DATA SCIENCE II

Alumno: Yesid García

Comisión: 75690

“Clasificación Nutricional en la Industria Alimentaria”

CONTENIDO

- Introducción.
- Descripción del problema.
- Objetivo.
- Fuente.

- Preguntas de Interes
- Hipótesis.
- Conclusiones.
- Oportunidades e Insights Clave.

INTRODUCCIÓN

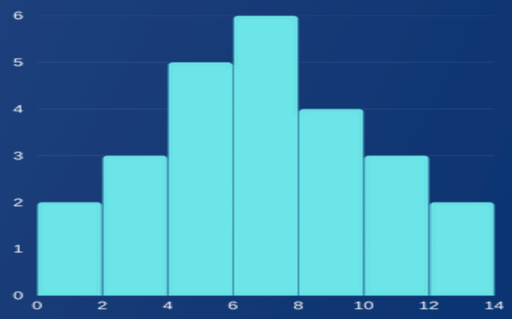
La clasificación precisa de productos alimenticios, basada en su perfil nutricional, es fundamental para la toma de decisiones estratégicas en la industria. Este enfoque permite una segmentación de mercado efectiva, un análisis competitivo profundo y el desarrollo de nuevas líneas de productos que respondan a las tendencias de salud y consumo.

Aplicaciones Clave:

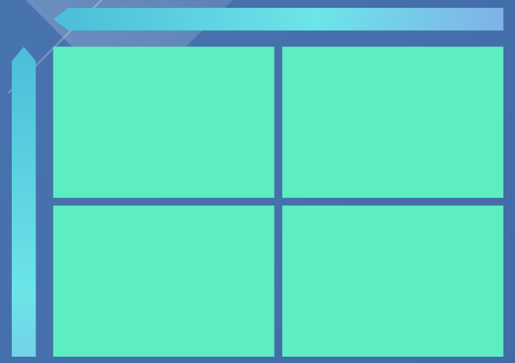
- Segmentación de Mercado: Identificar nichos y adaptar ofertas.
- Análisis Competitivo: Comparar productos y categorías.
- Desarrollo de Nuevas Líneas: Responder a tendencias de salud y consumo.

Nutrientes Clave:

- Energía (nutriments.energy100g)
- Grasas (nutriments.fat100g)
- Azúcares (nutriments.sugars100g)
- Sal (nutriments.salt100g)
- Proteínas (nutriments.proteins100g)
- Exceso de Grasas y Azúcares (a determinar)



DESCRIPCIÓN DEL PROBLEMA



El problema principal que aborda este proyecto, según la descripción proporcionada en el cuaderno de Colab, es la necesidad de clasificar de manera precisa productos alimenticios basándose únicamente en su perfil nutricional. Esto surge de la importancia de tener una clasificación correcta en la industria alimentaria para diversas aplicaciones, como:

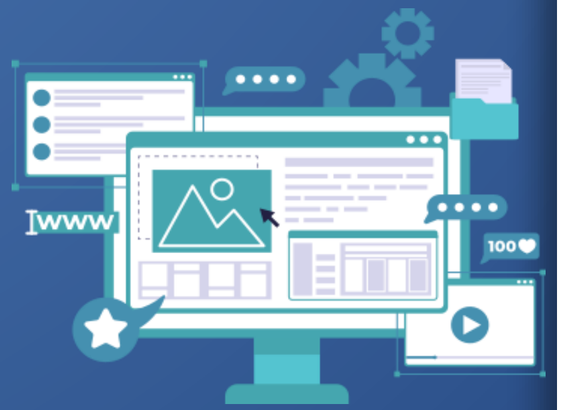
- **Segmentación de mercado.**
- **Análisis competitivo.**
- **Desarrollo de nuevos productos.**

El proyecto busca resolver este problema utilizando las variables nutricionales como base para predecir la categoría del producto, evaluando así la viabilidad de este enfoque para identificar patrones nutricionales típicos por segmento y evaluar oportunidades comerciales.

OBJETIVO PRINCIPAL: Predecir Categorías por Nutrientes

1. **Identificar Perfiles:** Reconocer perfiles nutricionales típicos por segmento.
2. **Evaluar Oportunidades:** Analizar oportunidades comerciales y patrones de consumo.
3. **Facilitar Segmentación:** Mejorar la segmentación del mercado.

Este análisis busca predecir automáticamente la categoría de un producto alimenticio (snacks, bebidas, lácteos, etc.) utilizando únicamente sus valores nutricionales. Esto nos permitirá identificar perfiles nutricionales típicos por segmento, evaluar oportunidades comerciales y detectar patrones de consumo, facilitando la segmentación del mercado.



Objetivos Específicos del Análisis

1. **Identificar Perfiles Nutricionales Característicos:** ¿Qué nutrientes definen a cada tipo de producto? Análisis de tendencias centrales y distribución de nutrientes por categoría.
2. **Desarrollar y Evaluar un Modelo de Clasificación Automática:** ¿Se puede predecir la categoría con un modelo automático usando solo los nutrientes? Selección de características, algoritmo, entrenamiento y evaluación de precisión.
3. **Comparar Perfiles Nutricionales entre Categorías:** ¿Qué categorías tienen un mejor o peor perfil nutricional general? Visualizaciones comparativas de nutrientes y cálculo de puntuaciones.
4. **Analizar Concentración de Azúcares y Grasas:** ¿Qué tipos de productos tienden a concentrar más azúcares o grasas?

python™

FUENTE DE DATOS

Origen: Los datos provienen de la API pública de **Open Food Facts**, una base de datos colaborativa con más de un millón de productos.

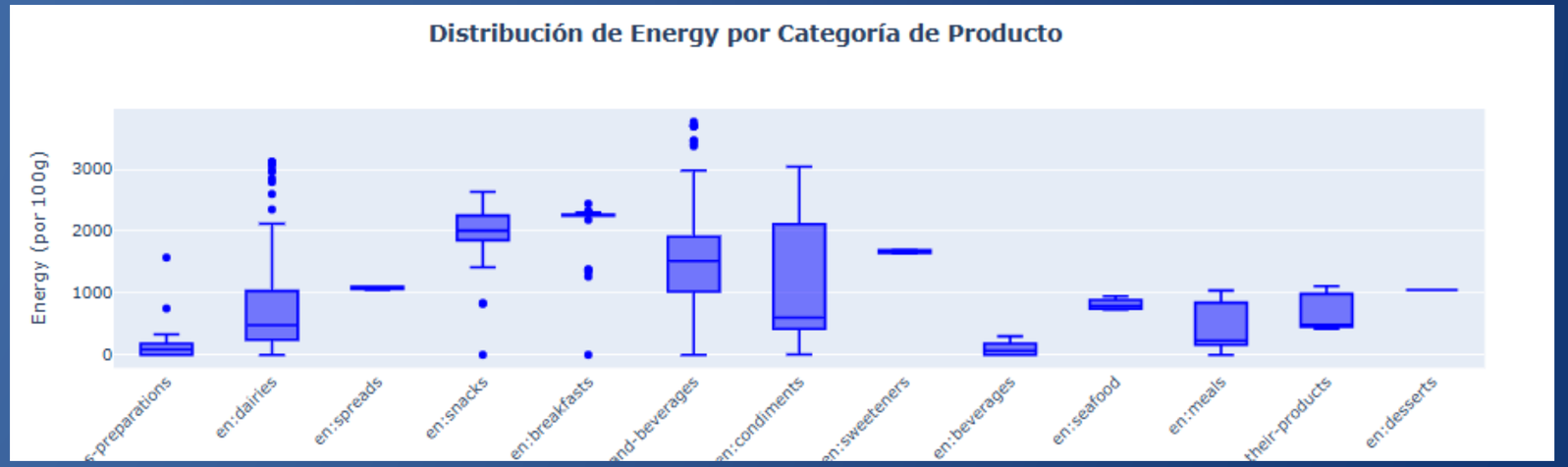
LINK:
https://www.google.com/url?q=https%3A%2F%2Fworld.openfoodfacts.org%2Fcgi%2Fsearch.pl%3Faction%3Dprocess%26json%3Dtrue%26page_size%3D100

Variables Extraídas: Categorías de productos c(ategoriestags), Nutrientes por 100g: energía, grasas, azúcares, proteínas, sal, fibra, Marca (para análisis secundarios).

