DIPLOMATURA EN CIENCIA DE DATOS, APRENDIZAJE AUTOMÁTICO Y SUS APLICACIONES

FAMAF Facultad de Matemática, Astronomía, Física y Computación

UNC

Córdoba Technology Cluster

CCAD Centro de Computación de Alto Desempeño de la Universidad Nacional de Córdoba

"Ciencia de datos aplicada al estudio de la Obesidad y otras enfermedades crónicas en Córdoba"

Por:

Basmadjian, Osvaldo Martín

Fernández, María Emilia

Romero, Fernando

Mentoras:

Laura Aballay

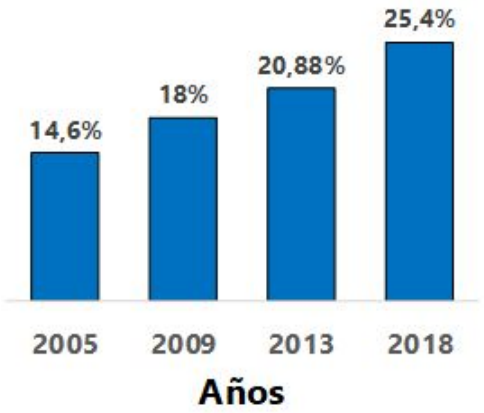
Eugenia Haluszka

DESCRIPCIÓN DEL PROBLEMA: obesidad

PREVALENCIA DE LA PATOLOGÍA

EN 40 AÑOS SU PREVALENCIA SE TRIPLICÓ EN TODO EL MUNDO (OMS, 2017)

Factores de Riesgo en Argentina

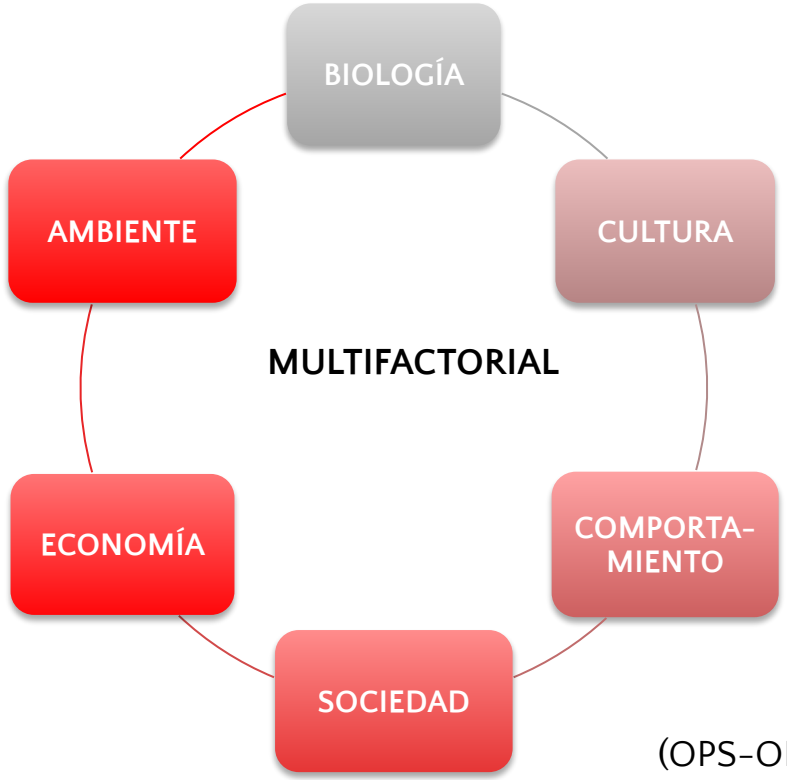


(Encuesta Nacional de Factores de Riesgo, 2019)

EN CÓRDOBA

Más del 50% las personas presentan exceso de peso, 25% del total obesidad

ETIOLOGÍA Y PATOFISIOLOGÍA



(OPS-OMS, 2016)

¿Cómo abordar estos factores de manera integral?

