PROYECTO DE CIENCIAS DE DATOS

Michelle Borjon Arriola
A01638100

EVIDENCIA 2

Matemáticas y Ciencias de Datos para la Toma de Decisiones

08/06/2020

INTRODUCCIÓN

De acuerdo con IBM Analytics, la ciencia de datos puede ser denominada como *"el proceso por el que se descubren conocimientos ocultos de enormes cantidades de datos estructurados y no estructurados, mediante la utilización de métodos como la estadística, el aprendizaje automatizado, la minería de datos y el análisis predictivo"*. Para poder realizar este proceso es importante estudiar la procedencia de los datos, qué representan y cómo pueden ser utilizados de tal manera en que se conviertan en información valiosa para la construcción y desarrollo de un proyecto.

La ciencia de datos puede ayudar a una persona para implementar un nuevo plan de negocios en su empresa, ver qué tan factible y conveniente es el desarrollo de un nuevo producto o la introducción de un nuevo servicio. Asimismo, puede ser de gran ayuda para construir modelos matemáticos que predigan los resultados de las ventas, ganancias, pérdidas, entre otras cosas más.

De acuerdo con lo mencionado en la situación problema y la Organización para la Cooperación y el Desarrollo Económico (ODCE), México ocupa el 2do lugar en obesidad en adultos. Se proyecta que para 2030 la obesidad en México aumente hasta un 39%, siendo que los habitantes con menor nivel educativo tienen mayor riesgo de padecerla. Si la población se educa en el conocimiento de la información nutrimental de los alimentos que consume, esto podría ayudar a tomar mejores decisiones sobre cómo nos alimentamos. Para ello, fue necesaria la construcción de un modelo matemático que permitiese calcular la cantidad de calorías que contiene un alimento en función de sus nutrientes.

La Ciencia de Datos toma un papel de suma importancia en la construcción de este modelo, esto se debe a que el ciclo de vida del Proceso de Ciencia de Datos en Equipo (TDSP) está diseñado para los proyectos de ciencia de datos que implementan modelos de aprendizaje o hacen uso de inteligencia artificial para realizar un análisis predictivo. El objetivo que se tiene es hacer avanzar un proyecto hacia un punto en específico haciendo uso de los conjuntos de datos que se recopilan.

El TDSP describe las fases principales por las que debe pasar un proyecto, las cuales se mencionaran a continuación:

1. Conocimiento del negocio
2. Adquisición y comprensión de los datos
3. Modelado
4. Implementación
5. Aceptación del cliente

ENTENDIMIENTO DEL NEGOCIO

En esta primera fase, lo más importante fue entender la problemática; por lo que, se debió hacer una investigación para conocer el negocio y para definir los objetivos del proyecto que dieron solución al problema.

Primero se identificó al cliente que, en este caso, eran los adultos que habitan en México, principalmente aquellos que cuentan con un nivel educativo menor al de la media. Una vez teniendo al cliente en mente, se debió de trabajar con él y con las partes interesadas en llevar a cabo el proyecto. Asimismo, fue necesario buscar y recopilar los datos pertinentes que pudieran ser de ayuda para resolver la problemática.

Identificar las principales variables que el análisis debe predecir fue sumamente importante, ya que estas pueden determinar el éxito del proyecto. En este modelo, las principales variables son los alimentos que consumen las personas y la cantidad de cada nutriente que estos contienen; puesto que para que el modelo que nos indique la cantidad de calorías consumidas con un margen de error mínimo es importante tomar en cuenta los nutrientes que se consumen.

Lo siguiente fue realizarse las 5 preguntas que se muestran a continuación, y determinar cuál de ellas es la que permitirá el éxito del proyecto:

- ¿Cuánto? / ¿Cuántos? (Regresión)
- ¿Qué categoría? (Clasificación)
- ¿Qué grupo? (Agrupación de clústeres)
- ¿Es extraño? (Detección de anomalías)
- ¿Qué opción se debe elegir? (Recomendación)

Para resolver la situación problema se ha empleado la regresión lineal del conjunto de datos que personalmente recopilé para crear una función que indique cuántas calorías se están consumiendo con respecto a los nutrientes; por lo tanto, es la primera pregunta la que ha permitido llevar a cabo el proyecto.

