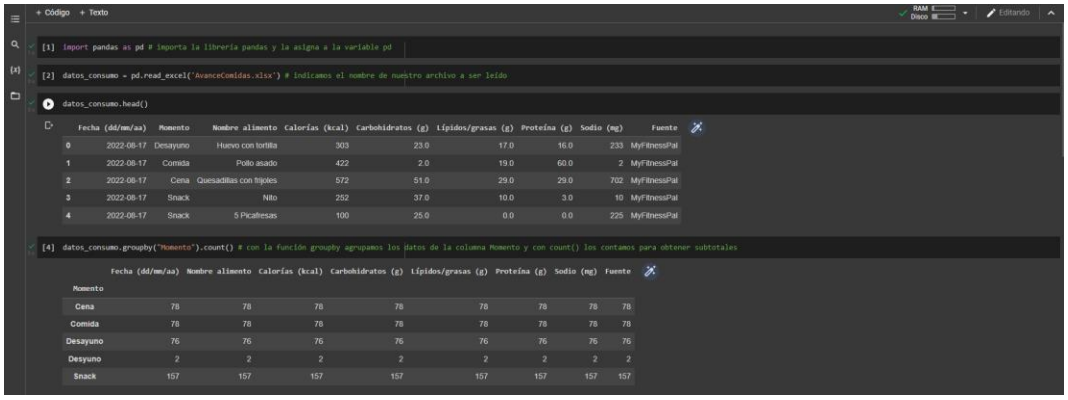
Héctor Gutierrez
A01253031
7 de Noviembre del 2022

Fase 3: Preparación de los datos

Parte 1: Selección, limpieza y preparación de los Datos en Python

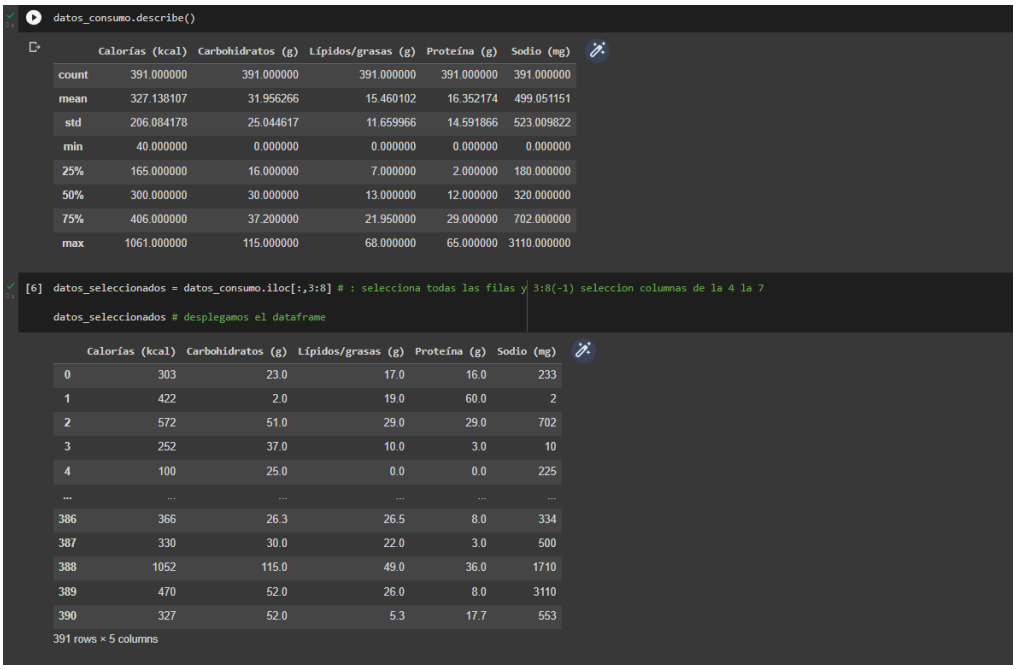


```
[1] import pandas as pd # importa la librería pandas y la asigna a la variable pd

[2] datos_consumo = pd.read_excel('AvanceComidas.xlsx') # indicamos el nombre de nuestro archivo a ser leído

datos_consumo.head()
```

