BOOTCAMP DATA SCIENCE

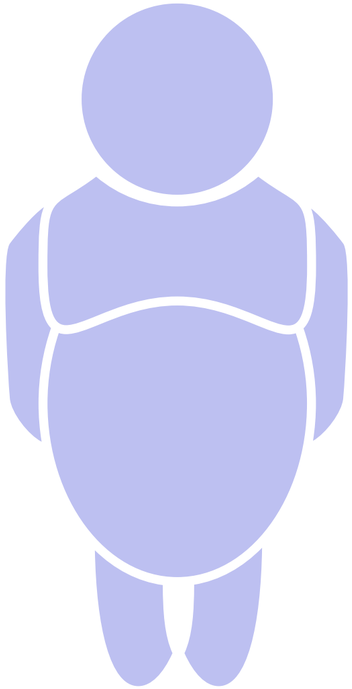
PROYECTO FINAL

USO Y ANÁLISIS DEL DATA SET PARA LA ESTIMACIÓN DE LOS NIVELES DE OBESIDAD

POR MARÍA JOSÉ LEÓN C

INTRODUCCIÓN

INTRODUCCIÓN



$$Mass\ body\ index = \frac{Weight}{height * height}$$

LA **OMS** INDICA QUE LA OBESIDAD Y EL SOBREPESO ES ACUMULACIÓN EXCESIVA DE GRASA EN DETERMINADAS ZONAS DEL CUERPO QUE PUEDE SER PERJUDICIAL PARA LA SALUD

MINISTERIO DE SALUD COLOMBIANO REPORTA QUE LA PREVALENCIA DE PERSONAS CON EXCESO DE PESO EN COLOMBIA, ES DEL 56,4 %, POR LO QUE SE HA CONVERTIDO EN UN PROBLEMA EN SALUD PÚBLICA EN EL PAÍS

CONTEXTO DE LA DATA

EXTRAÍDA DEL ARTÍCULO "***DATASET FOR ESTIMATION OF OBESITY LEVELS BASED ON EATING HABITS AND PHYSICAL CONDITION IN INDIVIDUALS FROM COLOMBIA, PERU AND MEXICO***" DE FABIO MENDOZA ET. AL, 2019.

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	Female	21.000000	1.620000	64.000000	yes	no	2.0	3.0	Sometimes	no	2.000000	no	0.000000	1.000000	no	Public_Transportation	Normal_Weight
1	Female	21.000000	1.520000	56.000000	yes	no	3.0	3.0	Sometimes	yes	3.000000	yes	3.000000	0.000000	Sometimes	Public_Transportation	Normal_Weight
2	Male	23.000000	1.800000	77.000000	yes	no	2.0	3.0	Sometimes	no	2.000000	no	2.000000	1.000000	Frequently	Public_Transportation	Normal_Weight
3	Male	27.000000	1.800000	87.000000	no	no	3.0	3.0	Sometimes	no	2.000000	no	2.000000	0.000000	Frequently	Walking	Overweight_Level_I
4	Male	22.000000	1.780000	89.800000	no	no	2.0	1.0	Sometimes	no	2.000000	no	0.000000	0.000000	Sometimes	Public_Transportation	Overweight_Level_II
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2106	Female	20.976842	1.710730	131.408528	yes	yes	3.0	3.0	Sometimes	no	1.728139	no	1.676269	0.906247	Sometimes	Public_Transportation	Obesity_Type_III
2107	Female	21.982942	1.748584	133.742943	yes	yes	3.0	3.0	Sometimes	no	2.005130	no	1.341390	0.599270	Sometimes	Public_Transportation	Obesity_Type_III
2108	Female	22.524036	1.752206	133.689352	yes	yes	3.0	3.0	Sometimes	no	2.054193	no	1.414209	0.646288	Sometimes	Public_Transportation	Obesity_Type_III
2109	Female	24.361936	1.739450	133.346641	yes	yes	3.0	3.0	Sometimes	no	2.852339	no	1.139107	0.586035	Sometimes	Public_Transportation	Obesity_Type_III
2110	Female	23.664709	1.738836	133.472641	yes	yes	3.0	3.0	Sometimes	no	2.863513	no	1.026452	0.714137	Sometimes	Public_Transportation	Obesity_Type_III

PREGUNTA A RESOLVER

¿ES POSIBLE ESTIMAR LOS NIVELES DE OBESIDAD EN LAS PERSONAS SEGÚN SU GÉNETICA Y HÁBITOS?