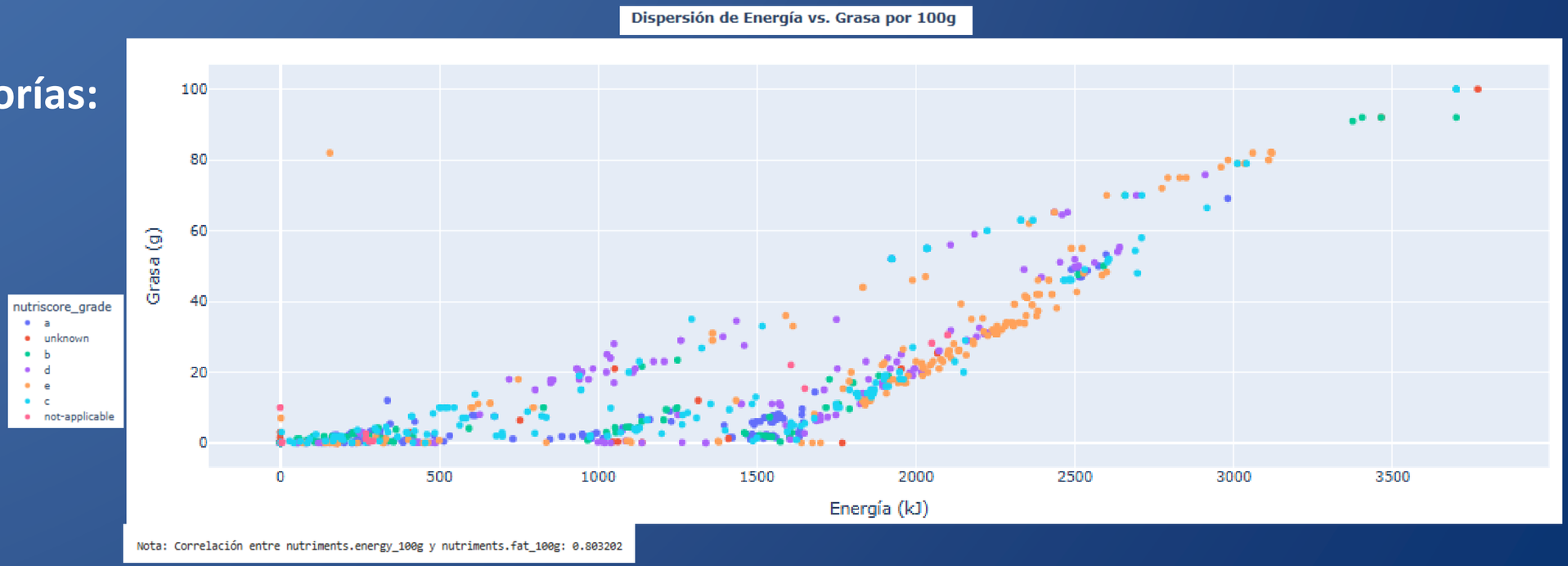
Tamaño de la Fuente: 713 filas x 23 columnas.

PREGUNTAS DE INTERÉS Y PATRONES NUTRICIONALES

1. **Nutrientes por Categoría:** Los perfiles nutricionales son distintivos y pueden usarse para clasificación.

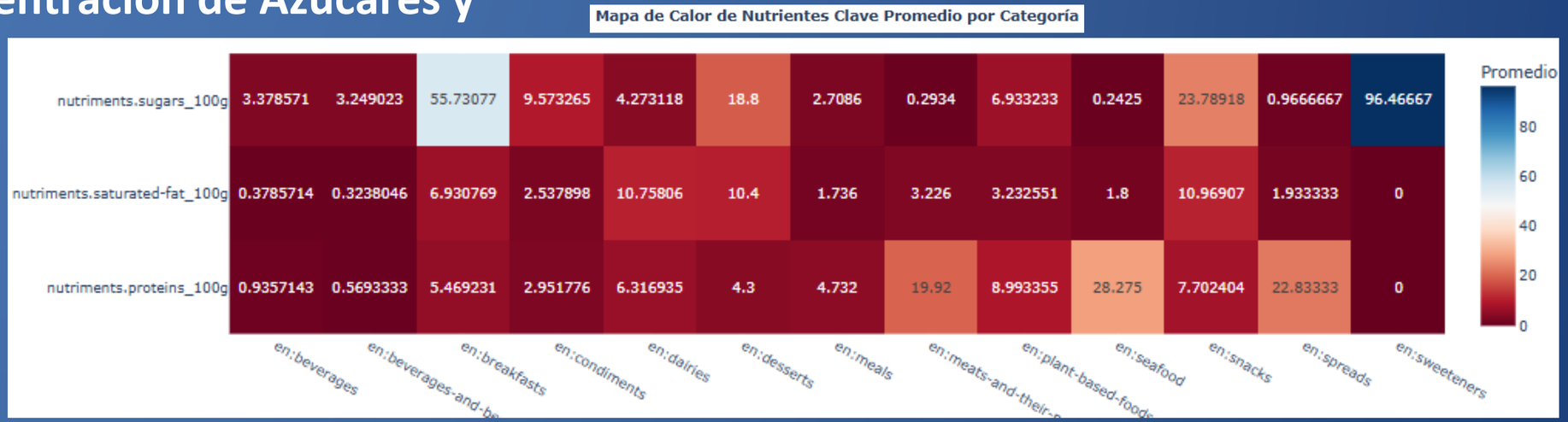


2. **Relación de Categorías:** Las mejor relación encontrada es la siguientes.

