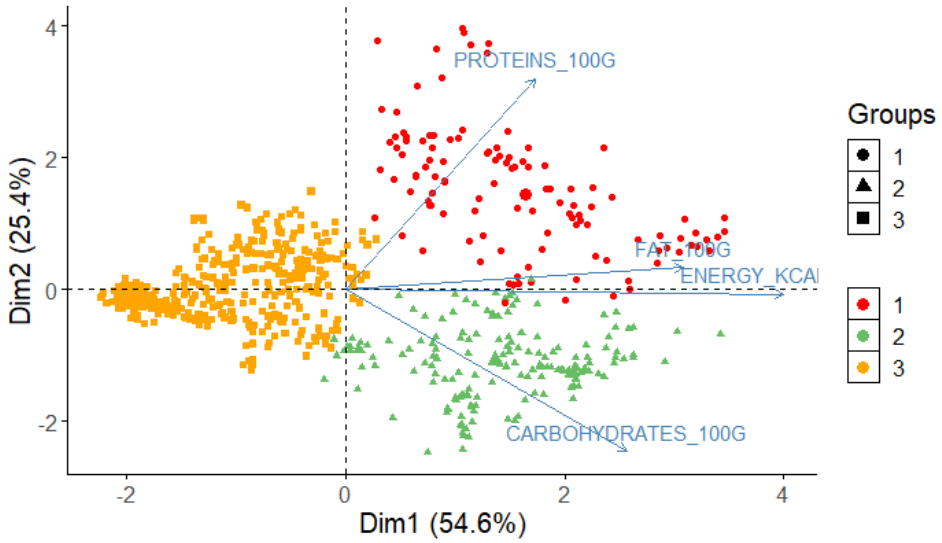


Gráfico 19 - Soja



El gráfico previo ilustra los productos en el plano, se observa primero como las variables ENERGY y FAT se encuentran altamente correlacionadas con el eje 1, por otra parte, CARBOHYDRATES y PROTEINS están correlacionada con el eje 1 y 2, por lo que son explicadas parcialmente por el eje 1 y 2. Productos cercanos a la dirección de los vectores indican valores altos en esas variables, productos opuestos a la dirección de los vectores indican bajos valores en esas variables. En este sentido, se puede identificar lo siguiente:

Grupo 1. Productos de soja con valores altos en energía, grasas y proteínas, pero con valores bajos en carbohidratos.

Grupo 2. Productos con valores bajos en proteínas, grasas, carbohidratos y energía.

Grupo 3. Productos con valores altos en energía y carbohidratos, pero con valores bajos en bajos en grasas y proteínas.

Gráfico 20 - Soja

Cluster 1 - soja:

Nombre del Producto	Energía(kcal)	Grasas	Carbohidratos	Proteínas
SOYGO	83	4.0	8.0	5.0
LE ZUPPE A VAPORE DELICATA SOIA EDAMAME RISO PISELLI PATATE	81	4.1	7.7	2.8
SOYGURT	81	1.9	11.0	3.8

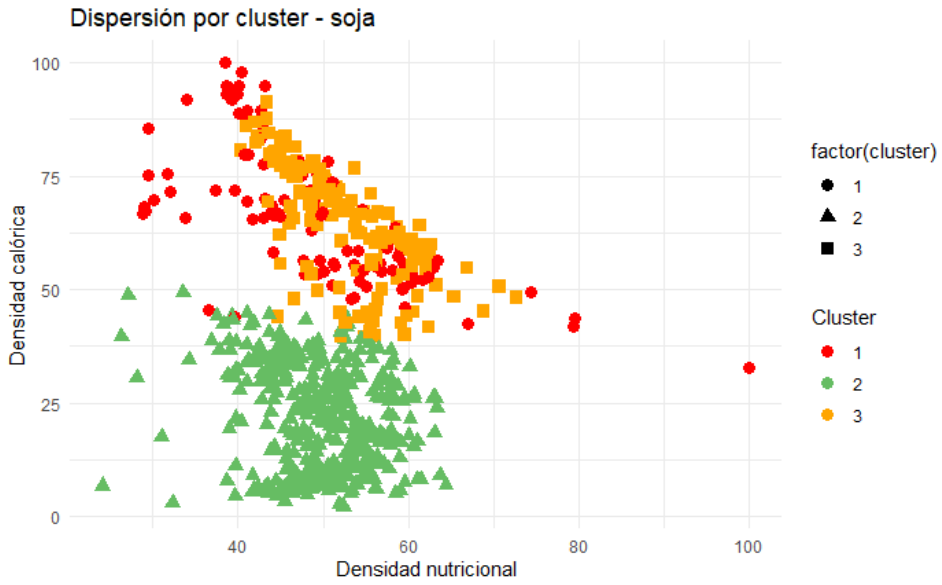
Cluster 2 - soja:

Nombre del Producto	Energía(kcal)	Grasas	Carbohidratos	Proteínas
COTOLETTE DI SOIA BIOLOGICHE	235.0000	9.00000	22.00000	15.0000
SWEET SOICY CHICKEN	232.2581	10.32258	23.22581	11.6129
CROCCHETTE DI SOIA BIO	238.0000	9.30000	20.00000	16.0000

Cluster 3 - soja:

Nombre del Producto	Energía(kcal)	Grasas	Carbohidratos	Proteínas
SOYLENT	450	21.66667	46.66667	21.66667
BROAD BEANS SOY	454	19.60000	41.80000	23.50000
SOY PLUS SIN LACTOSA VAINILLA	455	18.20000	48.50000	21.20000

Gráfico 21 - Soja



Los resultados del análisis de clustering para el producto soya pueden no ser contundentes debido a la alta variabilidad de datos, la dimensionalidad de los datos, la sensibilidad a los hiperparámetros, la presencia de datos ruidosos, la falta de interpretación en la clasificación de productos que entran en el modelo. Es importante tener en cuenta estas limitaciones al interpretar los resultados del análisis de clustering y considerar otras técnicas analíticas o abordajes específicos para abordar estas complejidades en los datos de producción de soya.