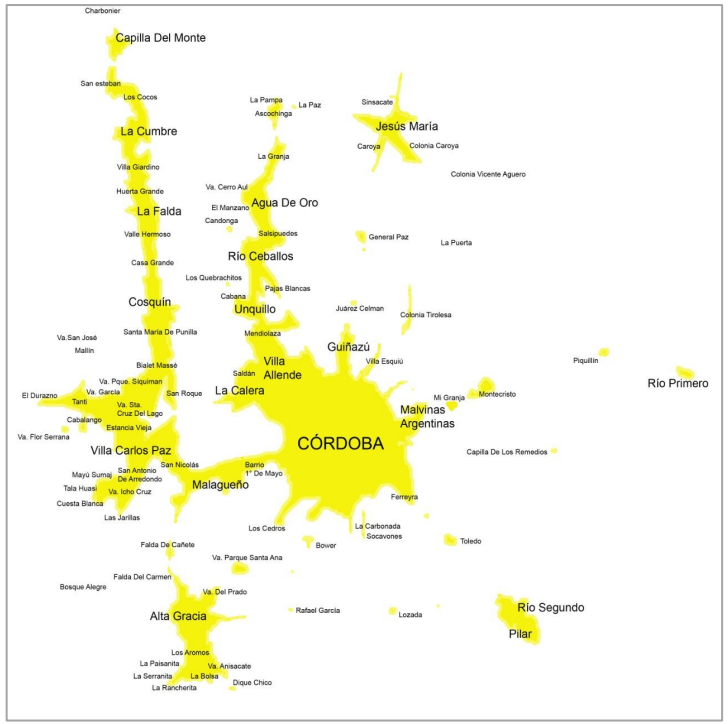
PRÁCTICO 1: Análisis y visualización de datos (1/4)

Objetivo general y alcances:

- Conocer las principales características de nuestra población, y su distribución entre subgrupos.
- Plantear posibles relaciones entre variables a través de visualizaciones adecuadas.

EL DATASET

Población de estudio
Adultxs del Gran Córdoba

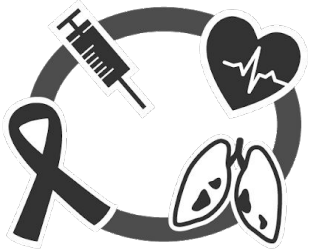


Encuestas
(N= 4292)

Variables
(N=239)