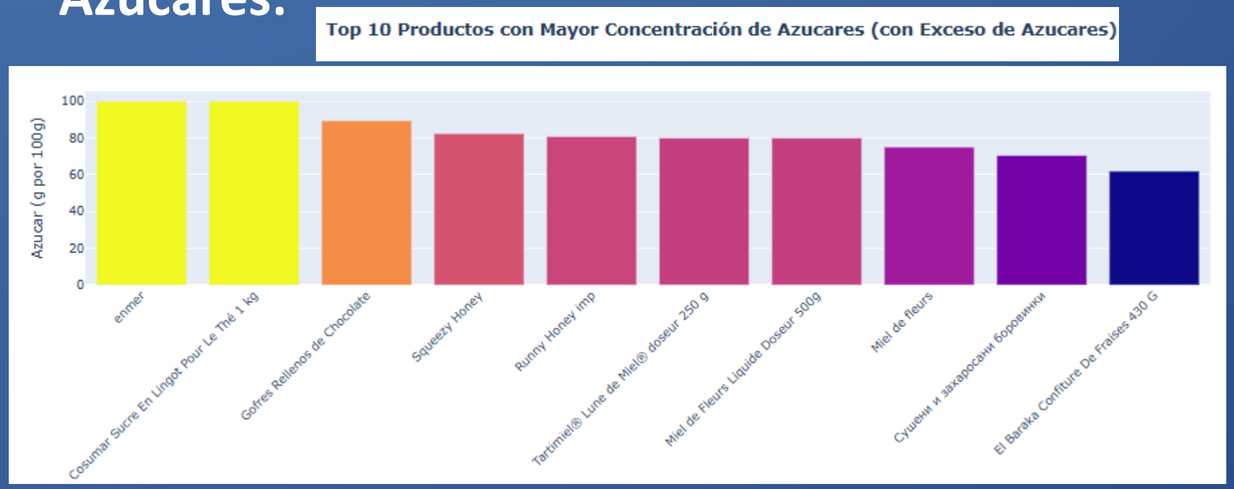


PREGUNTAS DE INTERÉS Y PATRONES NUTRICIONALES

3. Concentración de Azúcares y Grasas:

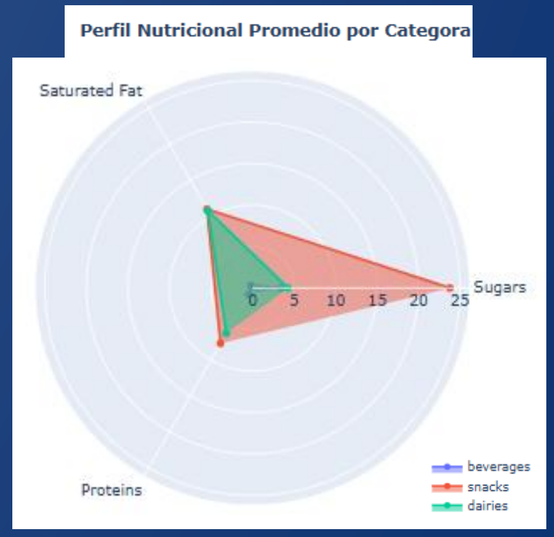


4. Productos con mayor concentración de Azúcares:



5. Oportunidades para innovar en snacks o postres más saludables:

Reducción de Azúcares Añadidos y Grasas Saturadas, Incremento del Contenido de Fibra y Proteínas, Uso de Ingredientes Naturales y Menos Procesados.



HIPÓTESIS PLANTEADAS

Hipotesis 6: Todos los productos con Nutri Score D y E poseen alergenos.

Hipotesis 1: Los productos pueden ser correctamente clasificados en categorías usando su perfil nutricional.

Hipotesis 5: Los productos más saludables como Snacks y bebidas tienden a concentrarse en ciertas categorías específicas y tienen un Nova Group por debajo de 3.

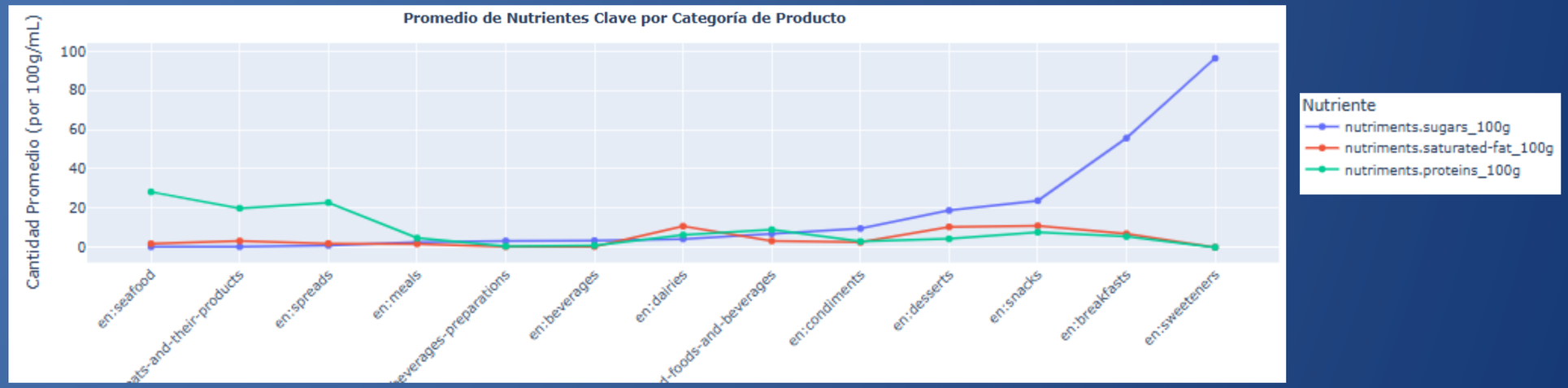
Hipotesis 2: Las bebidas tienen menor contenido de proteínas y grasas que los snacks o postres.

Hipotesis 4: Algunas marcas como Sidi Ali presentan más oportunidades de mejora nutricional azucares que otras marcas.

Hipotesis 3: Las categorías presentan perfiles nutricionales distintos y predecibles.

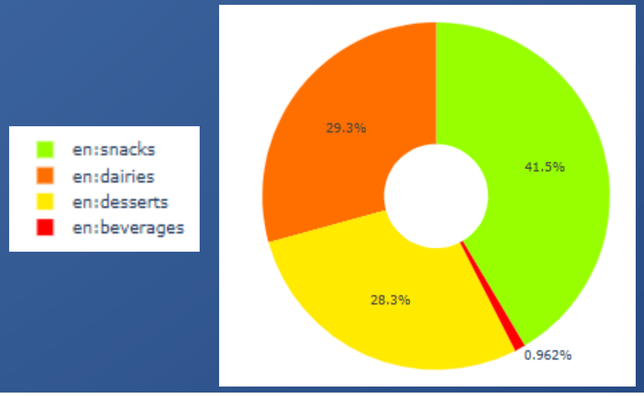
HIPÓTESIS 1: Los productos pueden ser correctamente clasificados en categorías usando su perfil nutricional.

Verdadero, El análisis exploratorio confirma una fuerte asociación entre el perfil nutricional y la categoría de un producto. Las diferencias en nutrientes entre categorías sugieren que cada tipo de producto tiene un "perfil nutricional característico".continuación de la siguiente visualización podremos tener la siguiente conclusión:

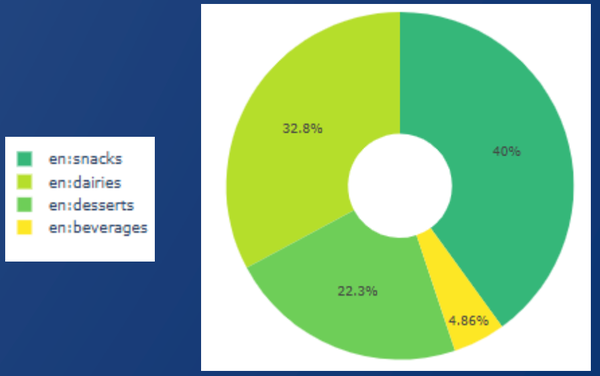


HIPÓTESIS 2: Las bebidas tienen menor contenido de proteínas y grasas que los snacks o postres.

