Documento Explicativo

Trabajo Grupal

Autor: Alejandro Barragán, Pedro Huerta Cicero, Lucía Sánchez

11 de Febrero de 2024

1. Contenido

En la carpeta se incluyen dos scripts con código:

1. Obesidad_streamlit.py
2. Trabajo_final.ipynb
3. Documento pdf explicativo

2. Librerías

Para poder ejecutar los scripts adjuntos se necesitan las siguientes librerías:

1. Pandas
2. Numpy
3. Matplotlib
4. Plotnine
5. Plotly
6. Os
7. Streamlit

3. Preprocesado

El notebook de jupyter contiene la carga, procesado y limpieza de los datos y las gráficas creadas para obtener insights acerca del dataset. Junto al código, se ha ido comentando por qué hemos dado cada paso, pero aquí se realizará una breve explicación del proceso que hemos llevado a cabo para obtener unos set de datos útil para la obtención de insights después.

Para el procesado de datos, hemos comenzado viendo que estructura tenía nuestro DataFrame con un head. La primera observación ha sido darnos cuenta de que los nombres de las columnas podían llegar a dificultarnos el trabajo pues algunos de ellos no eran para nada intuitivos. El primer paso ha sido entonces cambiar el nombre de estas.

Por la información ofrecida en la página de donde se ha extraído el set de datos, sabíamos que este no contaba con missing values, pero, para gozar de mayor rigurosidad, decidimos comprobarlo. Para ello, creamos otro data frame, eliminando todos los posibles valores nulos y comprobando que, efectivamente tiene los mismos registros que nuestro set original.

A continuación, nos dedicamos a observar si encontramos algún dato incorrecto o irregular en las diferentes variables. Comenzamos con las variables no numéricas. Hemos decidido categorizar una serie variables que de entrada eran numéricas como el tipo de obesidad, el consumo de alcohol o las comidas que se realizan entre comidas principales. Tener estas variables como categóricas es importante para poder realizar una visualización adecuada y estudiar la relación entre variables más adelante.

Con las variables numéricas, calculamos sus estadísticos básicos que nos podían dar una vista amplia de los datos que tenemos. Efectivamente, con ello nos dimos cuenta de que el set de datos contaba con una media de edad bastante baja. El set de datos tenía muy pocos registros de personas mayores de 40 años y, por tanto, por su poca representatividad, decidimos eliminarlos del estudio.

Seguimos observando las características de las variables y pasamos a analizar si, tanto el peso como la altura, también seguían una distribución lógica, para ello nos servimos de sus gráficos de densidad y parecía todo en orden.

Las variables numéricas contaban con una serie de ellas que deberían ser categóricas, pero, como una parte del set de datos se generó sintéticamente, estas variables tomaron valores numéricos y tenemos que categorizarlas de nuevo. Así contamos ya con un set de datos limpio para poder realizar nuestra visualización y conclusiones.