



Alumna: Rosa Vanessa Palacios Beltrán

Matricula: A01652612

Fecha de elaboración: 2 de Junio de 2020

Evidencia 2. Proyecto de Ciencia de Datos

Materia: Matemáticas y ciencia de datos
para la toma de decisiones (Gpo. 803)

Maestros: Ruth Josefina Sánchez Zambrano

Introducción

La ciencia de datos es una interpretación de datos para el entendimiento humano. Se encarga de transformar los datos en información y al mismo tiempo que la información obtenida se convierta en conocimiento. En campos interdisciplinarios como el análisis de negocios que incorporan la información, el modelo, las estadísticas, la analítica, y las matemáticas en un solo proceso, así es como la ciencia de datos trabaja. La información que se recolecta en algunos casos es muy detallada y su volumen es muy extensa.

Todos los días generamos información, estos datos en ciertas situaciones las empresas aprovechan la información tomándola para mejorar sus servicios o aumentar las ventas, esa es una forma de uso. Otras maneras que se le ha utilizado a la ciencia de datos es la forma de sacar ventajas con los datos o mejorar áreas como la salud, el cambio climático o en algo simple como en la toma de mejores decisiones todo esto basado en evidencias estadísticas y con la posibilidad de hacer predicciones futuras.

La intención de este proyecto es la aplicación de la ciencia de datos en los consumos alimenticios que he registrado durante estas semanas, y para que con mis datos tenga la información y a su vez el conocimiento para tomar una decisión sobre un cambio de mi persona en esta caso sobre mi consumo de alimentos y los nutrientes de estos. El proyecto está basado en la toma de una mejor decisión de nuestro consumo de alimentos ya que México está enfrentado un grado de obesidad por la alfabetización nutrimental, al realizar una interpretación de la información con ciencia de datos podemos hacer predicciones futuras y ver que cambios se pueden generar para una vida saludable y mitigar la obesidad.

Fase 1: Entendimiento del negocio

1. ¿Quién es el cliente?

En el caso de este proyecto yo misma soy el cliente ya que son mis datos por lo tanto mis resultados.

2. ¿Qué problemas estás tratando de resolver?

El problema ha resolver nace de que se esta enfrentando la obesidad en México con alfabetización nutrimental, y mediante la ciencia de datos poder tomar una decisión con base a si el consumos de cierta cantidad calórica, puedo tener cambios en mi masa corporal (peso) en un determinado tiempo.

3. ¿Qué solución o soluciones la Ciencia de Datos tratará de proveer?

La ciencia de datos con los análisis y estadísticas puede agilizar y facilitar el análisis de información, predicciones a futuros como enfermedades un ejemplo ataques cardiacos y así podría solucionar o prevenir enfermedades futuras de una persona.

En el caso del proyecto podemos prevenir enfermedades o predecir nuestro estado de salud conforme nuestro consumo alimenticio.

4. ¿Qué necesitas aprender para poder desarrollar la solución o soluciones?

Lo que necesito aprender para el desarrollo de la solución de problema es sobre la situación del problema en este caso es el enfriamiento hacia la obesidad en México con alfabetización nutrimental, saber sobre en que situación estamos, la solución para mitigar la obesidad implementado la ciencia de datos para generar consciencia sobre el consumo de alimentos, y en base a los datos tomar una solución por nuestra salud.

5. ¿Qué deberás hacer para desarrollar tu solución?

Lo primero que se debe hacer es entender los datos: la recolección de los datos, descripción de los datos, exploración de los datos y verificación de la calidad de los datos. Luego preparar los datos: selección de los datos, limpieza de los datos (entendiéndose como limpieza el proceso para eliminar errores en la información o excluir o remplazar datos necesarios), construcción de los datos (entendiéndose como la inclusión o agregación

de datos importantes para esta fase) e integración de los datos.