Verdadero, puede verse esa diferenciación y resultado:



Distribución Promedio de Grasa por Categoría (Bebidas, Snacks, Postres)



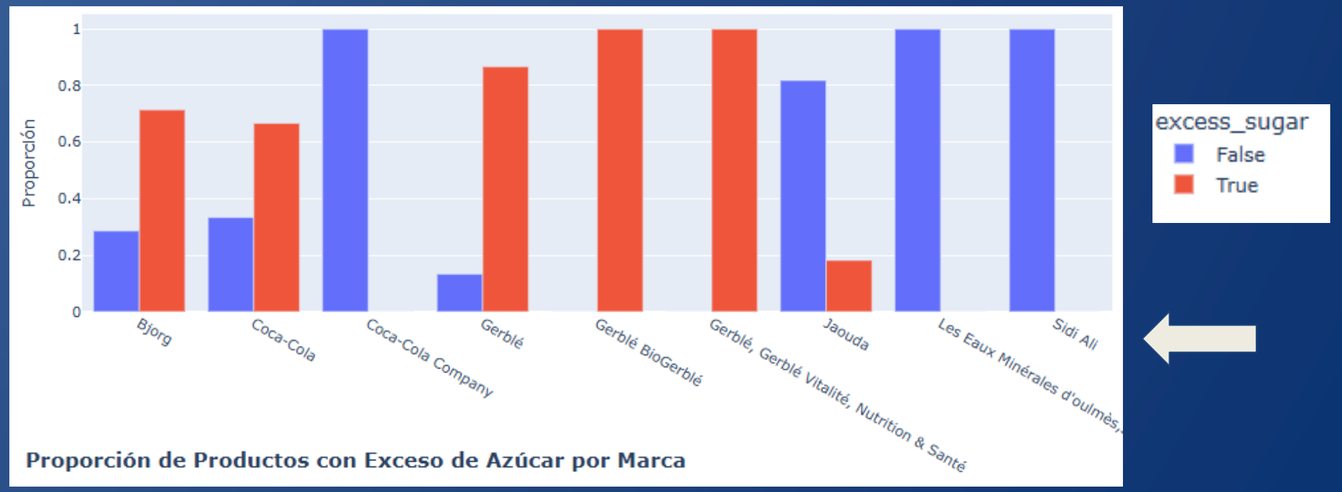
Distribución Promedio de Proteínas por Categoría (Bebidas, Snacks, Postres)

HIPÓTESIS 3: Las categorías presentan perfiles nutricionales distintos y predecibles.

Verdadero, los hallazgos en el EDA apoyan fuertemente este resultado de NUTRIENTES CLAVE POR CATEGORÍA:.

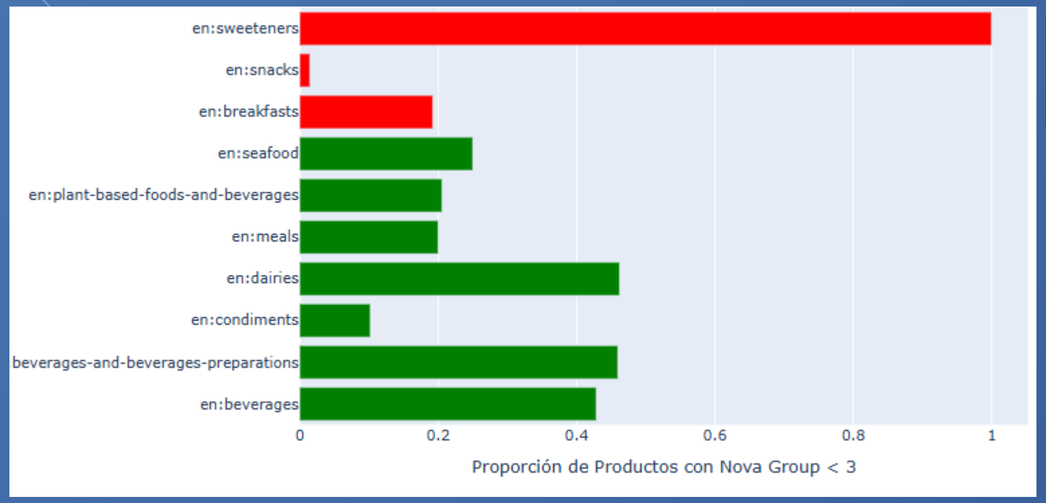
HIPÓTESIS 4: Sidi Ali presenta más oportunidades de mejora nutricional en azúcares que otras marcas.

Falso, Falso, Sidi Ali de hecho se encuentra entre las marcas con menores contenidos de azúcar y sin excesos de azúcares por normativa establecida por debajo de otras marcas que deberían tener prioridad en estas mejoras como Gerbié, Coca-Cola ó Bjorg.

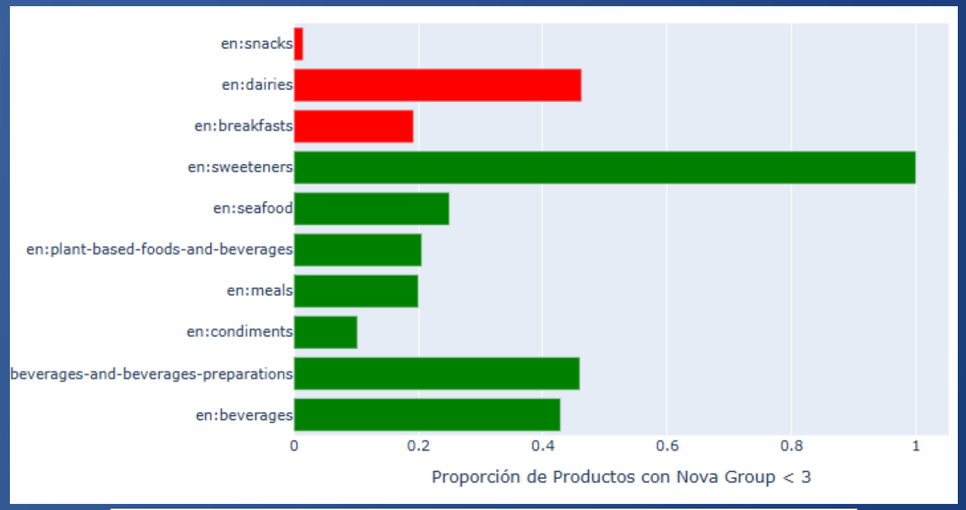


HIPÓTESIS 5: Los productos más saludables como Snacks y bebidas tienden a concentrarse en ciertas categorías específicas y tienen un Nova Group por debajo de 3.

Falso, Al agruparse, no se cumple esta condición en ningún contexto.



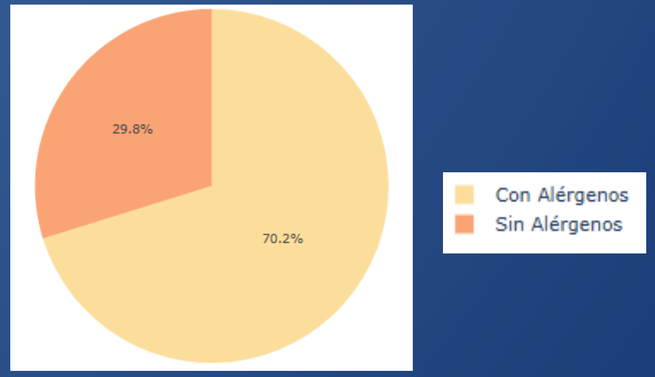
Proporción de Categorias con Nova Group < 3 Exceso de Azucares



Proporción de Categorias con Nova Group < 3 Exceso de G. Sat

HIPÓTESIS 6: Todos los productos con Nutri Score D y E poseen alergenos

Falso, Apenas el 70.2% de los productos con Nutri Score D y E tienen alergenos, lo que aún está lejos de crear una tendencia real hacia esta Hipótesis con el catálogo actual de productos:



Proporción de Productos con Nutri Score D o E y Alérgen

CONCLUSIONES INICIALES

- **Clasificación por Perfil Nutricional:** Confirmamos que los productos tienen perfiles nutricionales únicos y predecibles. Nuestro modelo de clasificación automática, usando sólo datos nutricionales, puede predecir con éxito la categoría de un producto (validando H1).

