Curso Data Science: Proyecto Final

# **Modelo analítico de Machine Learning aplicado a la nutrición**



Integrantes:

* Flavio Galimberti
* Felipe Quesada
* Francisco Macchi
* Giuliana Zeolla

Fecha de entrega: 03 de Octubre de 2022

Índice de Referencia

[**Modelo analítico de Machine Learning aplicado a la nutrición**](#_heading=h.gktqypo5u1b) **1**

[**Descripción del caso de negocio**](#_heading=h.gjdgxs) **3**

[**¿Quiénes somos?**](#_heading=h.30j0zll) **3**

[**Problema abordado**](#_heading=h.1fob9te) **3**

[**Objetivos del modelo**](#_heading=h.2et92p0) **3**

[**Descripción de los datos**](#_heading=h.tyjcwt) **3**

[**Variables**](#_heading=h.3dy6vkm) **3**

[**Diccionario de datos**](#_heading=h.1t3h5sf) **6**

[**Pre procesamiento de datos**](#_heading=h.4d34og8) **15**

[**Hallazgos encontrados por el EDA**](#_heading=h.2s8eyo1) **16**

[**Análisis Univariado**](#_heading=h.17dp8vu) **16**

[**Análisis Bivariado**](#_heading=h.3rdcrjn) **18**

[**Análisis Multivariado**](#_heading=h.26in1rg) **21**

[**Algoritmo elegido**](#_heading=h.lnxbz9) **25**

[**Métricas de desempeño del Modelo**](#_heading=h.35nkun2) **25**

[**Iteraciones de optimización**](#_heading=h.44sinio) **25**

[**Futuras líneas**](#_heading=h.2jxsxqh) **26**

[**Conclusiones**](#_heading=h.z337ya) **26**

## **Descripción del caso de negocio**

### **¿Quiénes somos?**

Somos una consultora dedicada a ayudar a resolver los desafíos más complejos de las organizaciones utilizando el poder de la ciencia de datos.

Nos consideramos un equipo fan de la ciencia de datos y guiado por el propósito esta poderosa herramienta es accesible para todos.

### **Problema abordado**

Nos contrató un equipo de nutricionistas para diseñar un algoritmo que prediga si un alimento corresponde o no al grupo de las proteínas. Nuestro cliente trabaja con deportistas de alto rendimiento a nivel mundial, y buscan que sus pacientes tengan una aplicación que les permita determinar, al ingresar la tabla nutricional del alimento, si el mismo corresponde al grupo de las proteínas.

## **Objetivos del modelo**

El objetivo principal es crear un modelo de maching learnig que prediga al ingresar los valores de la tabla nutricional de un alimento si el mismo corresponde o no al grupo de las proteínas.

Los objetivos secundarios son:

* Ayudar a aquellas personas que quieran iniciar un estilo de vida más saludable y no tengan conocimientos de nutrición.
* Lograr que cualquier persona pueda determinar si el macro-nutriente predominante de un alimento es proteínas más allá de lo que indique la industria (en el paquete).

## **Descripción de los datos**

Trabajamos con un dataset de nutrición con una gran base de alimentos y sus correspondientes valores nutricionales.

Link : <https://www.kaggle.com/datasets/sathyakrishnan12/nutrition-datasets>

### **Variables**

* Name: descripción del alimento de la lista.
* serving\_size: Porción medida en gramos del alimento. La misma es siempre de 100gr.
* total\_fat: Gramos de grasas que contiene el alimento por cada porción (100 gr).
* saturated\_fat: Gramos de grasas saturadas que contiene el alimento por cada porción (100 gr).
* Colesterol: Miligramos de colesterol que contiene el alimento por cada porción (100 gr).
* Sodium: Miligramos de sodio que contiene el alimento por cada porción (100 gr).
* Choline: Miligramos de colina que contiene el alimento por cada porción (100 gr).
* Folate: Microgramos de folato que contiene el alimento por cada porción (100 gr).
* folic\_acid: Microgramos de ácido fólico que contiene el alimento por cada porción (100 gr).
* Niacin: Miligramos de niacina que contiene el alimento por cada porción (100 gr).
* pantothenic\_acid: Miligramos de ácido pantoténico que contiene el alimento por cada porción (100 gr).
* Riboflavin: Miligramos de riboflavina que contiene el alimento por cada porción (100 gr).
* Thiamin: Miligramos de tiamina que contiene el alimento por cada porción (100 gr).
* vitamin\_a: Cantidad de vitamina A expresada en unidad internacional que contiene el alimento por cada porción (100 gr).
* vitamin\_a\_rae: Microgramos de vitamina A que contiene el alimento por cada porción (100 gr).
* carotene\_alpha: Microgramos de caroteno alfa que contiene el alimento por cada porción (100 gr).
* carotene\_beta: Microgramos de caroteno beta que contiene el alimento por cada porción (100 gr).
* cryptoxanthin\_beta: Microgramos de criptoxantina beta que contiene el alimento por cada porción (100 gr).
* lutein\_zeaxanthin: Microgramos de luteína zeaxantina que contiene el alimento por cada porción (100 gr).
* Lucopene: Cantidad de licopeno que contiene el alimento por cada porción (100 gr). El valor es para todos los registros igual a 0.
* vitamin\_b12: Microgramos de vitamina B12 que contiene el alimento por cada porción (100 gr).
* vitamin\_b6: Miligramos de vitamina B6 que contiene el alimento por cada porción (100 gr).
* vitamin\_c: Miligramos de vitamina C que contiene el alimento por cada porción (100 gr).
* vitamin\_d: Cantidad de vitamina D expresada en unidad internacional que contiene el alimento por cada porción (100 gr).
* vitamin\_e: Miligramos de vitamina E que contiene el alimento por cada porción (100 gr).
* tocopherol\_alpha: Miligramos de tocoferol alfa que contiene el alimento por cada porción (100 gr).
* vitamin\_k: Microgramos de vitamina K que contiene el alimento por cada porción (100 gr).
* Calcium: Miligramos de calcio que contiene el alimento por cada porción (100 gr).
* Copper: Miligramos de cobre que contiene el alimento por cada porción (100 gr).
* Irom: Miligramos de hierro que contiene el alimento por cada porción (100 gr).
* Magnesium: Miligramos de magnesio que contiene el alimento por cada porción (100 gr).
* Manganese: Miligramos de manganeso que contiene el alimento por cada porción (100 gr).
* Phosphorous: Miligramos de fosforoso que contiene el alimento por cada porción (100 gr).
* Potassium: Miligramos de potasio que contiene el alimento por cada porción (100 gr).
* Selenium: Microgramos de selenio que contiene el alimento por cada porción (100 gr).
* Zink: Miligramos de zinc que contiene el alimento por cada porción (100 gr).
* Protein: Gramos de proteinas que contiene el alimento por cada porción (100 gr).
* Alanine: Gramos de alanina que contiene el alimento por cada porción (100 gr).
* Arginine: Gramos de arginina que contiene el alimento por cada porción (100 gr).
* aspartic\_acid: Gramos de ácido aspártico que contiene el alimento por cada porción (100 gr).
* Cystine: Gramos de cistina que contiene el alimento por cada porción (100 gr).
* glutamic\_acid: Gramos de ácido glutamico que contiene el alimento por cada porción (100 gr).
* Glycine: Gramos de glicina que contiene el alimento por cada porción (100 gr).
* Histidine: Gramos de histidina que contiene el alimento por cada porción (100 gr).
* Hydroxyproline: Gramos de hidroxiprolina que contiene el alimento por cada porción (100 gr).
* Isoleucine: Gramos de isoleucina que contiene el alimento por cada porción (100 gr).
* Leucine: Gramos de leucina que contiene el alimento por cada porción (100 gr).
* Lysine: Gramos de lisina que contiene el alimento por cada porción (100 gr).
* Methionine: Gramos de metionina que contiene el alimento por cada porción (100 gr).
* Phenylalanine: Gramos de fenilalanina que contiene el alimento por cada porción (100 gr).
* Proline: Gramos de prolina que contiene el alimento por cada porción (100 gr).
* Serine: Gramos de serina que contiene el alimento por cada porción (100 gr).
* Threonine: Gramos de treonina que contiene el alimento por cada porción (100 gr).
* Tryptophan: Gramos de triptófano que contiene el alimento por cada porción (100 gr).
* Tyrosine: Gramos de tirosina que contiene el alimento por cada porción (100 gr).
* Valine: Gramos de valina que contiene el alimento por cada porción (100 gr).
* Carbohydrate: Gramos de carbohidratos que contiene el alimento por cada porción (100 gr).
* Fiber: Gramos de fibras que contiene el alimento por cada porción (100 gr).
* Sugars: Gramos de azúcares que contiene el alimento por cada porción (100 gr).
* Fructose: Gramos de fructosa que contiene el alimento por cada porción (100 gr).
* Galactose: Gramos de galactosa que contiene el alimento por cada porción (100 gr).
* Glucose: Gramos de glucosa que contiene el alimento por cada porción (100 gr).
* Lactose: Gramos de lactosa que contiene el alimento por cada porción (100 gr).
* Maltose: Gramos de maltosa que contiene el alimento por cada porción (100 gr).
* Sucrose: Gramos de sacarosa que contiene el alimento por cada porción (100 gr).
* Fat: Gramos de grasas que contiene el alimento por cada porción (100 gr).
* saturated\_fatty\_acids: Gramos de Ácidos grasos saturados que contiene el alimento por cada porción (100 gr).
* monounsaturated\_fatty\_acids: Gramos de Ácidos grasos monoinsaturados que contiene el alimento por cada porción (100 gr).
* polyunsaturated\_fatty\_acids: Gramos de Ácidos grasos poliinsaturados que contiene el alimento por cada porción (100 gr).
* fatty\_acids\_total\_trans: Miligramos de ácidos grasos trans totales que contiene el alimento por cada porción (100 gr).
* Alcohol: Gramos de alcohol que contiene el alimento por cada porción (100 gr).
* Ash: Gramos de ceniza que contiene el alimento por cada porción (100 gr).
* Caffeine: Miligramos de cafeína que contiene el alimento por cada porción (100 gr).
* Theobromine: Miligramos de teobromina que contiene el alimento por cada porción (100 gr).
* Wáter: Gramos de agua que contiene el alimento por cada porción (100 gr).