Después la modelación de los datos: selección de la técnica de modelación, diseño de pruebas, construcción del modelo y evaluación del modelo. Por ultimo tendremos que desarrollar una evolución y despliegue de los resultados.

Fase 2: Entendimiento de los datos

1. ¿Cuáles son tus datos existentes, datos adquiridos y datos adicionales?

Los datos resultan de una variedad de fuentes 1) datos existentes esto quiere decir que hay una gran variedad de datos, ejemplos: los datos transaccionales, datos de encuestas, registro web, entre otros. Los datos existentes son competentes para satisfacer las necesidades requeridas. 2) datos adquiridos, involucra la posibilidad de poder a completar los datos con información externa con el propósito de enriquecer el análisis, por ejemplos: datos demográficos, datos de la competencia, datos de fuentes del gobierno. 3) datos adicionales si la información adquirida con los datos anteriores no es suficiente, la organización debe de generar información para a completar y hacer un análisis más efectivo, ejemplos: encuestas, entrevistas, acceso a información de redes sociales, etc.

1) datos existentes: nombre, edad, peso, estatura, calorías, carbohidratos, lípidos/grasas, proteína, sodio

2) datos adquiridos: El IMC con la calculadora que hicimos

3) datos adicionales: tal vez cuando veamos la regresión necesites datos

2. ¿Qué tipos de datos se analizarán?

Se analizaran tipos de datos continuo de proporción, valores cuantitativos.

El Análisis Exploratorio de Datos (EDA) el cual tiene como objetivo adquirir información sobre los datos. En la Fase 2 tiene como intención de explorar todos los datos, para obtener una mejor comprensión del negocio, así se podrá realizar una hipótesis que ayude en el proceso de moderación o para

la transformación de los datos.

La recolección de los datos, descripción, análisis exploratorio y auditoría.

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

Por nuestras bases de datos de alimentos los datos son: las calorías que es lo que nos indica como puede cambiar nuestra masa en cierto tiempo.

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

Uno de los datos que podría ser excluido por que a partir del análisis de regresión el sodio aporta calorías, pero si el consumo de este es muy elevado puede traer problemas de salud. Otro podría ser el momento del consumo del alimento si es el desayuno, snack, comida o cena, incluso la fecha parece irrelevante.

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

Actualmente estoy en la fila numero 327 y entre más datos tengamos mayor presión de los datos, pero hay que asegurar que los datos sean confiables.

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

No, nuestra base de datos tiene 8 columnas(no es muy grande) no afecta mucha porque a comparación de otras bases de datos donde los atributos son una cantidad mayor se vuelve complicado hacer la interpretación.

7. ¿De dónde se obtuvieron los datos?

El consumo de los alimentos (calorías) por medio de las etiquetas de los producto, myfitnesspal y calculadora nutricional.

¿Se están fusionando varias fuentes de datos?

Si es así, con las etiquetas de los productos, myfitnesspal y calculadora nutricional dependiendo del caso

¿Hay áreas que podrían plantear un problema al fusionar?

Si que no siempre las diferentes fuentes den el mismo dato y puede traer una variación de calorías

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

En caso de que el producto o alimento consumido no tenga algún dato de los atributos solicitados, lo manejo de dos maneras: el promedio de este dato o poniendo como 0 dependiendo el caso. De ser necesario se puede eliminar el registro.

9. ¿Los datos a usar son adecuados para hacer el análisis?

Si son adecuados porque es lo que nosotros definimos para solucionar la hipótesis

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

Actualmente tengo 327 filas en mi base de datos de cada atributo con la recolección de 1962 datos, y la calidad de aun no está muy clara pero hasta el momento no me ha generado algún error.

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

La relación de los datos con la hipótesis que quiero resolver es que si puede haber cambios en mi peso dependiendo mi consumo de calorías en el tiempo determinado

Fase 3: Preparación de datos

1. ¿Tengo que fusionar mis datos? Sí, aunque esto depende de la organización y los objetivos, pero en mi opinión si tenemos que fusionar los

datos para que al analizarlos se conecten y el resultado final sea mas amplio. Ya que al fusionar implicaría que juntaríamos dos o mas tablas o fuentes de información.