ANÁLISIS EXPLORATORIO

INTEGRIDAD DE LA DATA

EXISTENCIA DE VALORES NULOS

NO HABÍA COLUMNAS NULAS

EXISTENCIA DE VALORES COHERENTES

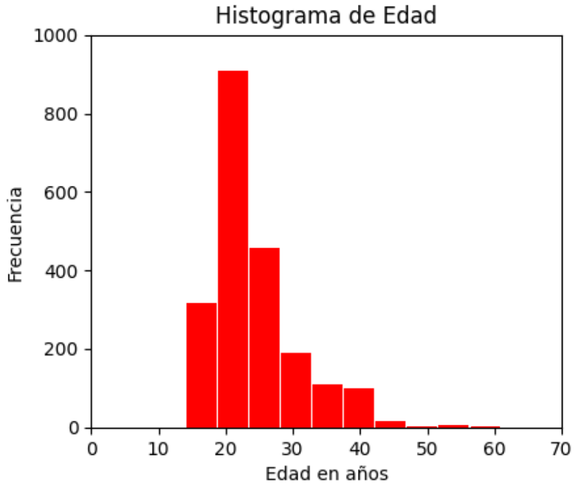
TODOS LOS DATOS DE LAS COLUMNAS TENÍAN VALORES QUE RESPONDÍAN A LA INFORMACIÓN QUE SUMINISTRABAN

