TECNOLÓGICO DE MONTERREY - CAMPUS GUADALAJARA

Entregable 2: Entendimiento de los datos

Materia: Matemáticas y ciencia de datos para la toma de decisiones (Gpo 800)

Fecha: 24 de Abril 2022

Profesor : David Rivera

Realizado por: Liliana Solórzano Pérez | A01641392

Primera Parte: Código

```
import pandas as pd
"""we import the pandas libary with the name pd instead
# we create a variable that will contain the data of our excel file
# we have to indicate that the data is in an excel file and use the function .read
# we have to be sure that the data file and the code are in the same folder
Note: be sure that the filename is short and easy"""
data = pd.read_excel('datos_nutricionales.xlsx')

# we use the function head to get the firts 5 register of our data sheet
data.head()

#we use shape to get to know the total number of colums and rows
data.shape #(rows, columns)
#Expected output value: (152, 9)

#we use columns to visualize the name of all the columns in our file
data.columns

#with the dtypes we can get to know the data type of our dataframe
data.dtypes
"""Los principales tipos de datos son:
object: cadena de texto
int64 : número entero
float64: número con decimal
datetime64: fecha y hora"""

#function info return the complete information of our dataframe
data.info()

#function describe() is used to obtain the description and the statistics
data.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 152 entries, 0 to 151
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Fecha (dd/mm/aa)    152 non-null    datetime64[ns]
 1   Momento             152 non-null    object
 2   Nombre alimento     152 non-null    object
 3   Calorías (kcal)     152 non-null    int64
 4   Carbohidratos (g)   152 non-null    float64
 5   Lípidos/grasas (g)  152 non-null    float64
 6   Proteína (g)        152 non-null    float64
 7   Sodio (mg)          152 non-null    float64
 8   Fuente              152 non-null    object
dtypes: datetime64[ns](1), float64(4), int64(1), object(3)
memory usage: 10.8+ KB
```

	Calorías (kcal)	Carbohidratos (g)	Lípidos/grasas (g)	Proteína (g)	Sodio (mg)
count	152.000000	152.000000	152.000000	152.000000	152.000000
mean	235.835526	33.151704	10.846382	17.553355	282.179342
std	201.735170	56.790944	12.156999	23.532098	464.565388
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	103.000000	1.960000	2.060000	1.540000	2.000000
50%	156.000000	11.030000	7.470000	12.580000	91.500000
75%	372.000000	47.795000	15.210000	19.610000	211.000000

Segunda Parte: ¿En qué consiste la Fase 2: Entendimiento de los datos?

Respuesta: En esta fase, se puede analizar un conjunto de datos desde una perspectiva más matemática, al igual que, mayormente enfocada a la programación. En esta fase, podemos desarrollar un pensamiento crítico enfocado al entendimiento de un programa, del mismo modo que el funcionamiento de algoritmos, librerías y funciones del lenguaje de programación Python 3. Es de vital importancia conocer a profundidad los datos con los que se están trabajando para así, proveer el funcionamiento de un programa de código de acuerdo a lo que se requiere.

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

Respuesta: Los datos registrados equivalen a los que se registraron con base en el consumo nutricional. Los datos adquiridos son los datos de estadística, como los que se realizaron en excel para obtener modelos de regresión, al igual que, saber cuales datos eran importantes para neustro

modelo matemático y descartar lo que no eran importantes, por otra parte, los datos adicionales o generados, equivalen a nuestro modelo matemático.

2. ¿Qué tipos de datos se analizarán?

Respuesta: Se analizarán datos de tipo 'float', 'int', 'object' y 'datetime'

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

Respuesta: De acuerdo con las actividades realizadas anteriormente, se concluyó que los atributos de las calorías, carbohidratos, lípidos, y proteínas son los relevantes para poder establecer un modelo matemático.

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

Respuesta: Como se mencionó anteriormente, uno de los atributos que puede ser excluido es el sodio, ya que, este no tiene gran impacto al momento de realizar un modelo matemático, de la misma forma que en las pruebas de estadística (regresión) se comprobó que el valor del sodio era mayor a nuestro valor crítico, es decir, se concluyó que no era significativo para nuestro modelo matemático.

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

Respuesta: En este caso se pueden llegar a hacer predicciones y algunas conclusiones generalizables, sin embargo, no se pueden hacer predicciones con tanta precisión, ya que, para esto, se necesitaría mayor cantidad de datos, ya que, entre mayor sea la cantidad de datos a analizar, mejor será la predicción estadística junto con su modelo.

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

Respuesta: Existen los atributos suficientes para realzar un modelo de fácil interpretación.

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?

Respuesta: Los datos se obtuvieron de una recopilación de estos día tras día de una fuente meramente objetiva. Como se ovserba, sí se pueden fusionar varias fuentes de datos, ya que, hay algunos datos repetidos, esto ocasionaría una tendencia en los datos, lo que podría presentar un problema en un futuro, ya que se busca obtener la mayor cantidad de diversos datos.

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

Respuesta: En este caso para manejar los valores faltantes de las fuentes de datos se pueden manejar con una regresión estadística. Con el análisis de una gráfica y determinando una

tendencia de estos mismos, para así, analizar los datos faltantes.

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

Respuesta: En el archivo que se tomó para la base de datos, se encuentran disponibles 152 filas con 9 columnas

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

Respuesta: La relación de los datos y la hipóteis del proyecto es demostrar el consumo nutricional de una persona promedio en México, ya que, en la mayoría de los casos, como se observó anteriormente, se consume una cantidad mayor de calorías a la que se debería de consumir normalemente. En pocas palabras se busca concientizar al ser humano sobre su consumo nutricioanl, aporeciando lo que aporta cada alimento a su organismo.

0 s completado a las 17:49