proyecto final (grupo 4)

Conjunto de datos de nutrición

Integrantes : Flavio Galimberti
Felipe Quesada
Francisco Macchi
giuliana zeolla

Profesor : Damián Dapueto
Tutor : Alejandro Pujol
Comisión : 29730

Fuente del conjunto de datos :
https://www.kaggle.com/datasets/sathyakrishnan12/nutrition-datasets

Fecha : 3 de Octubre del 2022

Institución : Coder House

Descripcion del caso de negocio

¿Quienes 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 en 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.

Contexto

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

Analisis de la informacion nutricional

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: Energética, Plástica o reparadora y Reguladora.

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

objetivo principal

Crear un modelo de aprendizaje automático que al ingresar los valores de la tabla nutricional de un alimento indique si corresponde o no al grupo de las proteínas.

Objetivos secundarios

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

Preguntas de la investigacion

- ¿Cuáles son los valores nutricionales de un alimento para que se consideren dentro del grupo de Proteínas?
- ¿Qué valores nutricionales tiene cada grupo?
- ¿Existen alimentos de origen vegetal que aporten como principal macronutriente proteína?

CONFLICTO DE DATOS

- Leemos el dataset y removemos el índice porque no nos aporta ninguna información.
- Nos dimos cuenta de que todas las variables son string y las pasaron a números eliminando los subíndices mg,g,iu,mcg de cada columna.
- Al tener todas las variables numéricas fáciles se nos hizo mucho mas poder trabajar con los datos y hacer un análisis más preciso.
- Convertiremos todo a la misma unidad de medida para no tener problemas con las transformaciones. Como tenemos gramos, miligramos y microgramos pasaremos todo a gramos.
- Estudiamos si hay variables que siempre o prácticamente siempre toman el mismo valor ya que toman siempre el mismo valor no aporta información al modelo.

Preprocesamiento de datos

- Convertiremos todo a la misma unidad de medida para no tener problemas con las transformaciones. Como tenemos gramos, miligramos y microgramos pasaremos todo a gramos. Renombramos las columnas ya que ahora todo se mide en gramos.
- Estudiamos si hay variables que siempre o prácticamente siempre toman el mismo valor ya que toman siempre el mismo valor no aporta información al modelo.
- Creamos un bucle que itera por todas las columnas y calcula la frecuencia de cada categoría. Como umbral seteamos 99,9%, es decir, si alguna de las categorías de las variables tiene una frecuencia mayor al 99,9% entonces esta variable siempre tiene el mismo valor y por ende no aporta mucha información al modelo.
- Las variables obtenidas no aportan información al modelo, por lo que procederemos a eliminarlas.
- Estudiamos si hay nulos en el dataset y nos dimos cuenta que la variable saturada_fat[g] presenta 1590 datos nulos, modificó esos registros ya que creo que no es conveniente rellenarlos por ningún método.
- Dejamos solo 6 decimales en todas las variables numéricas.

Preprocesamiento de datos

Nuevas columnas separadas por grupos Creación y subgrupos de alimentos

Para generar estos grupos nos basamos en los principales grupos nutricionales, más un grupo que llamamos golosinas.

- **Proteínas** = Carnes, Aves_de_Corral, Huevos, Pescados_mariscos, Productos_de_Soja, Frutos_secos
- **Lácteos** = Lácteos
- **Frutas** = Frutas
- **Vegetales** = Vegetales_verdes, Vegetales_rojos_y_naranjas, Vegetales_almidonados, Frijoles, Otros_vegetales
- **Granos** = Granos_enteros, Granos_refinados
- **Grasas_Aceites** = Aceites_grasas
- **golosinas** = golosinas

Uniones de datos de cada tipo de alimento (|)

- Primero realizamos el join de cada dato con lista | (o)
- Luego buscamos en cada elemento "name" del dataset si este contiene una de las palabras clave para cada grupo y subgrupo y se genera una nueva columna