3. Explicación del modelo propuesto en la empresa a aplicar en producción

En busca de mantener nuestra aplicación en constante evolución y alineada con nuestro compromiso de proporcionar información actualizada y valiosa a la comunidad vegana, hemos diseñado un plan de actualización que abarca diversas fases cruciales. Este plan no solo busca mantener la aplicación funcionando de manera óptima, sino también mejorar su calidad y relevancia continuamente. En este proceso, se integran acciones desde la revisión de código hasta la implementación de cambios en el ambiente de producción. A continuación, detallaremos las fases clave que conforman este plan y su importancia en el contexto de nuestra misión.

3.1 Fase de preparación

3.1.1 Revisión de código y herramientas: Tiene como objetivo verificar que todas las dependencias, como librerías en R, frameworks de FrontEnd y servidores, estén actualizados. Acciones:

- ➢ Hacer un inventario de todas las librerías y paquetes utilizados.
- ➢ Verificar las versiones actuales y actualizarlas si es necesario.
- ➢ Realizar pruebas de regresión para asegurar que las actualizaciones no rompan las funcionalidades existentes.

3.1.2 Backup de datos y código: Tiene como objetivo salvaguardar toda la información y el código existentes para evitar pérdidas. Acciones:

- ➢ Hacer una copia de seguridad de la base de datos actual.
- ➢ Archivar el código existente en un sistema de control de versiones (GitHub).

3.2 Fase de extracción y limpieza de datos

3.2.1 Descarga de datos nuevos: Tiene como objetivo obtener el dataset más actualizado de la base de datos pública de Open Food Facts. Acciones:

- ➢ Utilizar técnicas de web scraping para descargar los datos desde Open Food Facts.
- ➢ Verificar la integridad de los datos descargados.

3.2.2 Limpieza de datos: Tiene como objetivo asegurarse de que los datos estén en un formato utilizable. Acciones:

- ➢ Realizar tareas de limpieza, como la eliminación de registros duplicados, manejo de valores atípicos y asegurar la calidad de los datos.

3.3 Fase de clusterización en R (Backend)

3.3.1 Preparación de datos: Tiene como objetivo preparar el dataset para el análisis de clustering. Acciones:

- ➢ Importar los datos limpios en R.
- ➢ Realizar transformaciones adicionales, como la normalización.

3.3.2 Ejecución de clusterización: Tiene como objetivo aplicar el modelo de clusterización para clasificar los alimentos. Acciones:

- ➢ Utilizar algoritmo de clustering desarrollado.

3.3.3 Cálculo de Centroides: Tiene como objetivo identificar los valores centrales que caracterizan a cada cluster. Acciones:

- ➢ Calcular los centroides de cada cluster utilizando las funciones pertinentes en R desarrolladas.
- ➢ Almacenar estos centroides en un formato json para ser importado al FrontEnd.

3.4 Fase de actualización del software web (FrontEnd)

3.4.1 Actualización de nuevos Centroides: Tiene como objetivo actualizar el FrontEnd con la nueva información. Acciones:

- ➢ Importación de datos para actualizar la base de datos del FrontEnd.
- ➢ Verificar que los datos se hayan insertado correctamente.

3.4.2 Pruebas de Integración: Tiene como objetivo asegurar que los nuevos datos se integren y muestren correctamente en la interfaz de usuario. Acciones:

- ➢ Realizar pruebas de usabilidad y funcionalidad.
- ➢ Corregir cualquier error o problema de usabilidad que surja.

3.4.3 Despliegue: Tiene como objetivo implementar los cambios en el ambiente de producción. Acciones:

- ➢ Usar técnicas de CI/CD para desplegar los cambios.

Este plan de actualización representa un compromiso sólido con la mejora constante de nuestra aplicación y su contribución a la salud y bienestar de la comunidad vegana. A través de la preparación, extracción y limpieza de datos, clusterización en R y actualización del software web, estamos asegurando que nuestra plataforma siga siendo un recurso valioso y confiable. La sostenibilidad y eficacia de nuestra aplicación son fundamentales para mantener su relevancia a lo largo del tiempo, y este plan estructurado es el medio para lograrlo. En el punto número 6, profundizaremos aún más en cómo estas actualizaciones impactan positivamente en nuestra comunidad vegana.