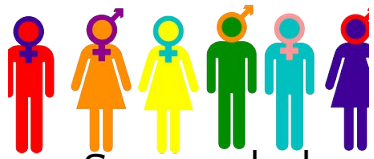
Alimentación

Nivel socio-económico



Antecedentes de enfermedades

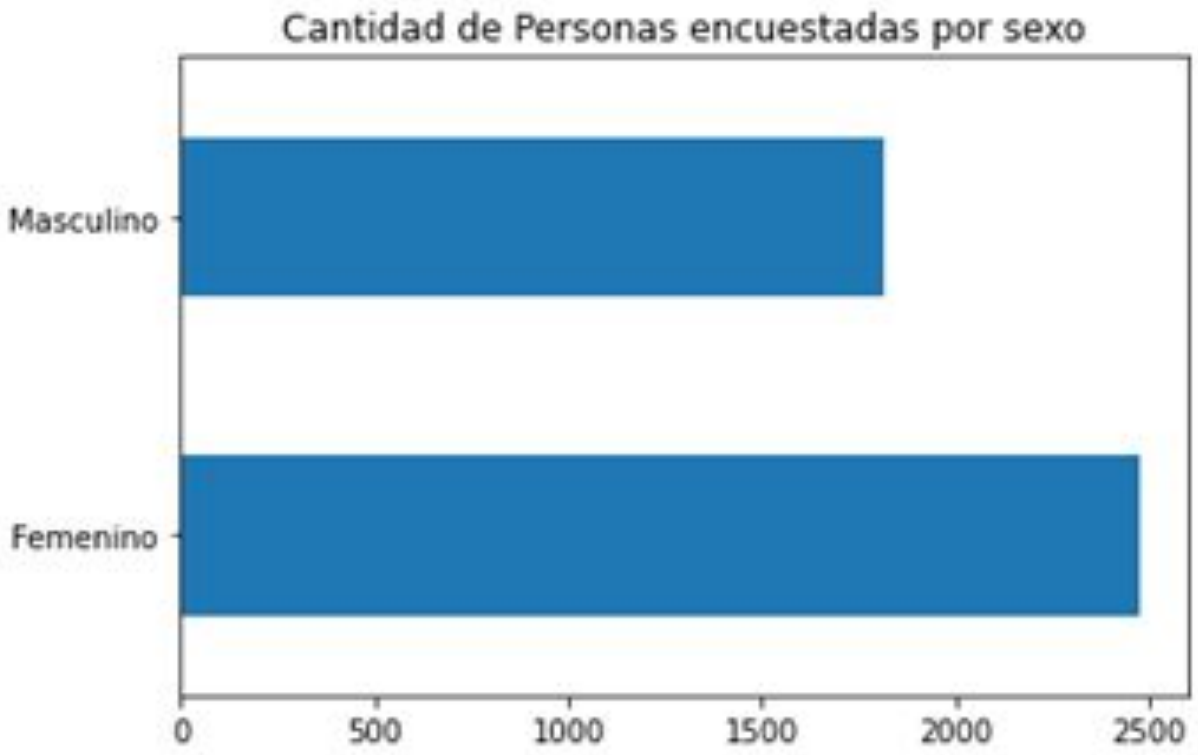
Nivel educativo



Sexo, edad

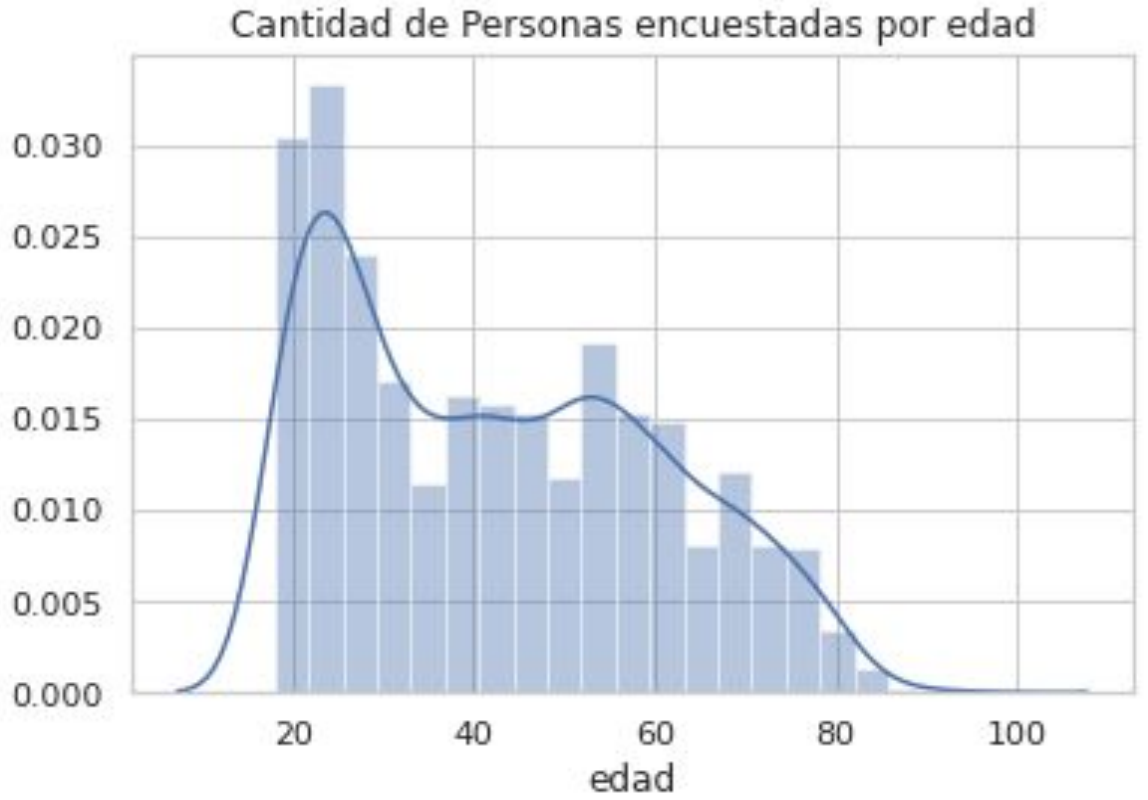
PRÁCTICO 1: Características generales de la población estudiada (2/4)

Barplot para el conteo de la variable categórica "sexo"



Hubo más mujeres encuestadas

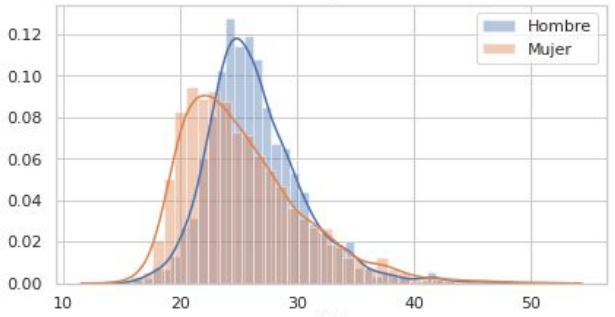
Histograma de probabilidad de la variable discreta "edad"