- **Diferencias Nutricionales entre Categorías:** Se observaron diferencias significativas en los perfiles nutricionales entre categorías. Por ejemplo, las bebidas tienen menos proteínas y grasas que los snacks o postres (validando H2 y H3).

- **Concentración de Azúcares y Grasas:** Identificamos las categorías con mayores concentraciones de azúcares y grasas, información crucial para la salud pública y patrones de consumo.

- **Impacto de la Normativa Argentina:** Un número considerable de productos excede los límites de azúcares y grasas saturadas según la Ley de Etiquetado Frontal de Argentina.

- **Calidad de los Datos:** Los datos de la API de Open Food Facts son adecuados para este análisis.

INSIGHTS CLAVE

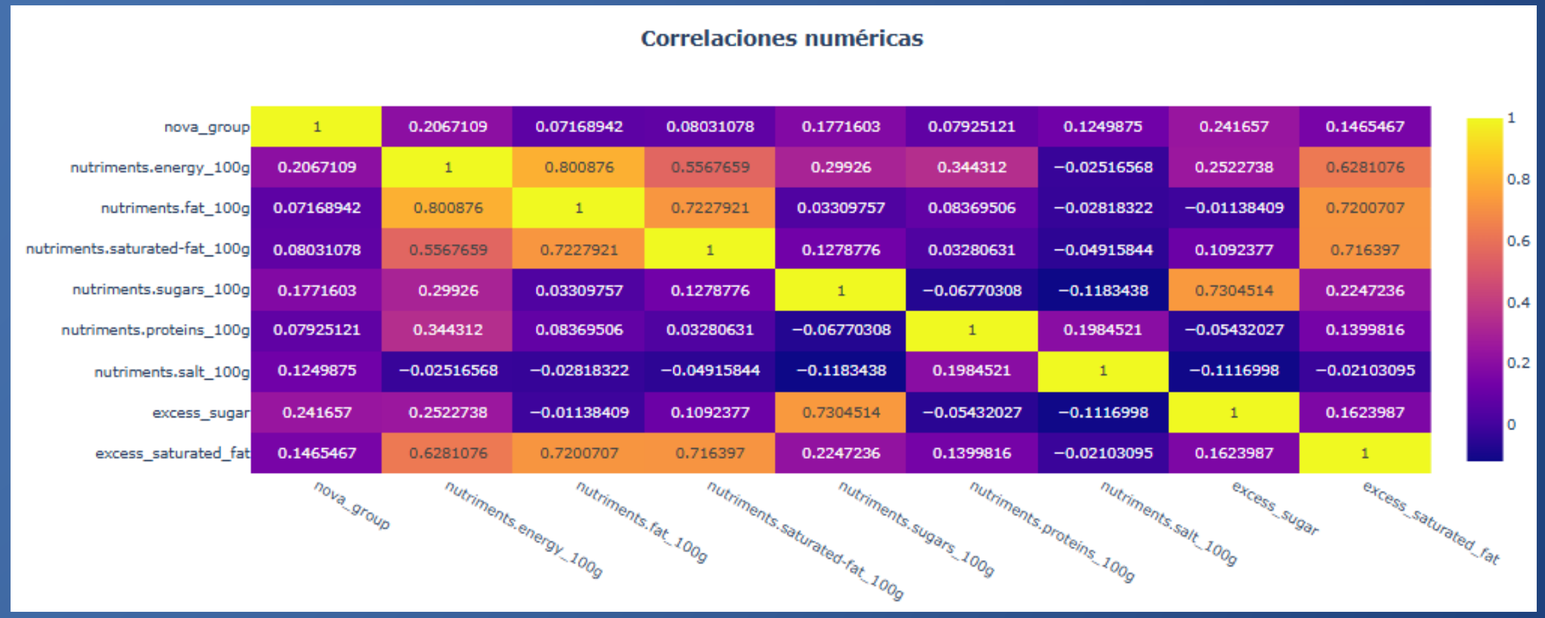
1. **Innovación Nutricional:** Hay una clara oportunidad para desarrollar versiones más saludables de productos, especialmente en categorías con altos niveles de azúcares y grasas como snacks y postres. Esto responde a la creciente demanda de opciones más sanas.

1. **Diversificación Estratégica:** Las categorías con perfiles nutricionales menos favorables son ideales para expandirse hacia líneas "light", "reducidas en" o "funcionales", alineándose con las preferencias del consumidor.

1. **Análisis por Marca:** Investigar la relación entre marcas y perfiles nutricionales podría revelar qué marcas tienen productos menos saludables o con más sellos de advertencia. Esto es útil para mejorar productos y analizar a la competencia.

1. **Futuras Investigaciones:** Los datos actuales son una base sólida para explorar hipótesis adicionales, como la conexión entre Nutri-Score, Nova Group y la presencia de alérgenos, aunque esto requiere un análisis más profundo.

MODELOS DE MACHINE LEARNING MÁS APROPIADOS



Al ver las mejores correlaciones a continuación, hemos decidido para este proyecto trabajar dos modelos:

- El primero desea predecir las **cantidades energéticas** basado en otras variables independientes de contenidos nutricionales, tal como lo vimos al contestar la pregunta de interés 2. Debido a que estamos hablando de una variable numérica, utilizaremos un modelo de Regresión Lineal Múltiple con el fin de poner a interactuar las variables mejor correlacionadas.

- El segundo modelo busca predecir si los productos tendrán o no **excesos de grasas saturadas** sin tener en cuenta el tipo de empaque. Dado que hablamos de un modelo de clasificación trabajaremos con varias opciones con el fin de escoger y optimizar la mejor opción.

MODELO DE REGRESIÓN LINEAL MULTIPLE

Para predecir la variable de contenido energético se implementó el siguiente conjunto de variables:

	nutriments.energy_100g	nutriments.fat_100g	nutriments.saturated-fat_100g	nutriments.sugars_100g	nutriments.proteins_100g	nutriments.salt_100g
0	2.0	0.0	0.0	1.4	0.0	0.000000
1	406.0	3.0	0.0	0.0	8.0	0.000000
2	0.0	0.0	0.0	0.0	0.0	0.065000
3	0.0	0.0	0.0	0.0	0.0	0.065000
4	0.0	0.0	0.0	0.0	0.0	0.000508

Filas y columnas:(711, 6)

Obtuvimos los siguientes coeficientes de regresión para cada variable independiente y predicciones con un 30% en el conjunto de prueba:

	Coefficient
nutriments.fat_100g	35.938497
nutriments.saturated-fat_100g	-12.194396
nutriments.sugars_100g	16.114688
nutriments.proteins_100g	43.288403
nutriments.salt_100g	-13.605071

	Actual	Predicted
694	2347.0	2448.715871
539	1725.0	2326.980767
289	1565.0	1023.812822
335	2180.0	2190.567246
595	248.0	450.895554
...	...	...
908	151.0	412.422663
522	1638.0	1147.727676
350	1711.0	1412.309665
188	946.0	1736.402104
36	0.0	260.633091

214 rows × 2 columns

MÉTRICAS DE REGRESIÓN LINEAL MÚLTIPLE

Realizamos los siguientes procedimiento para entrenar el modelo:

- **Métricas luego de entrenar el modelo:**

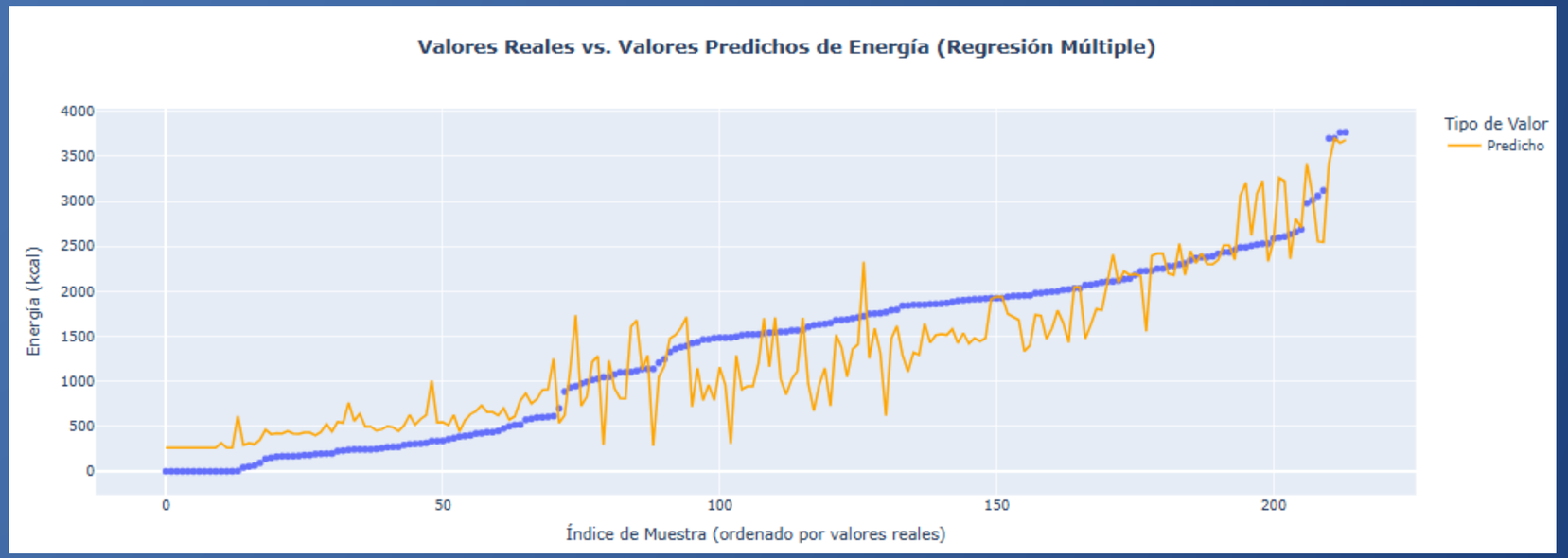
```
El resumen de la evaluación del modelo es el siguiente:
Mean Absolute Error: 316.6222803909272
Mean Squared Error: 146489.71046203267
Root Mean Squared Error: 382.7397424648147
R2 Score: 0.8256
```

El 82.6% de la variabilidad observada en los valores reales puede ser explicada por un modelo de regresión. Esto implica que las variables predictoras utilizadas en el modelo son muy efectivas para explicar las fluctuaciones en la variable que estás intentando predecir.

El 17.4% restante (100% - 82.6%) de la variabilidad no es explicada por el modelo y se atribuye a factores no incluidos en el modelo, al ruido o a la variabilidad inherente de los datos que el modelo no puede capturar.

un R2 de 0.826 es generalmente considerado un muy buen resultado.

MODELO DE REGRESIÓN LINEAL MÚLTIPLE



- **Validación Cruzada:** Al realizar el proceso de validación cruzada escogiendo 5 conjuntos diferentes de muestra al mismo porcentaje tenemos que:

```
Puntuaciones R2 de cada fold: [0.86245755 0.80789307 0.80150639 0.6960605  0.82460962]
Media de las puntuaciones R2: 0.7985
Desviación estándar de las puntuaciones R2: 0.0554
```

Revisando estos resultados notamos que el R2 inicial de 0.82 se obtuvo de una sola división de entrenamiento/prueba. La media de 0.7985 de la validación cruzada está muy cerca del R2 inicial lo cual es muy positivo para el modelo.

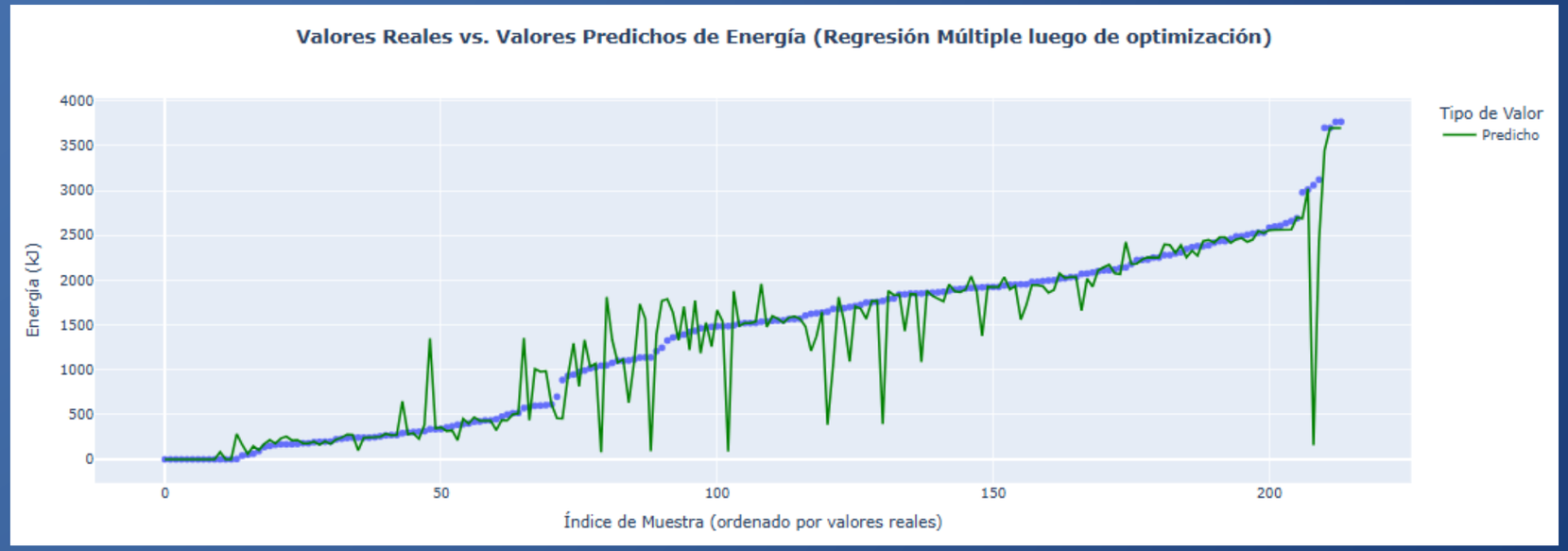
La validación cruzada, al promediar el rendimiento en múltiples subconjuntos de los datos, nos da una estimación mucho más robusta y confiable de cómo el modelo de regresión lineal múltiple se desempeñará con datos nuevos y no vistos. La cercanía entre ambos valores sugiere que este modelo es bastante consistente en su rendimiento.

