Proyecto de ciencia de datos

Naomi Padilla Mora

A01745914

08 de junio, 2021

Matemáticas y Ciencia de datos para lo toma de decisiones MA1042

Profesora Elisabetta Crescio

Introducción

La ciencia de datos es la ciencia que combina la estadística, los métodos científicos y análisis de datos con el objetivo de extraer información de valor de los datos con los que se trabaja, es decir, que convierte los datos en información útil. El rol que juega la ciencia de datos en la sociedad actual es sumamente importante, puesto que esta presente en la mayor parte de las tareas o proyectos de la sociedad. La ciencia de datos ayuda al desarrollo de las empresas, proyectos y a el desarrollo de la sociedad en general, ya que la aplicación de esta a bases de miles de datos puede optimizar el tiempo y las tareas ha realizar. Además, con las diversas técnicas de análisis de datos, se pueden realizar cosas tales como la predicción de datos en base a la creación de un modelo como una regresión lineal, exponencial, un árbol de decisión, entre muchos otros. Es por ello, que la intención de este proyecto consiste en un acercamiento a la ciencia de datos a través de la creación de una base de datos sobre registro alimenticio del estudiante y su análisis con diversos métodos de análisis de datos. A continuación, se presentarán las cuatro fases de la situación problema desarrollas a lo largo del curso y sus resultados.

Fase 1: Entendimiento del negocio

¿Quién es el cliente?

Basándonos en la finalidad que el análisis de datos en esta particular situación llegamos a un punto donde creemos que los principales clientes pueden ser personal de la salud, en especial los que están dedicados al estudio de la alimentación humana y los efectos que causa en el metabolismo y salud de cada persona. Es decir, creemos que los nutriólogos y bariatras quienes específicamente se dedican a la prevención de la obesidad en las personas. En México de acuerdo al Instituto Nacional de Estadística y Geografía (INEGI) en México existen 2.4 nutriólogos por cada 1000 habitantes, lo que representa un mercado muy grande a tan solo nivel nacional, a esto debemos agregar la cantidad de bariatras en el país y el mercado se vuelve aún más grande. Como cliente objetivo podemos ver de igual manera a las personas que quieren llevar un régimen alimenticio moderado.

¿Qué problemas estás tratando de resolver?

Además de facilitar al cliente su forma de presentar y crear dietas personalizadas, el objetivo es que estas dietas impacten de manera positiva a los pacientes ayudándolos a manejar su peso.

Otra área de oportunidad que pudimos identificar fue la dificultad y poca eficiencia de los registros de los nutriólogos, ya que muchas veces son muy tardados y tediosos de realizar, por lo tanto con este desarrollo de la aplicación podremos eficientar estos procesos y así poder facilitarle la tarea a los especialistas de la salud en la realización de las dietas e incluso en la personalización de las mismas con la facilidad que ofrecería la aplicación.

Hablando sobre un tema que es de suma importancia en México en el sentido de preocupación es la Obesidad y sobrepeso, ya que según cifras del INEGI, en 2018, de la población de 5 a 11 años, 18% tiene sobrepeso y va en incremento conforme aumenta la edad; 21% de los hombres de 12 a 19 años y 27% de las mujeres de la misma edad, presentan sobrepeso. En la población de 20 años o más, los hombres (42%) reportan una prevalencia más alta que las mujeres (37 por ciento). Por lo tanto, poder crear y desarrollar herramientas que ayuden a minimizar estas grandes problemáticas son de suma importancia para ayudar a toda la población afectada por estos problemas.

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

Facilidad para crear dietas personalizadas, con el fin de que las personas obtengan los nutrientes que necesitan mediante la realización de la dieta.

Modelo de predicción de calorías nos apoya para poder crear una dieta balanceada, donde las personas consuman los nutrientes necesarios para obtener una buena salud, además gracias a este modelo se logra identificar los productos que contienen más calorías, grasas, carbohidratos, sodio y la relación que existe entre estos, por lo que al analizar todos estos datos el modelo de predicción de calorías, en conjunto con la ciencia de datos nos lleva al resultado de encontrar cual es la relación existente y cuantas calorías aproximadamente deberíamos consumir para obtener todos los nutrientes para tener una buena salud.

