

Héctor Gutierrez

A01253031

12/10/2022

Fase 2: Entendimiento de los datos

Parte 1: Cargando mis datos en Python

```
import pandas as pd
datos_consumo = pd.read_excel("Comidas.xlsx")
datos_consumo.head()
datos_consumo.shape
datos_consumo.columns
```

```
[1] import pandas as pd

[2] datos_consumo = pd.read_excel("Comidas.xlsx")

[3] 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 Picaftesas            100             25.0              0.0            0.0        225  MyFitnessPal

[4] datos_consumo.shape

(286, 9)

[5] datos_consumo.columns

Index(['Fecha (dd/mm/aa)', 'Momento', 'Nombre alimento', 'Calorías (kcal)',
      'Carbohidratos (g)', 'Lípidos/grasas (g)', 'Proteína (g)', 'Sodio (mg)',
      'Fuente'],
      dtype='object')
```

```
datos_consumo.dtypes
datos_consumo.info()
```

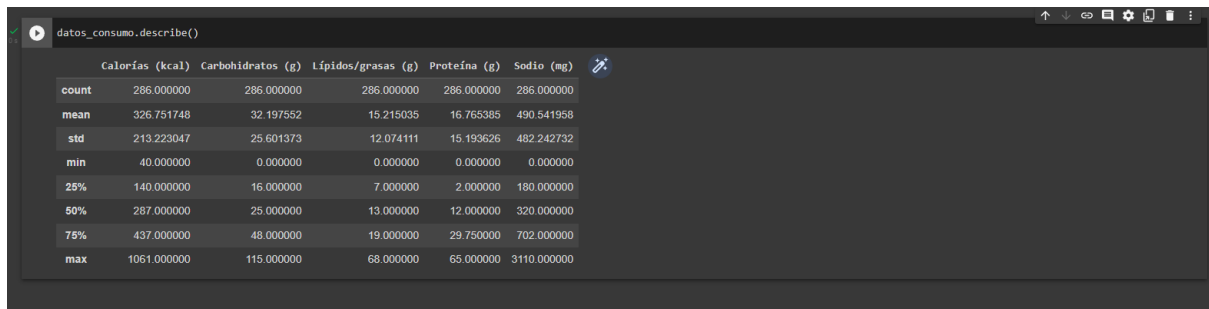
```
[6] datos_consumo.dtypes

Fecha (dd/mm/aa)    datetime64[ns]
Momento              object
Nombre alimento      object
Calorías (kcal)      int64
Carbohidratos (g)    float64
Lípidos/grasas (g)   float64
Proteína (g)         float64
Sodio (mg)           int64
Fuente              object
dtype: object

[7] datos_consumo.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 286 entries, 0 to 285
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype  
---  --
 0   Fecha (dd/mm/aa)      286 non-null   datetime64[ns]
 1   Momento                286 non-null   object  
 2   Nombre alimento        286 non-null   object  
 3   Calorías (kcal)        286 non-null   int64   
 4   Carbohidratos (g)      286 non-null   float64  
 5   Lípidos/grasas (g)     286 non-null   float64  
 6   Proteína (g)           286 non-null   float64  
 7   Sodio (mg)             286 non-null   int64   
 8   Fuente                 286 non-null   object  
dtypes: datetime64[ns](1), float64(3), int64(2), object(3)
memory usage: 20.2+ KB
```

```
datos_consumo.describe()
```



The screenshot shows a Jupyter Notebook interface with a dark theme. The top bar indicates the file name 'datos_consumo.describe()'. Below the notebook title, a table displays the statistical summary of the dataset. The table has five columns representing different nutritional components: Calorías (kcal), Carbohidratos (g), Lípidos/grasas (g), Proteína (g), and Sodio (mg). The rows represent various statistical measures: count, mean, std, min, 25%, 50%, 75%, and max.

| | Calorías (kcal) | Carbohidratos (g) | Lípidos/grasas (g) | Proteína (g) | Sodio (mg) |
|-------|-----------------|-------------------|--------------------|--------------|-------------|
| count | 286.000000 | 286.000000 | 286.000000 | 286.000000 | 286.000000 |
| mean | 326.751748 | 32.197552 | 15.215035 | 16.765385 | 490.541958 |
| std | 213.223047 | 25.601373 | 12.074111 | 15.193625 | 482.242732 |
| min | 40.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 140.000000 | 16.000000 | 7.000000 | 2.000000 | 180.000000 |
| 50% | 287.000000 | 25.000000 | 13.000000 | 12.000000 | 320.000000 |
| 75% | 437.000000 | 48.000000 | 19.000000 | 29.750000 | 702.000000 |
| max | 1061.000000 | 115.000000 | 68.000000 | 65.000000 | 3110.000000 |

Aquí primeramente lo que hace el código es importar la librería Pandas y se le asigna la variable `pd`, luego se crea la variable `datos_consumo` para cargar el archivo de Excel, indicando el nombre que se le puso en este caso fue “Comidas”. Después se utilizó la función `head` para mostrar los primeros datos del archivo para verificar que todo se cargo correctamente. Luego de ahí pase con la función `shape` para conocer la forma, total de filas y columnas de los datos. También utilice el método `columns` para saber todos los nombres de las columnas del Excel y `dtypes` para conocer los tipos de datos. La función `info()` me ayudo para conocer todos los datos y por ultimo `describe()` para conocer los datos como el promedio entre otros.