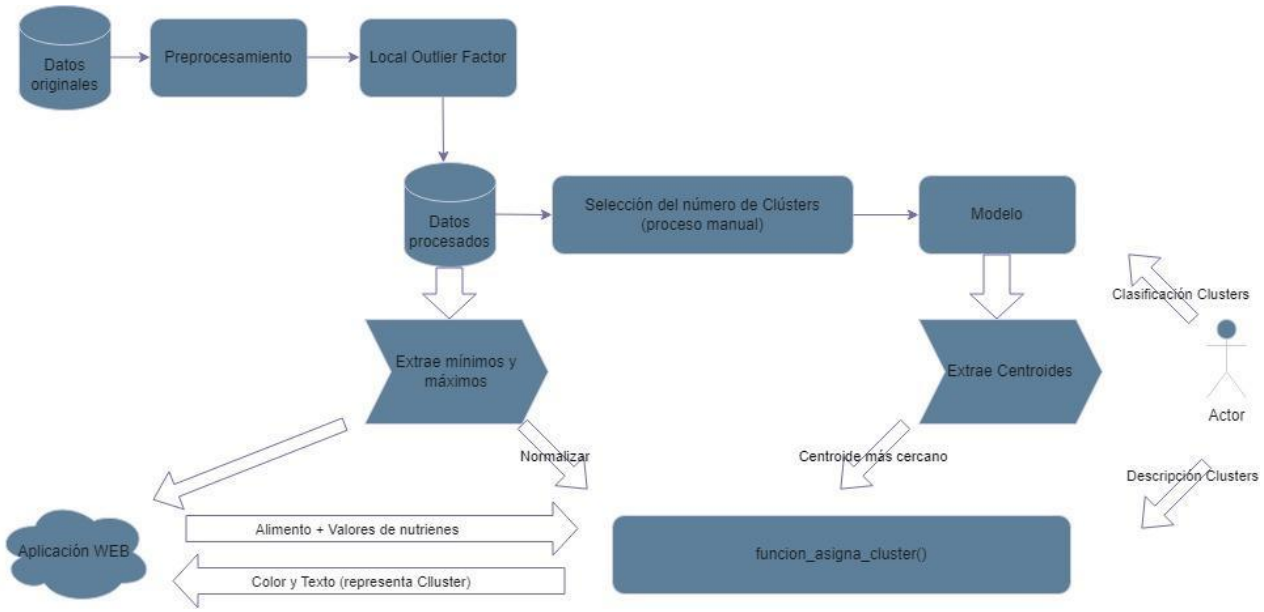
4. Grado de explicación de las soluciones a aplicar.

Se trata de una aplicación web que está disponible libremente para quienes deseen acceder a ella. Esta aplicación es responsiva, es decir, se adapta a las características del browser que está en uso, y posee una interfaz intuitiva que permite a los usuarios interactuar eficazmente, ingresando los valores de los nutrientes seleccionados para el estudio, los cuales deberían ser obtenidos desde la etiqueta descriptiva que posee el mismo alimento que se desea consultar ("información nutricional"), de una forma sencilla y rápida, ya que sólo debe seleccionar el alimento específico que desea consultar y desplazar el valor en las barras parametrizadas que se presentan para cada nutriente. Estos valores se le confirman al usuario en un tabla resumen que se encuentra disponible y que se actualiza cada vez que el usuario realiza una modificación en alguno de las barras indicadoras de cada nutriente o en la selección del nutriente propiamente tal. Finalmente, se le entrega la evaluación del producto consultado, esta consiste en una especificación de las características nutricionales del clúster al cual pertenece el mencionado producto, junto con una respuesta binaria, dependiente de la evaluación realizada donde el cuadro que contiene las características nutricionales se presenta en color verde, si corresponde al clúster más eficiente nutricionalmente o en color rojo, si pertenece a cualquiera de los otros clúster que posee el producto consultado, o lo que es equivalente, si no pertenece al clúster más eficiente nutricionalmente. La determinación del clúster al cual pertenece el alimento consultado se realiza comparando éste, con las coordenadas de los centroides de la totalidad de los clúster existentes para el tipo de alimento seleccionado en la consulta, siendo finalmente catalogado como perteneciente al cluster de cuyo centroide se encuentra más cercano.

Retomando la descripción de las características de la respuesta que recibe el usuario de la aplicación, se debe destacar que no existe posibilidad de error en el mensaje entregado, ya que en caso que éste decida mover alguno de los parámetros seleccionados, es decir, alguno de los valores con los cuales se hizo la consulta previa o el tipo de producto consultado, el resultado de la evaluación desaparece de inmediato, lo que le genera la certeza que la evaluación entregada es correcta.

La aplicación debe estar disponible para cualquier usuario que desee usarla y la versión productiva, necesariamente debería tener un contador de visitas y/o de evaluaciones realizadas, ya que este dato es fundamental para poder establecer la real potencialidad de transformarla en un producto comercial, donde exista una transacción económica por cada consulta o por una membresía del tipo que fuera.

A continuación se presenta imagen de "Esquema representativo de la forma en la cual se genera la Aplicación Web".



Esquema representativo de la forma en la cual se genera la Aplicación Web

5. Estimación de los beneficios económicos respecto a los costes o inversión a aplicar.

Se debe indicar en este punto que en una primera etapa, el servicio será absolutamente gratuito, por lo que sólo en caso que logre generar el interés todos los stakeholders, o al menos en particular de la comunidad vegana, se evaluaría la posibilidad real de que haya un cobro por el uso del servicio. Se estima que, habiendo innumerables servicios disponibles gratuitos hoy en la web, se genera en el 'inconsciente colectivo' la percepción que todos los servicios web deben ser sin costo para el usuario final. No obstante, es importante establecer que si la propuesta de servicio de información acerca de la eficiencia alimenticia de los principales alimentos veganos tiene una demanda real, a partir de ese momento se debería evaluar si habría una disposición a pagar por éste de parte de los distintos stakeholders.

