Tecnológico de Monterrey

- Evidencia 2: Proyecto de Ciencia de Datos
- Edgar Adolfo Ochoa Mendoza
- A00344644
- Matemáticas y ciencia de datos para la toma de decisiones (Gpo 800)
- 25/01/2022

Proyecto de Ciencia de Datos

Introducción

La ciencia de datos es una disciplina que combina múltiples campos como la estadística, la matemática, desarrollo de software, análisis de datos y el método científico para darle un valor e interpretación a un conjunto de datos.

Una forma de describir a un científico de datos es como una “Persona que sabe más de estadística que cualquier programador y que a la vez, sabe más de programación que cualquier estadístico” (Josh Wills, jefe del área de ingeniería de datos en Slack).

Esta ciencia se crea debió a la necesidad de analizar grandes volúmenes de datos derivados de la digitalización que estamos viviendo lo que conocemos como “big data”.
Esto abre la puerta para analizar y crear muchas soluciones novedosas en todos los campos. Por ejemplo, en la medicina la posibilidad de acumular una gran cantidad de imágenes médicas como radiografías, a permitido crear programas capaces de entrenarse con esas imágenes para detectar enfermedades mirando nuevas radiografías. En seguridad a permitido mejorar los algoritmos que detectan fraudes en nuestras cuentas de banco. Incluso en temas más vánales y cotidianos nos facilita la vida con algoritmos que te recomiendan películas y series basado en tu historial de vistas, rutas de tráfico más cortas, los mejores restaurantes y lugares turísticos de una ciudad.

Para este proyecto la ciencia de datos se enfoca en resolver uno de los mayores problemas de salud pública del país; la obesidad, México es el segundo país con más obesidad en el mundo por lo tanto este proyecto busca aportar por medio de la alfabetización nutrimental. Se realizará un análisis de datos para averiguar si existe relación entre la cantidad de calorías ingeridas y la masa corporal.

Proyecto de Ciencia de Datos

Fase 1: Entendimiento del negocio

Esta es una de las 6 fases de la metodología para análisis de datos; CRISP-DM (Cross-industry standard process for data mining por sus siglas en inglés). Consiste en identificar los objetivos del negocio, evaluar la situación, definir objetivos para el análisis de datos y desarrollar un plan de negocio.

Es decir, lo más importante ante cualquier proyecto es identificar el objetivo, pues con un objetivo claro también se hace más fácil identificar que se necesita para llegar a ese objetivo, para el caso de los proyectos de análisis de datos nos permite saber que datos son los que debemos reunir, porque medio lo haremos, que tratamiento recibirán y como vamos a procesarlos esto da como resultado un plan.
En particular el objetivo del proyecto de la materia es: “enfrentar la obesidad en México con alfabetización nutrimental”

Evaluando la situación notamos que somos el segundo país con más obesidad del mundo
Por lo que por medio del análisis de datos buscaremos si existe una relación existe entre el consumo de calorías y los cambios en mi masa corporal en un determinado tiempo

Pare esto será necesario registrar la información nutricional (carbohidratos, lípidos, proteínas, sodio, calorías) de los alimentos que consumo durante el tiempo que dure la materia y comparar mi masa corporal inicial con la final.

En nuestro proyecto

1. ¿Quién es el cliente?

Yo mismo soy el cliente por que el objetivo del Proyecto es “aplicar Ciencia de Datos a los consumos alimenticios que tú tienes en un lapso de 4 semanas para

poder predecir un cambio en tu persona.” Por lo tanto, debo conocer mi nombre, edad, actividad fisca, intensidad de la actividad, mi alimentación, etc.

2. ¿Qué problemas estás tratando de resolver?

México pertenece a las naciones con mayor obesidad en adultos en el mundo, esto de acuerdo con la Organización para la Cooperación y el Desarrollo.
Al año 2015 los países con mayor obesidad son: Estados Unidos con 38.2%, México con 32.4%, Nueva Zelanda con 30.7%.
Por lo tanto, este proyecto esta orientado ayudarme a entender como mi alimentación afecta mi masa corporal, por lo tanto, la hipótesis a Aceptar o rechazar es; si consumo 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 debe proveer una manera de analizar los datos que recolectare para buscar si existe o no relación entre las calorías consumidas y un cambio en la masa corporal.

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

A utilizar un sistema para registrar la información nutricional (carbohidratos, lípidos, proteínas, sodio, calorías) de los alimentos que consumo, aprender a utilizar herramientas de análisis de datos como las que proporciona Excel, aprender estadística para poder interpretar los resultados y generar un bueno modelo, también debo aprender programación para poder automatizar algunas tareas del proceso o resolver problemas más específicos que no se puedan hacer en otras herramientas como Excel.

5. ¿Qué deberás hacer para desarrollar tu solución?

Registrar la información nutricional (carbohidratos, lípidos, proteínas, sodio, calorías) de los alimentos que consumo, limpiar los datos, realizar un análisis de regresión para obtener un modelo que pueda predecir las calorías de un alimento basado en la cantidad de carbohidratos, lípidos, proteínas en gramos y sodio en mg. Interpretar los resultados, optimizar el modelo y programarlo en algún lenguaje como Python.