### **Diccionario de datos**

**calories (calorías):** Representan la energía que se encuentra en los alimentos. El cuerpo tiene una demanda constante de energía y utiliza las calorías de los alimentos para seguir funcionando.

*Vitaminas*

1. **Liposolubles:** Las vitaminas liposolubles, A, D, E y K, se consumen junto con alimentos que contienen grasa. Son las que se disuelven en grasas y aceites. Se almacenan en el hígado y en los tejidos grasos.
2. **vitamin A (vitamina A):** es una vitamina liposoluble que se encuentra naturalmente presente en los alimentos. La vitamina A es importante para la visión normal, el sistema inmunitario y la reproducción. Además, la vitamina A ayuda al buen funcionamiento del corazón, los pulmones, los riñones y otros órganos. Existen dos tipos diferentes de vitamina A. El primer tipo, la vitamina A preformada (retinol), se encuentra en la carne vacuna, carne de ave, pescado y productos lácteos. El segundo tipo, la provitamina A (carotenoides), se encuentra en frutas, verduras y otros productos de origen vegetal. El tipo más común de provitamina A presente en los alimentos.
3. **vitamin A RAE (vitamina A preformada):** vitamina A preformada (retinol), se encuentra en la carne vacuna, carne de ave, pescado y productos lácteos. RAE por sus siglas en inglés es “Retinol Activity Equivalents”.
4. **carotene alpha (carotenoides alfa):** Provitamina A (carotenoides). Su actividad provitamínica es inferior a la del beta-caroteno. Sin embargo, su capacidad antioxidante es mayor.
5. **carotene beta (carotenoides beta):** Provitamina A (carotenoides). Está presente en las frutas, verduras y granos.
6. **cryptoxanthin beta (criptoxantina beta):** Provitamina A (carotenoides). Es un importante carotenoide provitamina A en la dieta, que se encuentra principalmente en frutas y verduras.
7. **lutein\_zeaxanthin (luteina y Zeaxantina):** Provitamina A (carotenoides). Están relacionadas con el betacaroteno y la vitamina A. Los alimentos ricos en luteína incluyen la yema de huevo, el brócoli, la espinaca, el pimiento anaranjado, el kiwi, las uvas, el jugo de naranja, las calabacitas y el zapallo. La absorción de la luteína es mayor cuando se toma con una comida alta en grasas. La zeaxantina está compuesta por los mismos elementos, y en las mismas proporciones que la luteína. Difiere en algunas propiedades a causa de una diferencia en la estructura molecular.
8. **lycopene (licopeno):** Provitamina A (carotenoides). Es una sustancia química que existe en forma natural y que es la responsable del color rojo de las frutas y verduras. El licopeno se encuentra en tomates, sandias, pomelos rosados, albaricoques, rosa mosqueta y guayabas.
9. **vitamin D (vitamina D):** es un nutriente necesario para la salud. Ayuda al cuerpo a absorber el calcio, una de las principales sustancias necesarias para tener huesos fuertes. Junto con el calcio, la vitamina D contribuye a prevenir la osteoporosis, una enfermedad que hace que los huesos se vuelvan más delgados y débiles y sean más propensos a fracturas. Además, al cuerpo le hace falta la vitamina D para otras funciones. Los músculos la necesitan para el movimiento y los nervios para transmitir mensajes entre el cerebro y otras partes del cuerpo. La vitamina D es indispensable para que el sistema inmunitario pueda combatir las bacterias y los virus que lo atacan.
10. **vitamin D (vitamina D):** es un nutriente necesario para la salud. Ayuda al cuerpo a absorber el calcio, una de las principales sustancias necesarias para tener huesos fuertes. Junto con el calcio, la vitamina D contribuye a prevenir la osteoporosis, una enfermedad que hace que los huesos se vuelvan más delgados y débiles y sean más propensos a fracturas. Además, al cuerpo le hace falta la vitamina D para otras funciones. Los músculos la necesitan para el movimiento y los nervios para transmitir mensajes entre el cerebro y otras partes del cuerpo. La vitamina D es indispensable para que el sistema inmunitario pueda combatir las bacterias y los virus que lo atacan.
11. **tocopherol Alpha (tocoferol alfa) =** vitamina E vitamin K (vitamina K): es un nutriente que el cuerpo necesita para estar sano. Es importante para la coagulación de la sangre y la salud de los huesos, y para otras funciones del cuerpo.
12. **Hidrosolubles:** Las vitaminas hidrosolubles son aquellas que se disuelven en agua.
13. **folate (folato):** es una vitamina B que se encuentra naturalmente presente en muchos alimentos. El organismo necesita folato para producir ADN y otros tipos de material genético. El folato también es necesario para la división celular en el organismo. Una forma de folato, llamada ácido fólico, se usa en los alimentos fortificados y en la mayoría de los suplementos dietéticos.
14. **folic acid (ácido fólico):** El ácido fólico es una vitamina B. Ayuda al organismo a crear células nuevas. choline (colina): es un nutriente que se encuentra en muchos alimentos. El cerebro y el sistema nervioso la necesitan para regular la memoria y el estado de ánimo, para el control muscular y otras funciones. También necesita colina para formar las membranas que rodean las células del organismo. El hígado puede producir una pequeña cantidad de colina, pero la mayor parte de la colina en el organismo proviene de los alimentos que se consumen.
15. **niacin (niacina):** también conocida como vitamina B3, ayuda a convertir los alimentos que se consumen en la energía que se necesita. La niacina es importante para el desarrollo y la función de las células en el organismo. pantothenic acid (ácido pantoténico): también llamado vitamina B5, ayuda a convertir los alimentos que consumimos en la energía que necesitamos. Es importante para muchas funciones en el organismo, especialmente la fabricación y descomposición de las grasas.
16. **riboflavin (riboflavina):** conocida también como la vitamina B2, es importante para el crecimiento, desarrollo y funcionamiento de las células del organismo. La riboflavina ayuda a convertir los alimentos que consume en energía que necesita.
17. **thiamin (tiamina):** conocida también como la vitamina B1, ayuda a convertir los alimentos que consume en energía que necesita. La tiamina es importante para el crecimiento, desarrollo y funcionamiento de las células del organismo.
18. **vitamin B12 (vitamina B12):** también conocida como cianocobalamina, es un nutriente que ayuda a mantener sanas las neuronas y los glóbulos sanguíneos. Además, contribuye a la elaboración del ADN, el material genético presente en todas las células. También previene un tipo de anemia, denominada anemia megaloblástica, que provoca cansancio y debilidad en las personas.
19. **vitamin B6 (vitamina B6):** también llamada piridoxina, se encuentra naturalmente presente en muchos alimentos. La vitamina B6 es esencial para que funcionen bien las enzimas (proteínas que regulan los procesos químicos del cuerpo). La vitamina B6 también influye en el desarrollo cerebral durante el embarazo y la infancia, al igual que el sistema inmunitario.
20. **vitamin C (vitamina C):** conocida como ácido ascórbico, es un nutriente hidrosoluble que se encuentra en ciertos alimentos. En el cuerpo, actúa como antioxidante, al ayudar a proteger las células contra los daños causados por los radicales libres. Los radicales libres son compuestos que se forman cuando el cuerpo convierte los alimentos que consumimos en energía. Las personas también están expuestas a los radicales libres presentes en el ambiente por el humo del cigarrillo, la contaminación del aire y la radiación solar ultravioleta. Además, el cuerpo necesita vitamina C para producir colágeno, una proteína necesaria para la cicatrización de las heridas. La vitamina C también mejora la absorción del hierro presente en los alimentos de origen vegetal y contribuye al buen funcionamiento del sistema inmunitario para proteger al cuerpo contra las enfermedades.