Desarrollar todo este estudio requirió de conocimientos en modelos matemáticos, regresión lineal, uso de medidas de tendencia central y de dispersión, hacer uso de la tecnología. En un principio se trabajó principalmente con Excel para construir un modelo matemático que nos

permita calcular la cantidad de calorías de un alimento; sin embargo, en Excel se deben de introducir los datos manualmente, seleccionar celdas para realizar una fórmula, seleccionar la herramienta necesaria para hacer la regresión o un gráfico. Por lo que, me vi en la necesidad de usar otro tipo de programa que me permitiese agilizar este proceso sin necesidad de tener que hacerlo todo personalmente y que, más bien, sea un proceso automatizado. Este programa pude ser cualquier aplicación que permita programar, pero el curso se hizo uso de Python y sus librerías para hacer el estudio de todos los datos que hemos recopilado.

Una vez teniendo el modelo matemático, se deben realizar las comprobaciones necesarias para poder decir con confianza que nuestra solución es correcta y puede ser aplicada con varios clientes y no solo con uno de ellos. De esta forma se da una solución a la problemática a la que se enfrenta México, así como hace conciencia en las personas para que no consuman más calorías de las que deben de acuerdo con su índice de masa, su edad, la actividad física realizada por el individuo, su metabolismo basal, su facto de lesión y el efecto termogénico de los alimentos, entre otras.

ENTENDIMIENTO DE LOS DATOS

Los datos que fueron recopilados para realizar el estudio eran numéricos principalmente, aunque también se hizo recopilación de nombres, que son strings. En el archivo de Excel se fueron adjuntando los nombres del alimento junto con la cantidad de calorías y nutrientes que contenían, así como la fecha y hora en que fueron consumidos. Cada uno de los datos viene de la misma fuente, pues fui yo quien personalmente se encargó de adjuntarlos y organizarlos.

Para realizar la regresión lineal en Excel, pudo observarse que aquellas columnas de nutrientes con un valor crítico f mayor a 0.05 debían ser excluidas, pues pueden hacer del modelo uno inválido. En cambio, aquellas columnas de nutrientes con un valor crítico menor a 0.05 son bastante prometedoras para la construcción de nuestro modelo matemático. En el primer modelo que se obtuvo, era necesario eliminar sodio porque no cumplía con la condición del valor crítico; sin embargo, en el modelo final de Excel pudo apreciarse que

todos los nutrientes contaban con un valor critico f menor a 0.05, por lo que ninguno fue excluido del modelo.

Los datos con los que se cuentan son de una sola persona, son datos que yo he ido recopilando a lo largo del semestre, por lo que, si lo que se busca es hacer una conclusión generalizada los datos con los que cuento no me servirían mucho, puesto que una conclusión generalizada tendría que tomar en cuenta los conjuntos de datos de más personas. Por otro lado, los datos con los que cuento si pudiese ayudarme a hacer predicciones sobre mi propia alimentación.

Debido a que antes de consumir los alimentos estos no fueron pesados se cuenta con una aproximación de los gramos consumidos por alimento, por lo que la calidad de los datos puede no ser la mejor; sin embargo, es bastante fácil acceder a los nutrientes de un alimento, puesto que, si la comida que se ingirió no contenía la información nutrimental, esta puede encontrase en Internet e incluso viene la opción de especificar cuántos gramos fueron.

Los datos son adecuados a mi persona, pero puede que para otras personas no lo sean. Con esto me refiero a que cada uno de nosotros, de acuerdo con nuestras cualidades físicas y a nuestra actividad física, debemos consumir cierta cantidad de calorías. Puede darse el caso que personas consuman más, o menos, calorías de las que deben; por lo que realizar un análisis de datos personal podría resultar de gran ayuda. Por otra parte, debido a que todos los alimentos cuentan con su información nutrimental puede ser que mis compañeros y yo tengamos un resultado parecido, asimismo puede ser que obtengamos resultados muy diferentes, pero eso dependerá de los datos que cada uno de nosotros hemos recopilado.