	Fecha (dd/mm/aa)	Momento	Nombre alimento	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)	Fuente
0	2022-08-17	Desayuno	Huevo con tortilla	303	23.0	17.0	16.0	233	MyFitnessPal
1	2022-08-17	Comida	Pollo asado	422	2.0	19.0	60.0	2	MyFitnessPal
2	2022-08-17	Cena	Quesadillas con frijoles	572	51.0	29.0	29.0	702	MyFitnessPal
3	2022-08-17	Snack	Nito	252	37.0	10.0	3.0	10	MyFitnessPal
4	2022-08-17	Snack	5 Picafresas	100	25.0	0.0	0.0	225	MyFitnessPal

```
[4] datos_consumo.groupby("Momento").count() # con la función groupby agrupamos los datos de la columna Momento y con count() los contamos para obtener subtotales
```

Momento	Fecha (dd/mm/aa)	Nombre alimento	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)	Fuente
Cena	78	78	78	78	78	78	78	78
Comida	78	78	78	78	78	78	78	78
Desayuno	76	76	76	76	76	76	76	76
Desyuno	2	2	2	2	2	2	2	2
Snack	157	157	157	157	157	157	157	157



```
datos_consumo.describe()
```

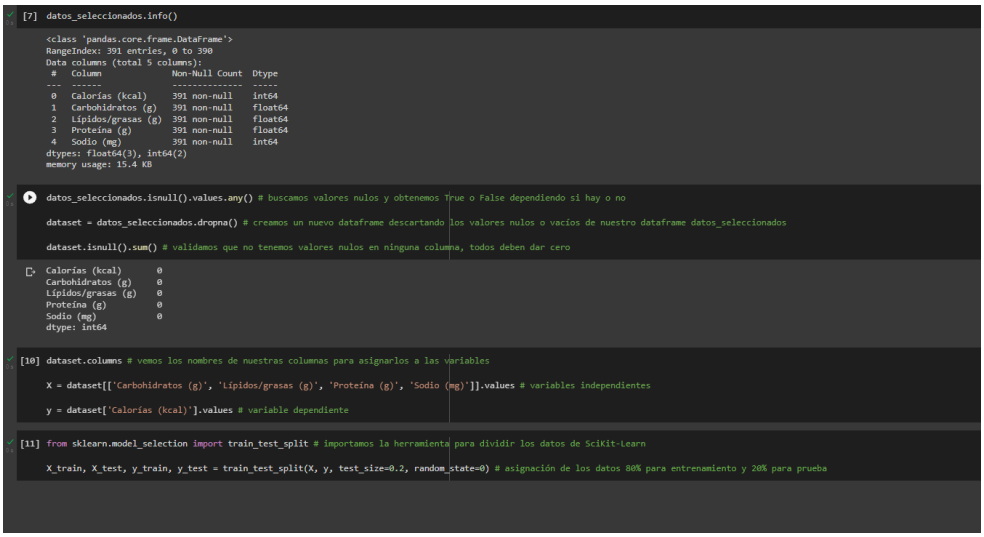
	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)
count	391.000000	391.000000	391.000000	391.000000	391.000000
mean	327.138107	31.956266	15.460102	16.352174	499.051151
std	206.084178	25.044617	11.659966	14.591866	523.009822
min	40.000000	0.000000	0.000000	0.000000	0.000000
25%	165.000000	16.000000	7.000000	2.000000	180.000000
50%	300.000000	30.000000	13.000000	12.000000	320.000000
75%	406.000000	37.200000	21.950000	29.000000	702.000000
max	1061.000000	115.000000	68.000000	65.000000	3110.000000

```
[6] datos_seleccionados = datos_consumo.iloc[:,3:8] # : selecciona todas las filas y 3:8(-1) seleccion columnas de la 4 la 7

datos_seleccionados # desplegamos el dataframe
```