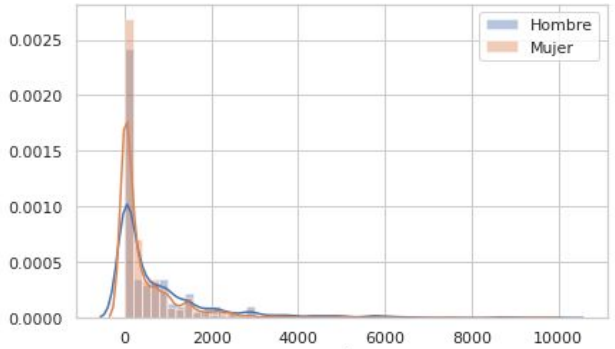
Adultxs entre 18 y 93 años
La mayoría de los encuestadxs tiene entre 20-25 años

PRÁCTICO 1: Análisis univariado (3/4)

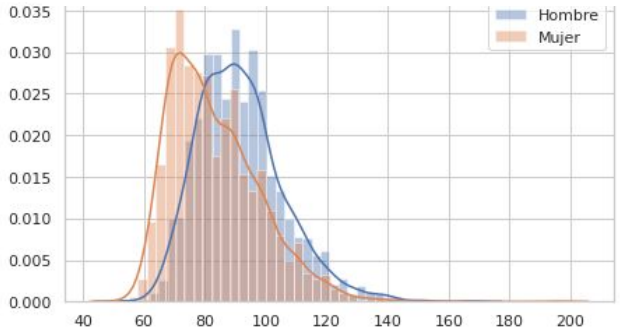
Histogramas para las variables cuantitativas continuas



Índice de masa corporal (IMC)

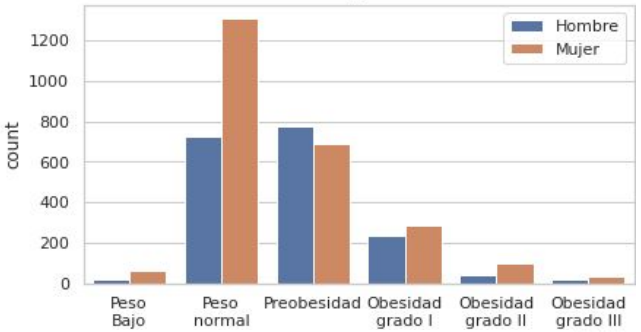


Nivel de actividad física (mets)

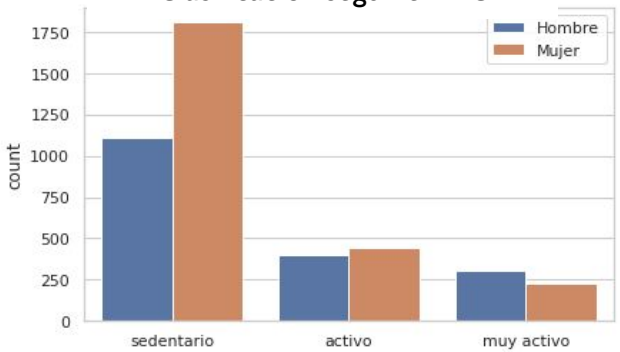


Circunferencia de cintura (cc)

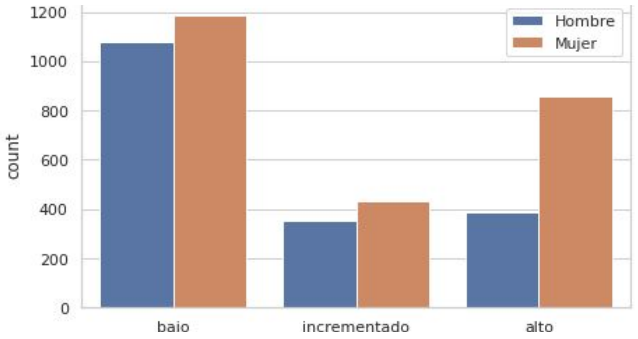
Barplots para las versiones discretas de las mismas variables