Fase 2: Describiendo mis datos

La fase 2 de la metodología para análisis de datos; CRISP-DM "entendimiento de los datos" tiene el objetivo de poder comprender la información que una organización posee para posteriormente determinar si es necesario hacer ajustes, por ejemplo, adquirir información de otras fuentes; definir si la información se puede utilizar para aplicar un proyecto de ciencia de datos e incluso modificar parte de esta información.

Las etapas de esta fase son:

• Recolección de los datos.

Hace referencia a la forma en que se obtiene los datos, estos pueden ser datos existentes, adquiridos o adicionales

• Descripción de los datos.

Para describir datos hay que enfocarse en dos aspectos fundamentales, la cantidad y la calidad de los datos. En general una mayor cantidad de datos genera mejores modelos, pero alarga el tiempo de procesamiento. Se debe tener en cuenta que no solo por ser muchos datos se puede asegurar un mejor modelo, es ahí donde entra la calidad de los datos.

• Exploración de los datos.

Esta parte sirve para tener un mejor entendimiento del negocio y formular hipótesis

• Verificación de la calidad de los datos

Es en esta etapa donde se comprueba la calidad de los datos ya que esta rara vez son perfectos, para esto existen diferentes técnicas y herramientas que indican el grado de calidad

En nuestro proyecto

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

   Los datos existentes son mi nombre, apellido, peso, edad, estatura.

Los datos adquiridos provienen de las aplicaciones y sitios web que utilice para llevar registro de mi consumo nutricional.

Los datos adicionales son encuestas, entrevistas, información de redes sociales, que se utilizan en caso de que los dos tipos de datos anteriores no sean suficientes. Hasta el momento en este proyecto no han sido necesarios.

2. ¿Qué tipos de datos se analizarán?

   Al analizar mi conjunto de datos en Python con la función “dtypes” se observa que estoy trabajando con datos categóricos como la variable “momento”, datos numéricos discretos como la cantidad calorías; y otros datos numéricos continuos como la cantidad de carbohidratos en gramos de cada alimento, la cantidad de lípidos, la cantidad de proteínas o la cantidad de mg de sodio en cada alimento

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

   Al analizar mi conjunto de datos en Python con la función “columns” se observan nueve columnas distintas de las cuales hasta el momento solo 4 han sido de interés para el objetivo del proyecto; la columna de calorías, y las columnas de carbohidratos, lípidos y proteína. La columna de sodio en un inicio parecía prometedora, pero tras realizar un análisis de regresión en los datos esta resulto ser no significativa

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

   La fecha, el momento, el nombre del alimento, la fuente y ahora sabemos que también el sodio, son variables que no tienen influencia directa en el análisis que se realizó sin embargo están ahí para cumplir otras funciones por ejemplo la fuente da legitimidad al origen de los datos, pero no tiene influencia en la masa corporal

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

   Hasta el momento cuento con un total de 64 datos en la base, estos han sido suficientes para crear un modelo que prediga la cantidad de calorías en un alimento basado en sus macronutrientes sin embargo aún son pocos

comparados con el objetivo y el modelo no ha sido lo suficientemente probado para poder obtener conclusiones certeras.

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

   Tras el análisis de regresión se encontró que 3 columnas son las significativas por lo que el modelo resultante es fácil de interpretar

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?

Se usan varias fuentes para obtener los datos para formar la base de datos, pero no se hace fusión de datos.

Las principales fuentes de datos en mi proyecto han sido dos aplicaciones móviles “MyFitnessPal” y “Fatsecret” y en ocasiones se han obtenido datos directamente de la etiquta de información nutricional de un alimento o de la página de su fabricante. En general esto ha funcionado bien, pero en ocasiones para la cantidad de sodio es necesario corroborar información en por lo menos dos fuentes pues la app “MyfitnessPal” muestra varios alimentos con 0 mg de sodio cuando en realidad el resto de las fuentes muestran que si tiene sodio. Por su parte “Fatsecret” por lo general muestra cantidades de sodio por encima de cualquier otra fuente.

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

   Si dado que no estoy realizando fusión de datos entre distintas fuentes, utilizo solo la que tenga la información completa del alimento y verifico que las cantidades sean similares en una tercera fuente de datos, si no es posible encontrar o verificar esta información tomo la decisión de no registrar ese alimento para evitar errores.

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

   La función “shape” en el conjunto de datos muestra 64 renglones y 9 columnas lo que da un total de 576 datos accesibles, en cuanto a su calidad aun no pasan por un debido proceso de limpieza de datos pero durante el análisis de regresión se detectaron y corrigieron algunas inconsistencias mayores

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

la hipótesis del proyecto es; si consumo cierta cantidad calórica, ¿puedo tener cambios en mi masa corporal (peso) en un determinado tiempo?, estos datos son justamente eso; un registro en el tiempo de lo que he consumido, eso combinado con los datos existentes iniciales como mi peso, edad y estatura. Nos permitirán aceptar o rechazar esta hipótesis cuando vuelva a medir mi masa corporal al final del proyecto.

Fase 3 Preparación de los datos

La fase 3 de la metodología CRISP-DM es la etapa de mayor trabajo en un proyecto de ciencia de datos se estima que la preparación de los datos toma generalmente entre el 50 y el 70% del tiempo y esfuerzo del proyecto.