MODELO DE REGRESIÓN LINEAL MÚLTIPLE

- **Optimización de Hiperparametros:** Al realizar este procedimiento utilizando el método Random Search tenemos los siguientes resultados finales mejorados.

```
El resumen luego de optimizar el modelo es el siguiente:
Mejores parámetros encontrados: {'learning_rate': np.float64(0.19615146512071296), 'max_depth': 7, 'min_samples_leaf': 1, 'min_samples_split': 14, 'n_estimators': 378
Mejor puntuación (neg_mean_squared_error) de validación cruzada: -79337.96723970698

Error Cuadrático Medio (MSE) en el conjunto de prueba: 117334.2951
Coeficiente de Determinación (R²) en el conjunto de prueba: 0.8603
```



El MSE en el conjunto de prueba es ligeramente mayor que el MSE de validación cruzada (121060 vs 78247). Esto es un comportamiento normal, ya que el modelo se está enfrentando a datos completamente nuevos. La diferencia no es drásticamente grande, lo que es una buena señal. El valor del R2 de 0.8559 (o 85.59%) es bastante alto. Esto indica que el modelo es capaz de explicar aproximadamente el 85.59% de la variabilidad en la variable objetivo del conjunto de prueba. En muchos contextos de regresión, un R2 por encima de 0.80 se considera muy bueno, sugiriendo que el modelo tiene una fuerte capacidad predictiva.

MODELO DE CLASIFICACIÓN

Para predecir la variable que determina exceso de grasas saturadas, se implementó el siguiente conjunto de variables mejor correlacionadas:

	excess_saturated_fat	nutriments.fat_100g	nutriments.saturated-fat_100g	nutriments.energy_100g	nutriments.sugars_100g	nutriments.proteins_100g	nutriments.salt_100g	nova_group
0	False	0.0	0.0	2.0	1.4	0.0	0.000000	-1
1	False	3.0	0.0	406.0	0.0	8.0	0.000000	3
2	False	0.0	0.0	0.0	0.0	0.0	0.065000	-1
3	False	0.0	0.0	0.0	0.0	0.0	0.065000	1
4	False	0.0	0.0	0.0	0.0	0.0	0.000508	-1

Implementamos los siguientes modelos para revisar cual cumple preliminarmente con las mejores métricas: **regresión logistica, kmeans, árbol de decisión y SVM.**

NOTA: K-Means es un algoritmo de clustering (aprendizaje no supervisado), por lo que se usó un enfoque diferente para este algoritmo de clustering.

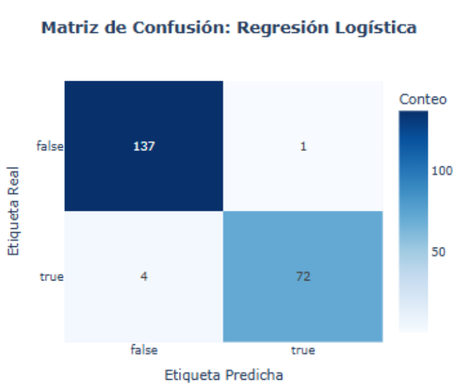
En esta etapa escogimos la Regresión Logistica y el SVM como modelos a la siguiente etapa. Ambos modelos muestran un rendimiento muy sólido. Sus métricas de Accuracy, Precision, Recall y F1-Score están por encima del 96-97%, lo que indica que clasifican muy bien tanto las clases positivas como las negativas, con pocos errores de cualquier tipo.

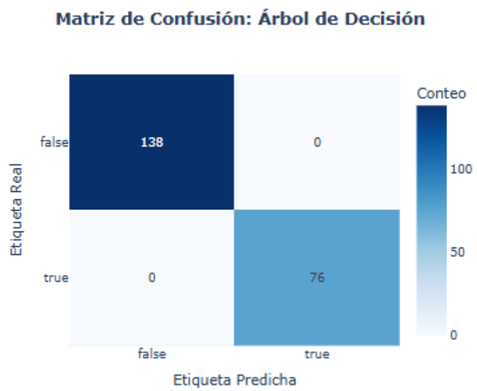
MÉTRICAS DE CLASIFICACIÓN

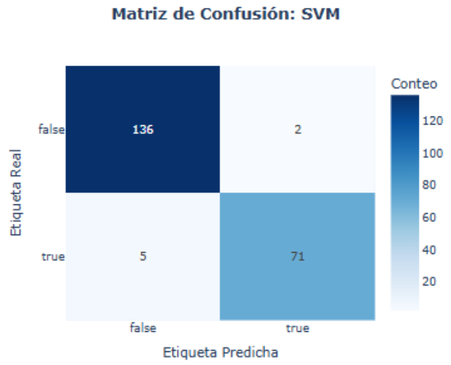
A continuación se entregan resultados que demuestran la preselección de los modelos para validación.:

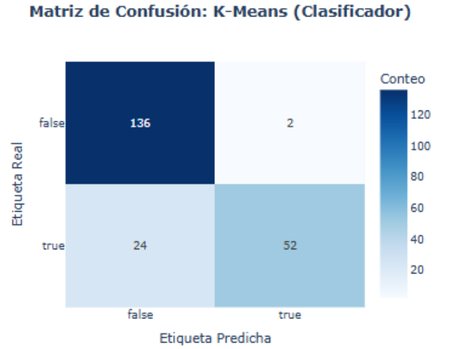
- **Métricas luego de entrenar los modelos:**

```
--- Resumen de Métricas de Clasificación ---
------------------------------------------------------------------
Modelo                    | Accuracy   | Precision  | Recall     | F1-Score
------------------------------------------------------------------
Regresión Logística       | 0.9766     | 0.9768     | 0.9766     | 0.9765
Árbol de Decisión         | 1.0000     | 1.0000     | 1.0000     | 1.0000
SVM                       | 0.9673     | 0.9674     | 0.9673     | 0.9671
K-Means (Clasificador)    | 0.8785     | 0.8901     | 0.8785     | 0.8727
------------------------------------------------------------------
```









MODELO DE CLASIFICACIÓN

- **Validación Cruzada de los modelos:** mantuvimos el método k-fold para 5 muestras por ser de una buena recomendación. Luego de hacer dicha validación estos fueron los resultados:

```
--- Resumen de Resultados de Validación Cruzada (K=5) ---
-------------------------------------------------------------------
Modelo                    | Accuracy   | Precision  | Recall     | F1-Score
-------------------------------------------------------------------
Regresión Logística       | 0.9733     | 0.9739     | 0.9733     | 0.9731
SVM                       | 0.9606     | 0.9615     | 0.9606     | 0.9605
-------------------------------------------------------------------
```

Por tanto, para fines practicos, vamos a usar el modelo de regresión logística durante el procedimiento de ensamble. Sin embargo, el modelo ya tiene un puntaje bastante alto por lo que no se considera este ensamble.

- **Regresión Logística (individual, sin Bagging):** En la validación cruzada, obtuvimos un Accuracy promedio de 0.9719.
- **Optimización y ensamble:** BaggingClassifier con Regresión Logística Optimizada nos llevará a un Accuracy de 1.0000 en el conjunto de prueba. Esto sugiere que el Bagging (y potencialmente la optimización de hiperparámetros) ha logrado mejorar significativamente el rendimiento del modelo en este conjunto de prueba específico, llevándolo a la perfección.

Podemos seguir trabajando con el modelo optimizado sin ensamble para poder garantizar una variabilidad en los datos predichos, pero por ser solo dos opciones de respuesta de tipo Booleano, la regresión logística sigue siendo una mejor opción de las predicciones.