Clasificación según el IMC

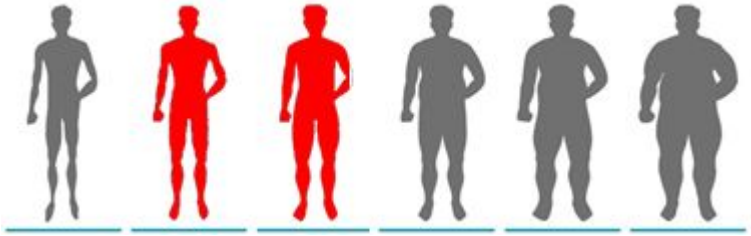


Clasificación según actividad física



Clasificación según cc

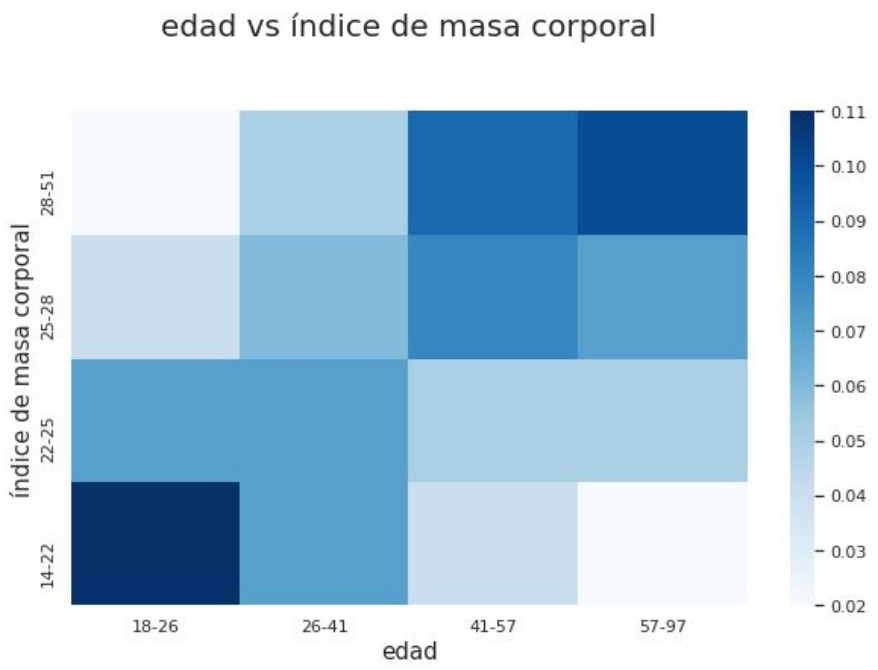
$$\text{IMC} = \frac{\text{peso [kg]}}{\text{estatura}^2\ [\text{m}^2]}$$



¡Hubiera sido mejor construir los barplots en porcentaje para facilitar la comparación entre sexos!

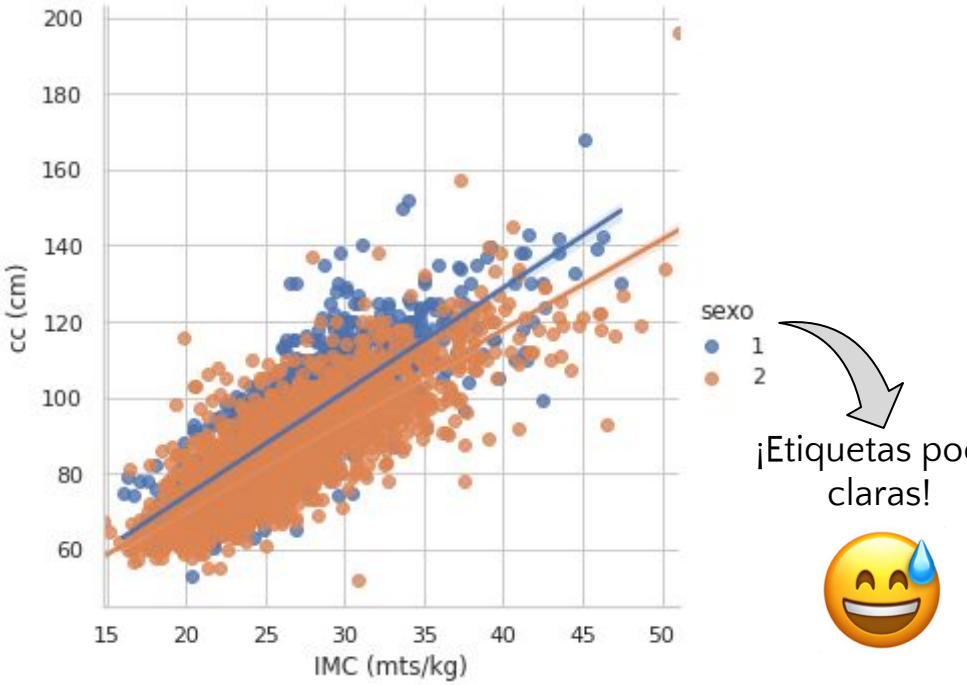
PRÁCTICO: Análisis bivariado (4/4)