PREPARACIÓN DE LOS DATOS

Esta fase se enfoca en los ajustes que tuvieron que realizarse en el conjunto de datos ya recopilado para obtener un mejor modelo matemático, por lo que se mencionará punto por punto qué tipo de ajustes se hicieron o no y su justificación.

Fusión de datos

Al ser yo la única persona que agregó datos al documento de Excel, no ha habido ninguna fusión de varias fuentes de datos porque los datos con los que cuento están completos, es

decir, no me vi en la necesidad de buscar datos ajenos en otras fuentes de información puesto que mis datos son un registro de mi alimentación diaria.

A pesar de que fusionar datos puede llegar a cubrir la necesidad de tener una fuente de datos completa para los usuarios, no es el caso en este proyecto. Además de ello, cada uno de los nutrientes representa una variable independiente, mientras que las calorías son la variable dependiente; por lo tanto, no deben fusionarse entre sí.

Subconjuntos

En el Excel no se realizaron subconjuntos de datos. Esto se debe a que, en caso de hacer un subconjunto, estaría excluyendo datos que pueden ser de suma importancia en la creación de mi modelo matemático. Asimismo, si selecciono solo una parte de mi conjunto de datos, podría afectar en el desarrollo del modelo y obtener como resultado un modelo no muy preciso.

Sin embargo, en el código de Python si se hizo un subconjunto en el cual excluimos los nombres de los alimentos, la hora y fecha cuando fueron consumidos. La razón detrás de este subconjunto es que para realizar la regresión lineal en Python debemos de seleccionar solo aquellas columnas que son meramente necesarias, las cuales eran las calorías y los nutrientes.

Agregar más datos

Conforme pasaba el tiempo fue necesario agregar más datos, pues la recopilación de datos iniciaba en la Semana 2 y terminaba en la Semana 14, por lo que constantemente se tuvieron que agregar más datos hasta que el periodo de recopilación terminase. Como he estado mencionando, con el primer modelo obtuve muy poca precisión y eso se debía a que tenía muy pocos datos en los cuales basarme. En cambio, en el modelo actual de Excel la precisión incrementó debido a que tiene casi el cuádruple de datos en comparación con los que se contaban inicialmente.

Por otro lado, también podrían agregarse más variables independientes al modelo, pero eso podría dar como resultado un modelo más preciso o, en su defecto, un modelo con menor precisión; sin embargo, las indicaciones dadas para dar resolución a la situación problema solo nos indica ciertos nutrientes que debemos tomar en cuenta, por lo que no podemos agregar más variables independientes.

Atributos de los datos

Personalmente no me vi en la necesidad de cambiar los atributos de mis datos, puede que en un principio mi modelo no era tan preciso y eso era un poco preocupante, pero al añadir más datos de alimentos, la precisión mejoró bastante.

Ordenar datos

Tener un orden permite que la comida registrada coincida con la fecha y hora de su consumo, así como con las calorías y la cantidad de nutrientes que aporta; además en el archivo de Excel proporcionado como base solicitaba realizar el registro la fecha y hora de nuestro consumo. Personalmente, considero que es mucho mejor analizar un conjunto de datos ordenado a uno desordenado, esto se debe a que es mucho más fácil analizar e interpretar datos ordenados.

Por otro lado, en Python no es tan necesario que los datos de calorías y nutrientes coincidan con el nombre del alimento, su hora y fecha, puesto que al final solo haremos uso de los primeramente mencionados.

Eliminación o reemplazo de datos

En el primer modelo que obtuve era necesario que se eliminarán tanto lípidos como el sodio, después de hacer un pequeño ajuste, solo se eliminaba el sodio porque su probabilidad era mayor a 0.05. Sin embargo, en mi modelo actual todas las variables independientes tienen una probabilidad menor a 0.05, por lo que no es necesario eliminar ninguna variable para crear el modelo matemático. Asimismo, ninguno de los datos en el registro de alimentos está en blanco, por lo que tampoco hay datos que necesiten ser reemplazados.