6. Grado de aplicación real de la solución planteada en la empresa

La implementación de la solución será realizada utilizando un conjunto de prácticas de MLOps para favorecer su puesta en producción y reentrenamiento, de esta forma asegurar tener un modelo actualizado y coherente.

Teniendo en consideración las tres canalizaciones de MLOps: datos, ML y software, a continuación se presenta un planteamiento conceptual de su implementación.

6.1 Canalizaciones MLOPs

6.1.1 Canalización de Datos (Data Pipeline):

- ➢ Recolección: Obtención de datos nutricionales de fuentes confiables, como bases de datos especializadas, investigaciones científicas y registros alimenticios. En particular, en el presente trabajo se utiliza la base de datos de alimentos del sitio web https://world.openfoodfacts.org/, una base de datos colaborativa, gratuita y abierta de productos alimenticios de todo el mundo.
- ➢ Preprocesamiento: Limpieza y normalización de datos, eliminando duplicidades, corrigiendo errores, y homogeneizando las unidades de medida.
- ➢ Características: Extracción de características clave de los alimentos, para lo cual se utilizó el método estadístico “Principal Component Analysis" (PCA). En el presente trabajo corresponden a carbohidratos, proteínas, grasas y calorías.
- ➢ Descarte: Aplicación del algoritmo de Local Outlier Factor (LOF) para identificar y descartar datos no pertinentes.
- ➢ Almacenamiento: Una vez procesada, esta información se aloja en sistemas de almacenamiento que garanticen un fácil acceso y manipulación para las etapas siguientes.

6.1.2 Canalización de ML (Model Pipeline):

- ➢ Selección del Modelo: Considerada la problemática de generación de clusters con un algoritmo no supervisado, se optó por el algoritmo k-means, dada su eficacia en este tipo de desafíos.
- ➢ Determinación de Clusters: Se utilizó el método del gráfico elbow para identificar el número óptimo de clusters. Al graficar la variabilidad explicada en función del número de clusters, se seleccionó el "codo" como el punto óptimo.
- ➢ Entrenamiento: Con el número adecuado de clusters, se entrenó el modelo k-means con el conjunto de datos previamente procesado.

- ➢ Validación: Luego de entrenar, se evalúo el modelo para asegurar que las agrupaciones son coherentes y distintivas, en base a los aportes nutricionales de los alimentos que componen cada cluster.
- ➢ Generación de Función: A partir del modelo y sus centroides, se generó una función que, dada la información nutricional de un nuevo alimento, pueda asignarlo al centroide más cercano, es decir, clasificarlo en el cluster correspondiente.
- ➢ Persistencia: Se almacenó el modelo y la función para su uso posterior en el proceso de despliegue.

6.1.3 Canalización de Software (Software Pipeline):

- ➢ Desarrollo de la Aplicación Web: Se creó una aplicación web con una interfaz sencilla e intuitiva que permite a los usuarios ingresar la información nutricional de un alimento. Lo cual fue descrito en detalle el punto 4 “Grado de explicación de las soluciones a aplicar”.
- ➢ Integración: A la aplicación ya desarrollada, se integra la función previamente generada. Así, al ingresar los datos de un determinado alimento, la aplicación establece el cluster correspondiente, para luego basado en sus características, asignarle el color rojo o verde, según corresponda, con la respectiva caracterización del cluster.
- ➢ Despliegue: Una vez lista, la aplicación se despliega en un servidor que puede garantizar flexibilidad, escalabilidad y continuidad del servicio. En particular, en el presente trabajo se utiliza el servicio proporcionado en https://www.shinyapps.io/
- ➢ Monitoreo y Actualización: Para efectos de una correcta prestación del servicio, se considera fundamental el monitorizar la aplicación para detectar potenciales anomalías y garantizar su rendimiento óptimo. Con el feedback obtenido, se deberían realizar ajustes y actualizaciones según sea necesario.

Al abordar la agrupación de alimentos según su aporte nutricional con una perspectiva MLOps, aseguramos la precisión y eficacia de nuestro modelo, y también su aplicabilidad real. La solución propuesta permite agrupar una base de datos existente y establecer cluster a nuevos alimentos en tiempo real, todo ello accesible desde una plataforma web. Esta sinergia entre datos, machine learning y software potencia las capacidades y el alcance de la solución, proporcionando una herramienta robusta y dinámica.

6.2 Recursos Humanos:

Se establece que los roles necesarios para completar el proyecto son los que se describen a continuación:

- ➢ Equipo de Desarrollo de Software: Profesionales con habilidades en desarrollo web en Shiny y backend con R, utilizando R-Studio, responsables de construir, mantener y optimizar la aplicación web y los servicios asociados.