La diferencia entre ese 50 y 70% reside en la calidad con la que se realizaron las dos etapas anteriores, mientras mejor se hayan realizado la etapa 1 y 2 menor tiempo llevara realizar la etapa 3.

Las etapas de esta fase son:

• Selección de los datos.

Consiste en seleccionar los datos relevantes. En general, hay dos formas de seleccionar datos: Selección de elementos y selección de atributos.

• 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.

• Integración de los datos.

Esto se realiza cuando se tienen múltiples fuentes de datos para poder responder al mismo conjunto de preguntas que se puede hacer una organización

Hay dos métodos básicos de integración de datos.

Fusión de datos: Implica la fusión de dos conjuntos de datos con registros similares, pero con atributos diferentes.

Agregar datos: Implica integrar dos o más conjuntos de datos con atributos similares, pero registros diferentes.

En nuestro proyecto

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

   Necesito para empezar la columna que contiene las calorías porque esta es la variable respuesta, después es necesario seleccionar las variables de carbohidratos, lípidos, proteínas y sodio porque son nutrientes y podrían estar relacionadas con la variable respuesta; las calorías.

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

   Utilice la función “blanco” en Excel para detectar si había espacios en blanco en mi base de datos, no se encontró ninguno por lo que no fue necesario eliminar ninguna entrada o remplazar valores. En caso de haber encontrado espacios en blanco habría sido necesario revisar que ese espacio en blanco no correspondiera a un cero, pues durante el llenado de la base de datos había algunos elementos que incluían el cero como valor en los nutrientes, por lo que hubiera sido necesario remplazar, El hecho de no encontrar espacios en blanco puede indicar que realice un correcto llenado de la base de datos

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

   Si es posible, por medio de uno de estos dos métodos:

   Generar registros y Derivara atributos

   Para el caso del proyecto lo ideal y lo que se ha estado haciendo es generar registros nuevos, porque cada vez que agregamos nuevos registros y sobre todo cuando son alimentos que no se habían mencionado nunca en base de datos, permiten al modelo tener más experiencia para crear un modelo más preciso y que abarque más variedad de alimentos.

   Pero para este caso no creo al menos no conozco un dato nuevo que se pueda generar a partir de hacer cálculos con los datos existentes, si estuviéramos trabajando con peso y estatura podríamos generar el IMC.

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

   Anqué si utilicé fuentes diferentes no fue necesario fusionar datos por que los alimento que tuvieran el mismo nombre no los volvía a buscar si no que reutilizaba la primera entrada con ese nombre, además durante proceso me di cuenta que no podía mezclar datos para una sola entrada de distintas fuentes porque estas tenían diferentes formas de medir los carbohidratos, calorías, lípidos, proteínas, sodio y porciones por lo que el modelo era impreciso.

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

   No, porque no tengo ningún dato que se necesite categorizar ni necesite una secuencia específica para el primer análisis de regresión solo necesite tener claro cuáles eran mis variables de entrada y cuál es mi variable de salida

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

   Si, por que de este modo nos aseguramos de que el modelo puede predecir cosas nuevas basadas en su experiencia y no que las prediga solo por que las conoce es decir por que estaban en su conjunto de datos. Por eso era también muy importante variar los alimentos que se registraban en la base de datos

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

   El primer análisis de regresión dio como resultado un r^2 de 0.05 muy inferior al objetivo por lo que fue necesario revisar la base de datos, ahí se encontró algunos datos erróneos por ejemplo un error de dedo que agrego un cero a las calorías o algunas entradas que mezclaron información de calorías y nutrientes de distintas fuentes para una misma entrada. Esto se corrigió desde la fase 1.

   Se siguieron agregando datos para mejorar la precisión del modelo durante las 3 semanas siguientes,

   No fue necesario integrar datos, aunque si se utilizaron diferentes fuentes para tener más variedad de alimentos, tampoco fue necesario crear atributos

Fase 4

Código

Se uso la función “import” para importar la librería panda una librería creada por un tercero que nos ayudara a manejar archivos Excel en Python y la nombramos pd.
Abrimos el archivo de Excel con la función read de la librería pandas y guardamos los datos en "datos_consumo".
A continuación se utiliza la función head() para verificar que se haya cargado correctamente el archivo

```
import pandas as pd

datos_consumo = pd.read_excel("Datos_sema_3.xlsx")
datos_consumo.head()
```

	Fecha (dd/mm/aa)	Momento	Nombre alimento	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)	Fuente
0	2022-01-07	Desayuno	Brocoli	26	4.6	0.4	2.7	33.0	MyFitnessPal App
1	2022-01-07	Desayuno	Chocomilk	152	27.0	3.4	3.7	184.0	MyFitnessPal App
2	2022-01-07	Desayuno	Huevo	155	0.6	11.0	13.0	124.0	MyFitnessPal App
3	2022-01-07	Desayuno	Atun	110	0.0	2.2	22.6	240.0	MyFitnessPal App
4	2022-01-07	Comida	Sopa	194	0.0	8.4	0.0	322.0	MyFitnessPal App

Con la función groupby agrupamos los datos de la columna Momento y con count() los contamos para obtener subtotales

```
datos_consumo.groupby("Momento").count()
```