Acomodado de datos

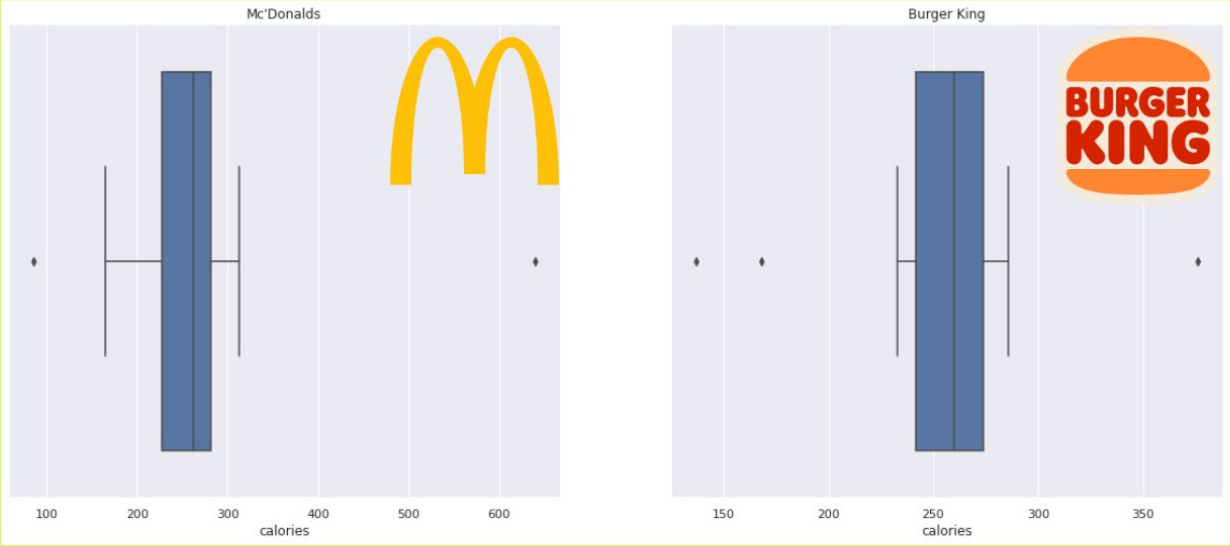
- Generamos diccionarios para los grupos y subgrupos creados anteriormente
- Generamos una columna con los principales grupos
- Hay alimentos que no pudimos agruparlos porque su etiqueta tiene descripción ambigua; es decir, puede pertenecer a mas de un grupo. Este tipo de alimentos no nos interesa para el modelo, por lo que vamos a proceder a eliminarlos del conjunto de datos.

Análisis Univariado

Calculo de información estadística y genérica de cada columna de un dataframe.

Algunas pruebas de filtros por nombre

- Hicimos una comparativa de la distribución de calorías con boxplot de los menús de McDonald y Burger King.



- Las calorías de Burger King están más concentradas que las de McDonald.

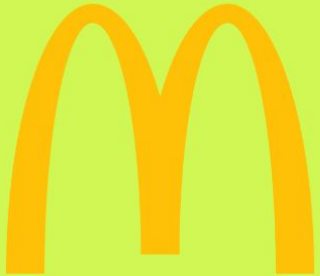
Análisis Univariado

Hicimos consultas para el estudio de outliers:

- El alimento que salió con más de 600 calorías en McDonald’s.
- Los alimentos que salieron con menos de 200 calorías en Burger King.
- El alimento que salió con más de 350 calorías en Burger King.

resultados de las consultas:

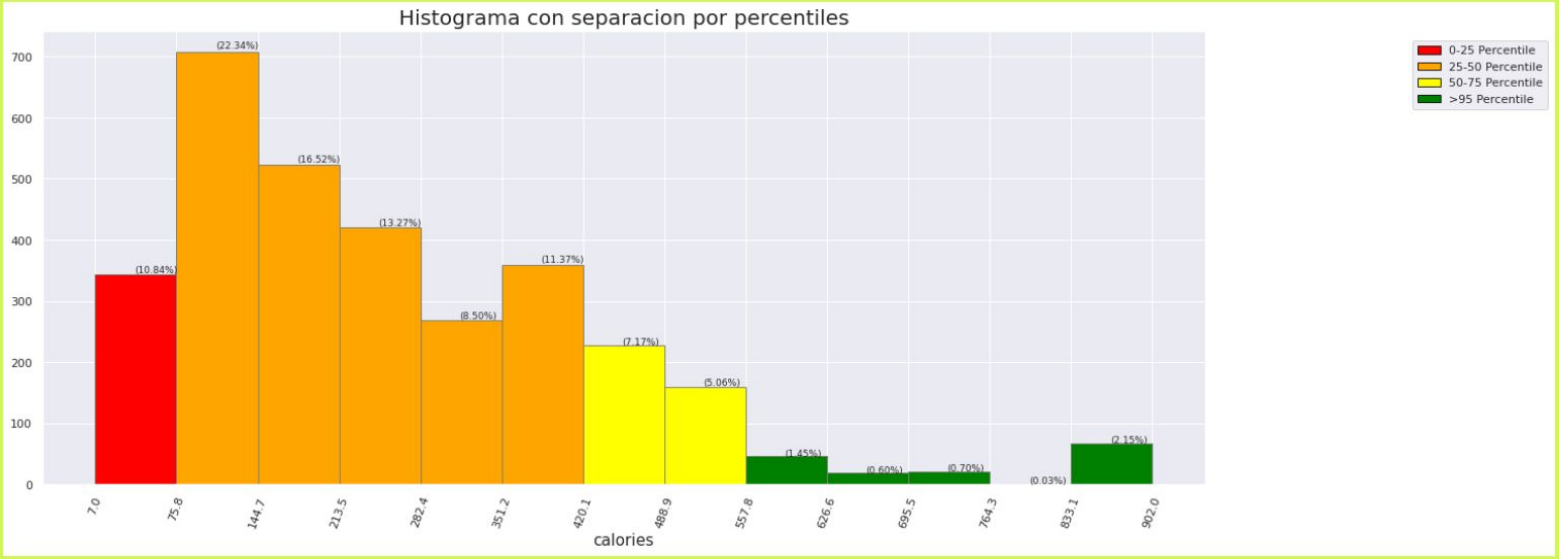
- El alimento de McDonald's que tiene mas de 600 calorias corresponden a las nueces que se utilizan para el helado.
- En Burger King, los alimentos con menos de 200 calorías son el batido de vainilla y una hamburguesa vegie.
- Y el alimento de Burger King con mas de 350 calorias es una CROISSAN con Salchicha y Queso



