

A1.1 Aprendizaje estadístico-automático

Los niveles de obesidad de la población son cada vez más preocupantes, sobre todo en países latinoamericanos. Por lo mismo, científicos de la Universidad de la Costa en Colombia se dieron a la tarea de recolectar información relacionada a este tema para individuos de Colombia, Perú, y México. La base de datos original se encuentra en el UCI Machine Learning Repository, pero una versión simplificada la encontrarás disponible con el nombre "A1.1 Obesidad.csv" en la misma página de la plataforma donde descargaste esta plantilla.

La base de datos cuenta con la siguiente información:

- **Sexo.** Se describe como femenino `Female` o masculino `Male`.
- **Edad.** Se describe como un número entre `14` y `61`.
- **Estatura.** Se describe como un número, en metros.
- **Peso.** Se describe como un número, en kilogramos.
- **FamiliarConSobrepeso.** Describe si algún familiar ha sufrido sobrepeso `yes` o `no`.
- **ComeMuchasCalorias.** Describe si el individuo come comidas con alto contenido calórico de forma frecuente `yes` o `no`.
- **ComeVegetales.** Indica si el individuo nunca come vegetales en sus comidas `1`, si lo hace algunas veces `2`, o si lo hace siempre `3`.
- **Fumador.** Indica si la persona es fumadora activa `yes` o `no`.
- **ConsumoDeAgua.** Indica si la persona toma menos de un litro de agua al día `1`, entre uno y dos litros de agua al día `2`, o más de dos litros de agua al día `3`.
- **NivelDeObesidad.** Se calcula a partir del índice de masa corporal (peso dividido entre estatura al cuadrado), y se categoriza como: bajo peso `Insufficient_Weight` para valores menores a 18.5, peso normal `Normal_Weight` para valores entre 18.5 y 24.9, sobrepeso tipo I `Overweight_Level_I` y sobrepeso tipo II `Overweight_Level_II` para valores entre 25.0 y 29.9, obesidad tipo I `Obesity_Type_I` para valores entre 30.00 y 34.9, obesidad tipo II `Obesity_Type_II` para valores entre 35.0 y 39.9, y obesidad tipo III `Obesity_Type_III` para valores superiores a 40.0.

¿Qué información te interesaría revisar si formaran parte del equipo de investigación? ¿Qué información te hubiera gustado recopilar y que actualmente no se encuentra en el estudio?


Desarrolla los siguientes puntos en una Jupyter Notebook, tratando, dentro de lo posible, que cada punto se trabaje en una celda distinta. Los comentarios en el código siempre son bienvenidos, de preferencia, aprovecha el markdown para generar cuadros de descripción que ayuden al lector a comprender el trabajo realizado.

1 Importa los datos del archivo "Obesidad.csv" a tu ambiente de trabajo en una Jupyter Notebook y muestra en consola un resumen, imprimiendo las primeras 10 filas de datos. Recuerda que es conveniente que el archivo "csv" y tu libreta estén en el mismo directorio.

```
In [9]: import pandas as ps
import numpy as np
import matplotlib.pyplot as plt
datos=ps.read_csv('Obesidad.csv')
datos.head(10)
```

```
Out[9]:
```

| | Sexo | Edad | Estatura | Peso | FamiliarConSobrepeso | ComeMuchasCalorias | ComeVegeta |
|---|--------|------|----------|------|----------------------|--------------------|------------|
| 0 | Female | 21.0 | 1.62 | 64.0 | yes | no | |
| 1 | Female | 21.0 | 1.52 | 56.0 | yes | no | |
| 2 | Male | 23.0 | 1.80 | 77.0 | yes | no | |
| 3 | Male | 27.0 | 1.80 | 87.0 | no | no | |
| 4 | Male | 22.0 | 1.78 | 89.8 | no | no | |
| 5 | Male | 29.0 | 1.62 | 53.0 | no | yes | |
| 6 | Female | 23.0 | 1.50 | 55.0 | yes | yes | |
| 7 | Male | 22.0 | 1.64 | 53.0 | no | no | |
| 8 | Male | 24.0 | 1.78 | 64.0 | yes | yes | |
| 9 | Male | 22.0 | 1.72 | 68.0 | yes | yes | |



2 Imprime en consola o genera un cuadro de descripción que muestre un mensaje donde indiques qué otra variable hubieras decidido medir si formarás parte del grupo de investigación, indicando claramente si la información recopilada se consideraría como cuantitativa o cualitativa.

```
In [10]: print("Frecuencia de ejercicio (cualitativa)")
```

Frecuencia de ejercicio (cualitativa)

3 Si tú fueras el líder del proyecto y quisieras realizar un estudio de inferencia, ¿qué variable de la base de datos definirías como la salida o respuesta? Aquella variable para la que te interesa encontrar asociaciones específicas. Imprime el promedio de dicha variable, si es cuantitativa, o la cantidad de personas que pertenecen a cada categoría, si es cualitativa.