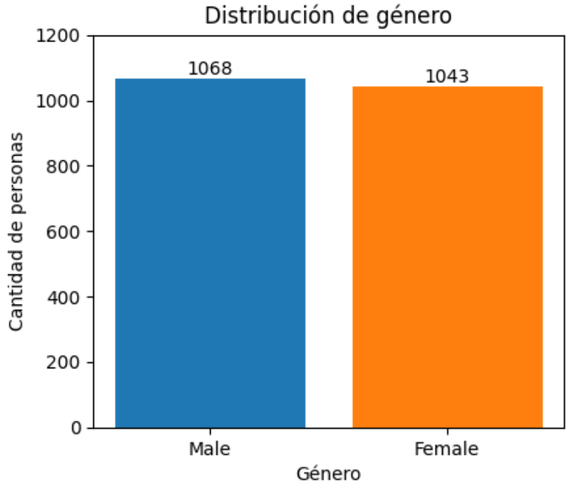
TIPOS DE DATOS

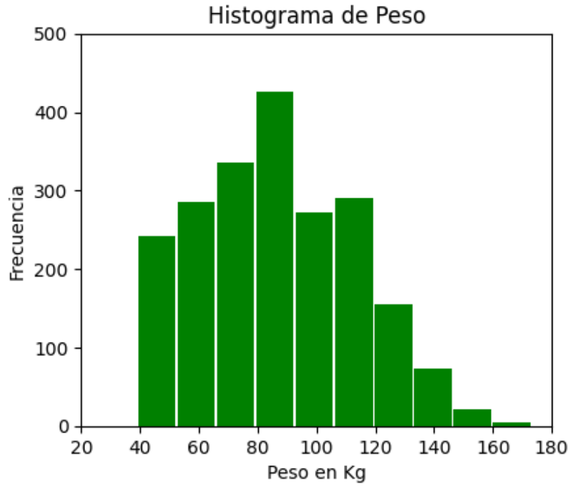
OBJECT / FLOAT

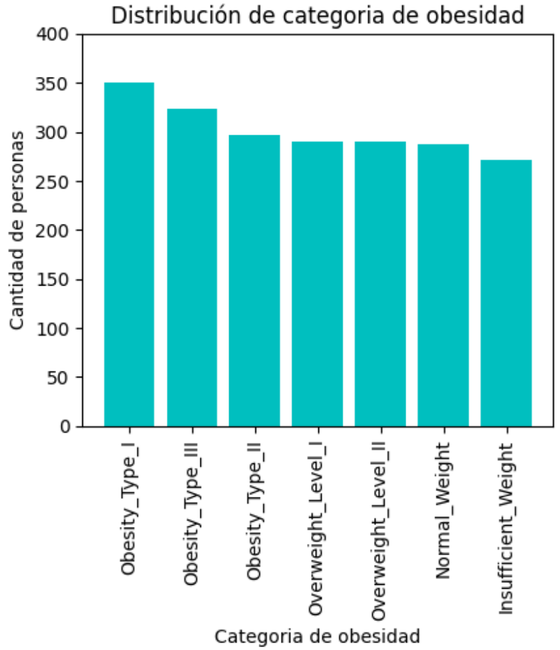
LOS DATOS SUFRIERON UN PRE PROCESAMIENTO ANTES DE SU PUBLICACIÓN POR LO CUAL LA DATA ESTÁ LIMPIA, COHERENTE Y BALANCEADA.