Otra solución la cual se planea resolver gracias a la ciencia de datos, es decir el modelo de predicción de calorías es, el hecho de que con esto se le facilitará al nutriólogo especialista el realizar una dieta para sus clientes, además de que esto también apoyara a los clientes del nutriólogo facilitandoles el registrar sus comidas y llevar una mejor cuenta de sus alimentos y nutrientes.

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

Conocimiento sobre herramientas estadísticas, tales como la media, mediana, correlación, y regresión lineal. Además conocimientos básicos de la herramienta Excel como el uso de las

fórmulas y la creación de diversos tipos de gráficos, interpretación de datos e interpretación de gráficas, puesto que esto permitirá interpretar los resultados obtenidos de los usuarios.

¿Qué deberás hacer para desarrollar tu solución?

Con el fin de presentar una manera más eficaz y controlada sobre el control alimenticio que llevan las personas, nuestra propuesta es el desarrollo de un app donde los clientes del personal de la salud, enfocándonos más en nutriólogos y bariatras, sean capaces de registrar los alimentos y bebidas que consumen diariamente, de tal manera que, el especialista pueda llevar a cabo un análisis no sólo de manera remota sin la necesidad de que sus pacientes asistan personalmente, sino que también de un modo eficiente, pues la aplicación contará con herramientas de estudio y observación, así como estadísticas de regresión lineal y la correlación entre los datos, además de la interpretación de información y gráficas que la misma ofrecerá.

Fase 2: Entendimiento de los datos

¿Qué tipos de datos se necesitaron?

Se necesitaron datos nutricionales para la creación del *dataframe*, respecto a la bitácora realizada diariamente de las comidas consumidas, en base a la categoría de estas.

¿De dónde se obtuvieron los datos?

Los datos se obtuvieron de la aplicación *Lose It!*

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

Sí, los datos son adecuados para el análisis, no hay datos vacíos, ni valores negativos que afecten al análisis de estos.

Fase 3: Preparación de los datos

Para la realización del análisis de los datos uno de los ajustes más importantes fue utilizar la base de datos *nutrimentales* actualizada, con un tamaño de 318 filas por 8 columnas sin embargo, a través de la herramienta de programación Python se modificó este *dataframe* de manera que el programa solo contemplara 318 filas y 5 columnas, las cuales son aquellas que representan las variables que generarán el modelo de regresión, estas columnas son *calorías*, *grasa*, *proteínas*, *carbohidratos* y *sodio*. Representando la primera fila del *dataframe*, por lo que se cuenta con 317 datos nutrimentales.

Fase 4: Modelación de los datos

Regresión lineal de Sklearn

La regresión lineal de Sklearn se encarga de ajustar los datos a un modelo lineal para minimizar la suma residual de los cuadrados entre los objetos del conjunto de datos y los objetos predichos por la aproximación lineal. Para la correcta aplicación de este modelo al *dataset* se importaron las librerías correspondientes y se realizó el código de manera eficiente. Los resultados obtenidos son los siguientes:

Mean score	0.4984666003558421
Puntaje de entrenamiento	0.7482943790323618
Puntaje del test	0.7306197369598382
MAE	59.46375366707714