- Para calcular el promedio puedes usar la función "mean()" de pandas.
- Para calcular la cantidad de personas que pertenecen a cada categoría puedes usar la función "value_counts()" de pandas.

```
In [11]: sexo=datos.Sexo.value_counts()
print(f"Mujeres: {sexo.Female}\tHombres: {sexo.Male}\n")
```

```

edadMean=datos.Edad.mean()
print(f"Promedio de Edad: {round(edadMean)} años\n")
estaturaMean=datos.Estatura.mean()
print(f"Promedio de Estatura: {round(estaturaMean, 2)} m\n")
pesoMean=datos.Peso.mean()
print(f"Promedio de Peso: {round(pesoMean, 2)} kg\n")
familia=datos.FamiliarConSobrepeso.value_counts()
print(f"Tienen familiares con Sobrepeso\tSi: {familia.yes}\tNo: {familia.no}\n")
calorias=datos.ComeMuchasCalorias.value_counts()
print(f"Come muchas calorías;\tSi: {calorias.yes}\tNo: {calorias.no}\n")
vegetales=round(datos.ComeVegetales).value_counts()
print(f"Ingesta de vegetales en sus comidas:\n\tNunca: {vegetales[1]}\n\tAlgunas ve
fuma=datos.Fumador.value_counts()
print(f"Fumadores: {fuma.yes}\tNo fumadores: {fuma.no}\n")
agua=round(datos.ConsumoDeAgua).value_counts()
print(f"Consumo de agua al día:\n\tMenos de un litro: {agua[1]}\n\tEntre uno y dos
imc=datos.NivelDeObesidad.value_counts()
print(f"IMC (índice de masa corporal):\n\tBajo: {imc.Insufficient_Weight}\n\tNormal

```

Mujeres: 1043 Hombres: 1068

Promedio de Edad: 24 años

Promedio de Estatura: 1.7 m

Promedio de Peso: 86.59 kg

Tienen familiares con Sobrepeso Si: 1726 No: 385

Come muchas calorías; Si: 1866 No: 245

Ingesta de vegetales en sus comidas:

 Nunca: 102

 Algunas veces: 1013

 Siempre: 996

Fumadores: 44 No fumadores: 2067

Consumo de agua al día:

 Menos de un litro: 485

 Entre uno y dos litros: 1110

 Dos litros: 516

IMC (índice de masa corporal):

 Bajo: 272

 Normal: 287

 Sobrepeso I: 290

 Sobrepeso II: 290

 Obesidad I: 351

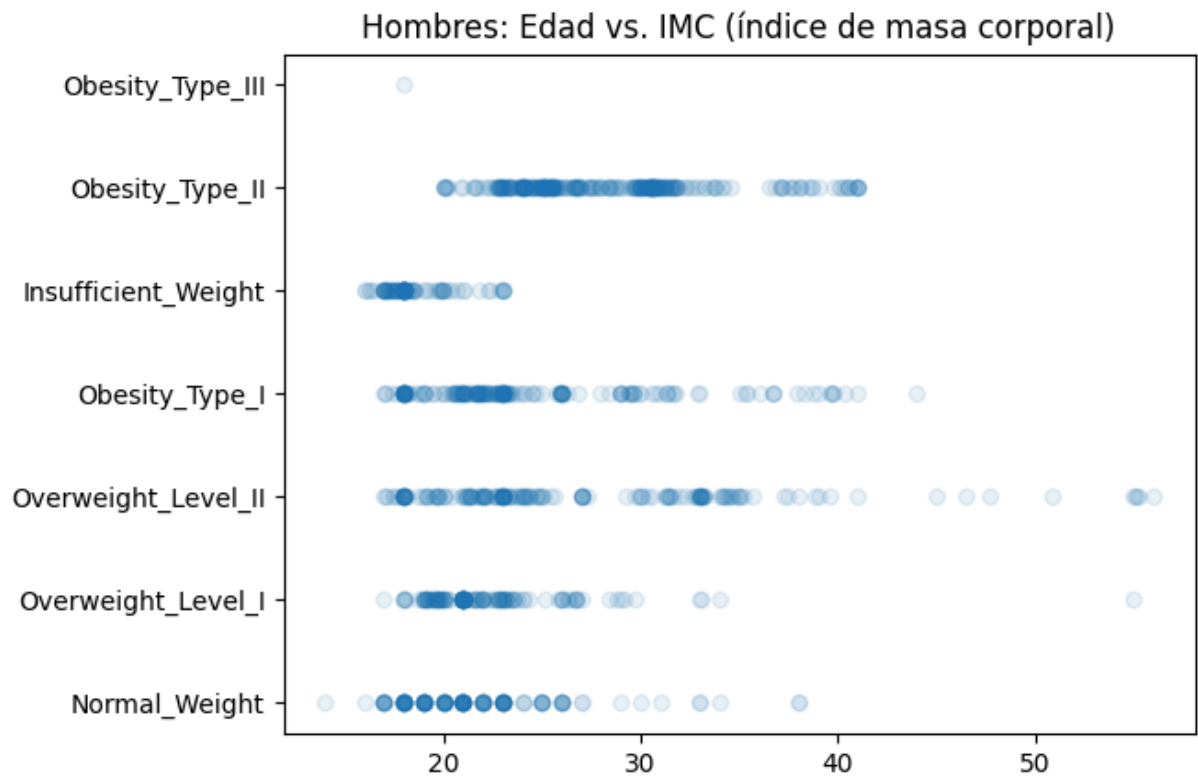
 Obesidad II: 297

 Obesidad III: 324

4 Genera una gráfica, exclusivamente para la población masculina o femenina (cada quién decidirá su población de interés), que muestre en el eje "Y" la variable de salida (de acuerdo a lo que definiste en el punto anterior), y en el eje "X" alguna variable cuantitativa que les

llame la atención. Asegúrate de usar un valor de transparencia menor a 0.1. Recuerda la función mágica que debes agregar para que las imágenes se visualicen sin problemas en una Jupyter Notebook.

```
In [12]: xDatos=datos.Edad[datos.Sexo=="Male"]
yDatos=datos.NivelDeObesidad[datos.Sexo=="Male"]
plt.scatter(xDatos, yDatos, alpha=0.1)
plt.title('Hombres: Edad vs. IMC (índice de masa corporal)')
plt.show()
%matplotlib inline
```



Firma de Honor: Doy mi palabra que he realizado esta actividad con integridad académica