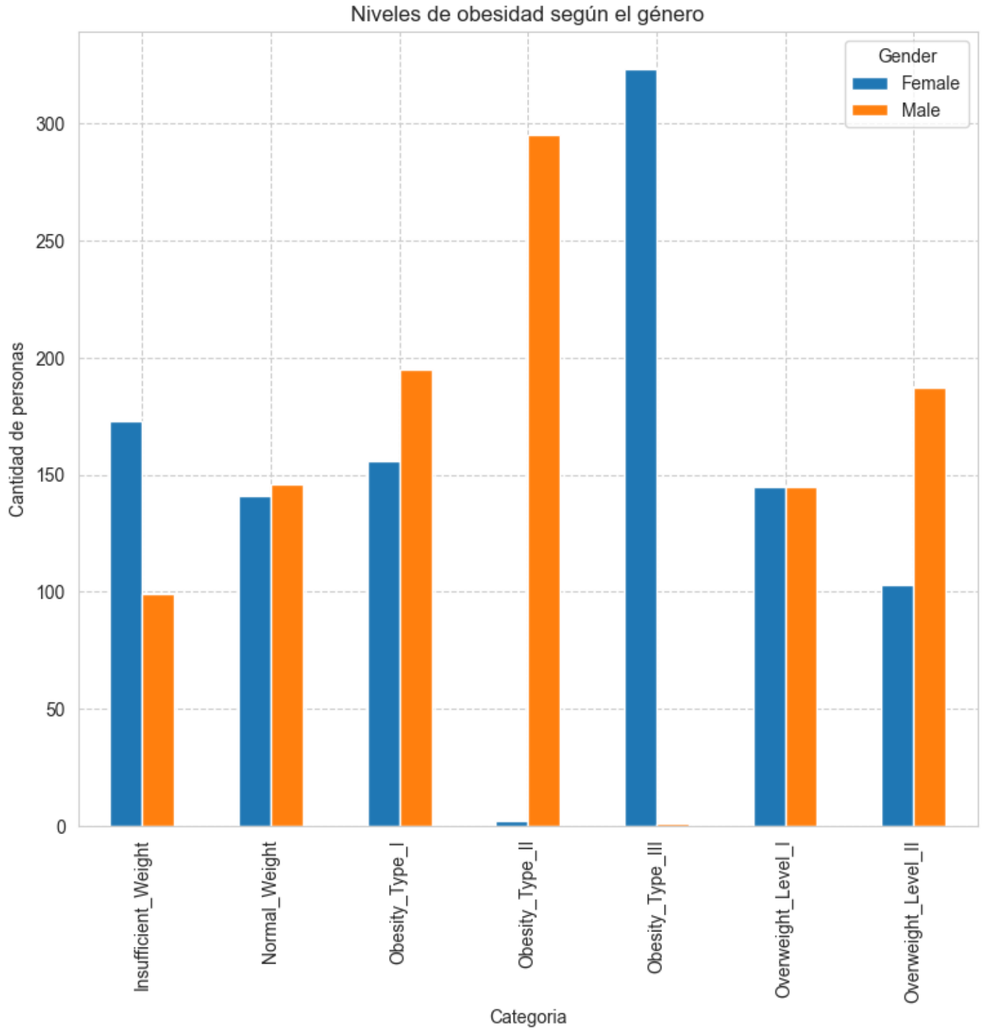
VISUALIZACIONES EXPLORATORIAS



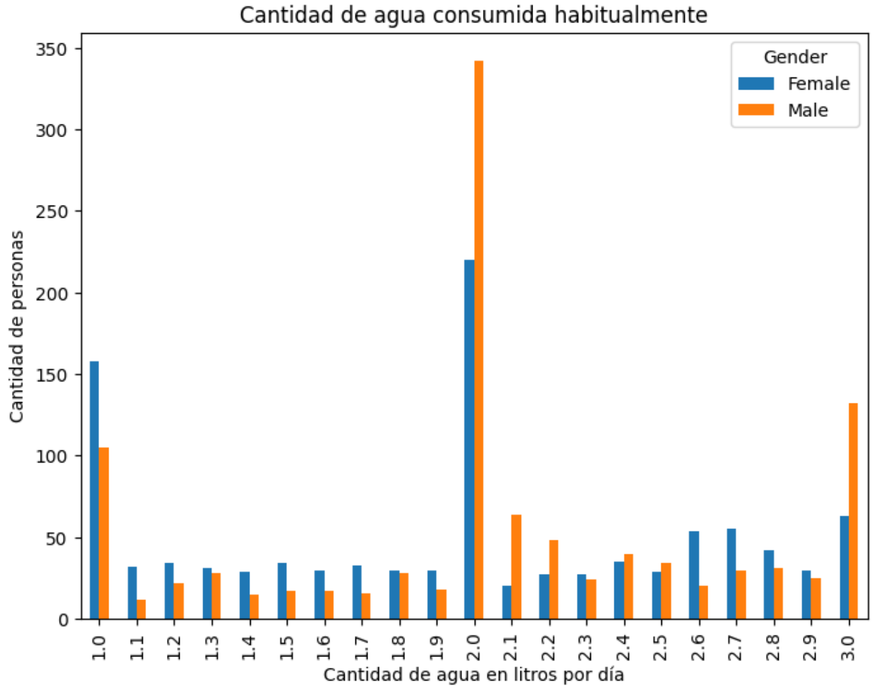


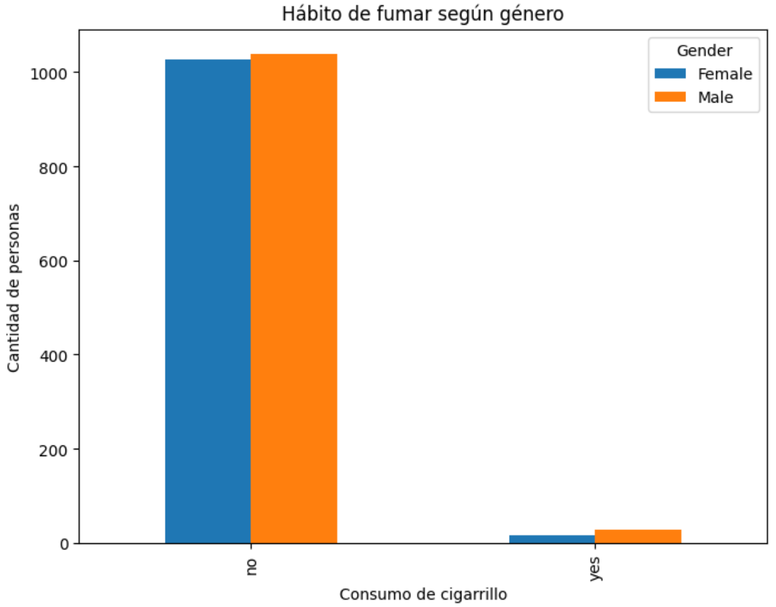


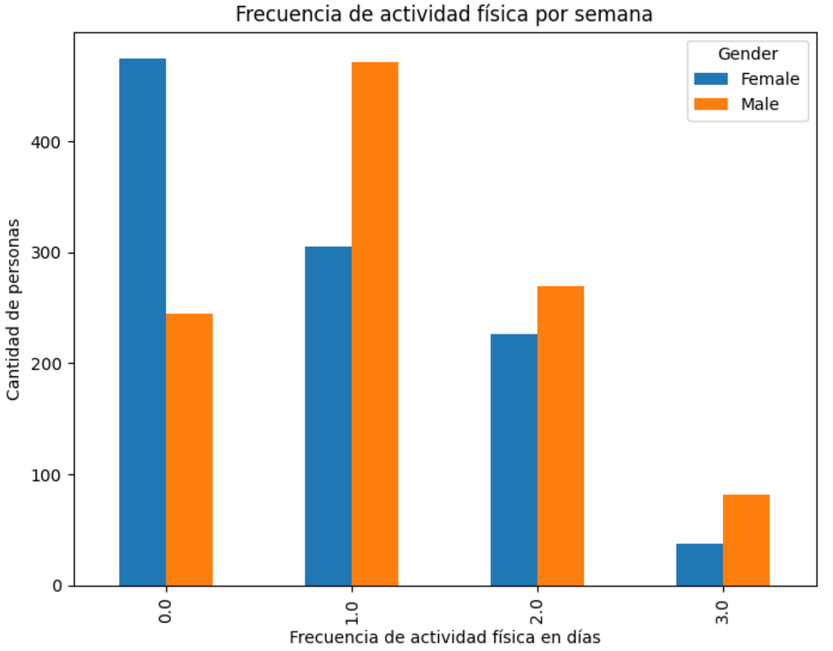




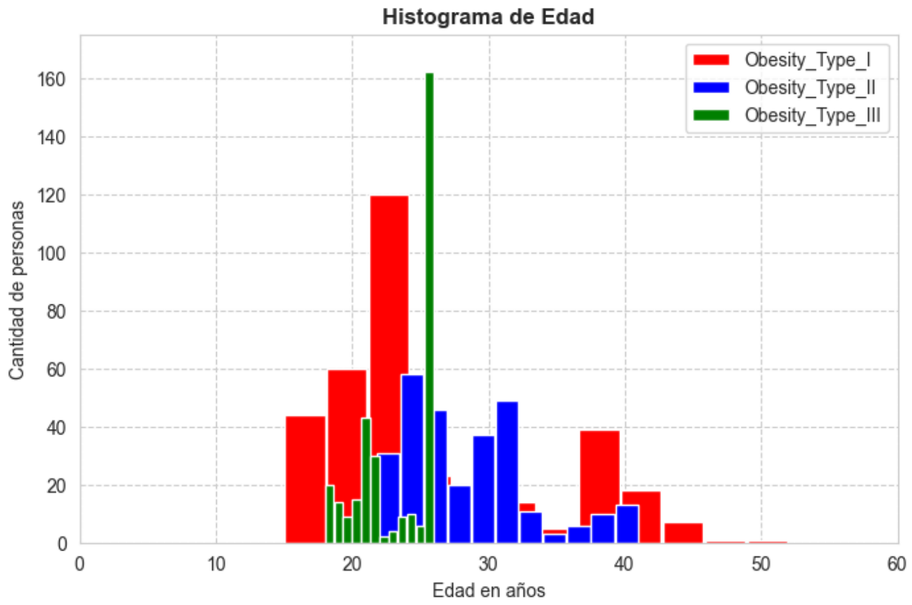