Parte 2: Describiendo mis datos

En resumidas cuentas, lo que nos dice es que la recolección de datos es una parte esencial de la ciencia de datos, ya que esto es lo que le da valor al análisis que se quiere realizar, tenemos 3 tipos de datos, existentes, que son los datos que se tienen en este caso por ejemplo mi peso, edad y altura, luego le siguen los datos adquiridos, que son los que se encuentran en Excel, los carbohidratos, proteínas y así. Por ultimo los datos adicionales son los que como el nombre lo dice son adicionales y en algún momento nos podrían ayudar, como lo pueden ser encuestas, revistas, etc.

Hay casos en el cual no se pueden analizar bien los datos porque se producen errores, los errores más comunes son los que están en blanco o que no tienen una respuesta concreta, errores tipográficos, errores de medición como lo puede ser que se ingresa cm en algo que pide Celsius y así hay muchos de los errores que se pueden cometer. Entre más datos se obtiene más preciso el modelo puede llegar a

ser, dentro de estos datos existen de tipo numérico, categórico o booleano. Después sigue analizar los datos y se realiza una hipótesis que ayude a dar forma en la transformación de los datos.

La presentación de los datos en el cual se visualice en realidad cual es el problema conducirá a decisiones mas claras y concisas. Los gráficos circulares o de dona sirven mucho para las comparaciones parciales, el diagrama de Sankey es una forma muy visual de poder determinar el flujo de volumen de información.

1. ¿Cuáles son tus datos existentes (registrados), datos adquiridos (datos externos) y datos adicionales (datos generados)?

Mis datos existentes son los datos que se tienen en este caso por ejemplo mi peso, edad y altura. Los datos adquiridos son los que se encuentran en Excel, los carbohidratos, proteínas entre otros. Y por último los datos adicionales son los que como el nombre lo dice son adicionales y en algún momento nos podrían ayudar, como lo pueden ser encuestas, revistas, etc.

2. ¿Qué tipos de datos se analizarán?

Los tipos de datos que se analizaran son de tipo continuo de proporción, ya que los datos son numéricos como por ejemplo la altura y el peso. Además, un valor cero que es la ausencia de la característica y la razón entre dos valores resulta significativa.

3. ¿Qué atributos (columnas) de la base de datos parecen más prometedores?

Yo considero que los datos que son mas prometedores son los nutrientes, ya que de ahí fue donde se realizó el análisis. Los nutrientes son los datos que mas peso tiene sin dejar afuera a las calorías que también son importantes, pero a mi parecer los nutrientes son los datos mas valiosos por el hecho de que en base a eso nos puede mostrar una respuesta mas

detallada, ya que con la ayuda de los nutrientes se puede sacar las calorías, como se había realizado en un ejercicio pasado.

4. ¿Qué atributos parecen irrelevantes y pueden ser excluidos?

Los atributos que son irrelevantes podrían ser la fuente de donde se realizó la extracción de los datos, ya que de ahí se hizo toda la extracción.

5. ¿Hay datos suficientes (filas) para sacar conclusiones generalizables o hacer predicciones precisas?

Ya que se tienen muy pocos datos, no creo que se puedan sacar conclusiones generalizables o predicciones precisas porque lo recomendable fue de obtener 300 filas, por lo cual me quede muy corto.

6. ¿Hay demasiados atributos para realizar un modelo que sea fácil de interpretar?

No creo que existan demasiado atributos para poder realizar un modelo fácil de interpretarlos. Los atributos son fáciles de interpretar, ya que son claros y concisos, a lo mejor la fecha podría ser algo difícil de interpretar, pero fuera de ahí ya no.

7. ¿De dónde se obtuvieron los datos? ¿Se están fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema al fusionar?

Los datos se obtuvieron de la aplicación myfitnesspal, que es una aplicación que la utiliza mucha gente que tiene una dieta estricta y necesita tener un registro de los nutrientes que ingiere cada día. Ya que solo se utilizó esa fuente no se combinó con otra.

8. ¿Hay algún plan para manejar los valores faltantes en cada una de las fuentes de datos?

El plan que se podría ejecutar en caso de haya un valor faltante podría ser asignándole 0, ya que la mayoría de los valores son numéricos y

también existen alimentos que en veces no contiene ciertos tipos de nutrientes como el agua.

9. ¿Cuántos datos están accesibles o disponibles y cómo está la calidad de los mismos?

Gracias a la función shape se tienen 2,288 datos disponibles. Los datos accesibles vienen siendo la fecha, momento, nombre del alimento, calorías, carbohidratos, lípidos/grasas, proteína y sodio. Dentro de las filas en cada columna se pueden ver el número de nutrientes que contiene cada alimento.

10. ¿Cuál es la relación de los datos y la hipótesis del proyecto?

La relación que existe entre los datos y la hipótesis del proyecto es que en base los datos que se obtuvieron se podría de alguna manera comprobar si estoy ingiriendo los nutrientes suficientes con base a los datos existentes que se tienen.

Google Colab: <https://colab.research.google.com/drive/1JPOFrd-95MBw5eyqZqiiW3VUO1wGO7p3?usp=sharing>