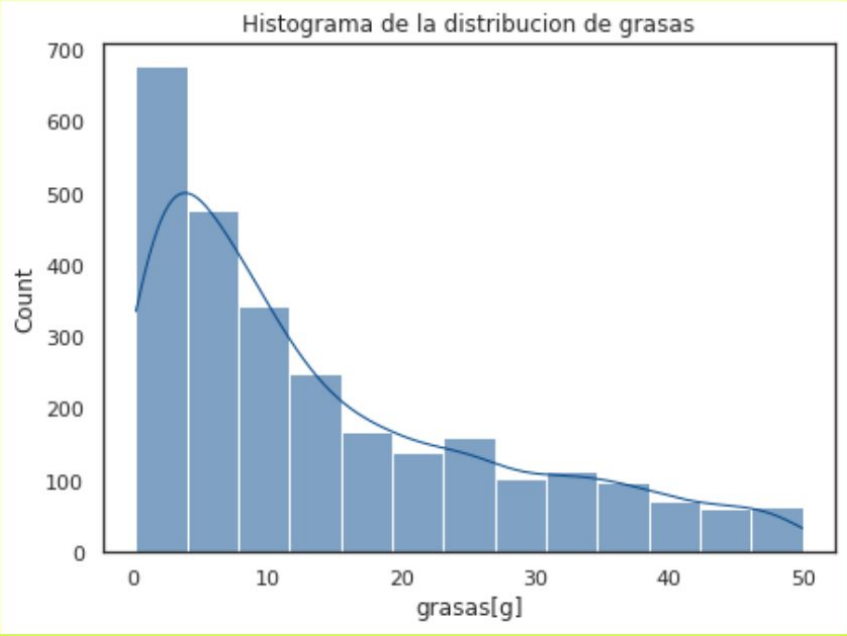
contra

Análisis Univariado

- Hicimos un gráfico que nos indica la cantidad de calorías por percentil.
- Gran parte de los alimentos del conjunto de datos se encuentra debajo de las 400 calorías, pero se observa un 2,15% que tienen más de 850 calorías.



Análisis Univariado

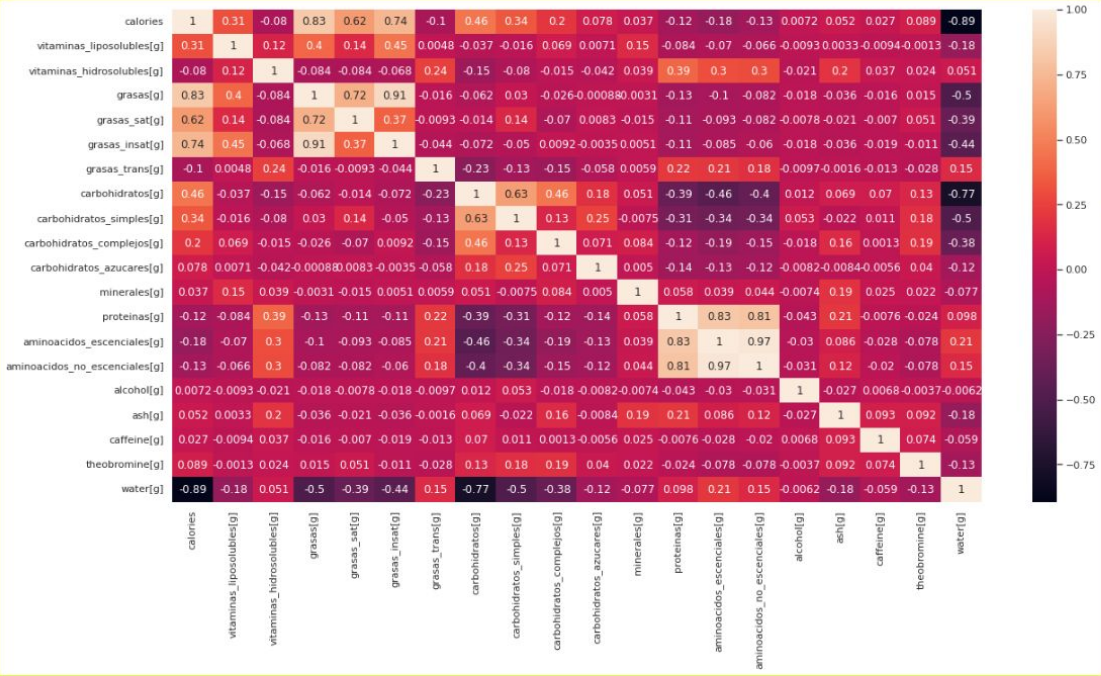


- También analizamos la distribución de grasas.
- Podemos observar que una gran cantidad de alimentos son bajos en grasas, por debajo de los 10 gramos por porción