VISUALIZACIONES EXPLORATORIAS

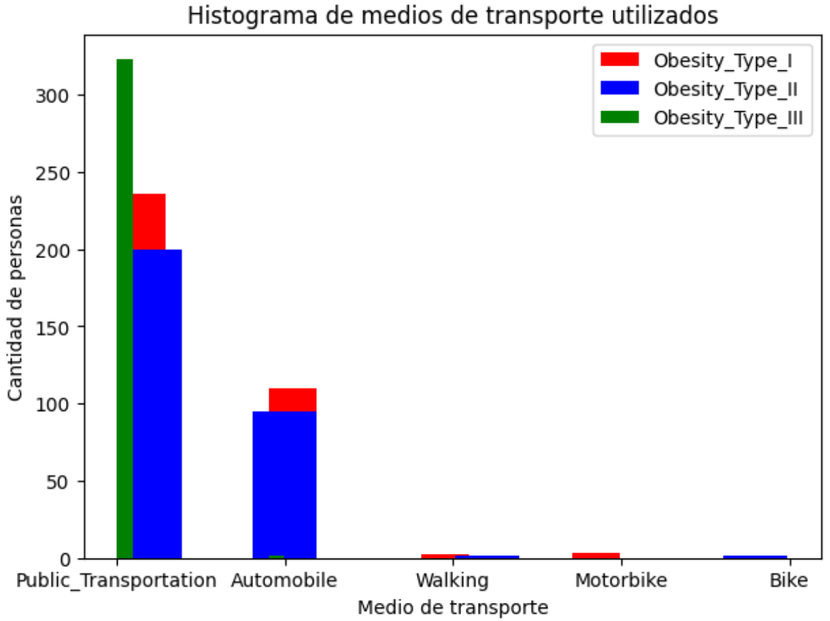


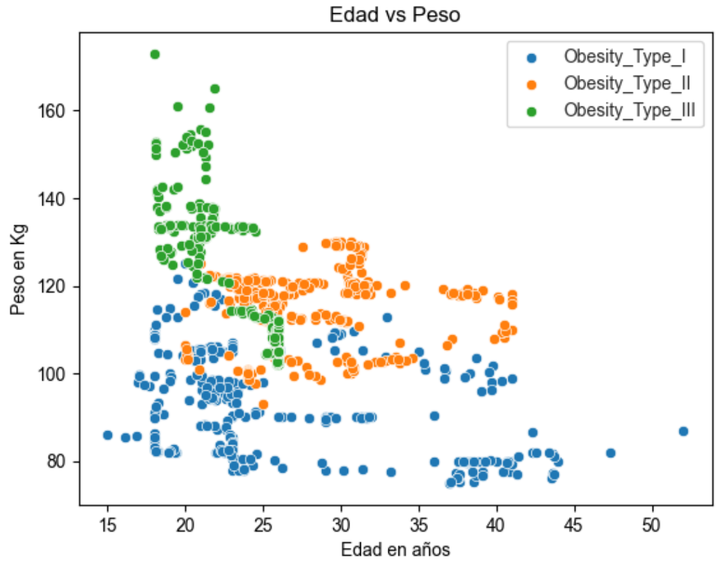




VISUALIZACIONES EXPLORATORIAS







INGENIERÍA DE ATRIBUTOS

USO DE LA FUNCIÓN GET_DUMMIES PARA VARIABLES CATEGORICAS

CAEC_no	SMOKE_yes	SCC_yes	CALC_Frequently
False	False	False	False
False	True	True	False
False	False	False	True
False	False	False	True
False	False	False	False

Insufficient_Weight	Normal_Weight	Obesity_Type_I
False	True	False
False	True	False
False	True	False
False	False	False
False	False	False