- ➢ Ingenieros de Datos: Se encargan de gestionar, limpiar y estructurar los datos. Trabajan estrechamente con el equipo de ciencia de datos para asegurar que la información esté lista para ser procesada y analizada.
- ➢ Científicos de Datos: Especialistas en k-means y otros algoritmos de aprendizaje automático. Se encargan de la creación, entrenamiento y optimización del modelo.
- ➢ Ingenieros de MLOps: Facilitan la integración de las canalizaciones de Datos, ML y Software. Son responsables de la Integración Continua, Entrega Continua y Entrenamiento Continuo, garantizando que el modelo esté siempre actualizado y que la aplicación funcione sin problemas.
- ➢ Testers/QA Engineers: Especialistas en pruebas que garantizan la calidad de la aplicación y el modelo, detectando errores y asegurando que todo funcione según lo esperado.
- ➢ Equipo de Operaciones/DevOps: Garantizan que la infraestructura (servidores, bases de datos, etc.) esté optimizada, segura y lista para despliegues y escalabilidad.

6.3 Estrategia de Puesta en Producción para la Solución de MLOps

6.3.1 Integración Continua (IC):

La puesta en marcha comienza con una robusta práctica de IC. Cada vez que se introduce un cambio en el modelo o en el código de la aplicación, se activa automáticamente un proceso de construcción y prueba.

- ➢ Automatización de Pruebas: Se desarrollan pruebas unitarias y de integración para validar los cambios en la funcionalidad y en el modelo de aprendizaje automático.
- ➢ Retroalimentación Inmediata: Si alguna prueba falla, se notifica al equipo de desarrollo para una corrección rápida, asegurando que solo los cambios validados pasen al siguiente paso.

6.3.2 Entrega Continua (EC):

Después de la IC, los cambios validados se mueven a un entorno de pre-producción.

- ➢ Despliegue Automatizado: Uso de herramientas de despliegue automático para llevar los cambios a un entorno de pre-producción o test.
- ➢ Pruebas de Aceptación Automatizadas: Antes de realizar cualquier cambio en el ambiente de producción, se ejecutan pruebas de aceptación para asegurar la calidad y la correcta funcionalidad.
- ➢ Validación Manual: Se realiza una última revisión por parte de un equipo especializado para garantizar que todo está en orden y listo para la producción.

6.3.3 Despliegue:

Se realiza un despliegue controlado, esto permite que los usuarios experimenten la nueva versión del modelo. Se monitorea el rendimiento y luego se procede al despliegue completo.

6.4 Estrategia de Entrenamiento para la Solución de MLOps

6.4.1 Monitorización de Datos:

Dado que los patrones en los datos nutricionales pueden cambiar con el tiempo, se monitorearán semanalmente los datos de alimentos del sitio web https://world.openfoodfacts.org/. Si el incremento de alimentos veganos objeto de este trabajo es igual o superior al 5% de los alimentos en la base de datos actual, se evalúan la cantidad de alimentos outlier en éstos, comparados con la base de datos actual, si la cantidad de outliers supera el 10%, se procederá a un preprocesamiento de datos y reentrenamiento del modelo, y se luego se continúa con las siguientes etapas ya descritas. En cualquier otro caso o escenario, se descarta el reentrenamiento del modelo.

6.4.2 Entrenamiento Continuo (ET):

- ➢ Reentrenamiento: Dependiendo de los resultados del punto precedente se reentrena el modelo con nuevos datos.
- ➢ Validación Automática: Una vez reentrenado, el nuevo modelo se compara automáticamente con el anterior para validar su rendimiento.
- ➢ Generación de Función: A partir del nuevo modelo y sus centroides, se genera una nueva función que, dada la información nutricional de un nuevo alimento, pueda asignarlo al centroide más cercano, es decir, clasificarlo en el cluster correspondiente.
- ➢ Control de Versiones: Se mantiene un registro de todas las versiones del modelo, asegurando que se pueda volver a una versión anterior en caso de problemas.

6.4.3 Despliegue:

Se realiza un despliegue controlado, esto permite que los usuarios experimenten la nueva versión del modelo. Se monitorea el rendimiento y luego se procede al despliegue completo.

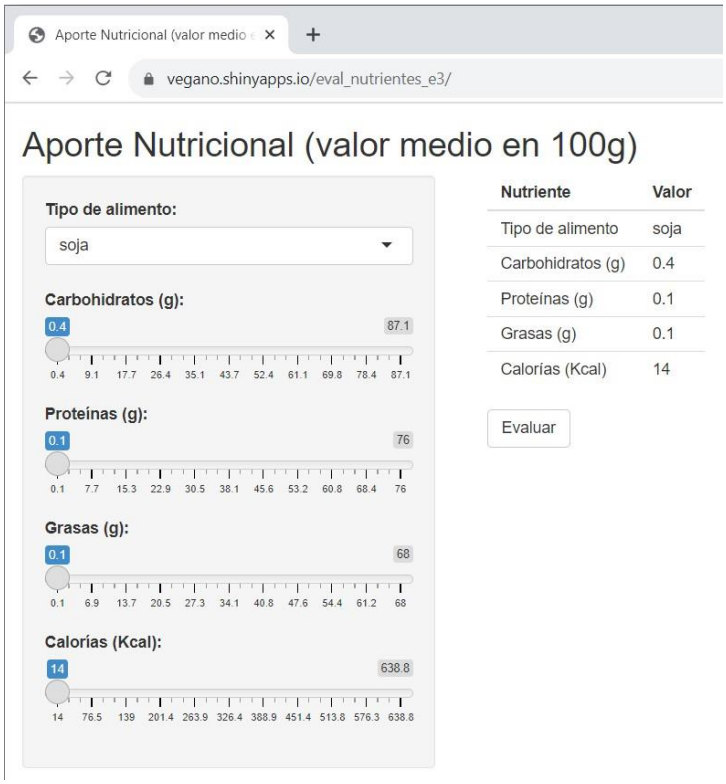
Finalmente, indicar que las estrategias de puesta en producción y entrenamiento, están diseñadas para proporcionar una transición suave y sin interrupciones de los cambios y actualizaciones al entorno de producción. Al seguir estos enfoques, se garantiza que la solución MLOps esté siempre optimizada, actualizada y lista para satisfacer las necesidades cambiantes de los datos nutricionales de los alimentos.