Análisis Bivariado

Implica el análisis de dos variables, con el propósito de determinar la relación empírica entre ellas

- Copiamos el dataset heredado del Análisis Univiariado
- Realizamos las conexiones del dataset

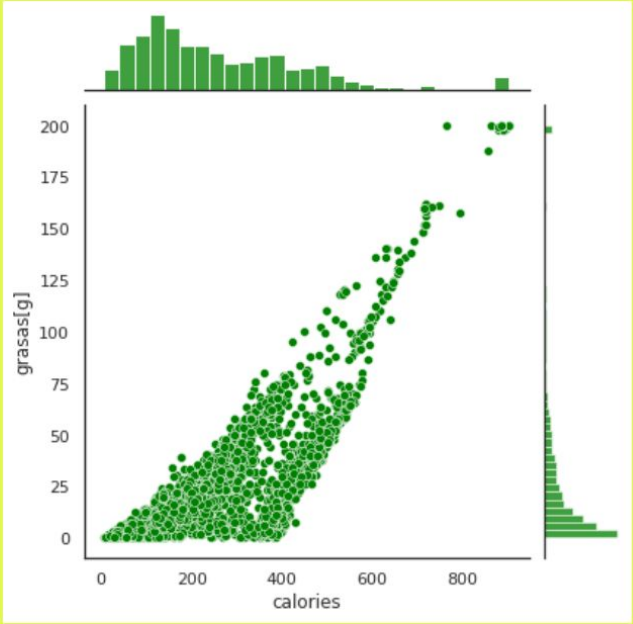


- grasas[g] esta carbonada correlacionada con grasas_sat[g] y grasas_insat[g]. Vamos a eliminar estas dos últimas del modelo.
- carbohidratos[g] esta carbonilla correlacionada con carbohidratos_simples[g]. Vamos a eliminar carbohidratos_simples[g] del modelo.
- proteínas[g] esta carbónica correlacionada con aminoácidos_esenciales[g] y aminoácidos_no_esenciales[g]. Vamos a eliminar estas dos últimas del modelo.

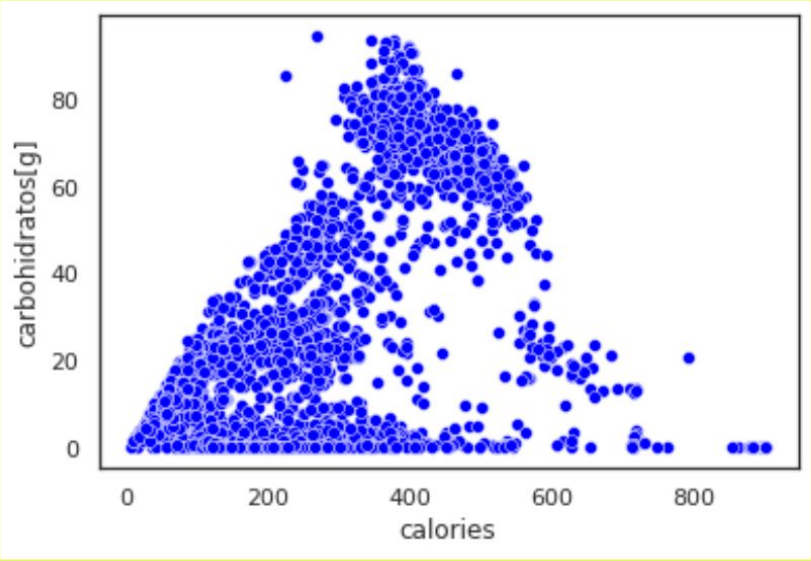
Graficos

Grasas y calorías

- A medida que consumimos un alimento con mayor cantidad de grasas por porción, la de calorías consumidas es también mayor que en aquellos alimentos que tienen menos grasas totales por porción.
- observamos aquellos alimentos que tienen mayor cantidad de grasas por porción (>100) y nos encontramos con los aceites tanto de origen animal como vegetal.