Heatmap para la relación entre dos variables categóricas



↑Edad ⟶ ↑IMC

Scatterplot para la relación entre dos variables continuas



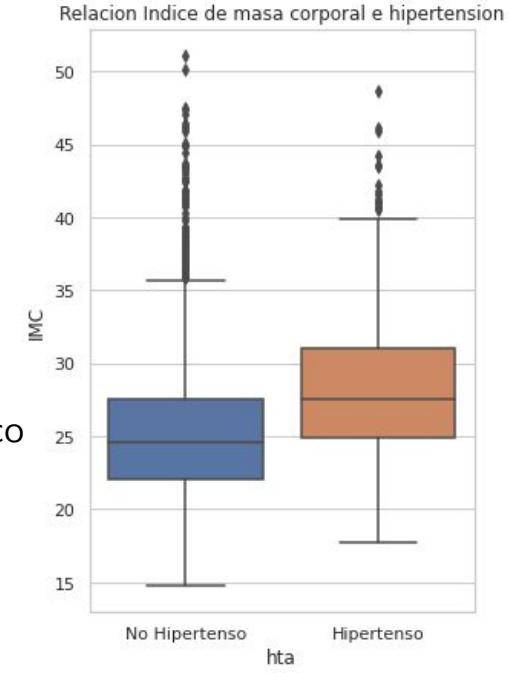
¡Etiquetas poco claras!

↑Circunferencia de cintura (cc) ⟶ ↑IMC

↓ Correlación en mujeres

(↑riesgo cardiovascular a == cc similares a los hombres)

Boxplot entre una variable categórica y una continua



↑IMC ⟶ ↑HTA

PRÁCTICO 2: Curación de datos (1/3)

Objetivos y alcances:

Generar un set de datos apto para ser utilizado para el entrenamiento y aprendizaje de modelos en las materias siguientes.

Codificación de variables → **LabelEncoder**

Generación de etiquetas de variables más descriptivas y normalizadas

tipo	→	act_fis_tipo
Nombre	→	act_fis_nombre
ubic	→	ubic_tumor

PRÁCTICO 2: Curación de datos (2/3)

Codificación de variables sensibles

HashingEncoder

DNI: 10523… → DNI: ‘0322599fc63702617f294d5f702e0a01’

Análisis e imputación de variables observaciones faltantes

Imputación de variables categóricas faltantes a partir de variables continuas derivadas

Análisis de integridad de los datos

normalización de observaciones inconsistentes en la variable “cancer” →

Imputación según:

- Valor en variables de tumor benigno o maligno
- Tipo de cáncer

PRÁCTICO 2: Curación de datos (3/3)

Identificación y eliminación de outliers

↑ p75 + 1.5* rango IC
↓ p25 + 1.5* rango IC

IMC
mets (actividad física)
fgrs2 (grasas insaturadas)

Eliminamos valores extremos de obesidad

Normalización de los tipos de datos

Verificación de tipo de datos correcto para variables continuas, categóricas y de tiempo.
Reducción al tipo de dato que ocupe el menor tamaño posible

Optimizamos el tamaño del dataset

PRÁCTICO 3: Aprendizaje Supervisado(1/4)

Objetivos:

- Definir la variable target, en este caso IMC (Índice de masa corporal).
- Determinar los factores que puedan ayudar a predecir la presencia de obesidad en la población.
- Predecir de forma automática la presencia de obesidad en todo el dataset completo, como así discriminando por sexos.

Para llevar a cabo el objetivo, determinamos la Target como:

- Presencia de obesidad = IMC > 29.9 kg/m2
- Ausencia de obesidad = IMC < 29.9 kg/m2

PRÁCTICO 3: Aprendizaje Supervisado(2/4)

Selección de features:

- Del total de 239 features, seleccionamos un total de entre 15 y 20 features para realizar el análisis y predicción.