6.5 Aplicación propiamente tal

Para acceder a la aplicación de evaluación de eficiencia nutricional de alimentos veganos, se debe hacer a través del siguiente link:

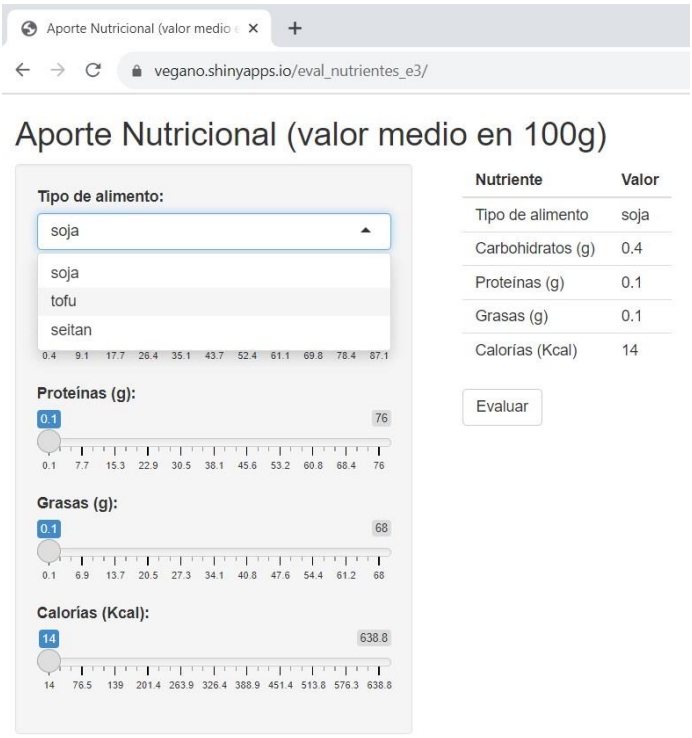
https://vegano.shinyapps.io/eval_nutrientes_e3/

A continuación se presenta imagen de “Pantalla Inicio” de la aplicación web.



Pantalla Inicio

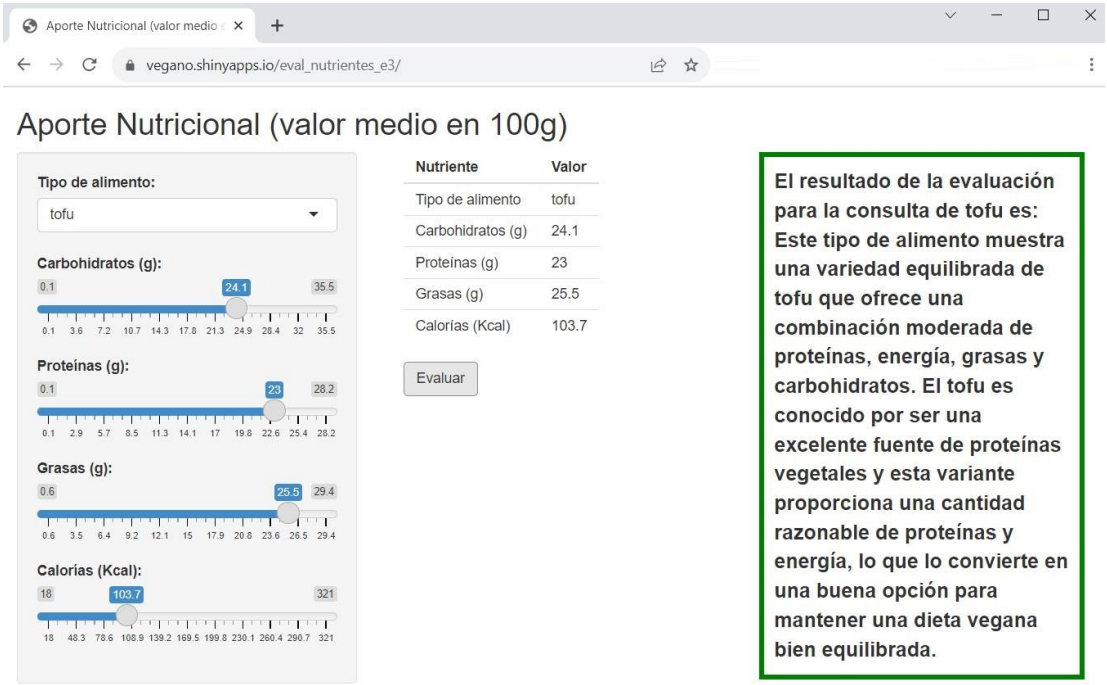
A continuación se presenta imagen de “Selección de Tipo de Alimento”.