MODELACIÓN DE LOS DATOS

En esta última fase se hizo el modelo matemático haciendo uso del programa Python, en el cual, por medio de un código obtuvimos la ecuación de regresión lineal, así como los coeficientes de correlación que hay entre cada una de las variables.

Librerías

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from mpl_toolkits.mplot3d import Axes3D
```

Para la modelación del conjunto de datos de Excel en Python, fue necesario hacer uso de varias librerías del programa. A continuación, se hará mención de las librerías utilizadas para modelar la situación problema y se dará una breve explicación de cada una:

- *Pandas* – Permite almacenar y manipular datos tabulados en filas de observaciones y columnas de variables.
- *Seaborn* – Permite la visualización de datos en forma de gráficos estadísticos atractivos e informativos.
- *Matplotlib.pyplot* – Colección de funciones que permiten que matplotlib funcione como MATLAB para la creación de gráficos.
- *Sklearn.linear_model-LinearRegression* – Ajusta un modelo lineal con coeficientes para minimizar la suma residual entre las variables independientes y las variables dependientes.
- *Sklearn.model_selection-train_test_split* – Divide matrices en subconjuntos aleatorios para entrenamiento y para prueba.
- *Sklearn.metrics-r2_score* – Función para obtener el coeficiente de determinación de una regresión.
- *Mpl_toolkits.mplot3d-Axes3D* – Función que permite la proyección de un gráfico 3D.

Resultados

Para realizar el modelo de regresión lineal en Python, primero se importaron las librerías anteriormente mencionadas. Después de eso llamamos al archivo de Excel que contiene el historial de consumo, como este se encontraba en la misma carpeta que el código de Python no fue necesario introducir toda su ubicación, sino que bastó con introducir su nombre.

```
In [2]: datos_consumo = pd.read_csv("A01638100_ConsumoNutricional.csv")
        datos_consumo.head()
```

Out[2]:

	Fecha	Hora	Nombre	Calorias (cal)	Carbohidratos (g)	Proteína (g)	Lípidos (g)	Sodio (mg)
0	17/02/2020	05:45	Pan con leche	476.0	118.8	23.10	26.3	227.0
1	17/02/2020	10:15	Galletas de avena	745.0	22.3	3.80	9.7	31.0
2	17/02/2020	16:30	Sopa de municiones con arroz	388.0	60.6	6.70	1.3	864.0
3	18/02/2020	07:45	Cereal con leche	380.0	54.8	9.10	2.3	349.0
4	18/02/2020	12:55	Ensalada de pepino con zanahoria	76.0	21.9	3.05	0.5	67.0

Una vez que el archivo es leído por el código, podemos empezar a trabajar sobre él para obtener nuestro modelo matemático. Para eso, primero debemos determinar qué variables de del archivo utilizaremos, y cuáles serán variables independientes (x) y cuáles serán dependientes (y). Como lo que se quiere calcular es la cantidad de calorías a partir de sus nutrientes, las calorías tomarán el lugar de y, mientras que los nutrientes serán las variables independientes.

```
In [3]: columnas = ["Carbohidratos (g)", "Proteína (g)", "Lípidos (g)", "Sodio (mg)"]
        fila = "Calorias (cal)"

In [4]: x = datos_consumo[columnas]
        y = datos_consumo[fila]
```

A partir de esto, entrenamos y probamos los datos para obtener un modelo de regresión lineal. De esta forma, se podrán obtener los coeficientes de cada x, el intercepto y el coeficiente de determinación (R2).

```
In [8]: modelo_mat.coef_
Out[8]: array([2.05313307, 4.55827625, 7.40455212, 0.073316  ])

In [9]: modelo_mat.intercept_
Out[9]: 84.40888067991773
```

Como puede observarse en la imagen, cada uno de los coeficientes de la lista corresponden a cada una de las variables independientes. Asimismo, puede observarse el punto intercepto con eje y. Cabe mencionar, que estos datos se obtuvieron después varias iteraciones del código, aproximadamente 9, hasta obtener una R2 cercana a 1, ya que esto nos indica si nuestro modelo es significativo.