	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)
0	303	23.0	17.0	16.0	233
1	422	2.0	19.0	60.0	2
2	572	51.0	29.0	29.0	702
3	252	37.0	10.0	3.0	10
4	100	25.0	0.0	0.0	225
...	...	...	...	...	...
386	366	26.3	26.5	8.0	334
387	330	30.0	22.0	3.0	500
388	1052	115.0	49.0	36.0	1710
389	470	52.0	26.0	8.0	3110
390	327	52.0	5.3	17.7	553

391 rows × 5 columns

Héctor Gutierrez
A01253031
7 de Noviembre del 2022



```
[7] datos_seleccionados.info()

    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 391 entries, 0 to 390
    Data columns (total 5 columns):
     #   Column             Non-Null Count  Dtype
    ---  ------             --------------  -----
     0   Calorías (kcal)    391 non-null    int64
     1   Carbohidratos (g)  391 non-null    float64
     2   Lípidos/grasas (g) 391 non-null    float64
     3   Proteína (g)       391 non-null    float64
     4   Sodio (mg)         391 non-null    int64
    dtypes: float64(3), int64(2)
    memory usage: 15.4 KB

    datos_seleccionados.isnull().values.any() # buscamos valores nulos y obtenemos True o False dependiendo si hay o no

    dataset = datos_seleccionados.dropna() # creamos un nuevo dataframe descartando los valores nulos o vacíos de nuestro dataframe datos_seleccionados

    dataset.isnull().sum() # validamos que no tenemos valores nulos en ninguna columna, todos deben dar cero

    Calorías (kcal)      0
    Carbohidratos (g)    0
    Lípidos/grasas (g)   0
    Proteína (g)         0
    Sodio (mg)           0
    dtype: int64

[10] dataset.columns # vemos los nombres de nuestras columnas para asignarlos a las variables

    X = dataset[['Carbohidratos (g)', 'Lípidos/grasas (g)', 'Proteína (g)', 'Sodio (mg)']].values # variables independientes

    y = dataset['Calorías (kcal)'].values # variable dependiente

[11] from sklearn.model_selection import train_test_split # importamos la herramienta para dividir los datos de SciKit-Learn

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0) # asignación de los datos 80% para entrenamiento y 20% para prueba
```

Básicamente lo que hice en Google Colab fue seleccionar los datos que solo analizaremos para después ser limpiados usando varias funciones, que con estas nos podemos dar cuenta por ejemplo si hay datos con valores nulos. Después prepare los datos asignándole variables a mis datos X y Y. Ya por último se dividió los datos unos para el aprendizaje automático y los otros para la validación.

Parte 2: Preparación de los datos

La fase 3 consiste en la preparación de datos, que se estima que es la parte mas importante, ya que toma del 50 al 70% del tiempo y el esfuerzo del proyecto. Este mismo involucra las siguientes tareas: Selección de datos, Limpieza de datos, Generación de nuevos datos, Integración de datos, Formato de los datos y por último División de datos en conjuntos para entrenamiento y prueba.

Dentro de lo que es selección de datos es básicamente seleccionar los datos relevantes que se van a analizar, existen dos tipos de seleccionar datos. Después seguimos con la limpieza de datos que aquí implica un análisis mas detallado de los problemas en los datos que se han elegido. La generación de nuevos datos se

utiliza en casos que se tenga que crear una nueva columna que nos ayude a entender mas los datos. La integración de datos es fusionar los datos que tengan diferentes múltiples de fuentes para que puedan ser más fácil analizarlos y tener un resultado más amplio, hay 2 métodos básicos de integración de datos. Ya para terminar el formato de datos viene siendo ordenar los datos de alguna forma para después ejecutar el modelo, esto para poder ahorrar tiempo de procesamiento antes de modelar.