Selección de Tipo de Alimento

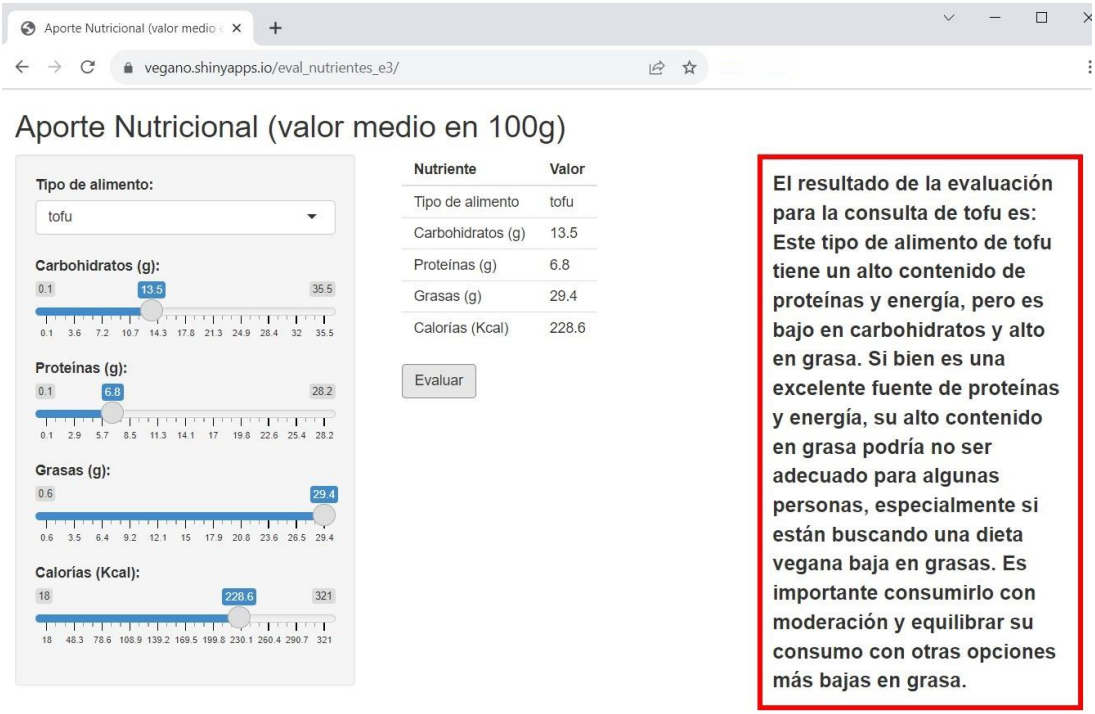
Posterior a la selección del tipo de alimento se deben ingresar los aportes nutricionales correspondientes al alimento a consultar y presionar “Evaluar”, luego de lo cual se obtendrá la respectiva respuesta de la evaluación nutricional del alimento, donde podría pertenecer al grupo de alimentos eficientes representado en un cuadro verde o no pertenecer al grupo de alimentos eficientes representado en un cuadro rojo.

A continuación se presenta imagen de “Respuesta perteneciente al grupo de alimentos eficiente”.



Respuesta perteneciente al grupo de alimentos eficiente

A continuación se presenta imagen de "Respuesta no perteneciente al grupo de alimentos eficiente".



Respuesta no perteneciente al grupo de alimentos eficiente

7. Reflexión final sobre problemas encontrados a lo largo del TFM y soluciones puestas en marcha.

A lo largo del desarrollo de este Trabajo Final de Máster (TFM), el equipo se enfrentó a diversos desafíos intrínsecos al procesamiento y análisis de grandes conjuntos de datos. Estos desafíos y sus respectivas soluciones se detallan a continuación:

1. **Adquisición de datos**: Una vez identificada la fuente primaria del conjunto de datos, en este caso, Open Food Facts, se experimentaron inconvenientes con la estabilidad de su API. Durante las peticiones de datos, particularmente cuando se aplicaron filtros específicos o se intentaba obtener el conjunto de datos completo, la API mostraba interrupciones recurrentes, resultando en la falta de retorno de información. Para sortear este obstáculo, se optó por descargar directamente el conjunto de datos en su formato CSV desde el recurso proporcionado por la propia plataforma [https://static.openfoodfacts.org/data/en.openfoodfacts.org.products.csv]. Esta decisión permitió tener acceso completo y sin interrupciones al dataset.

2. **Filtrado y preprocesamiento de datos**: Una vez en posesión del dataset, se enfrentó el reto de su dimensión y heterogeneidad. Con múltiples columnas, como "*product_name*", "*generic_name*", "*packaging*", "*categories*", "*origins*", "*labels*", "*emb_codes*" e "*ingredients_text*", fue imperativo definir una estrategia de filtrado coherente. Dado el enfoque definiido en los productos "Tofu", "Seitán" y "Soya", se determinó que la columna "product_name" sería la más adecuada para filtrar inicialmente estos productos de interés.

3. **Normalización lingüística**: Al analizar más detenidamente el campo "product_name", se identificó un desafío adicional: la variabilidad lingüística. Los nombres de los productos se presentaban en múltiples idiomas, contabilizando aproximadamente 30 diferentes. Para asegurar una captura completa y precisa de la data de los productos de interés, en este contexto multilingüe, se tradujeron las palabras "Tofu", "Seitan" y "Soya" a cada uno de estos idiomas. Esta acción permitió realizar un filtrado más efectivo y coherente del dataset.