Momento	Fecha (dd/mm/aa)	Nombre alimento	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)	Fuente
Cena	16	16	16	16	16	16	16	16
Comida	26	26	26	26	26	26	26	26
Desayuno	48	48	48	48	48	48	48	48
Snack	28	28	28	28	28	28	28	28

Usamos la función describe() para obtener la estadística descriptiva del conjunto de datos

```
datos_consumo.describe()
```

	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)
count	120.000000	120.000000	120.000000	120.000000	120.000000
mean	181.325000	15.242917	9.578083	6.370833	221.624167
std	176.935926	18.536276	17.995573	7.681316	297.313141
min	7.000000	0.000000	0.000000	0.000000	0.000000
25%	54.000000	1.000000	1.300000	1.000000	30.000000
50%	152.000000	6.500000	3.800000	3.700000	149.000000
75%	197.000000	27.000000	11.000000	10.167500	280.000000
max	844.000000	87.000000	93.750000	40.000000	1750.000000

El siguiente paso es seleccionar solo las variables que utilizaremos, es decir las relacionadas con calorias y nutrientes, para esto debemos utilizar la función iloc[] donde especificamos las filas y columnas de las que queremos obtener datos. Esta información se guarda en la variable datos seleccionados.

```
datos_seleccionados = datos_consumo.iloc[:,3:8]
datos_seleccionados
```

	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)
0	26	4.60	0.40	2.70	33.0
1	152	27.00	3.40	3.70	184.0
2	155	0.60	11.00	13.00	124.0
3	110	0.00	2.20	22.60	240.0
4	194	0.00	8.40	0.00	322.0
...	...	...	...	...	...
115	367	47.64	13.21	15.91	711.0
116	110	0.00	2.20	22.60	240.0
117	140	14.00	8.00	2.00	110.0
118	130	7.00	2.00	20.00	300.0
119	50	6.00	2.60	0.60	201.0

120 rows × 5 columns

Con la función info() vemos la información completa de los datos del nuevo dataframe por ejemplo el tipo de dato que utiliza cada variable.

```
datos_seleccionados.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 120 entries, 0 to 119
Data columns (total 5 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Calorías (kcal)     120 non-null    int64  
 1   Carbohidratos (g)   120 non-null    float64
 2   Lípidos/grasas (g)  120 non-null    float64
 3   Proteína (g)        120 non-null    float64
 4   Sodio (mg)          120 non-null    float64
dtypes: float64(4), int64(1)
memory usage: 4.8 KB
```

Para limpiar los datos utilizamos la función isnull().values.any() para buscar valores nulos

Y con dropna() creamos un nuevo data frame llamado “dataset” descartando los valores nulos. Para validr que efectivamente no hay datos nulos utilizamos la función .isnull().sum()

```
datos_seleccionados.isnull().values.any()
```

```
False
```

```
dataset = datos_seleccionados.dropna()
dataset.isnull().sum()
```

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

Después se asignan las variables independite y la variable respuesta a una lista cada uno.

Utilizando las herramientas de la biblioteca sklearn se separa el data set en datos de entrenamiento y datos de prueba , se genera un modelo de regresión linear y se entrena con los datos antes mencionados finalmente se obtienen los coeficientes del modelo.

```
X = dataset[['Carbohidratos (g)', 'Lípidos/grasas (g)', 'Proteína (g)'
, 'Sodio (mg)']].values
y = dataset['Calorías (kcal)'].values
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.
2, random_state=0)
from sklearn.linear_model import LinearRegression
modelo_regresion = LinearRegression()
modelo_regresion.fit(X_train, y_train)
x_columns = ['Carbohidratos (g)', 'Lípidos/grasas (g)', 'Proteína (g)'
, 'Sodio (mg)']
coeff_df = pd.DataFrame(modelo_regresion.coef_, x_columns, columns=['C
oeficientes'])
coeff_df
```

	Coeficientes
Carbohidratos (g)	3.738800
Lípidos/grasas (g)	8.718784
Proteína (g)	3.567116
Sodio (mg)	-0.006389

A continuación se evalua el modelo con los datos destinados a eso y se obtiene una muestra de esta validación

```
y_pred = modelo_regresion.predict(X_test)
validacion = pd.DataFrame({'Actual': y_test, 'Predicción': y_pred, 'Di
ferencia': y_test-y_pred })
muestra_validacion = validacion.head(25)
muestra_validacion
```

	Actual	Predicción	Diferencia
0	36	50.972746	-14.972746
1	367	366.208215	0.791785
2	509	502.112235	6.887765
3	329	326.676949	2.323051
4	824	818.475321	5.524679
5	130	141.764617	-11.764617
6	26	50.812956	-24.812956
7	296	293.657658	2.342342
8	26	50.812956	-24.812956
9	161	166.642402	-5.642402
10	19	38.761539	-19.761539
11	150	164.884753	-14.884753
12	155	164.436738	-9.436738
13	356	351.981231	4.018769
14	50	66.664261	-16.664261
15	155	164.436738	-9.436738
16	200	205.249587	-5.249587
17	40	33.728502	6.271498
18	170	181.717025	-11.717025
19	296	293.657658	2.342342
20	275	299.059482	-24.059482
21	238	243.041686	-5.041686
22	610	573.392256	36.607744
23	170	181.717025	-11.717025