2. ¿Tengo que hacer un subconjunto? Sí, igual que la pregunta anterior depende del objetivo y las necesidades, un subconjunto de datos lo usaría para dos conjuntos de datos y tener una forma seccionada de cómo se muestran los datos. Los subconjuntos serían el escoger por horario de cena o un subgrupo.
3. ¿Podrías agregar más datos? Si, porque al agregar datos implica tener que integrar dos o mas conjuntos de datos con atributos similares, registros diferentes. Los datos se integran en función de campos similares (como el nombre del producto o la duración del contrato). En este caso entre mas datos se tengan el modelo será mas confiable pero con los datos que tenemos podemos hacerlo.
4. ¿Necesitas hacer cambios en los atributos de los datos? Sí, en mi registro nutricional que creo los únicos datos que agregaría serian mas días para que los resultados sean mas amplios y precisos, en cuestión de atributos creo que es muy completo todo lo que hemos hecho aunque seria incluir información adicional no incluir mas datos, y así poder saber cuanto es el consumo semanal y que cada semana veamos el proceso para compara
5. ¿Es necesario ordenar los datos para el análisis? Sí , porque el orden de los datos a la hora de analizar seria muy complicado y tardado generar un código, tener limpieza en los datos implica un análisis mas detallado y algunos problemas serian al no hacerlo así serian: datos perdidos, errores en los datos, inconsistencias en la codificación de los datos y metadatos faltantes o malos.

6. ¿Hay que eliminar o reemplazar valores en blanco? Si, creo que en este punto eliminar o reemplazar valores en blanco podría volverse mas fácil, obtener los resultado de una forma simple pero a la vez completa, es necesario eliminar valores en blanco para que nuestro análisis se pueda realizar y no marque errores.
7. ¿Qué ajustes se tuvieron que hacer a los datos (agregar datos, integrar datos, modificar datos (remover o eliminar información)?
Agregar datos: Cada día agregar lo que comíamos y de igual manera poner datos alimenticios como las calorías, carbohidratos, sodio, lípido y proteínas.
Integrar datos: Al hacer las comprobaciones y los códigos realizados durante las actividades semanales fue necesario juntar los datos
Modificar datos: Remover datos si tuve que ajustar por ejemplo para hacer esta actividad la primera parte solo necesitaba los datos de la tabla de registro nutricional y no los analices que previamente hice.

Fase 4: Modelación de datos

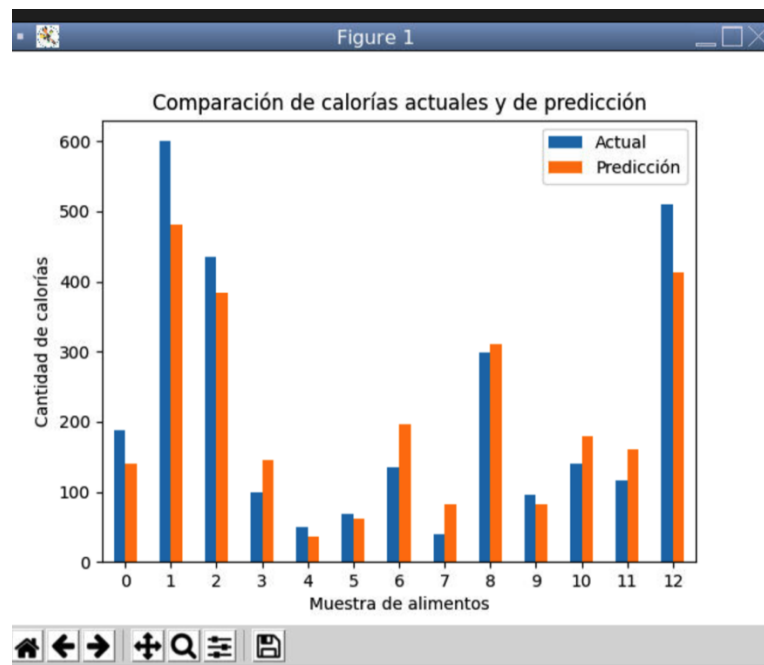
1. ¿Cuántas iteracciones/intentos realizaste para obtener el resultado?
Tuve que hacer 6 inténenos para obtener el gráfico con la comparación de las calorías actuales y las de predicciones.
2. ¿Qué problemas se presentaron y cómo los resolviste?
El problema de regresión que hubo fue de la variable de resultado que este caso respuesta Y, por otra parte la variable predictor o factor de riego la cual fue X quien afecta a Y, también las variables de carácter en la comunidad para el aprendizaje automático.
3. ¿Qué resultados arrojó el análisis?
Los datos del grafico hicieron una comparación de las calorías actuales y las calorías de predicciones, mostrando la cantidad de calorías y la muestra