4. **Exclusión de datos no pertinentes**: Una fase crítica del preprocesamiento fue identificar y excluir aquellos productos que no se alineaban con los objetivos específicos del estudio. Para lograr esto, se emplearon filtros negativos en el campo "product_name". Estas exclusiones no solo se basaron en términos en español, sino que se tradujeron a los aproximadamente 30 idiomas detectados en el dataset, asegurando una exclusión exhaustiva y precisa.

5. **Enriquecimiento y normalización del dataset**: En un esfuerzo por complementar la riqueza del análisis, se decidió integrar variables macroeconómicas, como el Producto Interno Bruto (PIB). Esta integración llevó al equipo a enfrentar el desafío de normalizar los nombres de los países a un estándar unificado, optando por la versión en inglés de cada nombre. Además, para facilitar futuros análisis y cruces de información, se incorporó el código ISO3 por país. Esta labor se realizó tanto en los datos provenientes de Open Food Facts como en los datos económicos obtenidos del Banco Mundial.

6. Tratamiento de datos faltantes: Una revisión meticulosa del dataset reveló una cantidad considerable de datos faltantes, especialmente en las columnas numéricas. Por tal motivo, el enfoque se centró en seleccionar aquellas columnas cuya suma total de registros fuese positiva. Para las columnas de tipo numérico con datos 'NaN', se optó por imputar un valor 0 (cero). Por otro lado, en cuanto a las columnas categóricas, se llevó a cabo una transformación a formato numérico para facilitar análisis posteriores.

7. Detección y tratamiento de Outliers: En proyectos de inteligencia artificial, la presencia de outliers puede distorsionar modelos, haciendo que estos se adapten a extremos y, por ende, reduciendo su capacidad de generalizar para datos nuevos o no vistos. Para abordar este problema, se implementó la técnica de Local Outlier Factor (LOF). A través de esta técnica, se identificaron y excluyeron el 2,5% extremo de los datos, permitiendo trabajar con el 97,5% restante, lo que ofrece una representación más homogénea y fiable del conjunto global.

8. Desbalance de clases en la Clusterización: El desbalance de clases en la clusterización ocurre cuando uno o más clusters acumulan un volumen significativamente mayor de observaciones que los demás. Este desbalance puede llevar a una interpretación errónea o poco útil del modelo, ya que clusters pequeños pueden estar representando patrones igualmente importantes que los grandes. Para este estudio se hace fundamental abordar este problema porque un desbalance pronunciado podría indicar que el algoritmo está siendo parcial a ciertas características y podría no captar la estructura subyacente de los datos de manera precisa. Se aplicaron técnicas de re-muestreo, para balancear el número de observaciones antes del proceso de clusterización.

9. Selección de características redundantes o irrelevantes: La inclusión de características irrelevantes o redundantes puede llevar al sobreajuste y a una disminución del rendimiento del modelo. Un modelo con demasiadas características irrelevantes no solo es computacionalmente más costoso, sino que también puede resultar en un modelo menos interpretable y más difícil de validar. Se realizó un análisis de componentes principales (PCA) para eliminar la multicolinealidad y reducir la dimensionalidad del conjunto de datos.

10. Actualización de librerías y dependencias: Las librerías y dependencias externas del proyecto pueden recibir actualizaciones que interrumpan la compatibilidad con el código existente. La interrupción en la compatibilidad puede llevar a fallos en el funcionamiento del modelo o de la aplicación, y puede requerir un tiempo considerable para resolver el problema. Por ello se buscó mantener un entorno que permita replicar fácilmente las condiciones originales del proyecto.

11. Desarrollo de Aplicación Web: La incorporación de una aplicación web para interactuar con usuario representó un desafío, debido que ninguno de los miembros del equipo tenía experiencia en el desarrollo de este tipo de aplicaciones. Por lo cual fue necesario realizar un proceso investigativo respecto a las posibilidades existentes y que se alinearan con nuestras necesidades, producto de lo cual determinamos que la biblioteca Shiny para R cumplía con nuestros requerimientos y optamos por utilizarla para el desarrollo de la aplicación web del presente trabajo.

A lo largo de la ejecución de este Trabajo Fin de Máster (TFM), se han enfrentado y superado diversos desafíos relacionados con el procesamiento y análisis de datos extensos. La adquisición de datos se resolvió mediante la descarga directa del conjunto de datos en formato CSV, superando las interrupciones de la API de Open Food Facts. El filtrado y preprocesamiento de datos se centraron en la columna "product_name" para enfocarse en los productos de interés y garantizar coherencia. La normalización lingüística y la exclusión de datos no pertinentes se llevaron a cabo en múltiples idiomas, asegurando la alineación con los objetivos de estudio. Además, se enriqueció y normalizó el dataset con variables macroeconómicas, se trató la falta de datos y se abordó el desbalance de clases en la clusterización. Finalmente, se destacó el desarrollo de una aplicación web utilizando la biblioteca Shiny para R como un logro significativo en el proyecto.

En resumen, este TFM demostró la importancia de abordar desafíos en la gestión de datos de manera efectiva para alcanzar los objetivos de investigación. Cada obstáculo se enfrentó con soluciones meticulosamente implementadas, lo que contribuyó al éxito del estudio y enriqueció la comprensión de la ciencia de datos y la inteligencia artificial. Estos aprendizajes son aplicables no solo a este trabajo, sino también a un contexto más amplio de investigación y aplicación de datos.