Finalmente se obtiene el r2

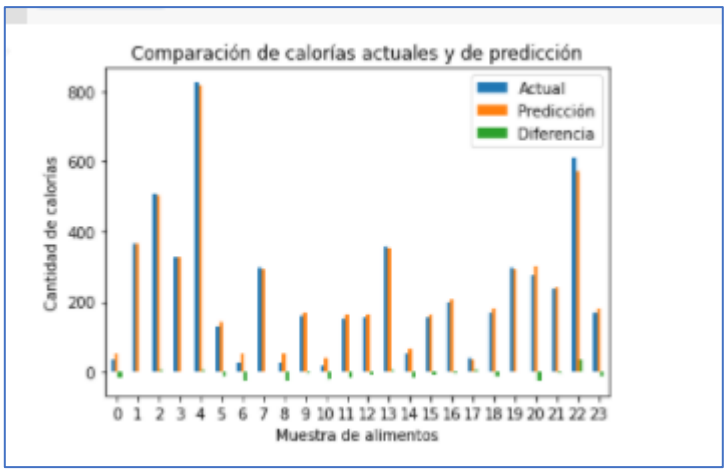
```
from sklearn.metrics import r2_score
r2_score(y_test, y_pred)
```

```
0.9943720585423431
```

Y con la librería de matplotlib se hace una grafica de la muestra anterior

```
import matplotlib.pyplot as plt
muestra_validacion.plot.bar(rot=0)
plt.title("Comparación de calorías actuales y de predicción")
plt.xlabel("Muestra de alimentos")
```

```
plt.ylabel("Cantidad de calorías")
plt.show()
```



Puede consultar este código en mi libreta de Google colab en este enlace.

Fase 4: modelación de los datos

"En este punto el trabajo duro comienza a dar sus frutos"[i]

En la cuarta fase (modelación de los datos) de la metodología CRISP-DM se utilizan herramientas tecnológicas como Excel, Python, etc para comenzar a modelar los datos.

Para esto se sigue un proceso de "prueba y error" es decir se sigue un modelo estándar como la regresión lineal, se aplica dicho proceso a un conjunto de datos y obtenemos un modelo, este no necesariamente será perfecto desde el primer instante por lo que en esta etapa debemos ser capaces de entender que significan los valores que obtenemos al realizar el modelo por ejemplo la probabilidad de una variable nos puede indicar si una variable es significativa para el modelo o no. El r^2 del modelo muestra que porcentaje del conjunto de datos utilizado puede ser explicado por el modelo. Conociendo esta información podemos cambiar la configuración del modelo para hacerlo más preciso, este proceso se puede repetir hasta estar satisfechos con el modelo o darnos cuenta de que estamos utilizando el enfoqué incorrecto. Esto puede ser debido a que estamos utilizando un modelo

inadecuado para la cantidad de datos que estamos trabajando, por que los objetivos del modelo no son claros, no se tiene el tipo adecuado de datos o se necesita más precisión de la que nuestro enfoque de modelo puede otorgar.

En nuestro proyecto

1. ¿Cuántos intentos o corridas realizaste para obtener los resultados sin errores? Porqué

   1 o 2, los errores que obtuve fueron en realidad relacionados con la sintaxis de mi código, pero en cuanto los datos no tuve ningún error en la base de datos por que fui muy cuidadoso al momento de llenar las entradas, cosa que me alegra pues me ahorro tiempo en esta etapa, solo fue necesario correr algunas pruebas para verificar que no hubiese espacios en blanco.

2. ¿Cómo los resolviste los problemas que se presentaron?

   Como lo mencioné, no tuve problemas con mis datos por lo que solo fue necesario preocuparme por errores de sintaxis en el código que resolví observando con detalle la guía proporcionada por el profesor.

   Pero debo mencionar que en fase anteriores cuando se realzo el primer análisis de regresión en Excel dio como resultado un r^2 de 0.05 muy inferior al objetivo por lo que fue necesario revisar la base de datos, ahí se encontró algunos datos erróneos por ejemplo un error de dedo que agrego un cero a las calorías o algunas entradas que mezclaron información de calorías y nutrientes de distintas fuentes para una misma entrada. Los datos eran tan pocos en aquel momento que fue posible identificar el error con la vista y solucionarlo a mano. Pero sentó un precedente para la forma en que debía seguir llenando mi base de datos.

3. ¿Qué resultados arrojó el análisis? Incluye imagen de cada resultado y explica cada uno de los resultados:

   Estadística descriptiva

	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)
count	120.000000	120.000000	120.000000	120.000000	120.000000
mean	181.325000	15.242917	9.578083	6.370833	221.624167
std	176.935926	18.536276	17.995573	7.681316	297.313141
min	7.000000	0.000000	0.000000	0.000000	0.000000
25%	54.000000	1.000000	1.300000	1.000000	30.000000
50%	152.000000	6.500000	3.800000	3.700000	149.000000
75%	197.000000	27.000000	11.000000	10.167500	280.000000
max	844.000000	87.000000	93.750000	40.000000	1750.000000

La imagen anterior muestra la estadística descriptiva de los alimentos que consumí y registré durante las 4 semanas anteriores en esta tabla encontramos el total de registros, la media, la desviación estándar y una descripción de los cuartiles con su respectivo mínimo, q1, q2, q3 y máximo para cada columna de datos.