ALGORITMO DE PREDICCIÓN

VARIABLES CATEGÓRICAS =ALGORITMOS DE CLASIFICACIÓN

RANDOM FOREST

REGRESIÓN LOGÍSTICA

X =VARIABLES NÚMERICAS Y BINARIAS
Y= NIVEL DE OBESIDAD

TEST SIZE: 30%

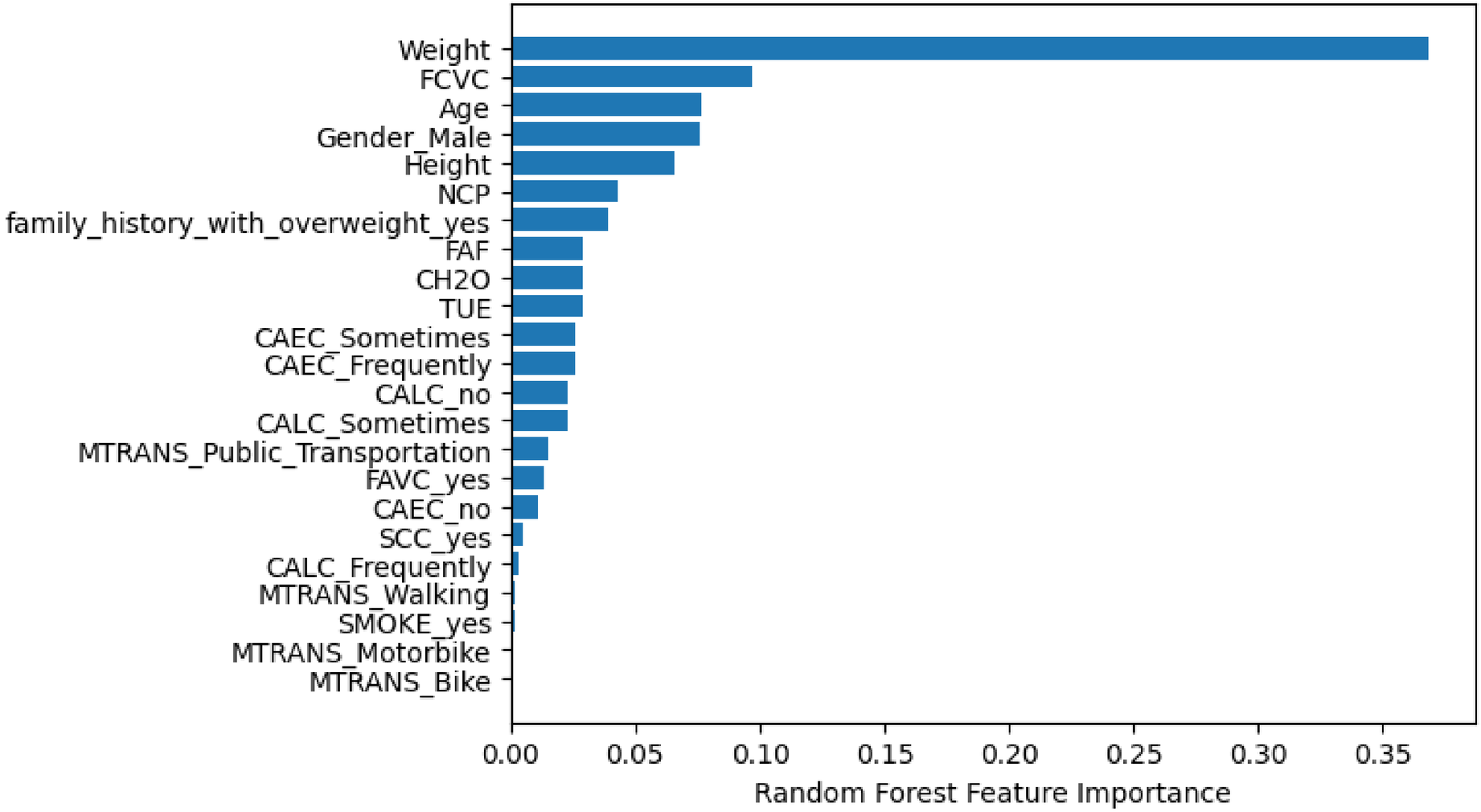
RANDOM FOREST

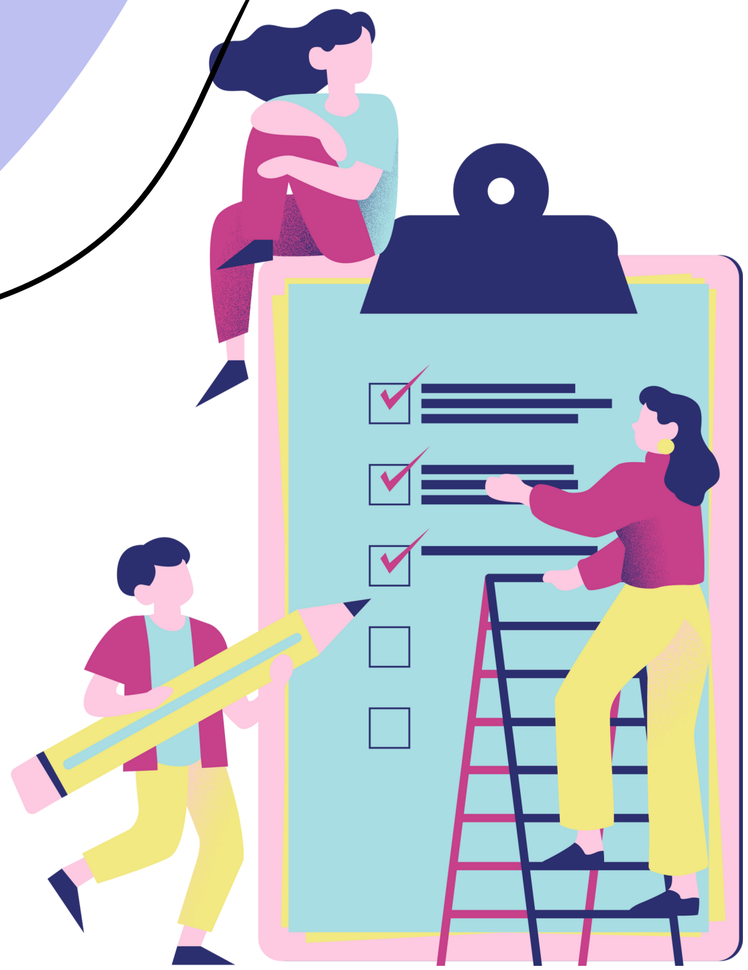
	precision	recall	f1-score	support
Insufficient_Weight	0.90	0.92	0.91	86
Normal_Weight	0.75	0.76	0.76	93
Obesity_Type_I	0.88	0.86	0.87	102
Obesity_Type_II	0.93	0.99	0.96	88
Obesity_Type_III	1.00	0.99	0.99	98
Overweight_Level_I	0.79	0.73	0.76	88
Overweight_Level_II	0.76	0.76	0.76	79
accuracy			0.86	634
macro avg	0.86	0.86	0.86	634
weighted avg	0.86	0.86	0.86	634

REGRESIÓN LOGÍSTICA

	precision	recall	f1-score	support
Insufficient_Weight	0.81	0.95	0.88	86
Normal_Weight	0.77	0.59	0.67	93
Obesity_Type_I	0.89	0.88	0.89	102
Obesity_Type_II	0.90	1.00	0.95	88
Obesity_Type_III	1.00	0.99	0.99	98
Overweight_Level_I	0.71	0.76	0.74	88
Overweight_Level_II	0.74	0.67	0.70	79
accuracy			0.84	634
macro avg	0.83	0.84	0.83	634
weighted avg	0.84	0.84	0.83	634

IMPORTANCIA DE LAS CARACTERÍSTICAS DEL RANDOM FOREST





CONCLUSIONES Y TRABAJO FUTURO