Tabla 1: Resultados regresión lineal de Sklearn

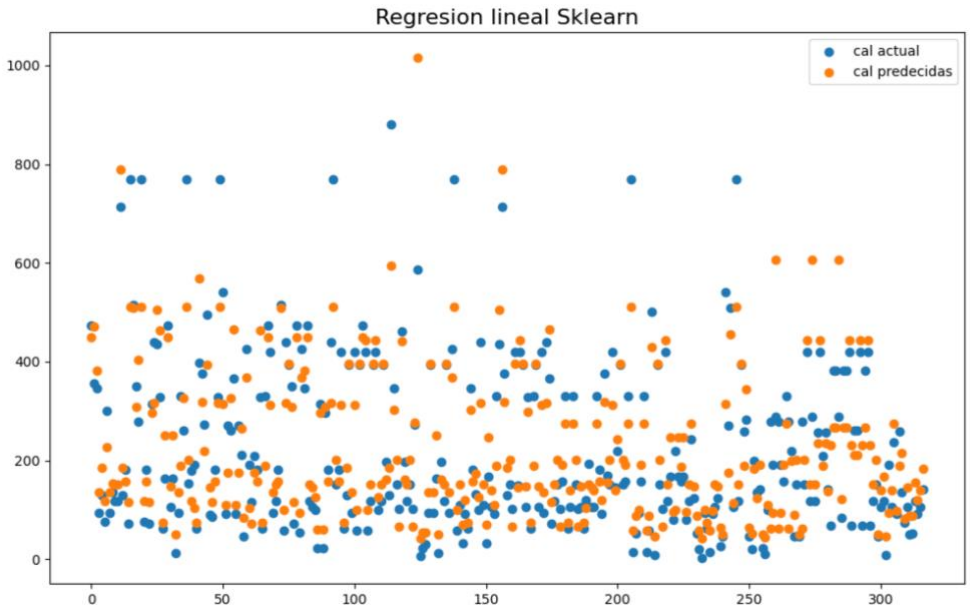

Ilustración 1: Gráfica con la Regresión lineal de Sklearn

Como se observa en la ***tabla 1*** y en la ***Ilustración 1*** el modelo no resulta muy eficiente. En la ***tabla 1*** se aprecia el valor de entrenamiento para ese ajuste aleatorio de datos es de 0.7482… se puede decir que el valor no es muy alto como se espera, incluso con un nuevo arreglo de datos como se comprueba en el valor del test, donde este incluso muestra que su eficiencia es menor, demostrando que además, el modelo presenta un poco de *Overfitting*. Aunque el modelo puede permitir seguir generando diversos arreglos aleatorios que pueden aumentar o disminuir los puntajes tanto en el entrenamiento como en el test, al observar el

mean score (o calificación promedio) podemos darnos cuenta en promedio de todos los arreglos posibles la eficiencia que se espera del modelo es de tan solo casi el 0.5, esto significa que el modelo no será eficiente para otros datos. Analizando estos valores se pueden dar las siguientes interpretaciones, puede que varios de los datos en *dataset* tengan valores muy alejados al resto, esto genera que la distancia de estos datos y la regresión lineal sea mayor, lo cual afecta la eficiencia del modelo, en otras palabras, el modelo que representa al comportamiento de los datos, no es un modelo lineal, es posible que sea porque la aplicación de la que se obtuvo la información nutrimental no está construida con modelos lineales o incluso, la forma de registrar ciertos alimentos más complejos en su información nutrimental, como una hamburguesa, tamales, etc., pudo desfavorecer al modelo.

Decision Tree Regressor

El Decision Tree Regressor es un método de aprendizaje supervisado utilizado para la clasificación y la regresión. Este se encarga de predecir una variable en base al aprendizaje de las características de los datos, generando a su vez una curva seno. Este fue el regresor que mejores resultados dio sobre el *dataset*. La aplicación de este se puede observar en dos pasos, la generación del modelo Decision Tree y la generación del mismo modelo pero de una manera optimizada en base al número de ramas del alcance óptimo del árbol de decisión según el *mean absolute error* (MAE) que generan. Los resultados obtenidos fueron los siguientes:

Puntaje de entrenamiento	0.9985597979267136
Puntaje del test	0.9494162771322203
MAE	12.778637566137567