Coeficientes de regresión

	Coeficientes
Carbohidratos (g)	3.738800
Lípidos/grasas (g)	8.718784
Proteína (g)	3.567116
Sodio (mg)	-0.006389

Estos son los coeficientes que componen el modelo matemático que obtuve, es decir son los numero por los que se deben multiplicar los carbohidratos lípidos, proteínas y sodio del alimento que estemos tratando para predecir sus calorías. Como podemos ver el coeficiente más pequeño es el sodio que de hecho tras un análisis de regresión en Excel fue descartado como variable significativa en modelos anteriores.

Valores actuales y de predicción

	Actual	Predicción	Diferencia
0	36	50.972746	-14.972746
1	367	366.208215	0.791785
2	509	502.112235	6.887765
3	329	326.676949	2.323051
4	824	818.475321	5.524679
5	130	141.764617	-11.764617
6	26	50.812956	-24.812956
7	296	293.657658	2.342342
8	26	50.812956	-24.812956
9	161	166.642402	-5.642402
10	19	38.761539	-19.761539
11	150	164.884753	-14.884753
12	155	164.436738	-9.436738
13	356	351.981231	4.018769
14	50	66.664261	-16.664261
15	155	164.436738	-9.436738
16	200	205.249587	-5.249587
17	40	33.728502	6.271498
18	170	181.717025	-11.717025
19	296	293.657658	2.342342
20	275	299.059482	-24.059482
21	238	243.041686	-5.041686
22	610	573.392256	36.607744
23	170	181.717025	-11.717025

Esta es una muestra de datos donde podemos ver la diferencia entre el valor real de calorías en un alimento y el valor pronosticado por nuestro modelo tras recibir como entrada los nutrientes de dicho alimento.

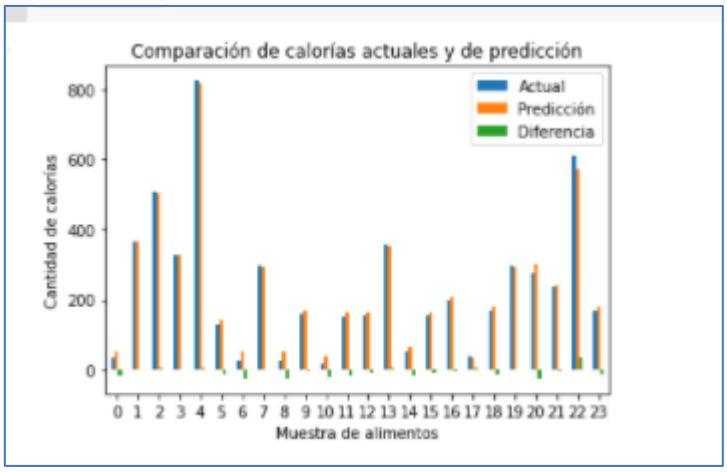
Coeficiente de determinación r2

```
0.9943720585423431
```

Este nos explica que porcentaje de los datos en el data frame pueden ser explicados por nuestro modelo. El objetivo que es que este numero sea lo

más cercano posible a 1, aunque para la materia los profesores han estipulado como mínimo un 0.7 por lo que puedo concluir que mi modelo cumple el objetivo.

Gráfica



Finalmente, esta grafica es una manera visual de representar la misma información que mostré en la tabla anterior “Valores actuales y de predicción” este tipo de graficas son útiles para explicar de una manera simple conclusiones sobre tu proyecto de análisis de datos a algún interesado.

4. ¿Cuáles son tus conclusiones de la modelación?

   Solo generar un modelo es en realidad un proceso relativamente simple de toda la metodología para realizar análisis de datos, el trabajo real es optimizarlo ya que es un proceso iterativo que además depende de la calidad con la que se hayan realizado las fases anteriores del proyecto.

Efectos del consumo calórico en el tiempo.

Código

Se uso la función “import” para importar la librería panda una librería creada por un tercero que nos ayudara a manejar archivos Excel en Python y la nombramos pd.
Abrimos el archivo de Excel con la función read de la librería pandas y guardamos los datos en "datos_consumo".
A continuación se utiliza la función head() para verificar que se haya cargado correctamente el archivo

```
import pandas as pd

datos_consumo = pd.read_excel("Datos_sema_3.xlsx")
datos_consumo.head()
```

	Fecha (dd/mm/aa)	Momento	Nombre alimento	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)	Fuente
0	2022-01-07	Desayuno	Brocoli	26	4.6	0.4	2.7	33.0	MyFitnessPal App
1	2022-01-07	Desayuno	Chocomilk	152	27.0	3.4	3.7	184.0	MyFitnessPal App
2	2022-01-07	Desayuno	Huevo	155	0.6	11.0	13.0	124.0	MyFitnessPal App
3	2022-01-07	Desayuno	Atun	110	0.0	2.2	22.6	240.0	MyFitnessPal App
4	2022-01-07	Comida	Sopa	194	0.0	8.4	0.0	322.0	MyFitnessPal App

Después creé una lista llamada “datos que contiene” exclusivamente las columnas de fecha y calorias del conjunto de datos.

```
datos = datos_consumo[["Fecha (dd/mm/aa)","Calorías (kcal)"]]
datos.head()
```