de alimentos, observe que la comparación entre la actual y la de predicción aunque medio cercanas nunca fue exacta o igualitaria, obtuve una desviación media al cuadrado de 55.37

4. ¿Qué aprendiste y cuáles son tus conclusiones de la modelación?

Lo que aprendí del modelo matemático fue que su función en cuestión de preparar los datos con forme a las características, hacer la modelación ayuda a obtener una visualización de los datos, pero hay que saber que modelación es la correcta realizar entorno a la relación, composición, comparación y distribución, al hacer el modelo podemos hacer detalle en el modelo estadístico que ayudara a resolver la hipótesis y hacer predicciones el cual es nuestro acoso predijo el la cantidad de calorías aunque no fue exacta pero si cercana al resultado, el modelo recogió los datos he hizo un proceso mediante formulas matemáticas.

5. Incluye captura de los resultados obtenidos en esta fase. Estadísticas, predicciones e imagen de la gráfica obtenida:



	Actual	Predicción		
0	187	140.829992		
1	600	480.715206		
2	435	384.263539		
3	100	146.074754		
4	50	35.980199		
5	69	61.029561		
6	135	195.892596		
7	40	82.816282		
8	298	309.746172		
9	96	82.816282		
10	140	178.926842		
11	117	160.218868		
12	510	413.523395		

[8 rows x 5 columns]

Coeficientes

Carbohidratos (g)

5.185956

Lípidos/grasas (g)

4.098372

Proteína (g)

-0.247439

Sodio (mg)