Tabla 2: Resultados del Decision Tree Regressor

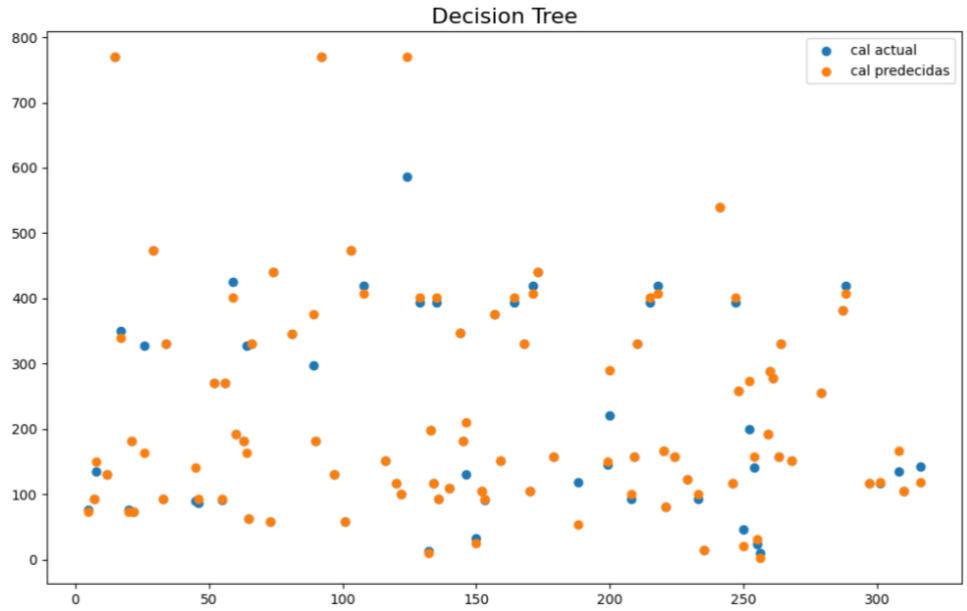

Ilustración 2: Gráfica de Decision Tree Regressor

Puntaje de entrenamiento	0.9975823619448028
Puntaje del test	0.9507951116253163
MAE	14.022421386483884