	Fecha (dd/mm/aa)	Calorías (kcal)
0	2022-01-07	26
1	2022-01-07	152
2	2022-01-07	155
3	2022-01-07	110
4	2022-01-07	194

A continuación, se realiza una suma de todas las calorías en la variable “datos” es decir que se realiza la suma de todas las calorías que registre en la base de datos.

```
suma_calorías = datos["Calorías (kcal)"].sum()
suma_calorías
```

```
21759
```

Después con la función nunique obtenemos el valor de cuatas fechas distintas hay en la base de datos, y almacenamos esos datos en las variables “días”

```
días = datos ["Fecha (dd/mm/aa)"].nunique()
días
```

```
14
```

se divide el valor de “suma_calorias” entre “días” para obtener el promedio de calorías que consumí por día y almacenarlo en la variable “calorias_promedio” después se muestra este valor con con contexto utilizando la función print .

```
calorías_promedio = suma_calorías/días
print("Tu promedio de calorías consumidas en", días,"días es:", calorí
as_promedio)
```

```
Tu promedio de calorías consumidas en 14 días es: 1554.2142857142858
```

A continuación, se le solicita al usuario su peso, altura, edad y genero

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

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

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

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

```
Ingresa tu peso en kilogramos: 83
Ingresa tu altura en centimetros: 179
Ingresa tu edad en años: 20
Ingresa tu género, Mujer/Hombre: Hombre
```

Con esta información podemos utilizar la formula correspondiente para saber cuántas calorías debería consumir al día el usuario y mostrarlo por pantalla

```
if(genero == "Mujer"):
   calorías_requeridas = 655+(9.56*peso)+(1.85*altura)-(4.68*edad)

elif(genero == "Hombre"):
  calorías_requeridas = 66.5+(13.75*peso)+(5*altura)-(6.8*edad)

print("Con base en tus datos, tu consumo de calorías al día debe ser d
e:", calorías_requeridas)
```

```
Con base en tus datos, tu consumo de calorías al día debe ser de: 1966.75
```

Se obtiene la diferencia entre el promedio consumido y el promedio recomendado

```
diferencia = calorías_promedio - calorías_requeridas

diferencia
```

```
-412.5357142857142
```

Finalmente se multiplica esta cantidad por 365 y se ajustan las unidades de medida para obtener cual será el cambio en la masa corporal dentro de un año.

```
efecto_anual = diferencia * 450/3500 * 365 /1000
```

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

```
Si continuas con el consumo calórico actual, en un año tu cambio de masa corporal sería aproximadamente de: -19.35971173469387 kg
```

Puede consultar este código en mi libreta de Google colab en este enlace.

Conclusiones

La hipótesis en torno a la cual a girado todo este proyecto es: ¿Si consumo cierta cantidad calórica, puedo tener cambios en mi masa corporal (peso) en un determinado tiempo?

Mis resultados muestran que efectivamente el consumo de calorías puede generar un cambio en mi masa corporal, sim embargo mi estimación de perder más de 19 kg en un año con mi dieta actual me parece irreal, esto debido a que de hecho estoy bajo supervisión de un nutriólogo para subir de peso, y esa es la tendencia que e seguido. Estos resultados discrepan de la realidad por la forma en que registre los datos. No registraba todo lo que comía en el día solo algunas cosas, por lo tanto al hacer el promedio de consumo calórico diario este es menor a real.

Regresión en Excel vs Python

Excel

Estadística Descriptiva

Calorías (kcal)		*Carbohidratos (g)*		*Lípidos/grasas*
Media	182.125	Media	20.053125	Media
Error típico	25.3833015	Error típico	3.87166606	Error típico
Mediana	152	Mediana	11.75	Mediana
Moda	26	Moda	4.6	Moda
Desviación es	143.589637	Desviación es	21.9014506	Desviación
Varianza de la	20617.9839	Varianza de la	479.673538	Varianza de
Curtosis	3.56678816	Curtosis	3.82246519	Curtosis
Coeficiente d	1.79625459	Coeficiente d	1.79986026	Coeficiente
Rango	584	Rango	87	Rango
Mínimo	26	Mínimo	0	Mínimo
Máximo	610	Máximo	87	Máximo
Suma	5828	Suma	641.7	Suma
Cuenta	32	Cuenta	32	Cuenta

Python

	Calorías (kcal)	Carbohidratos (g)
count	120.000000	120.000000
mean	181.325000	15.242917
std	176.935926	18.536276
min	7.000000	0.000000
25%	54.000000	1.000000
50%	152.000000	6.500000
75%	197.000000	27.000000
max	844.000000	87.000000

La estadística descriptiva muestra un resumen numérico de los datos, en este caso me gusto mas la que genera Excel, por que proporciona más información y además puede estar en cualquier idioma que maneje Excel, incluido español, aunque me gusto la presentación más limpia que se genera en Python

Coeficientes de regresión

	Coeficientes
Intercepción	0
Carbohidratos	4.04351404
Lípidos/grasas	9.83793982
Proteína (g)	3.67212638

	Coeficientes
Carbohidratos (g)	3.738800
Lípidos/grasas (g)	8.718784
Proteína (g)	3.567116
Sodio (mg)	-0.006389

Los coeficientes de regresión no son descomunalmente diferentes, la mayoría de la diferencia se debe a que los coeficientes generados en Python son resultado de un análisis de más datos que el generado por Excel, el modelo de Excel se configuro con intercepción igual a 0 y paso por un proceso de optimización donde se elimino la variable sodio.