```
In [11]: r2_score(y_test,pred)
Out[11]: 0.8987533139075233
```

Habiendo confirmado que el coeficiente de determinación es cercano 1, se puede proceder a crear un modelo matemático a partir del intercepto y coeficientes obtenidos. Por lo cual, el modelo final en Python queda de la siguiente manera:

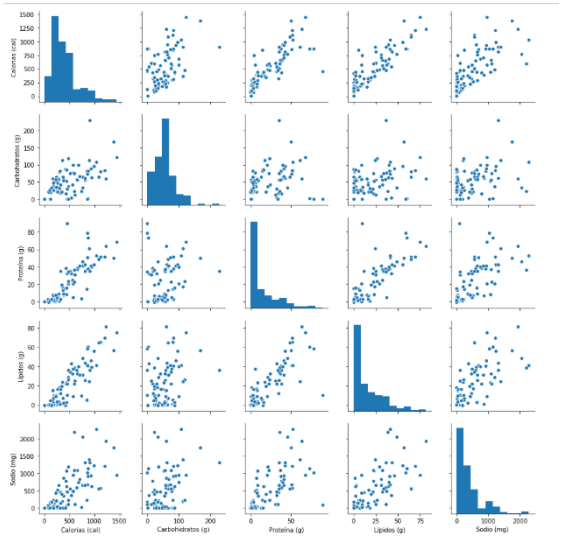
$$Calorías = 84.4088 + 2.0531(Carbohidratos) + 4.5582(Proteínas) + 7.4045(Lípidos) + 0.0733(Sodio)$$

Gráficas

Pair plot

Esta gráfica muestra la correlación que hay entre cada una de las variables, tanto independientes como dependientes. Hay más correlación entre las variables que presentan una gráfica con sus datos siguiente una línea recta; mientras que, la correlación es menor cuando los datos se encuentran más dispersos.

```
In [12]: sns.pairplot(datos_consumo[["Calorias (cal)","Carbohidratos (g)", "Proteína (g)","Lípidos (g)", "Sodio (mg)"]])
         plt.show()
```



Como podemos apreciar en el gráfico, cuando las variables coinciden consigo mismas los datos no se ven como puntos, esto es porque su coeficiente de correlación es igual a 1. En cambio, cuando coinciden con otras variables, podemos observar que algunas combinaciones tienen más correlación que otras. Esto lo podremos apreciar más a fondo en los siguientes dos gráficos.

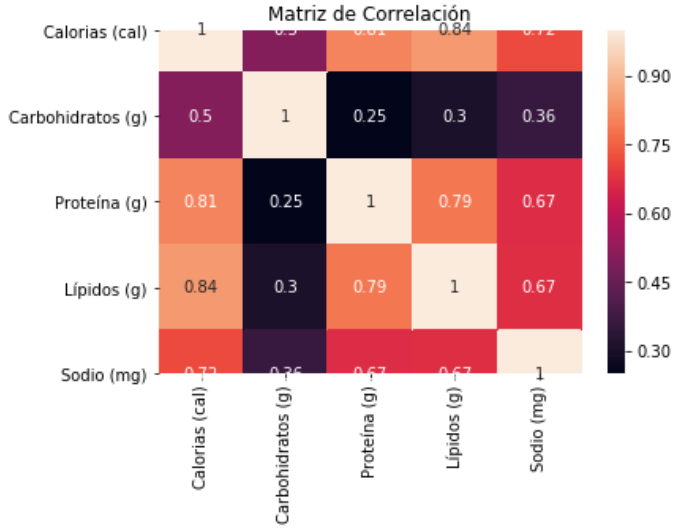
Matriz de Correlación y Heat Map

Estos dos gráficos van muy de la mano porque ambos son matrices que muestran la correlación de las variables, la diferencia que hay entre ellos es que mientras que la matriz de correlación se observa como una tabla, el heat map muestra mediante una variación de colores la correlación que hay entre cada variable.