CONCLUSIONES FINALES

Basado en el análisis exploratorio de datos, la validación de hipótesis y la experimentación con modelos de Machine Learning realizados en este proyecto, podemos establecer las siguientes conclusiones finales:

1. **Distintividad de Perfiles Nutricionales por Categoría:** El análisis exploratorio confirmó de manera robusta que las diferentes categorías de productos alimenticios poseen perfiles nutricionales significativamente distintos. Las visualizaciones de la distribución y los promedios de nutrientes como energía, grasas, azúcares y proteínas mostraron patrones claros y diferenciados para categorías como bebidas, snacks y lácteos. Esta distinción es fundamental y valida la hipótesis H3, que postula que las categorías presentan perfiles nutricionales distintos y predecibles.
2. **Viabilidad de la Clasificación Automática:** La capacidad de predecir la categoría de un producto utilizando únicamente sus valores nutricionales fue demostrada como plausible durante la fase de análisis exploratorio. La clara separación de los perfiles nutricionales observada en el EDA sugiere que modelos de clasificación automática.

3. Validación de Diferencias Nutricionales Específicas: La hipótesis H2, que sugería que las bebidas tendrían menor contenido de proteínas y grasas que los snacks o postres, fue validada por el análisis de los promedios de nutrientes por categoría y las visualizaciones comparativas. Esto confirma una diferencia nutricional específica y relevante entre estos grupos de productos.

4. Identificación de Productos con Exceso de Nutrientes Críticos: La creación de las columnas excess_sugar y excess_saturated_fat, basadas en la normativa argentina de etiquetado frontal, permitió cuantificar y visualizar los productos que superan los umbrales de azúcares y grasas saturadas. Esto identificó claramente qué tipos de productos tienden a concentrar mayores cantidades de estos nutrientes críticos, respondiendo a una de las preguntas de interés clave del proyecto.

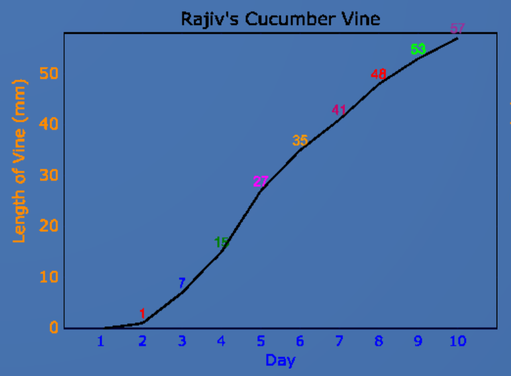
CONCLUSIONES FINALES

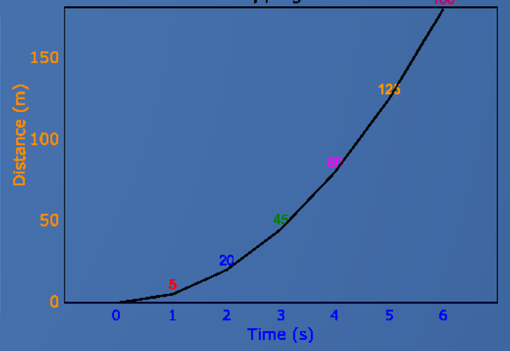
5. Potencial para Innovación y Diversificación de Productos: El análisis de los perfiles nutricionales y la identificación de excesos de nutrientes revelaron oportunidades significativas para la industria alimentaria. Las categorías con altos niveles de azúcares y grasas saturadas son candidatas ideales para el desarrollo de líneas "light" o reformulaciones que mejoren su perfil nutricional. Asimismo, existe potencial para diversificar la oferta con productos "funcionales" que incorporan ingredientes beneficiosos para la salud, particularmente en categorías como snacks, postres, bebidas y lácteos.

6. Evaluación de Hipótesis Adicionales: Las hipótesis H4 (sobre la marca Sidi Ali) y H5 (sobre la relación entre productos saludables, categorías específicas y Nova Group < 3) fueron refutadas por el análisis detallado. Esto demuestra la importancia de validar las suposiciones iniciales con datos concretos y ajusta la comprensión de las relaciones entre marcas, procesamiento de alimentos y perfiles nutricionales en este conjunto de datos. La hipótesis H6 (sobre la relación entre Nutri Score D/E y alérgenos) también fue refutada, indicando que no todos los productos con peor Nutri-Score contienen alérgenos.

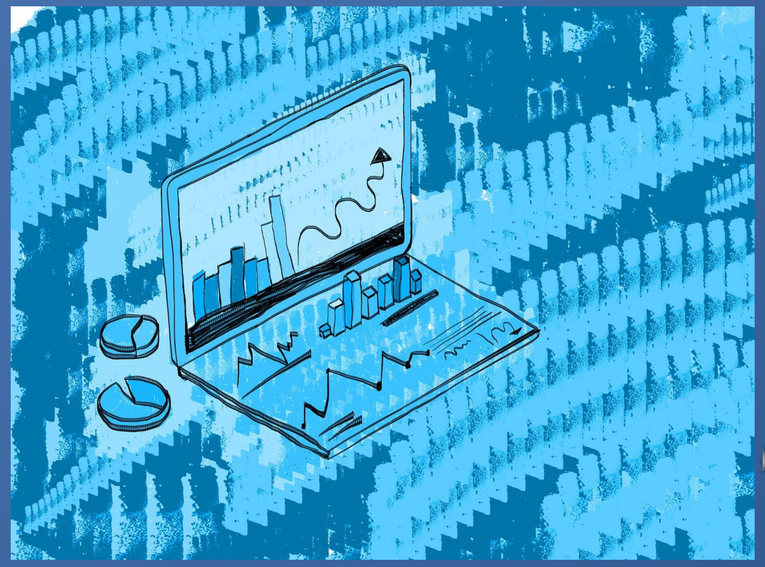
7. Selección y Evaluación de Modelos de Machine Learning: Se exploraron y evaluaron modelos de regresión y clasificación para abordar diferentes aspectos predictivos del proyecto:

- El Modelo de Regresión Lineal Múltiple para predecir niveles energéticos por 100 gr mostró un rendimiento sólido (R^2 ~0.86 después de la optimización), demostrando que la energía de un producto puede ser predicha con alta precisión a partir de otros nutrientes. La optimización con RandomizedSearchCV mejoró ligeramente el modelo base.
- Para el problema de Clasificación Binaria (predicción de exceso de grasas saturadas), se compararon varios algoritmos. La Regresión Logística y SVM mostraron un rendimiento consistentemente alto (Accuracy > 96%) en la validación cruzada. El modelo de Árbol de Decisión obtuvo métricas perfectas en el conjunto de prueba inicial, lo cual fue identificado como un posible indicio de sobreajuste, descartándolo para la implementación principal. El BaggingClassifier con la Regresión Logística optimizada alcanzó un rendimiento perfecto en el conjunto de prueba (Accuracy = 1.0000), validando la efectividad del ensamble en este caso particular, aunque se reconoció la posible simplicidad del problema de clasificación binaria debido a la alta correlación lineal entre la grasa saturada y su "exceso" definido normativamente.


Rajiv's Cucumber Vine
Length of Vine (mm)
Day
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
1
7
16
27
35
41
48
53
57


Dropping Ball
Distance (m)
Time (s)
0
50
100
150
0
1
2
3
4
5
6
5
20
45
80
125
180


A+




Thank You
ありがとう。
TACK
grazie
Bedankt
Danke Schön
merci
MERCI
OBRIGADO
gracias
GRAZIE
Tack
תודה
Merci
bedankt
GRACIAS
DANKE SCHÖN
obrigado
Gracias
thank you
BEDANKT
Bedankt
danke schön
شكراً
Grazie
Obrigado
tack
Merci
gracias
THANK YOU
Danke Schön
TACK
Grazie
OBRIGADO
MERCI
bedankt
obrigado
merci
gracias
Thank You
Bedankt
Danke Schön