Graficos Carbohidratos y Calorias

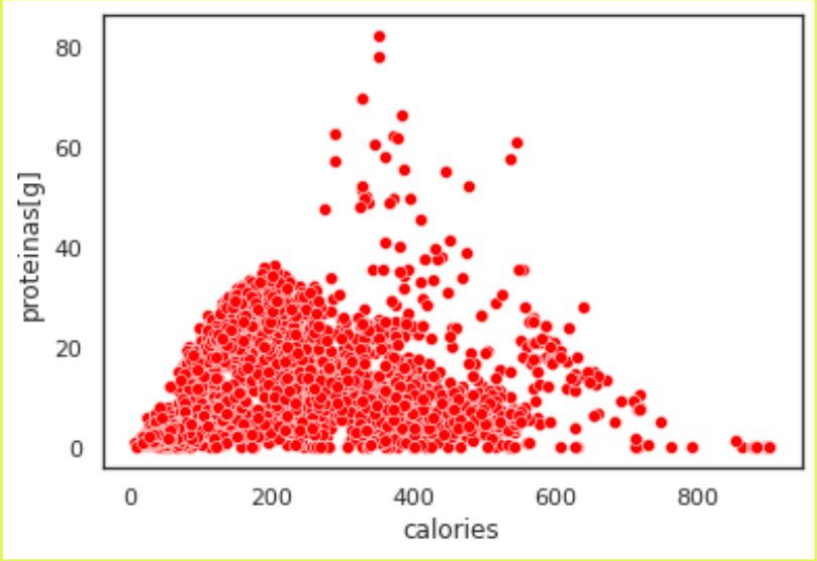


- En el caso de los carbohidratos la relacion no es lineal. Como podemos observar en el gráfico aquellos alimentos con mayor cantidad de carbohidratos por porción no son los que mayor aporte calórico hacen

Graficos

Proteínas y Calorías

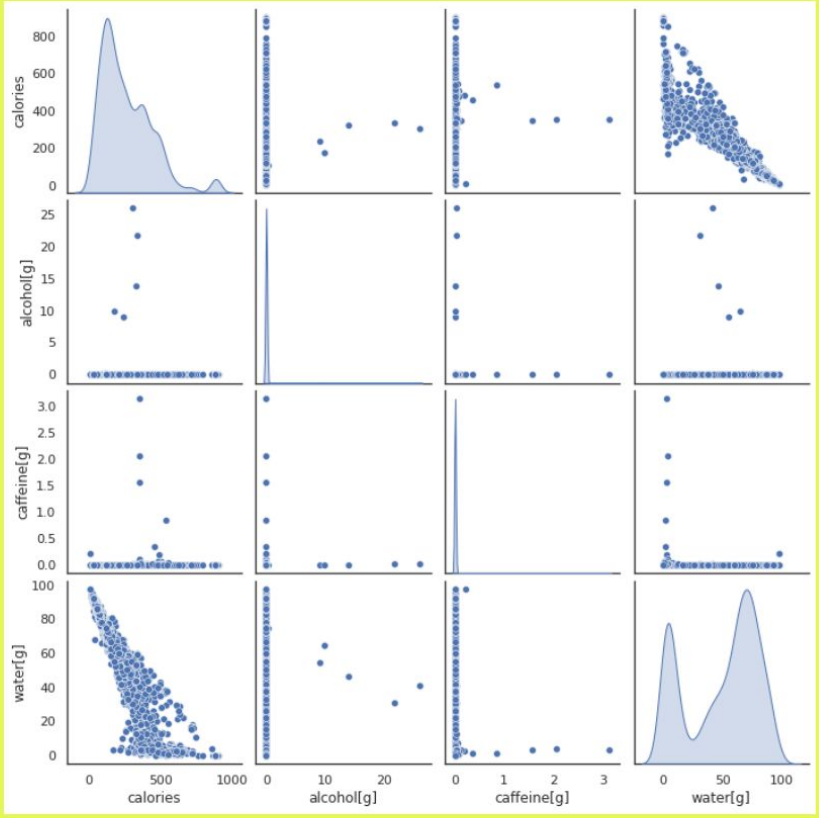
- Y en el caso de las proteínas, podemos llegar a una conclusión similar a la de los carbohidratos. Aquellos alimentos que aportan mayor cantidad de proteínas por porción no coinciden con los de mayor aporte calórico.



Graficos

Alcohol, cafeína y agua vs Calorías

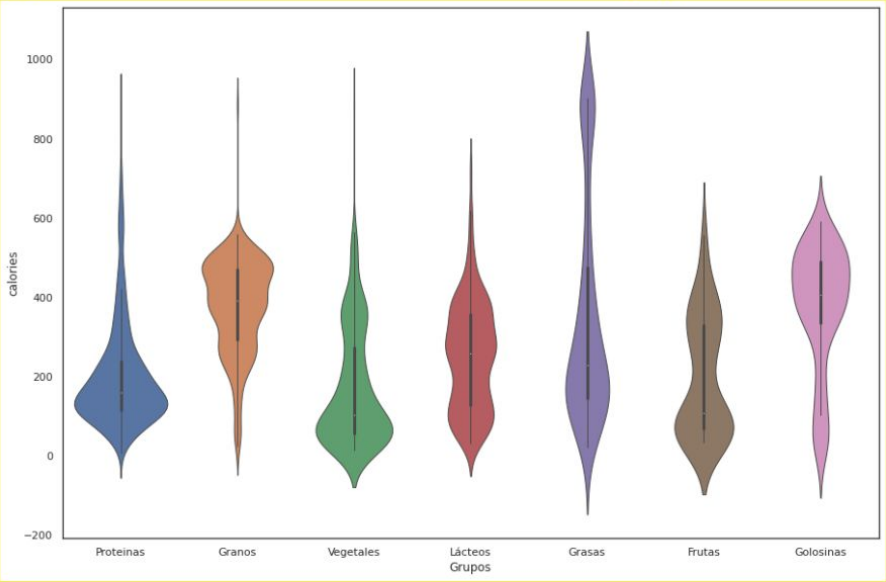
- alimentos que mayor cantidad de agua aportan por porción son los que menor cantidad de calorías aportan (y viceversa).
- Aquellos alimentos que aportan agua no aportan cafeína ni alcohol.
- En cuanto a las calorías, los alimentos que aportan cafeína y alcohol aportan prácticamente las mismas calorías.