**Minerales**

1. **calcium (calcio):** es un mineral presente en muchos alimentos. El cuerpo necesita el calcio para mantener los huesos fuertes y llevar a cabo muchas funciones importantes. Casi todo el calcio se almacena en los huesos y los dientes, donde apoya su estructura y rigidez. El cuerpo también necesita calcio para que los músculos se muevan y los nervios transmitan mensajes del cerebro a distintas partes del cuerpo. Además, el calcio ayuda a que la sangre circule a través de los vasos sanguíneos por todo el cuerpo y a liberar hormonas y enzimas que influyen casi todas las funciones del cuerpo.
2. **copper (cobre):** es un mineral que necesitamos para mantenernos sanos. El organismo usa cobre para llevar a cabo muchas funciones importantes, como producir energía, tejidos conectivos y vasos sanguíneos. El cobre también ayuda a mantener el sistema nervioso y el sistema inmunitario y activa los genes. El organismo también necesita cobre para el desarrollo del cerebro.
3. **irom (hierro):** es un mineral necesario para el crecimiento y desarrollo del cuerpo. El cuerpo utiliza el hierro para fabricar la hemoglobina, una proteína de los glóbulos rojos que transporta el oxígeno de los pulmones a distintas partes del cuerpo, y la mioglobina, una proteína que provee oxígeno a los músculos. El cuerpo también requiere hierro para elaborar hormonas y tejido conectivo.
4. **magnesium (magnesio):** es un nutriente que el cuerpo necesita para mantenerse sano. El magnesio es importante para muchos procesos que realiza el cuerpo. Por ejemplo, regula la función de los músculos y el sistema nervioso, los niveles de azúcar en la sangre, y la presión sanguínea. Además, ayuda a formar proteína, masa ósea y ADN (el material genético presente en las células).
5. **manganese (manganeso):** es un mineral que necesita el organismo para mantenerse sano. El organismo usa el manganeso para producir energía y proteger las células. El organismo también necesita manganeso para fortalecer los huesos, para la reproducción, la coagulación sanguínea y para mantener un sistema inmunitario sano.
6. **phosphorus (fosforo):** es un mineral que se encuentra en cada una de las células de nuestro organismo. La mayor parte del fósforo está en los huesos y los dientes, y otra parte en los genes. El organismo necesita fósforo para producir energía y llevar a cabo muchos procesos químicos importantes.
7. **potassium (potasio):** un mineral que se encuentra en muchos alimentos. El organismo necesita potasio para casi todo su funcionamiento, incluso para el buen funcionamiento del riñón y del corazón, la contracción muscular y la transmisión nerviosa.
8. **selenium (selenio):** es un nutriente que el cuerpo necesita para mantenerse sano. El selenio es importante para la reproducción, la función de la glándula tiroidea, la producción de ADN y para proteger al cuerpo contra infecciones y el daño causado por los radicales libres.
9. **zink (zinc):** es un nutriente que las personas necesitan para estar sanas. El zinc se encuentra en las células de todo el cuerpo. Ayuda al sistema inmunitario a combatir bacterias y virus que invaden al cuerpo. El cuerpo también necesita zinc para fabricar proteínas y el ADN, el material genético presente en todas las células. Durante el embarazo, la infancia y la niñez, el organismo requiere zinc para crecer y desarrollarse bien. El zinc también favorece la cicatrización de las heridas y el funcionamiento normal del sentido del gusto y el olfato.
10. **sodium (sodio):** El cuerpo necesita una determinada cantidad de sodio para funcionar adecuadamente, pues ayuda en la función de los nervios y los músculos. El sodio también ayuda a mantener un equilibrio adecuado de los líquidos. Los riñones controlan la cantidad de sodio presente en el cuerpo; si tiene demasiado y los riñones no pueden eliminarlo, se acumula en la sangre.

**Proteinas**

Las proteínas son los pilares fundamentales de la vida. Cada célula del cuerpo humano las contiene. La estructura básica de la proteína es una cadena de aminoácidos. Es necesario consumir proteínas en la dieta para ayudarle al cuerpo a reparar células y producir células nuevas. El cuerpo humano necesita un gran número de aminoácidos en cantidades suficientemente grandes para mantener una buena salud. Los aminoácidos se encuentran en fuentes animales tales como las carnes, la leche, el pescado y los huevos. También se encuentran en fuentes vegetales como la soja (soya), los fríjoles, las legumbres, la mantequilla de nueces y algunos granos (como el germen de trigo y la quínoa).

1. **Aminoácidos esenciales:** no los puede producir el cuerpo y deben ser proporcionados por los alimentos.
2. **aminoácidos esenciales ramificados (BCAA):** los aminoácidos esenciales ramificados (BCAA) son nutrientes esenciales que el cuerpo obtiene de proteínas ingeridas en la comida. En especial carnes, lácteos y legumbres. “Ramificados” se refiere a la estructura química de los aminoácidos. La combinación de estos tres aminoácidos esenciales (valina, isoleucina y leucina) compone casi la tercera parte de los músculos esqueléticos en el cuerpo humano y desempeñan un papel muy importante en la síntesis de proteínas.
3. **valine (valina):** aminoácido esencial BCAA.
4. **isoleucine (isoleucina):** aminoácido esencial BCAA.
5. **leucine (leucina):** aminoácido esencial BCAA arginine (arginina): aminoácido esencial en los alimentos que contienen proteínas, como la carne, el pollo, el pescado, los huevos, los productos lácteos y las legumbres. Una dieta balanceada provee cerca de 4 a 5 gramos al día.
6. **Lysine (lisina):** es un aminoácido esencial componente de las proteínas sintetizadas por los seres vivos.
7. **methionine (metionina):** es encontrada en carne, pescado y lácteos.
8. **phenylalanine (Fenilalanina):** es uno de los aminoácidos esenciales que utilizan las células para sintetizar proteínas.
9. **threonine (treonina):** es uno de los aminoácidos esenciales que utilizan las células para sintetizar proteínas.
10. **tryptophan (triptófano):** El L-triptófano es un aminoácido esencial. El L-triptófano se ingiere como parte de la dieta y puede estar presente en los alimentos con proteína.
11. **Histidine (histidine):** aminoácido esencial.
12. **Aminoácidos no esenciales:** son producidos por el cuerpo a partir de los aminoácidos esenciales o en la descomposición normal de las proteínas.
13. **alanine (alanina):** Beta-alanina es un aminoácido no esencial que se encuentra en alimentos como la carne, el pollo y el pescado. Las personas consumen cerca de 1 gramo de beta-alanina al día, dependiendo de su dieta. El organismo utiliza beta-alanina para producir carnosina en los músculos esqueléticos.
14. **aspartic acid (ácido aspartico):** Es uno de los aminoácidos no esenciales. Ayuda a que cada célula del cuerpo funcione, a la producción y liberación de hormonas y al funcionamiento normal del sistema nervioso.
15. **Cystine (cistina):** aminoacido no esencial importante para el desarrollo del tejido conectivo.
16. **glutamic acid (ácido glutamico):** La glutamina y el ácido glutámico son dos aminoácidos no esenciales que están estrechamente relacionados de forma química y son interconvertibles, lo que significa que cada uno puede convertirse en el otro por sí solos según los requerimientos del organismo. Ambos compuestos tienen una cadena de base similar de moléculas.
17. **hydroxyproline ( hidroxiprolina):** aminoácido no esencial.
18. **proline (prolina):** aminoácido no esencial.
19. **Serine (serina):** aminoácido no esencial.
20. **Tyrosine (tirosina):** aminoácido no esencial glicina.

**Carbohidratos**

Son moléculas de azúcar. Junto con las proteínas y las grasas, los carbohidratos son uno de los tres nutrientes principales que se encuentran en alimentos y bebidas. Su cuerpo descompone los carbohidratos en glucosa. La glucosa, o azúcar en la sangre, es la principal fuente de energía para las células, tejidos y órganos del cuerpo. La glucosa puede usarse inmediatamente o almacenarse en el hígado y los músculos para su uso posterior. Se dividen en dos grupos:

1. **Simples**

**Azúcares:** También se llaman carbohidratos simples porque se encuentran en su forma más básica. Pueden agregarse a los alimentos, como el azúcar en dulces, postres, alimentos procesados y refrescos. También incluyen los tipos de azúcar que se encuentran naturalmente en frutas, verduras y leche.

1. **Complejos**

**Almidones:** Son carbohidratos complejos que están hechos de muchos azúcares simples unidos. Su cuerpo necesita descomponer los almidones en azúcares para usarlos como energía. Los almidones incluyen pan, cereal y pasta. También incluyen ciertas verduras, como papas, guisantes y maíz Fibra (fiber): También es un carbohidrato complejo. Las dietas altas en fibra tienen muchos beneficios para la salud. Pueden ayudar a prevenir problemas estomacales o intestinales, como el estreñimiento. También pueden ayudar a bajar el colesterol y azúcar en la sangre. La fibra se encuentra en muchos alimentos que provienen de plantas, como frutas, verduras, nueces, semillas, frijoles y granos integrales.