0.084962

Efecto del consumo calórico en el tiempo

1. Mi código para realizar la estimación con los datos que registre:

```
import pandas as pd # importando pandas y asignarla a la variable
pd

import numpy as np # importando numpy y la asignamos a la variable
np

dataset = pd.read_csv('RegistroNutricional - Sheet1.csv') #
indicamos el nombre de nuestro archivo a ser leído
datos_consumo = dataset.iloc[:,3] # : indica todas las filas y 3
indica la columna de las calorías
print(datos_consumo) # imprimiendo los datos seleccionados
suma_calorias = np.sum(datos_consumo) # suma los datos
print("Total de calorías consumidas:", suma_calorias) # imprime el
total de calorías
dias = int(input("Ingresa el total de días de consumo: "))

peso = int(input("Ingresa tu peso en kilogramos: "))

altura = int(input("Ingresa tu altura en centímetros: "))

edad = int(input("Ingresa tu edad en años: "))
```

```

genero = input("Ingresa tu género, Mujer/Hombre: ")

calorias_promedio = suma_calorias/dias # total de calorías
consumidas entre el número de días que tomó consumirlas
print("Tu promedio de calorías consumidas en", dias,"días es:",
calorias_promedio)
if(genero == "Mujer"):
    calorias_requeridas = 655+(9.56*peso)+(1.85*altura)-(4.68*edad)
# fórmula para estimar calorías requeridas en mujer

elif(genero == "Hombre"):
    calorias_requeridas = 66.5+(13.75*peso)+(5*altura)-(6.8*edad) #
fórmula para estimar calorías requeridas en hombre

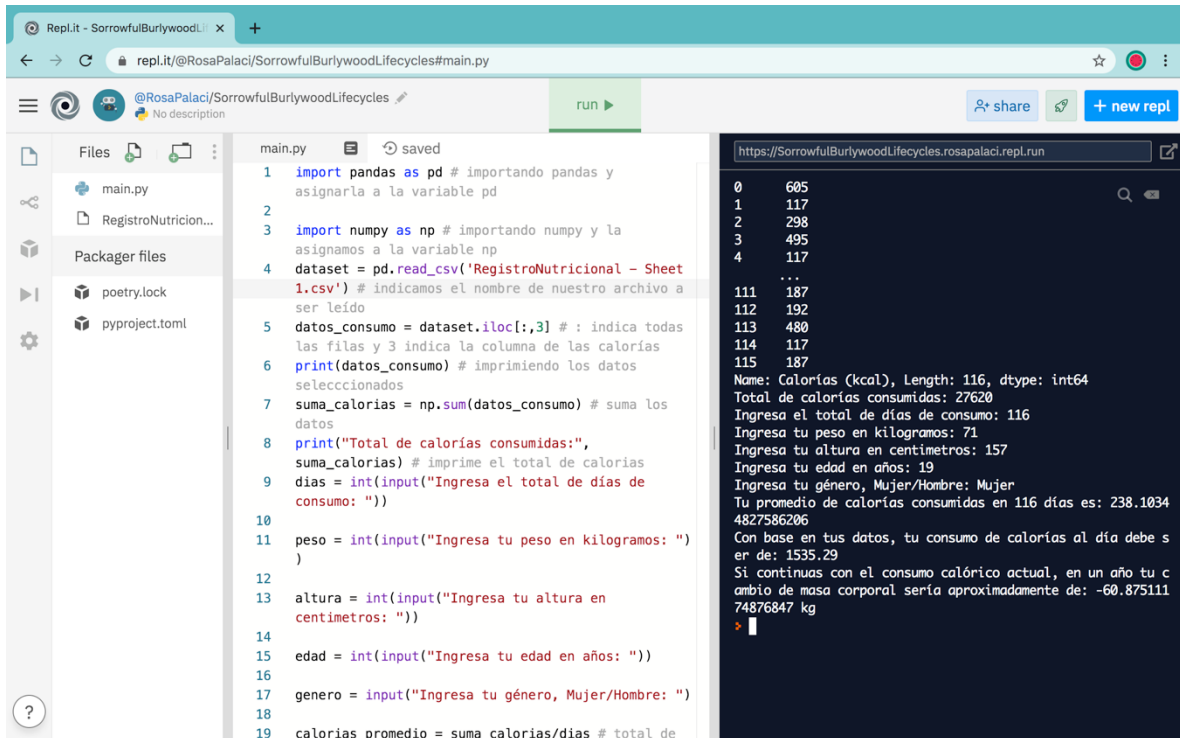
print("Con base en tus datos, tu consumo de calorías al día debe
ser de:", calorias_requeridas)
diferencia = calorias_promedio - calorias_requeridas
efecto_anual = diferencia * 450/3500 * 365 /1000 # realiza la
proporción, se multiplica por 365 (días) y se divide entre 1000
(gramos) para obtener kilogramos

print("Si continuas con el consumo calórico actual, en un año tu
cambio de masa corporal sería aproximadamente
de:",efecto_anual,"kg")