Análisis multivariado

observación y análisis simultáneos de más de una variable

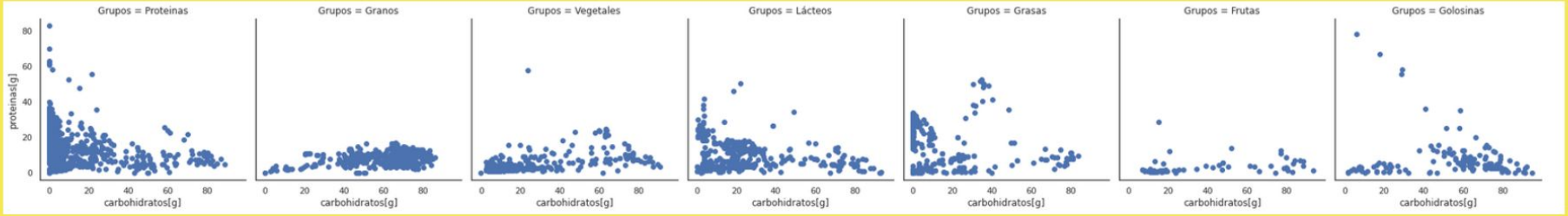
Diagrama de violín la distribución y densidad de probabilidad de calorías en cada uno de los grupos



- 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 reparte entre 100 calorías y 400 calorías.
- La porción de 100 gramos de grasas se reparte 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

Análisis multivariado

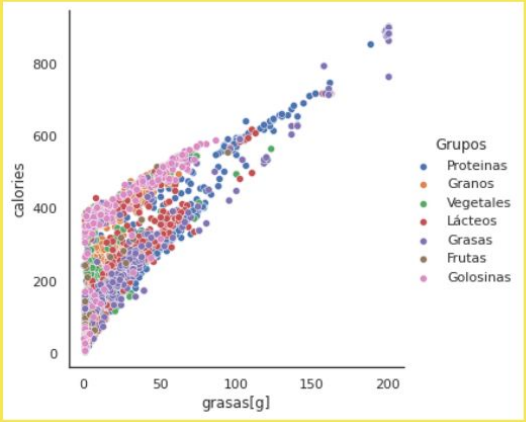
Carbohidratos y proteínas en los distintos grupos de alimentos (FaceGrid de la librería Seaborn)



- 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.

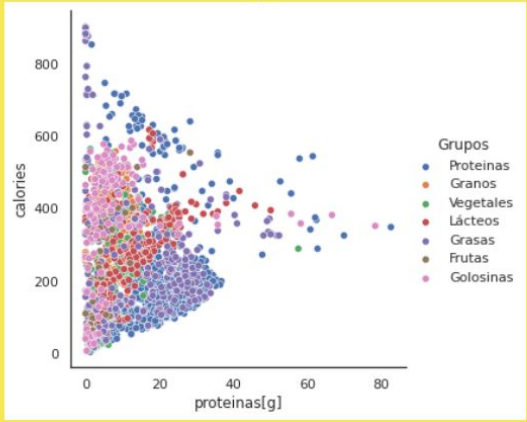
Análisis multivariado conexiones entre las calorias y los tres grandes grupos de macro-nutrientes

calorías y grasas



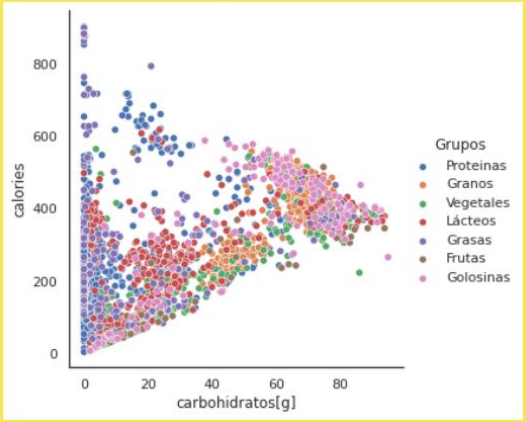
- En todos los grupos vemos una confirmación positiva entre las grasas y las calorías. A mayor aporte de grasas, mayor aporte de calorías.

calorias y proteinas



- En este caso no se observa una confirmación 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.