**Tipos de azúcares:**

* **Fructose (fructosa):** es el azúcar que está en forma natural en todas las frutas. También se llama levulosa o azúcar de las frutas.
* **Galactose (galactosa):** miembro del grupo de carbohidratos conocidos como azucares simples. Se la encuentra normalmente en combinación con otros azucares, por ejemplo con lactosa (azúcar de la leche).
* **Glucose (glucosa):** se encuentra en las frutas en pequeñas cantidades. También es un almíbar formado del almidón de maíz.
* **Lactose (lactosa):** es el azúcar que se encuentra en la leche. Está compuesto de glucosa y galactosa.
* **Maltose (maltosa):** azúcar de la malta, se produce durante el proceso de fermentación. Se encuentra en la cerveza y el pan.
* **Sucrose (Sacarosa):** azúcar de mesa común, se produce a partir del jugo bajo en azúcar de la remolacha o la caña de azúcar. Se compone aproximadamente de una mitad de glucosa y una mitad de fructosa. Incluye azúcar sin refinar, azúcar granulada, azúcar moreno, azúcar de pastelería y azúcar turbinado.

**Grasas**

Son un tipo de nutriente que se obtiene de la alimentación. Es esencial comer algunas grasas. Todas las grasas están compuestas de ácidos grasos saturados e insaturados. Se denominan saturadas o insaturadas dependiendo de cuánta cantidad de cada tipo de ácido graso contienen.

El colesterol es una sustancia cerosa y parecida a la grasa que se encuentra en todas las células de su cuerpo. Su cuerpo necesita algo de colesterol para producir hormonas, vitamina D y sustancias que le ayuden a digerir los alimentos.

1. **Grasas saturadas**

Elevan el nivel de colesterol LDL ("malo"). Esto lo pone en riesgo de sufrir un ataque cardíaco, un accidente cerebrovascular u otros problemas de salud mayores. Es solida a temperatura ambiente.

1. **Grasas insaturadas**

**monounsaturated fatty acids (grasas monoinsaturadas):** Es una de las grasas saludables, junto con las grasas poliinsaturadas. Las monoinsaturadas tienen forma líquida a temperatura ambiente, pero comienzan a endurecerse cuando se enfrían. Las grasas monoinsaturadas se encuentran en alimentos de plantas como nueces, aguacates y aceites vegetales.

**polyunsaturated fatty acids (grasas poliinsaturadas):** Es una de las grasas saludables, junto con la grasa monoinsaturada. Se encuentra en alimentos vegetales y animales, tales como el salmón, los aceites vegetales y algunas nueces y semillas. Comer cantidades moderadas de grasa poliinsaturada (y monoinsaturada) en lugar de grasas trans y saturadas puede tener beneficios para su salud.

1. **Grasas trans**

**fatty acids total trans (grasas trans):** son grasas perjudiciales que se forman cuando el aceite vegetal pasa por un proceso llamado hidrogenación. Esto hace que la grasa de endurezca y solidifique a temperatura ambiente. Las grasas hidrogenadas o "grasas trans", a menudo se utilizan para conservar algunos alimentos frescos por mucho tiempo. Es solida a temperatura ambiente. Estas grasas no saludables pueden aumentar su riesgo de presentar cardiopatías.

**Otros**

1. **Alcohol:** Las frutas y dulces con relleno de alcohol son un ejemplo de alimentos con etanol. Otros productos alimenticios como los budines de ciruela y las tartas de frutas pueden contener etanol si se usan bebidas alcohólicas destiladas como saborizantes y conservadores.
2. **Ash (ceniza):** es la porción del alimento o material orgánico que queda luego de ser quemado a muy altas temperaturas. Incluye potasio, sodio, calcio y magnesio en grandes cantidades, y en menor medida aluminio, hierro, cobre, manganeso o zinc, arsénico, iodo.
3. **Caffeine (cafeína):** es un estimulante que se puede encontrar en bebidas (como el café, el té, y las bebidas energizantes) y en hierbas (como guaraná y nuez de cola). La cafeína también se agrega a algunos suplementos dietéticos. Cantidades moderadas de cafeína podrían aumentar sus niveles de energía y reducir la fatiga por varias horas.
4. **Theobromine (teobromina):** es un alcaloide perteneciente al grupo de las metilxantinas, y por tanto, posee una estructura similar a la de la cafeína. Se encuentra fundamentalmente en los granos de cacao.
5. **Water (agua):** es una combinación de hidrógeno y oxígeno. Es la base de los líquidos corporales. El agua constituye más de dos terceras partes del peso del cuerpo humano. Todas las células y órganos necesitan agua para funcionar. Sirve como lubricante, es un componente de la saliva y los líquidos que rodean las articulaciones. Regula la temperatura corporal a través de la transpiración. También ayuda a prevenir y aliviar el estreñimiento al movilizar el alimento a través de los intestinos.

**Nutrientes**

Los nutrientes son compuestos químicos contenidos en los alimentos que aportan a las células todo lo que necesitan para vivir. Los nutrientes realizan 3 tipos de funciones en las células:

1. **Energética:** aportan energía para el funcionamiento celular. Necesitamos nutrientes energéticos para poder hacer todas nuestras actividades. Ejemplo: para caminar o correr hay que mover las piernas y esto se consigue cuando se contraen las células de algunos músculos, pero para que esto ocurra las células musculares necesitan energía que la obtienen de algunos nutrientes.
2. **Plástica o reparadora:** proporcionan los elementos materiales necesarios para formar la estructura del organismo en el crecimiento y la renovación del organismo. En época de crecimiento el tamaño de nuestro cuerpo aumenta unos centímetros al año y esto solo es posible si se aporta la materia necesaria para que las células puedan dividirse y aumentar el número de ellas. También durante toda la vida se están reponiendo células que mueren por ejemplo células de la piel, glóbulos rojos o células destruidas en una herida, para lo cual es imprescindible aportar materia al organismo.
3. **Reguladora:** controlan ciertas reacciones químicas que se producen en las células. Para que todo funcione bien en nuestro organismo necesitamos de unos nutrientes que hacen que esto sea posible.

Los tipos de nutrientes son Glúcidos, Lípidos, Proteínas, Vitaminas, Agua y Sales minerales. Cada uno cumple unas funciones distintas, aportando los elementos necesarios para nuestras células.

* **Glúcidos:** (también llamados Hidratos de Carbono o Azúcares) principalmente de función energética. Aportan energía a las células.
* **Lípidos:** también de función principal energética (aportan una reserva de energía, siempre serán utilizados en primer lugar los glúcidos como aporte de energía).
* **Proteínas:** de función principal plástica. Aportan elementos regeneradores para la célula. Vitaminas: función reguladora. Aportan elementos que regulan el buen funcionamiento de todos los elementos y procesos en la célula.
* **Sales minerales:** reguladora y plástica.
* **Agua:** tiene muchas funciones específicas.

### **Pre procesamiento de datos**

1. **Conversión de variables**

Nuestras variables estaban en distintas unidades de medida por lo que decidimos pasar todo a gramos para no tener problema en las transformaciones.

1. **Variables desbalanceadas**

Encontramos que la variable “serving\_size[g]” y “lucopene” tomaban siempre el mismo valor por lo cual fueron eliminadas del dataset ya que no aportan información al modelo.

1. **Estudio de datos nulos**

La variable “saturated\_fat[g]” presentaba 1590 datos nulos. Como es un dato que puede inducirnos al error al momento de la predicción optamos por no usar ninguna de las técnicas de llenado de datos y eliminamos esos registros del dataset.

1. **Creación de nueva columna**

Como nuestro objetivo es predecir si un alimento corresponde o no al grupo de los macro-nutrientes proteínas, armamos un diccionario con los distintos grupos y subgrupos a los que puede pertenecer un alimento. Luego, mediante la función “contains” creamos nuestra variable target que contiene siete valores. Al final, vamos a tratarla como una variable booleana ya que lo que nos interesa es saber si el alimento es o no una proteína.

Por otro lado, según la metodología utilizada para la creación de grupos nos quedaron registros con valores nulos. Esto se debe a que el dataset contenía alimentos tales como “pollo a la naranja” y la función no supo clasificarlo en el subgrupo de carnes o en el de frutas. Y como se debe a un alimento que según los agregados tiene otros macronutrientes presentes (que no estarían en un plato simple de pollo) preferimos no tenerlos en cuenta en nuestra base, por lo que fueron eliminamos del dataset.

1. **Agrupación de variables**

Como detallamos en el diccionario de datos, muchas variables estaban relacionadas entre sí, como por ejemplo “vitamin\_a” y “vitamin\_a\_rae” pertenecen al grupo de las vitaminas liposolubles; por lo que decidimos agruparlas ya que al estudiar la correlación vimos que estaban fuertemente correlacionadas y no queríamos eliminar ninguna del modelo.

El dataset final que usamos como input del modelo consta de 21 columnas y 3165 registros.