```

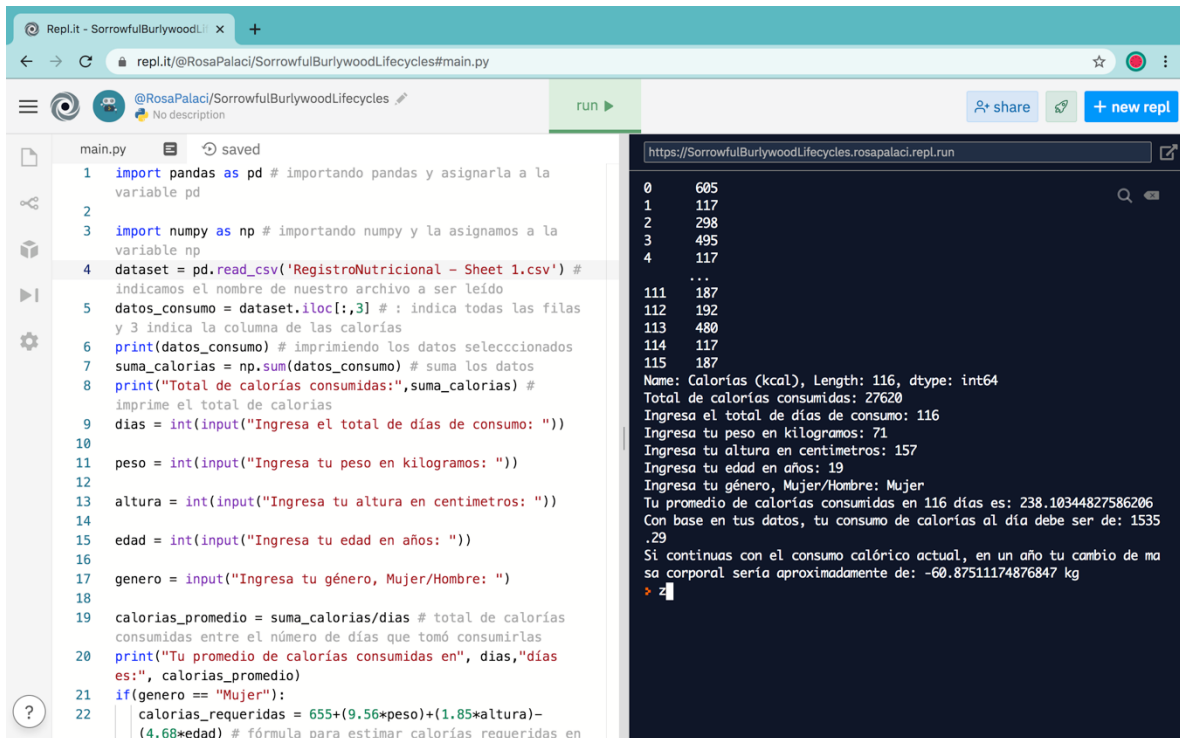
2. Las capturas de pantalla del programa mostrando los resultados:

Evidencia 2. Proyecto de Ciencia de Datos



```
1 import pandas as pd # importando pandas y
  asignarla a la variable pd
2
3 import numpy as np # importando numpy y la
  asignamos a la variable np
4 dataset = pd.read_csv('RegistroNutricional - Sheet
  1.csv') # indicamos el nombre de nuestro archivo a
  ser leído
5 datos_consumo = dataset.iloc[:,3] # : indica todas
  las filas y 3 indica la columna de las calorías
6 print(datos_consumo) # imprimiendo los datos
  seleccionados
7 suma_calorias = np.sum(datos_consumo) # suma los
  datos
8 print("Total de calorías consumidas:",
  suma_calorias) # imprime el total de calorías
9 dias = int(input("Ingresa el total de días de
  consumo: "))
10
11 peso = int(input("Ingresa tu peso en kilogramos: "))
12
13 altura = int(input("Ingresa tu altura en
  centímetros: "))
14
15 edad = int(input("Ingresa tu edad en años: "))
16
17 genero = input("Ingresa tu género, Mujer/Hombre: ")
18
19 calorias_promedio = suma_calorias/dias # total de
```