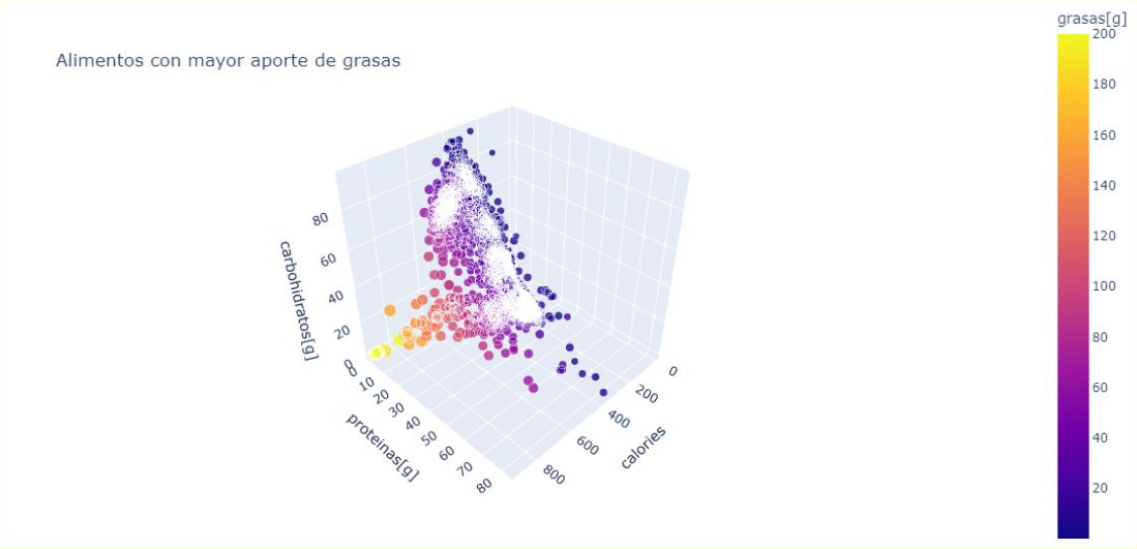
calorias y carbohidratos



- 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.

Análisis multivariado

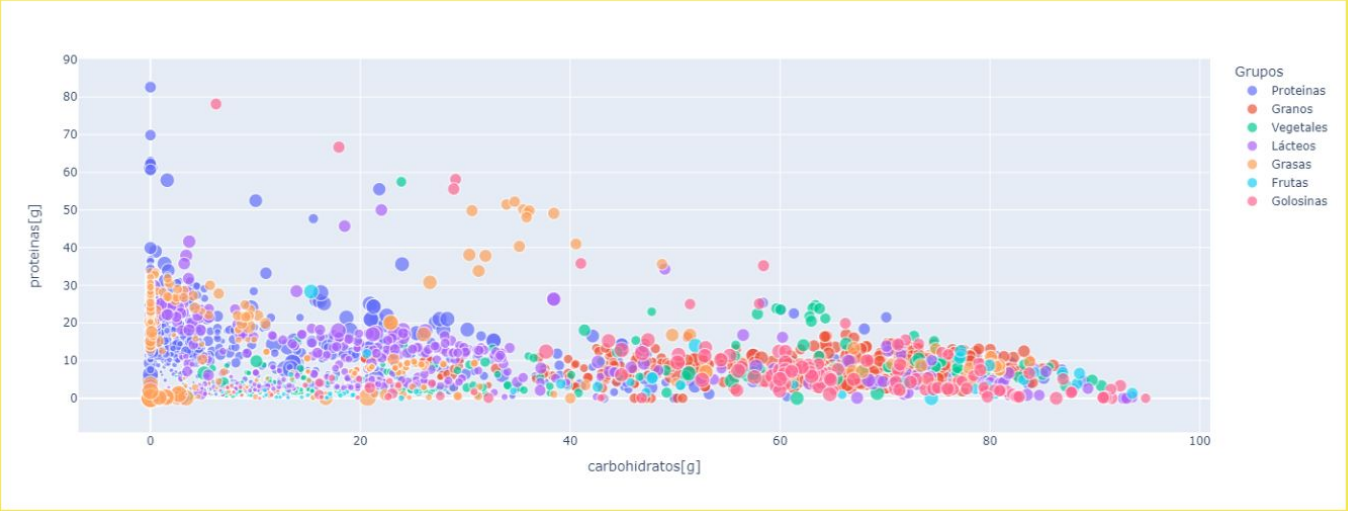
Ploty 3D para visualizar si habia muchos alimentos con mayor cantidad de grasas



- En nuestro conjunto de datos tenemos menor de alimentos que aportan muchas grasas (color amarillo). Esto nos permite decir que el conjunto de datos es adecuado para el proyecto ya que aquellos alimentos con mayor aporte de grasas no van a ser ejemplos de proteínas