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 calories 3165 non-null int64

1 Grupos 3165 non-null object

2 vitaminas\_liposolubles[g] 3165 non-null float64

3 vitaminas\_hidrosolubles[g] 3165 non-null float64

4 grasas[g] 3165 non-null float64

5 grasas\_sat[g] 3165 non-null float64

6 grasas\_insat[g] 3165 non-null float64

7 grasas\_trans[g] 3165 non-null float64

8 carbohidratos[g] 3165 non-null float64

9 carbohidratos\_simples[g] 3165 non-null float64

10 carbohidratos\_complejos[g] 3165 non-null float64

11 carbohidratos\_azucares[g] 3165 non-null float64

12 minerales[g] 3165 non-null float64

13 proteinas[g] 3165 non-null float64

14 aminoacidos\_escenciales[g] 3165 non-null float64

15 aminoacidos\_no\_escenciales[g] 3165 non-null float64

16 alcohol[g] 3165 non-null float64

17 ash[g] 3165 non-null float64

18 caffeine[g] 3165 non-null float64

19 theobromine[g] 3165 non-null float64

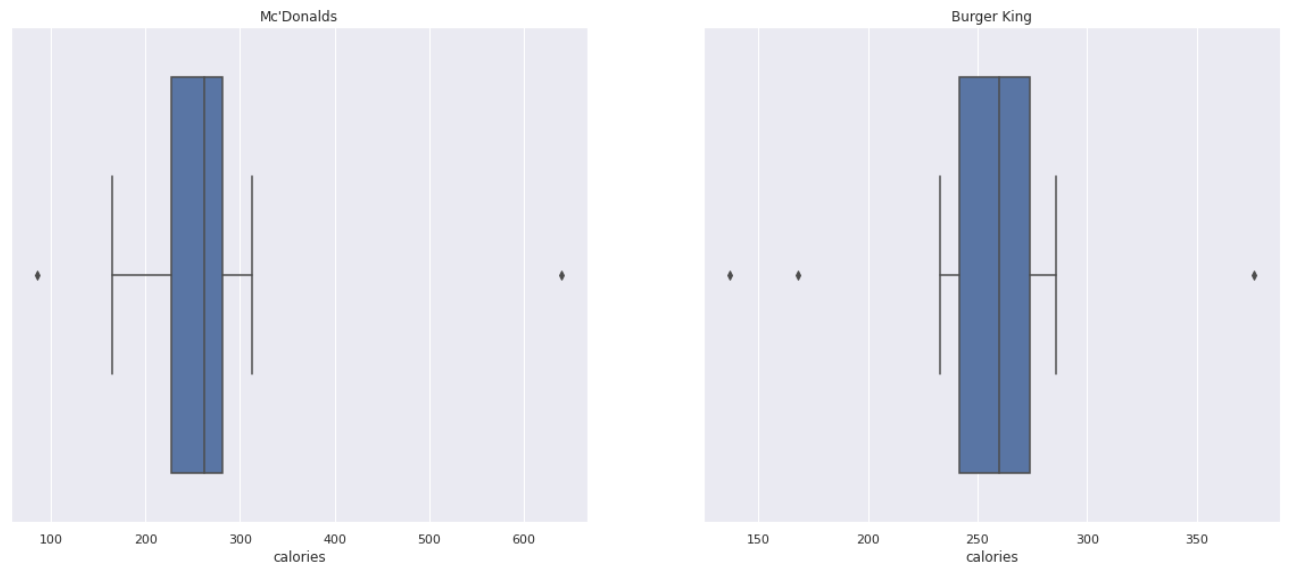
20 water[g] 3165 non-null float64

## **Hallazgos encontrados por el EDA**

### **Análisis Univariado**

Realizamos algunas pruebas de filtros por nombre ya que contamos con menú de las conocidas cadenas de comida rápida que están en muchos lugares del mundo: McDonalds y Burger King.

Hicimos una comparativa de la dispersión de las calorías de ambos menús utilizando un boxplot:

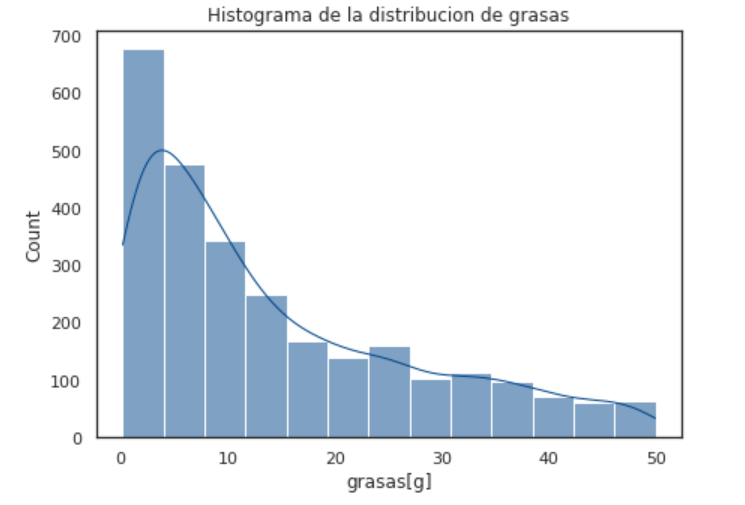


Nos encontramos con que las calorías del menú de Burger King estaban más concentradas. Además, estudiamos los outliers de cada cadena:

1 – McDonals tenía un alimento con menos de 100 calorías y otro con más de 600 (era el maní que se utiliza para los helados)

2 – En Burger encontramos dos alimentos con menos de 200 calorias, correspondientes al batido de vainilla y a un tipo de hamburguesa vegie. Y un aimento con más de 350 calorias que corresponde a una croissan con salchicha y queso.

Además estudiamos el histograma de la distribución de grasas en el dataset.

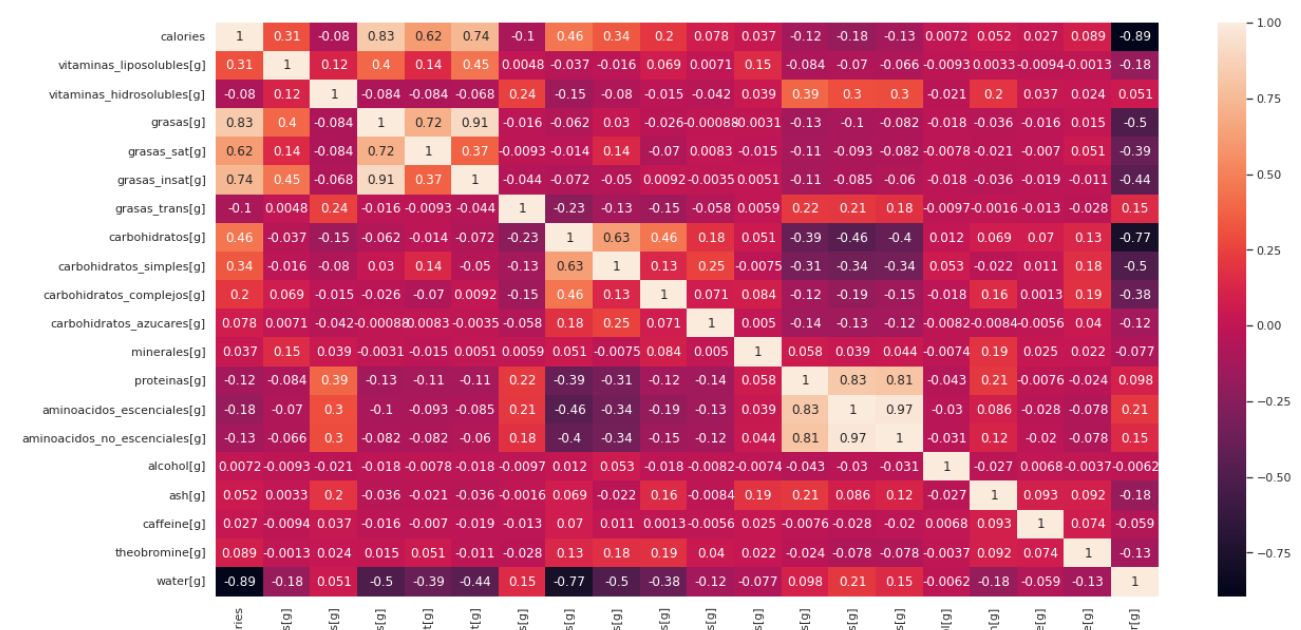


Como se puede observar, la mayor cantidad de alimentos se encuentra por debajo de los 10 gramos de grasas por porción.