0 605
1 117
2 298
3 495
4 117
...
111 187
112 192
113 480
114 117
115 187
Name: Calorías (kcal), Length: 116, dtype: int64
Total de calorías consumidas: 27620
Ingresa el total de días de consumo: 116
Ingresa tu peso en kilogramos: 71
Ingresa tu altura en centímetros: 157
Ingresa tu edad en años: 19
Ingresa tu género, Mujer/Hombre: Mujer
Tu promedio de calorías consumidas en 116 días es: 238.1034
4827586206
Con base en tus datos, tu consumo de calorías al día debe ser de: 1535.29
Si continúas con el consumo calórico actual, en un año tu cambio de masa corporal sería aproximadamente de: -60.87511174876847 kg



```
1 import pandas as pd # importando pandas y asignarla a la
  variable pd
2
3 import numpy as np # importando numpy y la asignamos a la
  variable np
4 dataset = pd.read_csv('RegistroNutricional - Sheet 1.csv') #
  indicamos el nombre de nuestro archivo a ser leído
5 datos_consumo = dataset.iloc[:,3] # : indica todas las filas
  y 3 indica la columna de las calorías
6 print(datos_consumo) # imprimiendo los datos seleccionados
7 suma_calorias = np.sum(datos_consumo) # suma los datos
8 print("Total de calorías consumidas:", suma_calorias) #
  imprime el total de calorías
9 dias = int(input("Ingresa el total de días de consumo: "))
10
11 peso = int(input("Ingresa tu peso en kilogramos: "))
12
13 altura = int(input("Ingresa tu altura en centímetros: "))
14
15 edad = int(input("Ingresa tu edad en años: "))
16
17 genero = input("Ingresa tu género, Mujer/Hombre: ")
18
19 calorias_promedio = suma_calorias/dias # total de calorías
  consumidas entre el número de días que tomó consumirlas
20 print("Tu promedio de calorías consumidas en", dias, "días
  es:", calorias_promedio)
21 if genero == "Mujer":
22     calorias_requeridas = 655+(9.56*peso)+(1.85*altura)-
  (4.68*edad) # fórmula para estimar calorías requeridas en
```

0 605
1 117
2 298
3 495
4 117
...
111 187
112 192
113 480
114 117
115 187
Name: Calorías (kcal), Length: 116, dtype: int64
Total de calorías consumidas: 27620
Ingresa el total de días de consumo: 116
Ingresa tu peso en kilogramos: 71
Ingresa tu altura en centímetros: 157
Ingresa tu edad en años: 19
Ingresa tu género, Mujer/Hombre: Mujer
Tu promedio de calorías consumidas en 116 días es: 238.1034827586206
Con base en tus datos, tu consumo de calorías al día debe ser de: 1535.29
Si continúas con el consumo calórico actual, en un año tu cambio de masa corporal sería aproximadamente de: -60.87511174876847 kg

```

0      605
1      117
2      298
3      495
4      117
...
111     187
112     192
113     480
114     117
115     187
Name: Calorías (kcal), Length: 116, dtype: int64
Total de calorías consumidas: 27620
Ingresa el total de días de consumo: 116
Ingresa tu peso en kilogramos: 71
Ingresa tu altura en centímetros: 157
Ingresa tu edad en años: 19
Ingresa tu género, Mujer/Hombre: Mujer
Tu promedio de calorías consumidas en 116 días es: 238.10344827586206
Con base en tus datos, tu consumo de calorías al día debe ser de: 1535
.29
Si continuas con el consumo calórico actual, en un año tu cambio de ma
sa corporal sería aproximadamente de: -60.87511174876847 kg

```

- El proceso para estimar el efecto del consumo calórico en el tiempo, primero tenemos que obtener los datos de la bitácora de Excel pero solo el registro nutricional y exportarlo en CSV y cargar el archivo en repl.it una vez hecho esto, use las librerías pandas para cargar los datos en un DataFrame y la librería numpy para realizar operaciones con los datos. Importe pandas y le asigne la variable pd mientras que importe numpy con la variable np. Cree una variable llamada datos_consumo y le asigne el DataFrame que contiene los datos que leí, utilice iloc para los elementos de acuerdo a su ubicación. Llame la función para imprimir los datos seleccionados después utilice np.sum() de numpy para calcular el total de calorías consumidas, sumando los datos e imprimiendo el total de calorías.