- Las features seleccionadas mostraron alguna relación con la variable target.

- Seleccionamos diferentes features dependiendo sobre qué dataset se quería trabajar, sea el completo, la población de mujeres o de hombres.

- En general las features seleccionadas tienen que ver con:
  - Los niveles de actividad física.
  - La alimentación.
  - Enfermedades previas presentes en la persona.

PRÁCTICO 3: Aprendizaje Supervisado(3/4)

Modelos de Clasificación:

- Los modelos que consideramos para realizar el entrenamiento del dataset fueron:
  - Descenso de Gradiente Estocástico.
  - Naive Bayes.
  - Regresión Logística.
  - Árboles de decisión.
  - K Vecinos más cercanos.

- En base a todos esos modelos, se realizó el entrenamiento y el ajuste de hiperparámetros, tomando siempre como objetivo la métrica “Accuracy”

PRÁCTICO 3: Aprendizaje Supervisado(4/4)

Resultados:

	Descenso de Gradiente Estocastico				Arbol de Desicion	
	Poblacion Total		Poblacion Mujeres		Poblacion Hombres	
	Train	Test	Train	Test	Train	Test
Accuracy	**0.82**	**0.91**	**0.82**	**0.91**	**0.90**	**0.90**
Precision	0.34	0.73	0.34	0.73	0.70	0.67
Recall	0.20	0.60	0.20	0.60	0.66	0.73
F1 Score	0.25	0.66	0.25	0.66	0.68	0.70
True Negatives	1438	370	1438	370	965	237
False Negatives	214	27	214	27	66	13
True Positives	53	40	53	40	129	36
False Positives	103	15	103	15	55	18

- Tanto para la población completa como para las mujeres, se observó una mejora significativa en la accuracy del modelo.
- Mientras que para hombres, no fue así, llevándonos a pensar que accuracy no haya sido la mejor métrica para este conjunto de datos.

PRÁCTICO 4: Aprendizaje Supervisado(1/4)

Objetivos:

- Definir la variable target, en este caso HTA (Presencia de Hipertensión Arterial).
- Determinar los factores que puedan ayudar a predecir la presencia de hipertensión en la población.
- Predecir de forma automática la presencia de hipertensión en todo el dataset completo, y además sólo en mujeres.

La variable target ya se encontraba clasificada en:

- Presencia de hipertensión. HTA = 1.
- Ausencia de hipertensión. HTA = 0.

PRÁCTICO 4: Aprendizaje Supervisado(2/4)

Selección de features:

- Del total de 239 features, seleccionamos 27 para este análisis.
- Las features seleccionadas mostraron alguna relación directa con la variable target.
- Para este análisis decimos seleccionar las mismas features para el análisis de la población entera como de mujeres.
- En general las features seleccionadas tienen que ver con:
  - Alimentación.
  - Estrés.
  - Niveles de actividad física.

PRÁCTICO 4: Aprendizaje Supervisado(3/4)

Modelos de Clasificación:

- Los modelos que consideramos para realizar el entrenamiento del dataset fueron:
    - Descenso de Gradiente Estocástico.
    - Multinomial Naive Bayes.
    - Perceptron.
    - Árboles de decisión.
    - K Vecinos más cercanos.
    - Dummy Classifier
- Además utilizamos los siguientes meta-modelos:
    - Random Forest.
    - Voting Classifier.
- En base a todos esos modelos, se realizó el entrenamiento y el ajuste de hiperparámetros, tomando siempre como objetivo la métrica “Accuracy”

PRÁCTICO 4: Aprendizaje Supervisado(4/4)

Resultados:

	Random Forest			
	Poblacion Mujeres		Poblacion Todos	
	Train	Test	Train	Test
Accuracy	0.77	0.77	0.71	0.72
Precision	0.82	0.81	0.76	0.77
Recall	0.87	0.87	0.83	0.84
F1 Score	0.84	0.84	0.79	0.80
True Negatives	288	289	439	447
False Negatives	41	40	88	80
True Positives	63	66	139	134
False Positives	76	73	119	124

- Cometimos un error en seleccionar accuracy como métrica
- F1 score sería una mejor elección, considerando la distribución desigual de la variable target.

GRACIAS POR SU ATENCIÓN!!!