VER TABLA DE FRECUENCIAS

### **Análisis Bivariado**

Dentro de este apartado estudiamos la correlación de nuestras variables:

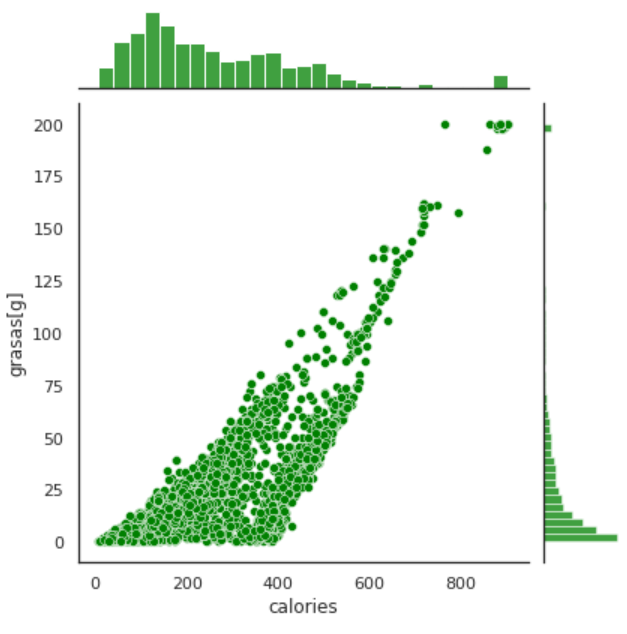


Y llegamos a las siguientes conclusiones:

* grasas[g] está fuertemente correlacionada con grasas\_sat[g] y grasas\_insat[g]. Eliminamos estas dos últimas del modelo.
* carbohidratos[g] está fuertemente correlacionada con carbohidratos\_simples[g]. Eliminamos carbohidratos\_simples[g] del modelo.
* proteínas[g] está fuertemente correlacionada con aminoacidos\_esenciales[g] y aminoacidos\_no\_esenciales[g]. Eliminamos estas dos últimas del modelo.

Continuamos con el estudio de la correlación entre algunas variables:

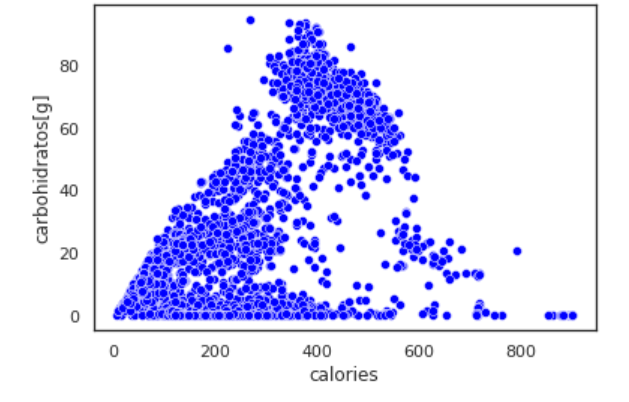
1. **Grasas y Calorías:**



A medida que consumimos un alimento con mayor cantidad de grasas por porción, la antidad de calorias consumidas es tambien mayor que en aquellos alimentos que tienen menos grasas totales por porción.

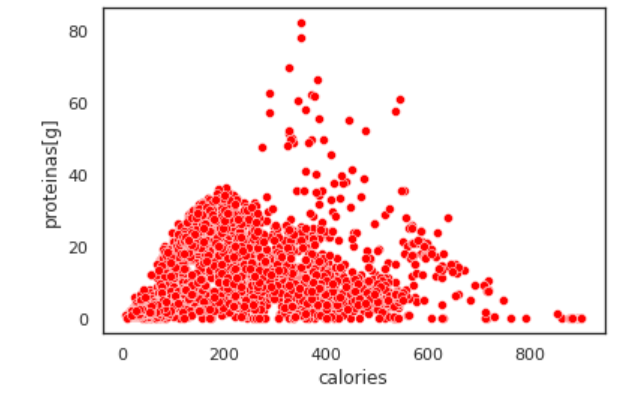
Además observamos aquellos alimentos que tenían mayor cantidad de grasas por porción (>100) y nos encontramos con los aceites tanto de origen animal como vegetal.

1. **Carbohidratos y Calorías:**



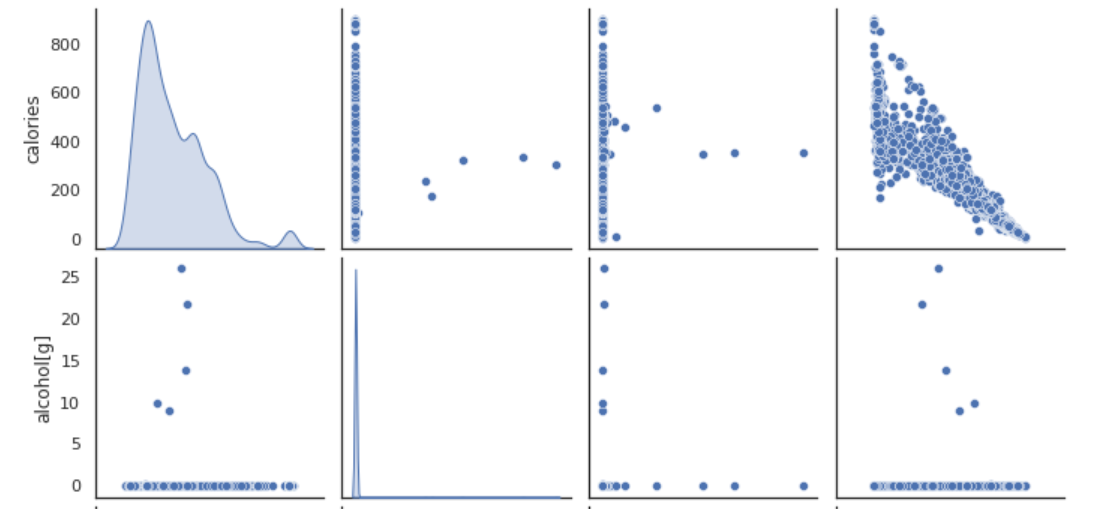
En el caso de los carbohidratos la relación no es lineal. Como podemos observar en el gráfico aquellos alimentos con mayor cantidad de carbohidratos por porción no son los que mayor aporte calórico hacen.

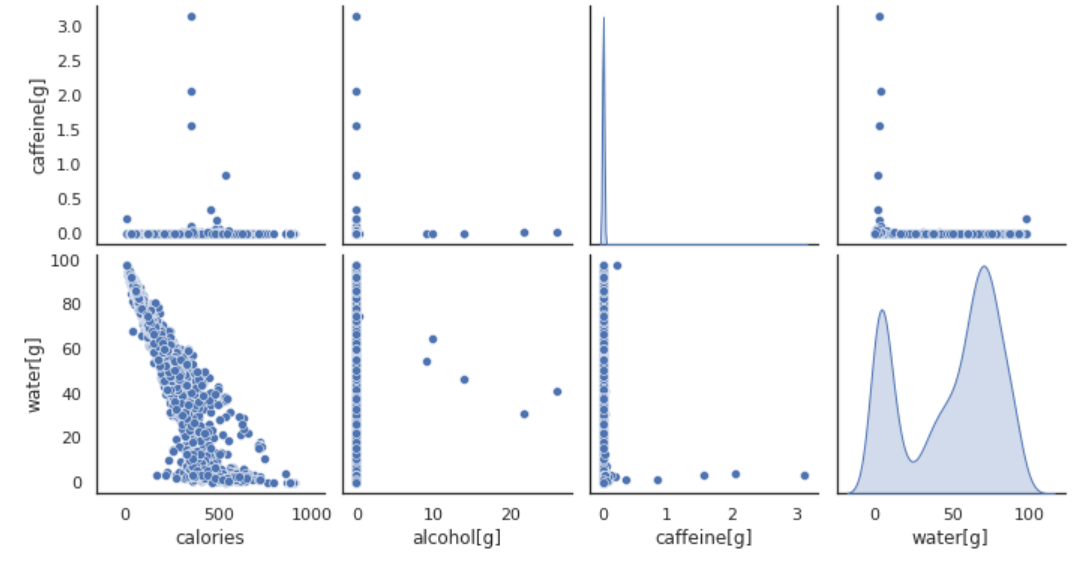
1. **Proteínas y Calorías:**



Y en el caso de las proteínas, podemos llegar a una conclusión similar a la de los carbohidratos. Aquellos alimentos que aportan mayor cantidad de proteínas por porción no coinciden con los de mayor aporte calórico.

1. **Alcohol, cafeína y agua vs Calorías:**



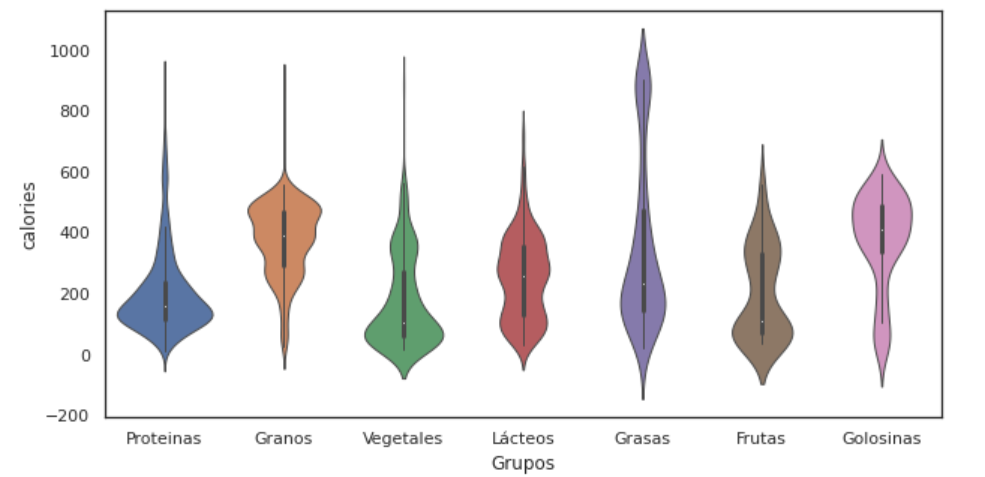


Acá observamos que aquellos alimentos que mayor cantidad de agua aportan por porción son los que menor cantidad de calorías aportan (y viceversa). Y los que aportan agua no aportan cafeína ni alcohol.