Definí las variables para realizar el cálculo usando `input()` e `int()` para así habilitar la captura de datos e indicar que las variables son números enteros. Calculamos el promedio de calorías con el número total de calorías consumidas entre el número de días que tomo consumirlas. Use la ecuación de Harris-Benedict realice la estimación de calorías que necesitamos cada día según los datos.

4. Calcule la diferencia entre las calorías consumidas y las calorías requeridas. Por ultimo realice la proporción, multiplicando por 365 días y se divide entre 1000 gramos para obtener kilogramos.

Reflexión final (Conclusiones)

- La hipótesis inicial: ¿Si consumo cierta cantidad calórica, puedo tener cambios en mi masa corporal (peso) en un determinado tiempo? De acuerdo a mis resultados en la estimación fue un resultado negativo por lo cual significa que perdere mas en base a esto acepta la hipotesis pero hay factores que no tomamos en cuenta que podrian afectar algo el resultado final.
- ¿Qué procedimiento para realizar una regresión lineal te pareció mejor, el realizado en Excel en la semana 4 o en Python en la semana 15?
En mi opinión para realizar un regresión lineal con Python en la semana 15 es una mejor opción.
¿Por qué? Porque el trabajo realizado en Excel en la semana 4 pude observar si podemos obtener un modelo matemático para calcular las calorías de los alimentos en función de la cantidad de nutrientes que contiene, aunque veo muchas información pero en comparación con la actividad en Python en la semana 15 creo que hay una mejor organización y con mas facilidad de la obtención de información, es mas detallada.
- ¿Por qué es importante la Ciencia de Datos y la ética para el uso adecuado de los datos e información?
La ciencia de datos y la etica es importante porque asi podremos hacer un uso adecuado de los datos e información, ya que con la ciencia de datos podemos

abarcen todo tipo de técnicas para analizar datos a escala, los cuales se recopilan, se almacenan, se analizan y se interconectan con otros tipos de datos. La recopilación que se puede hacer, tiene evidentemente un impacto cultural, científico, político y económico según como usen esta información la ciencia, las empresas y los gobiernos. El impacto que tiene la ciencia de datos debe ser usada con la ética sobre su uso ya que puede haber una recopilación de datos sin consentimiento violando la privacidad del usuario. La responsabilidad sobre el manejo ético de información no sólo recae sobre los gobiernos, sino también sobre las empresas que se dedican a la ciencia de datos, y que deban plantearse tener una normativa.

Referencias

- Alison Bert. (2018). Tres maneras en que la ciencia de datos y la tecnología están ayudando a mejorar el mundo. 2/06/2020, de Elsevier Sitio web: <https://www.elsevier.com/es-es/connect/ciencia/tres-maneras-en-que-las-que-la-ciencia-de-datos-y-la-tecnologia-estan-ayudando-a-mejorar-el-mundo>
- Juan Pablo Mora. (2020). Que es la Ciencia de Datos, el aprendizaje automático (ML), el Big Data y cuales son sus usos?. 3/junio/2020, de Universidad Javeriana Sitio web: <https://www.javeriana.edu.co/documents/12847/10949798/Qu%C3%A9+es+la+ciencia+de+datos/aa1e64ec-0961-4ef9-a4e5-472fe5774c0b>
- Fatima Delgado M.. (2019). Los Desafíos Éticos de la Ciencia de Datos.. 06-06-2020, de Medium Sitio web: <https://medium.com/saturdays-ai/los-desaf%C3%ADos-%C3%A9ticos-de-la-ciencia-de-datos-25ce771d892e>