Tabla 3: Resultados de Decision Tree Regressor Optimizado

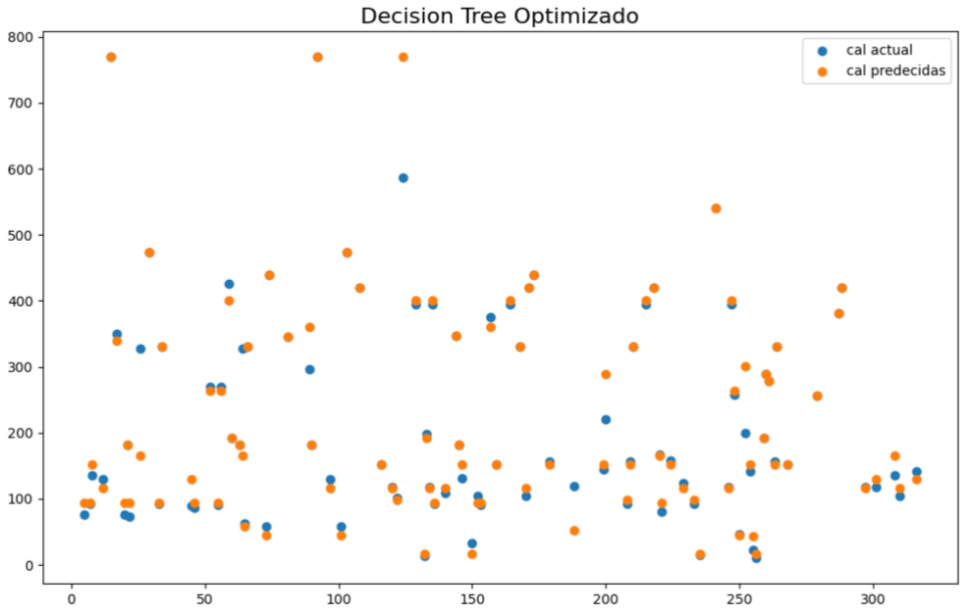

Ilustración 3: Gráfica de Decision Tree Regressor Optimizado

Como se observa en la ***tabla 2*** y en la ***Ilustración 2*** el modelo es mucho más eficiente no solo en el entrenamiento sino también en el test. En la ***tabla 2*** se observa claramente como aumentó el valor tanto del entrenamiento como del test, esto confirma que el modelo no se trata de un modelo lineal y es por ello que el anterior, utilizando la regresión lineal de Sklearn era deficiente. De igual forma, se puede observar como el MAE ahora es incluso menos de la tercera parte del MAE en Sklearn, esto se puede ver de manera más clara en el ***Ilustración 2*** donde se observa como los puntos naranjas (cal predichas) coinciden en mayor cantidad con los puntos azules (cal actuales). Incluso, se puede observar que el modelo presenta un poco de *Overfitting* ya que el puntaje en el entrenamiento es mayor al valor del test, sin embargo, este no es tan grande y por ello se le puede considerar un buen modelo o un modelo eficiente.

Por otro lado, lo que se observa en la ***tabla 3*** y la ***Ilustración 3*** es la representación de los resultados de la optimización del modelo de Decision Tree. Aunque a simple vista parece que esta "optimización" solo disminuyó el puntaje del entrenamiento y aumentó el MAE, lo mas destacable de este, es que aumentó, aunque solo un poco, el puntaje del test, la importancia de esto es que gracias a que el puntaje del test aumenta y el del entrenamiento disminuye, el nivel del *overfitting* también disminuye, además de que el MAE continúa siendo menor a la tercera parte de lo que se tiene en la regresión lineal de Sklearn. De igual forma, en la ***Ilustración 3*** se puede observar como los puntos naranjas coinciden en mayor medida con los puntos azules, un poco menos que en la ***Ilustración 2*** pero si de manera considerablemente más eficiente que en la ***Ilustración 1***, por que se puede señalar que esta presentación optimizada del Decision Tree también representa un modelo eficiente.

Comparación de los resultados con el método OLS

El método *Ordinary Least Squares* (OLS) consiste en una regresión lineal, este método se aplicó al *dataset* en la herramienta Excel y también en el software de programación Python a manera de comprobar los resultados. Los resultados obtenidos en ambos softwares son basicamente los mismos, esto de puede observar en la ***tabla 4***.

Regresión lineal Excel		OLS Python	
R^2	0.749903224777658	R^2	0.7499032247776583
R^2_{aj}	0.746696855864551	R^2_{aj}	0.7466968558645513
Intercepción	37.0103479750356	Intercepción	37.010348

Grasas	2.3382112078047	Grasas	2.338211
Proteínas	6.3570299105673	Proteínas	6.357030
Carbohidratos	3.76426322211652	Carbohidratos	3.764263
Sodio	0.0230731984381168	Sodio	0.023073

Tabla 4: Comparación OLS de Excel vs Python

De igual forma, en base a los datos de la ***tabla 4***, podemos saber que el modelo lineal no es muy eficiente, puesto que, como ya se menciono, el *dataset* no está representado por un modelo lineal y por lo que al evaluarlo con este método no es tan bueno.

OLS			R. L. de Sklearn	Decision Tree	D. Tree Optimizado
R^2	0. 749903	Train	0.7482943790323	0.9985597979267	0.9975823619448
R^2_{aj}	0. 746696	Test	0.7306197369598	0.9494162771322	0.9507951116253

Tabla 5: Comparación método OLS, Sklearn, Decision Tree Regressor

En base a lo que se observa en la ***tabla 5***, podemos ver como los resultados del método OLS coinciden o son muy similares a los que se obtienen con el método de Sklearn, esto porque ambos consisten en una regresión lineal a grandes rasgos, incluso en ambos se concuerda en el grado de eficiencia que presenta este modelo. Por otro lado, se puede observar como el modelo incrementa considerablemente su eficiencia con los métodos de Decision Tree Regressor, tanto el normal como el optimizado y como ya se mencionó y se observa en la ***tabla 2*** y ***tabla 3*** los MAE no son muy grandes. Analizando estos datos como lo obtenido en las regresiones lineales a comparación de lo obtenido con el método de Decision Tree, el aplicar *Machine Learning* con métodos como el Decision Tree resulta muy benefactor para el modelo puesto que el método permite encontrar el modelo que describe de manera correcta el comportamiento de los datos, permitiendo que este sea más eficiente para el arreglo de datos y otros datos. Se puede concluir que todos los modelos probados a lo largo de este trabajo son de gran utilidad y su eficiencia depende del *dataset* y de como este se recolectó, en este caso, al trabajar con alimentos complejos el registro de la información nutrimental se vio afectado, sin embargo, gracias a la variedad de métodos o de regresores que se aprendieron fue posible seleccionar uno de ellos que beneficiara al modelo. Además, como se observa en este caso, no todos los modelos son lineales y eso no significa que no sean