```
matriz_corr= datos_consumo[["Calorias (cal)","Carbohidratos (g)", "Proteína (g)","Lípidos (g)", "Sodio (mg)"]].corr()
matriz_corr
```

	Calorias (cal)	Carbohidratos (g)	Proteína (g)	Lípidos (g)	Sodio (mg)
Calorias (cal)	1.000000	0.499336	0.808550	0.843330	0.715901
Carbohidratos (g)	0.499336	1.000000	0.250132	0.303827	0.361179
Proteína (g)	0.808550	0.250132	1.000000	0.788273	0.667462
Lípidos (g)	0.843330	0.303827	0.788273	1.000000	0.669117
Sodio (mg)	0.715901	0.361179	0.667462	0.669117	1.000000

```
sns.heatmap(matriz_corr, annot=True)
plt.title("Matriz de Correlación")
plt.show()
```



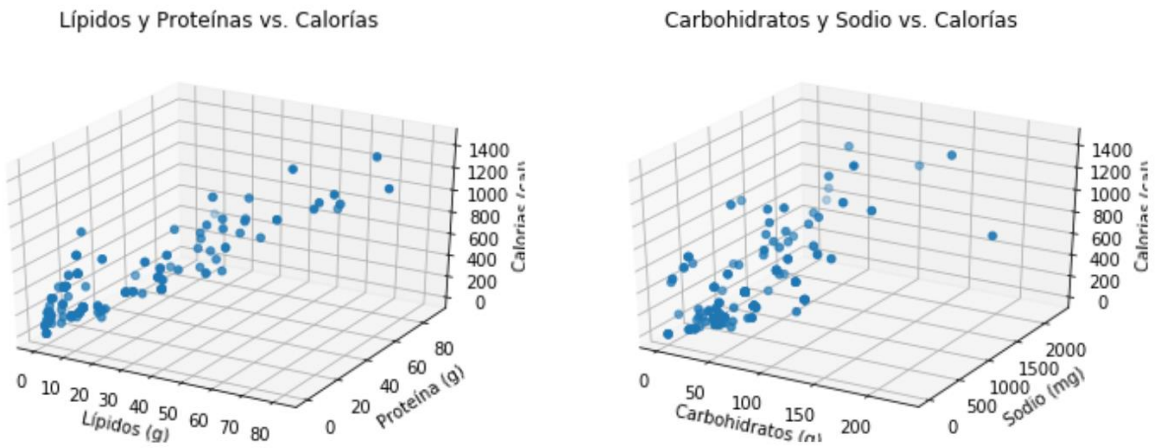
Como puede observarse, los coeficientes de determinación que presenta cada gráfico son los mismos. La diferencia es que la matriz nos muestra más decimales de lo que nos deja ver el heat map.

Scatter plot 3D

Un scatter plot es un diagrama de dispersión, en este caso, se utilizó un diagrama en 3D para ver la interacción entre las variables. Debido a que solo pueden seleccionarse 3 de las variables para un diagrama, decidí realizar dos diagramas.

```
figura1 = plt.figure()
axis = figura1.add_subplot(111, projection="3d")
axis.scatter(xs=datos_consumo["Lípidos (g)"], ys=datos_consumo["Proteína (g)"], zs=datos_consumo["Calorias (cal)"])
plt.suptitle("Lípidos y Proteínas vs. Calorías")
axis.set_xlabel("Lípidos (g)")
axis.set_ylabel("Proteína (g)")
axis.set_zlabel("Calorias (cal)")
plt.show()
```