En cuanto a las calorías, los alimentos que aportan cafeína y alcohol aportan prácticamente las mismas calorías.

### **Análisis Multivariado**

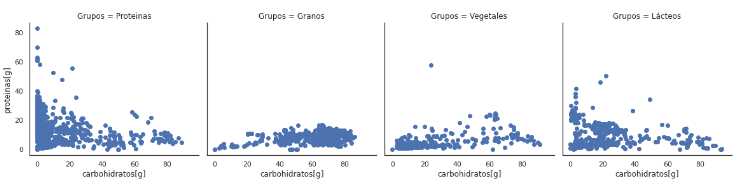
En este punto comenzamos realizando mediante un diagrama de violín la distribución y densidad de probabilidad de calorías en cada uno de los grupos:

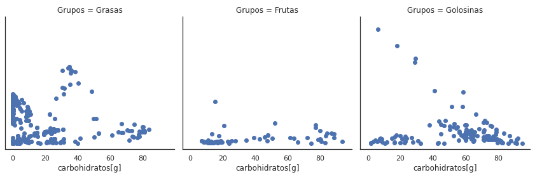


Y nuestras observaciones son las siguientes:

* La porción de 100 gramos de proteínas se concentra entre 100 y 200 calorías.
* La porción de 100 gramos de granos se concentra mayormente entre 300 y 500 calorías.
* La porción de 100 gramos de vegetales se concentra en 100 calorías.
* La porción de 100 gramos de lácteos se distribuye entre 100 calorías y 400 calorías.
* La porción de 100 gramos de grasas se distribuye entre 100 calorías y 400 calorías.
* La porción de 100 gramos de frutas se concentra en 100 calorías.
* La porción de 100 gramos de golosinas se concentra entre 350 y 450 calorías.

Luego, para analizar la relación entre los carbohidratos y proteínas en los distintos grupos de alimentos. Ahora elegimos el FaceGrid de la librería Seaborn:

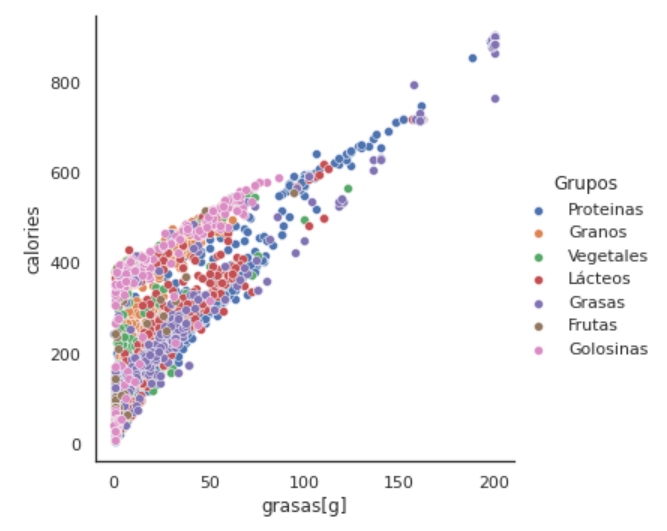




Y observamos que en frutas, vegetales, granos y golosinas hay un mayor aporte de carbohidratos que de proteínas. Mientras que en lácteos y grasas la distribución es más uniforme. En proteínas, el aporte del mismo nutriente es más notorio.

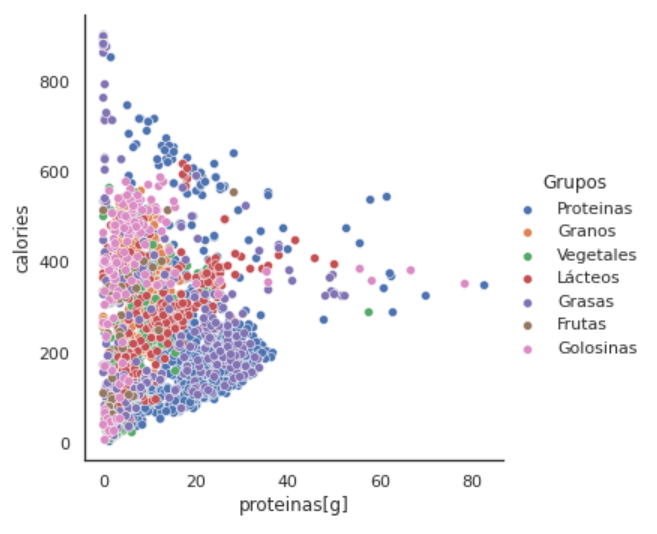
Ahora vamos a analizar la correlación entre las calorías y los tres grandes grupos de macro-nutrientes pero observando todos los grupos dentro del mismo gráfico:

1. **Calorías y grasas**



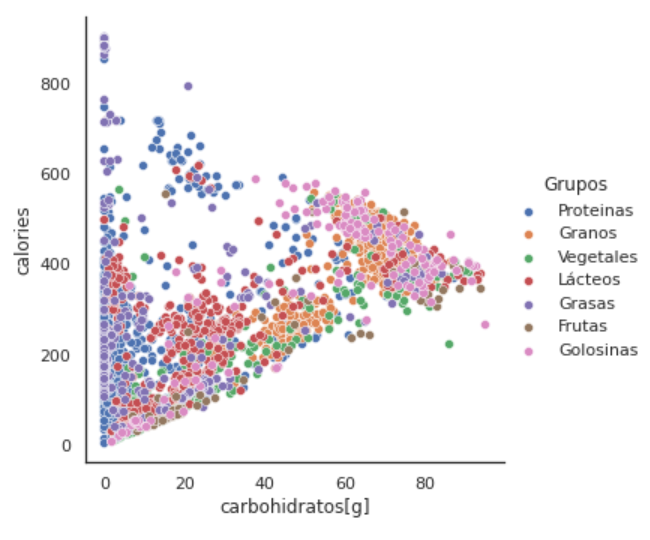
En todos los grupos vemos una correlación positiva entre las grasas y las calorías. A mayor aporte de grasas, mayor aporte de calorías.

1. **Calorías y proteínas**



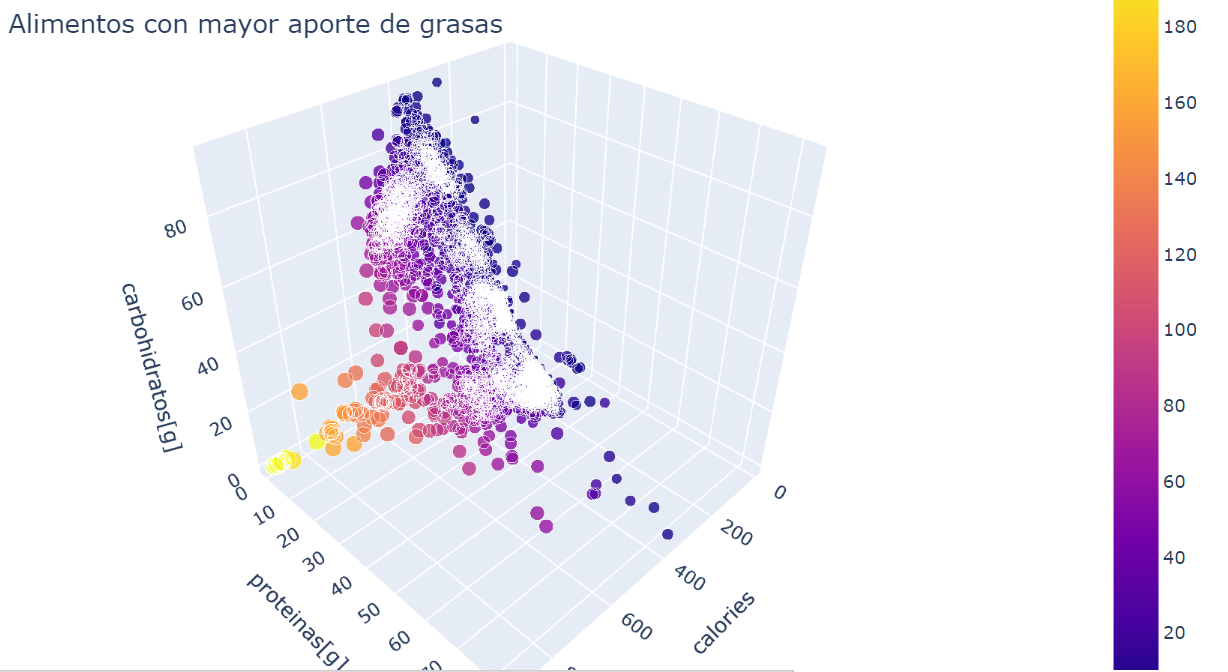
En este caso no se observa una correlación positiva entre las proteínas y las calorías. En aquellos alimentos donde se observa un mayor aporte de proteínas (grupos de proteínas y grasas), el aporte promedio de calorías es de 200 por porción.