buenos modelos, solo es necesario encontrar el método adecuado e incluso comparar los resultados puede resultar muy enriquecedor para analizar al modelo. De igual forma, se puede concluir que aunque el modelo presente *Overfitting* no implica que sea un mal modelo, dependiendo del nivel del *Overfitting*, ya que al ser poco, la eficiencia del modelo no se ve realmente muy afectada.

Fase 5: Reflexión final

En conclusión, el desarrollo de este trabajo ejemplifica un breve uso y significado de la ciencia de datos, comprobando su importancia y como es que está esta presente en la vida diaria. Es gracias a la ciencia de datos que esta clase de análisis es posible y permite a los científicos de datos identificar problemáticas y a su vez solucionarlas a base de información que muchas veces se tiene al alcance de todos. La ciencia de datos es importante porque nos permite comprender el funcionamiento, en mayor parte estadístico, de la información que nos rodea, es gracias a la ciencia de datos que es posible general algoritmos que representan el comportamiento de las cosas y generan patrones, como los algoritmos que utilizan las plataformas de *streaming* para recomendar películas según a las preferencias que observa en el usuario, o la manera en la que las aplicaciones de navegación recaudan los datos de tráfico de diversos usuarios a tal velocidad que generan la ruta más óptima para que el usuario llegue a su destino o en este caso, para realizar un modelo personalizado en base a los alimentos que consume el usuario y así poder predecir las calorías que se consumirán, teniendo una posible aplicación en la generación de un modelo de negocios como el que se propone en la ***Fase 1***. Por otra parte, es sumamente importante mencionar que como en muchas de las ciencias, en la ciencia de datos también está presente la ética, la cual es sumamente importante en las aplicaciones de la ciencia de datos porque se pueden presentar diversas situaciones o dilemas como el uso de base de datos de manera *ilegal* o considerada poco profesional, o incluso el uso de datos personales sin la autorización de los usuarios, como ocurrió en el caso *Cambridge Analytica* y *Facebook*. El campo de aplicación de la ciencia de datos es tan grande que incluso ciertos análisis pueden tener un gran impacto político y social, como se presenta en el ejemplo antes mencionado, es por ello que es de suma importancia realizar estos análisis de un forma éticamente profesional y con motivos que beneficien al desarrollo de la sociedad o diversos campos laborales, respetando la privacidad del usuario,

la manera de distribuir y recolectar la información y la forma de proteger los resultados de los mismos para impedir que estos puedan utilizarce de manera inapropiada.

Referencias

Amaya, L., & Berrío, G. (2021, 31 marzo). *Dilemas Éticos*. Ética psicológica. http://eticapsicologica.org/index.php/documentos/articulos/item/7-dilemas-eticos

BBC News Mundo. (2018, 21 marzo). *5 claves para entender el escándalo de Cambridge Analytica que hizo que Facebook perdiera US$37.000 millones en un día*. https://www.bbc.com/mundo/noticias-43472797

Blancas, E. (2015, 26 noviembre). *Data Science en México: retos y oportunidades*. Medium. https://medium.com/@edublancas/data-science-en-m%C3%A9xico-retos-y-oportunidades-4a68f683b777

Cwaik, J. (2017, 20 enero). *Los dilemas éticos del Big Data*. AméricaEconomía. https://www.americaeconomia.com/analisis-opinion/los-dilemas-eticos-del-big-data