Valores actuales y de predicción

Lechera		Kcal Reales
Carbohidratos	59	30
Lípidos/grasas	5	
Proteína (g)	6.4	

Coca Cola		Kcal Reales
Carbohidratos	27	10
Lípidos/grasas	0	
Proteína (g)	0	

Galletas Oreo		Kcal Reales
Carbohidratos	18.3	12
Lípidos/grasas	5.6	
Proteína (g)	1.2	

	Actual	Predicción	Diferencia
0	509	500.174592	8.825408
1	180	187.158973	-7.158973
2	238	245.557037	-7.557037
3	140	146.465704	-6.465704
4	370	345.954711	24.045289
5	196	201.702407	-5.702407
6	26	47.260659	-21.260659
7	728	724.922380	3.077620
8	36	50.866940	-14.866940
9	55	61.241145	-6.241145
10	824	818.236312	5.763688
11	54	71.163896	-17.163896
12	155	162.554930	-7.554930
13	26	47.260659	-21.260659
14	41	57.749771	-16.749771
15	170	171.690137	-1.690137
16	250	42.569920	207.430080
17	196	201.702407	-5.702407
18	26	47.260659	-21.260659
19	54	71.163896	-17.163896
20	54	71.163896	-17.163896
21	196	201.702407	-5.702407
22	150	160.595620	-10.595620
23	280	286.084784	-6.084784

Al momento de evaluar el modelo forzándolo a predecir valores encuentro más práctico Python pues con unas cuantas líneas de código puedes crear todo un conjunto de datos que evaluar mientras en Excel este trabajo se hizo a mano.

Coeficiente de determinación r2	
Coeficiente de 0.9631592	0.9943720585423431
Este incremento en el R^2 una vez mas se lo atribuyo a que el análisis en Python ya contaba con más datos.	

¿Por qué es importante la Ciencia de Datos y la ética para el uso adecuado de los datos e información?

Cada día más situaciones de nuestra vida cotidiana se digitalizan; hoy algo tan simple como ordenar un café puede realizarse de manera virtual dejando a su vez una serie de registro que aportan a lo que hoy conocemos como big data. La ciencia de datos nace de la necesidad de analizar los grandes volúmenes de datos derivados del big data.

Esto abre la puerta para analizar y crear muchas soluciones novedosas en todos los campos. Por ejemplo, en la medicina la posibilidad de acumular una gran cantidad de imágenes médicas como radiografías, ha permitido crear programas capaces de entrenarse con esas imágenes para detectar enfermedades mirando nuevas radiografías. En seguridad ha permitido mejorar los algoritmos que detectan fraudes en nuestras cuentas de banco. Incluso en temas más vánales y cotidianos nos facilita la vida con algoritmos que te recomiendan películas y series basado en tu historial de vistas, rutas de tráfico más cortas, los mejores restaurantes y lugares turísticos de una ciudad.

Pero con todos estos beneficios se abren también debates éticos pues para que los doctores puedan crear sus modelos para detectar enfermedades necesitan acceso a registros médicos como el tuyo y el mío. Y esta es información sensible, lo mismo pasa con los bancos, para que puedan detectar fraudes también deben estar al pendiente de tu ubicación tus hábitos de consumo, tus registros financieros. Para que las redes sociales sigan siendo gratuitas y ofreciéndote contendió personalizado analizan tus gustos para vender tus datos como consumidor. En un mundo ideal es un intercambio gusto; otorgamos nuestros

datos a cambio de beneficios el problema viene cuando estos datos acaban en las manos o usos equivocados, por ejemplo cuando alguna empresa de telefonía es hackeada y roban sus bases de datos, con esta información los criminales pueden hacer llamadas de extorción o suplantar tu identidad, por otro lado cuando una empresa se extralimita con su poder sobre tus datos como el caso de Facebook y Cambridge analítica acusados de utilizar información generada por los usuarios en redes para manipular la opinión política durante las elecciones.

Es aquí donde debemos tomar el camino de la ética para determinar quien debe cuidar nuestros datos y que pueden hacer con ellos.

Bibliografía

Tec de Monterrey. (s. f.). *Guia-para-implementacion-y-visualizacion-de-una-regresion-lineal-multiple-en-python*. Canvas. Recuperado 30 de enero de 2022, de https://experiencia21.tec.mx/courses/222770/pages/guia-para-implementacion-y-visualizacion-de-una-regresion-lineal-multiple-en-python?module_item_id=11377988

Oracle. (2020). *¿Qué es la ciencia de datos?* Recuperado 30 de enero de 2022, de https://www.oracle.com/mx/data-science/what-is-data-science/

[i] Tec de Monterrey. (s. f.-a). *Aprende-sobre-preparacion-fase-3-y-modelacion-fase-4-de-los-datos*. Canvas. Recuperado 30 de enero de 2022, de https://experiencia21.tec.mx/courses/222770/pages/aprende-sobre-preparacion-fase-3-y-modelacion-fase-4-de-los-datos?module_item_id=11377987