Análisis multivariado

Relación entre proteínas y carbohidratos

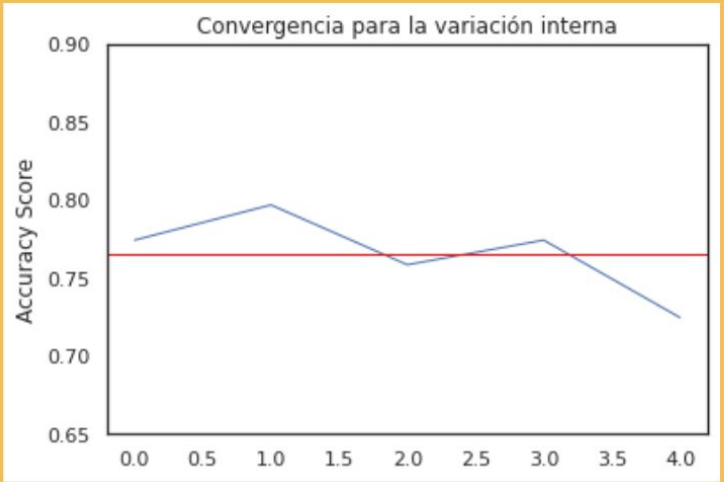


- En este grafico observamos que los alimentos que aportan mayor cantidad de proteinas no tienen un aporte significativo de carbohidratos, y viceversa. Por otro lado, aquellos que aportan mayor cantidad de proteínas cantidad aportan a su vez menor de calorías (tamaño de las bolitas). Y los que mayor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

Modelado

- Utilizamos el algoritmo de arboles de decision

Iteraciones de optimización

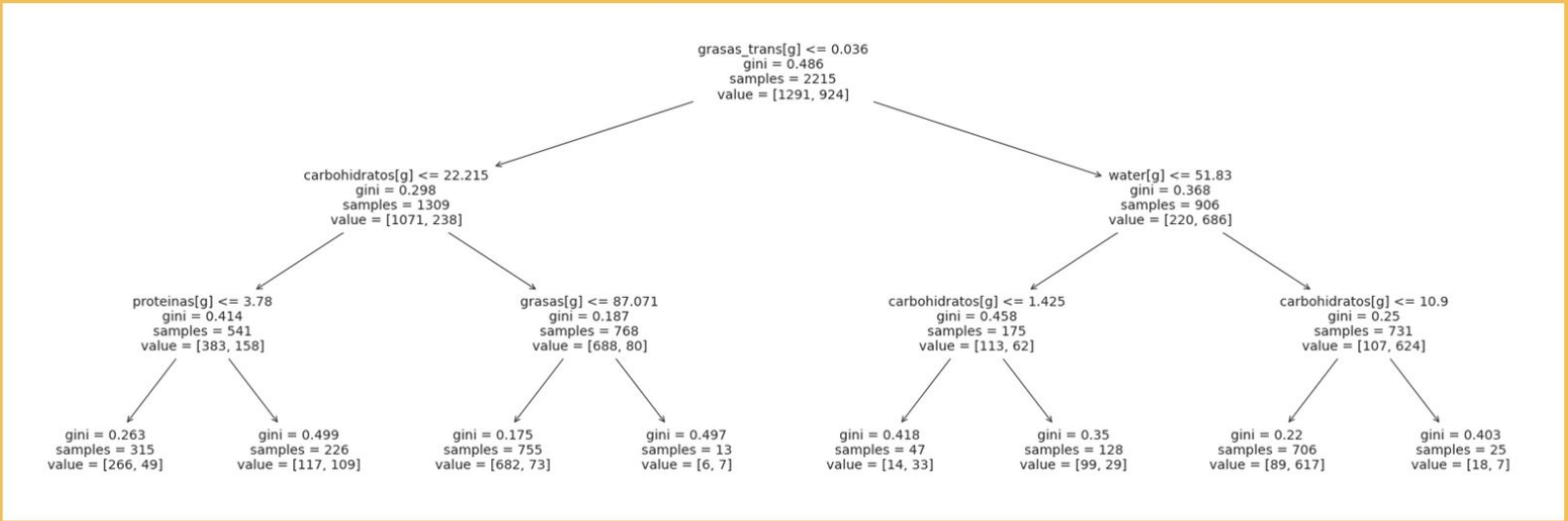


- Utilizamos la validación cruzada (K-Fold-CrossValidation) para mejorar nuestro modelo de árbol.
- Divide los datos de entrenamiento en k=5. Nuestros parámetros iniciales fueron dos variables y una rama
- En una primera instancia obtuvimos un error muy alto (sesgo=0.77) y una varianza muy baja, igual a 0.02.
- Nos encontramos en presencia de underfitting, ajustamos el modelo utilizando una rama adicional y todas las variables para mejorar esos resultados

Modelado

Iteraciones de optimización

- En esta instancia, la puntuación en el entrenamiento y la prueba fueron más cercanas por lo que nos encontramos ante un modelo mucho más robusto.



Futuras lineas

- Para mejorar los resultados de nuestro algoritmo deberíamos utilizar una base con alimentos naturales sin tanto proceso en el medio. Además, podemos implementar otros algoritmos para evaluar su desempeño.
- Por otro lado, como investigaciones futuras se podría lanzar un algoritmo que predice qué grupo corresponde al 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 sesgo 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.

CODER HOUSE