```
figura1 = plt.figure()
axis = figura1.add_subplot(111, projection="3d")
axis.scatter(xs=datos_consumo["Carbohidratos (g)"], ys=datos_consumo["Sodio (mg)"], zs=datos_consumo["Calorias (cal)"])
plt.suptitle("Carbohidratos y Sodio vs. Calorías")
axis.set_xlabel("Carbohidratos (g)")
axis.set_ylabel("Sodio (mg)")
axis.set_zlabel("Calorias (cal)")
plt.show()
```



Como se puede observar, las diferencias en el código son que se usaron diferentes variables para los ejes “x” y “y”, por lo que las gráficas se ven diferentes, pues la distribución de los datos está dada por los nutrientes.

De acuerdo con los resultados obtenidos, podemos decir que el modelo es significativo, pues en el archivo de Excel pudo observarse que el factor de probabilidad de los nutrientes era menor a 0.05, asimismo la R2 era cercana a 1. Con la modelación en Python, también ser observó que el coeficiente de determinación es cercano a 1 y, aunque el modelo difirió al de Excel porque en Python se obtuvo la intersección en el eje y , sigue siendo un modelo significativo.

REFLEXIÓN FINAL

En la actualidad, la ciencia de datos es una herramienta necesaria en todos los ámbitos profesionales, pues permite que podamos comprender y utilizar los datos recuperados del Big data para poder aplicarlos en la creación y desarrollo de proyectos. Es por esto, que en los últimos años las empresas se han visto en la necesidad de contratar más científicos de datos, por lo tanto, las universidades también han comenzado a implementar más cursos, carreras profesionales y especialidades que impartan esta materia para que sus estudiantes puedan obtener una ventaja al momento de postularse a un empleo.

Como ya se mencionó los conjuntos de datos frecuentemente son recuperados del Big data, muchos de estos conjuntos cuentan con datos de nuestro historial en línea, ya sean las

búsquedas que hacemos en algún navegador, nuestras redes sociales, todo esto deja un rastro de nuestra persona lo que puede ayudar a corporativos para crear productos o servicios que puedan llamar nuestra atención. Sin embargo, el hecho de que estos corporativos tengan acceso a nuestra información es un poco preocupante porque hay un límite que se llama privacidad. Es aquí donde entra la ética en la ciencia de datos, las empresas deben de preguntarse constantemente si están violando o no la privacidad del cliente al acceder a cierta información o rastro que deja en la red, pero eso no solo depende de ellos, sino que nosotros también debemos de trabajar en ello usando ciertas configuraciones para que no todos puedan acceder a nuestra información y publicaciones, o simplemente abstenernos de publicar algo realmente íntimo. En todo caso, está en nosotros proteger nuestra privacidad, aunque sea una pequeña parte de ella.

BIBLIOGRAFÍA

Ciencia de los Datos. (2020). IBM Analytics. Recuperado de https://www.ibm.com/analytics/mx/es/technology/data-science/

Colmenarejo, R. (s.f.). Ética y big data. Universitat Oberta de Catalunya. Recuperado de http://cv.uoc.edu/annotation/abf6be7bfa5bb7fc599e2dbc57fa2e8d/588056/PID_00243549/modul_1.html#w26aab5c13

Cwaik, J. (2017). Los dilemas éticos del Big Data. Análisis & Opinión. Recuperado de https://www.americaeconomia.com/analisis-opinion/los-dilemas-eticos-del-big-data

Hunter, J., Dale, D., Firing, E., Droettboom, M & Colaboradores. (2012). Mplot3d tutorial. Recuperado de https://matplotlib.org/mpl_toolkits/mplot3d/tutorial.html

Hunter, J., Dale, D., Firing, E., Droettboom, M & Colaboradores. (2012). Pyplot tutorial. Recuperado de https://matplotlib.org/tutorials/introductory/pyplot.html

Microsoft Azure. (2020). *El ciclo de vida del proceso de ciencia de datos en equipo.* Recuperado de https://docs.microsoft.com/es-es/azure/machine-learning/team-data-science-process/lifecycle

Microsoft Azure. (2020). *Fase de descripción de negocio del ciclo de vida del Proceso de ciencia de datos en equipo.* Recuperado de https://docs.microsoft.com/es-es/azure/machine-learning/team-data-science-process/lifecycle-business-understanding

Pandas Basics. (2020). Learn Python. Recuperado de https://www.learnpython.org/es/Pandas%20Basics

Sklearn.linear_model.LinearRegression. (2019). Scikit learn. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Sklearn.metrics.r2_score. (2019). Scikit learn. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html#sklearn-metrics-r2-score

Sklearn.model_selection.train_test_split. (2019). Scikit learn. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Tecnológico de Monterrey. (2020). *Situaciones Problema. Enfrentando la obesidad en México con alfabetización nutrimental.* Matemáticas y ciencias de datos para la toma de decisiones. Recuperado de https://experiencia21.tec.mx/courses/22056/pages/situaciones-problema

Waskom, M. (2020). Seaborn: statistical data visualization. Recuperado de https://seaborn.pydata.org/