1. ¿Qué datos hay que seleccionar? Por qué.

   Los datos que se tienen que seleccionar son los nutrientes (calorías, carbohidratos, lípidos/ grasas, proteínas y sodio), que son 5 columnas y 391 filas en total. Se seleccionaron estos porque son los datos mas relevantes y son los que mas nos brinda información.

2. ¿Hay que eliminar o reemplazar valores en blanco? Sí / No / Por qué.

   No, gracias a las siguientes funciones:

   - datos_seleccionados.isnull().values.any()

   - dataset = datos_seleccionados.dropna()

   - dataset.isnull().sum()

   Que fueron utilizadas en Google Colab pude observar que no existe ningún valor nulo, ya que en la primera función no me mostro ningún True demostrando de que si exista un valor nulo. Después se creo un nuevo dataframe con los mismos datos seleccionados y volvió a validar si había valores nulos y todos dieron 0, por lo cual no existe ninguno.

Héctor Gutierrez
A01253031
7 de Noviembre del 2022

3. ¿Es posible agregar más datos? Sí / No / Por qué.

   Si, yo considero que podría ser útil crear una nueva columna en la cual se cuenten las calorías que se quemaron en el día, ya que por lo general no siempre contamos con las calorías con las que ingerimos porque por lo general tendemos a hacer actividades físicas en las cuales perdemos calorías y nutrientes. Esto nos podría ayudar a tener un marco más claro del análisis que se está realizando.

4. ¿Hay qué integrar o fusionar datos de varias fuentes? Sí / No / Por qué.

   No, porque todos los datos fueron extraídos de la misma fuente, que es Myfittnespal. Gracias a esto no es necesario fusionarlos, por lo cual se puede analizar de una mejor manera y se puede tener un resultado mucho mas amplio. Esto fue considerado antes de realizar el análisis, ya que si se tuvo en cuenta de que la integración de datos puede volverse algo complejo si no se le dedica un buen tiempo para poder comprender muy bien los datos. En esta etapa se podría considerar que los datos mas relevantes son los nutrientes, pero si nos ponemos más específicos, son las calorías, ya que con esto yo considero que podría ser más fácil poder contestar preguntas o hipótesis del proyecto que se esta realizando.

5. ¿Es necesario ordenar los datos para el análisis? Sí / No / Por qué.

   Yo considero que no es necesario utilizar técnicas para realizar un formato u orden particular para los datos, porque el algoritmo que se tiene no considero que se requiera que los datos estén clasificados antes o después de ejecutar el modelo. Cabe recalcar que si se clasifican los datos me podría ayudar a ahorrar tiempo de procesamiento al hacer este paso antes de modelar, pero como se tienen pocos datos, yo considero que no es muy necesario este paso.

Héctor Gutierrez
A01253031
7 de Noviembre del 2022

6. ¿Tengo que hacer conjuntos de datos para entrenamiento y prueba? Sí / No / Por qué.

   Si, se asignaron 80% de los datos para entrenamiento y 20% para prueba, esto porque con los datos con los que partimos pueda entrenar a mi algoritmo, es decir, las características y las etiquetas. Mientras que por la parte de prueba, los datos que se seleccionaron se pueda comprobar si el modelo que se ha generado a partir de los datos de entrenamiento funciona o también si las respuestas predichas por el modelo para un caso totalmente nuevo son acertadas o no.

7. ¿Qué ajustes se tuvieron que hacer a los datos (agregar, integrar, modificar registros (filas), cambiar atributos (columnas)?

   No se realizó ningún ajuste a los datos, mas que eliminar columnas, que fueron la de la fecha, momento, nombre del alimento y la fuente, por el momento solo se dejaron los datos de los nutrientes, ya que como se había mencionado anteriormente, son los que yo considero que son los datos mas relevantes a mi parecer.

Link:
https://colab.research.google.com/drive/1s9B1PHSc1qafyQad2bGbGQjRvqxuXAzV?usp=sharing