1. **Calorías y carbohidratos**



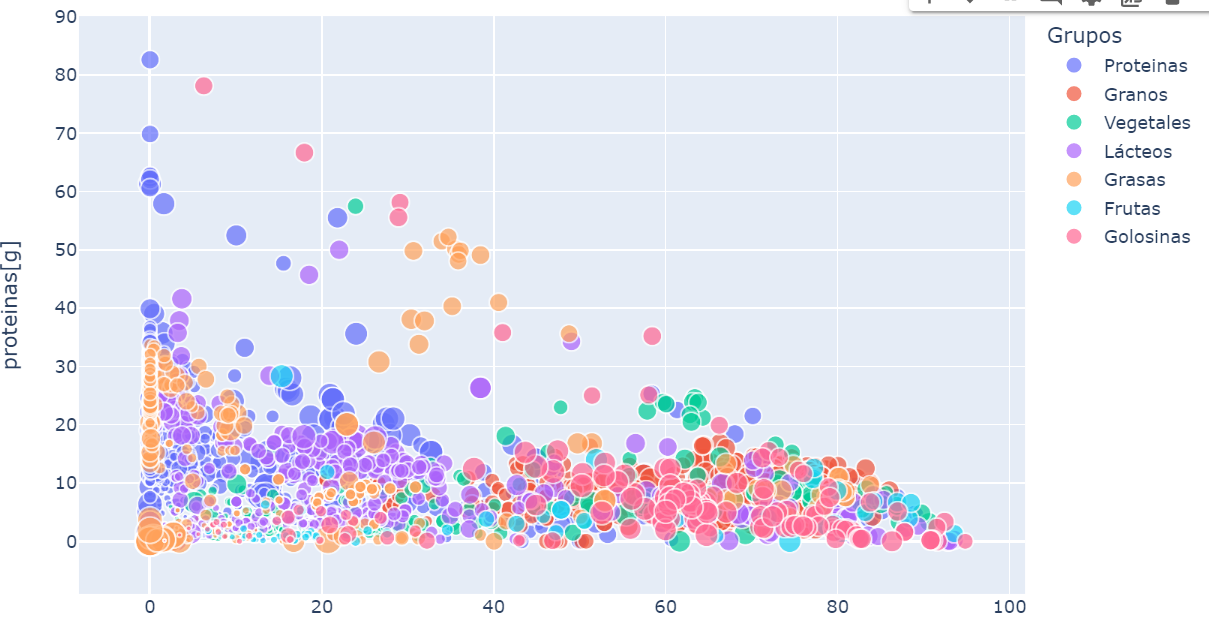
Y acá observamos una mayor dispersión. Las golosinas son el grupo que mayor cantidad de carbohidratos aportan llegando a un promedio de 550 calorías por porción.

Luego utilizamos un ploty 3D para visualizar si había muchos alimentos con mayor cantidad de grasas:



En este grafico 3D observamos que en nuestro dataset tenemos menor cantidad de alimentos que aporten muchas grasas (color amarillo). Esto nos permite decir que el dataset es adecuado para el proyecto ya que aquellos alimentos con mayor aporte de grasas no van a ser ejemplos de proteínas.

Y por último, realizamos un gráfico para observar la relación entre proteínas y carbohidratos en los distintos grupos de alimentos:



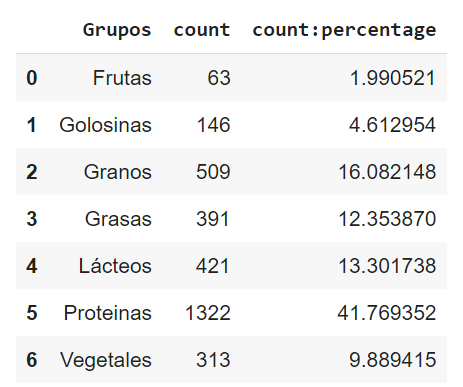


En este gráfico observamos que los alimentos que aportan mayor cantidad de proteínas no tienen un aporte significativo de carbohidratos, y viceversa. Por otro lado, aquellos que aportan mayor cantidad de proteínas aportan a su vez menor cantidad de calorías (tamaño de las bolitas). Y los que mayor cantidad de calorías aportan, no aportan proteínas ni carbohidratos. Por lo estudiado en anteriores gráficos podemos inferir que son los aceites y grasas animales que tienen un gran aporte de grasas (nutriente que no se observa en el presente gráfico).

## **Algoritmo elegido**

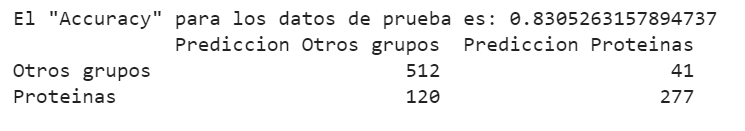
Utilizamos el algoritmo de árboles de decisión.

En primer lugar observamos si nuestro target estaba balanceado y nos encontramos con una respuesta positiva.



El porcentaje para entrenamiento y prueba fueron del 70% y 30% respectivamente.

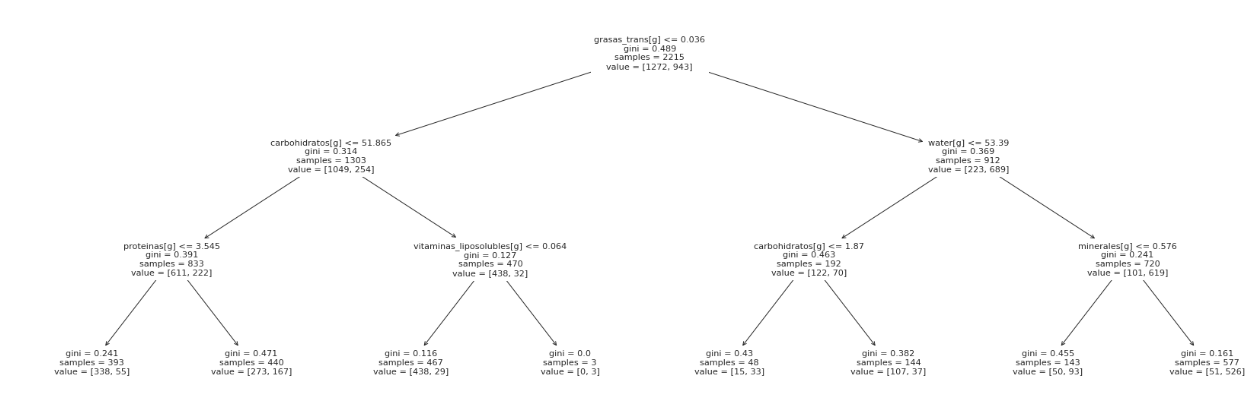
## **Métricas de desempeño del Modelo**

****

## **Iteraciones de optimización**

Utilizamos la validación cruzada (K-Fold-CrossValidation) para mejorar nuestro modelo de árbol.

Dividimos los datos de entrenamiento en k=5. Nuestros parámetros iniciales fueron dos variables y una rama. En una primera instancia obtuvimos un error muy alto (sesgo=0.77) y una varianza muy baja, igual a 0.02. Como nos encontrábamos en presencia de underfitting, ajustamos el modelo utilizando una rama adicional y todas las variables para mejorar esos resultados. En esta instancia el score en el entrenamiento y prueba fueron más cercanos por lo que nos encontramos ante un modelo mucho más robusto.



## **Futuras líneas**

Para mejorar los resultados de nuestro algoritmo deberíamos utilizar una base con alimentos naturales sin tanto proceso en el medio. Además, podríamos implementar otros algoritmos para evaluar su performance.

Por otro lado, como futuras investigaciones se podría lanzar un algoritmo que predice a qué grupo corresponde el alimento y no solo si corresponde o no al grupo de proteínas.

## **Conclusiones**

En el caso de nuestro problema de clasificación al implementar el modelo de árboles de decisión solo con dos variables y solo una rama de profundidad, obtuvimos un bias muy alto y una varianza muy baja; lo que se traduce en un problema de underfitting. Para evitar esto aplicamos K-Fold-CrossValidation y ajustamos los parámetros utilizando tres ramas de profundidad y todas las variables que habíamos determinado al final del data wrangling. El resultado fue un modelo mucho más robusto, ya que si bien el error no es muy bajo, es más parecido a la varianza.