

Efectos del tabaco y el alcohol en la salud

Proyecto de Fin de BootCamp

Esther Bellido González Julián Campillo Navarro Cristina Pastor López Andreu Sau Ramírez Paula Torres Cerdán

Índice



- 1. Objetivos
- 1. Consideraciones y aclaraciones
- 1. Limpieza de datos
- 1. EDA
- 1. Análisis predictivo
 - 5.1 Modelo de regresión logística
 - 5.2 Modelo de árbol
 - **5.3 Modelo XGBoost**
- 1 Conclusiones

Objetivos

N

Objetivos

- Efectos que tiene tanto fumar como beber en nuestra salud
- Realizar un análisis exploratorio
- Realizar un análisis predictivo
- Sacar conclusiones de nuestra base de datos y modelos realizados

Consideraciones y aclaraciones



Consideraciones y aclaraciones

- 1. Datos obtenidos de kaggle (Corea del Sur): https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset/data
- 1. Programas utilizados:
 - 1.R
 - 2.Python
 - 3.PowerBI
 - 4.Excel
- 4. Dimensiones DataSet original: (991346, 24) → casi 1 Millón de personas

Limpieza de datos



Limpieza de datos

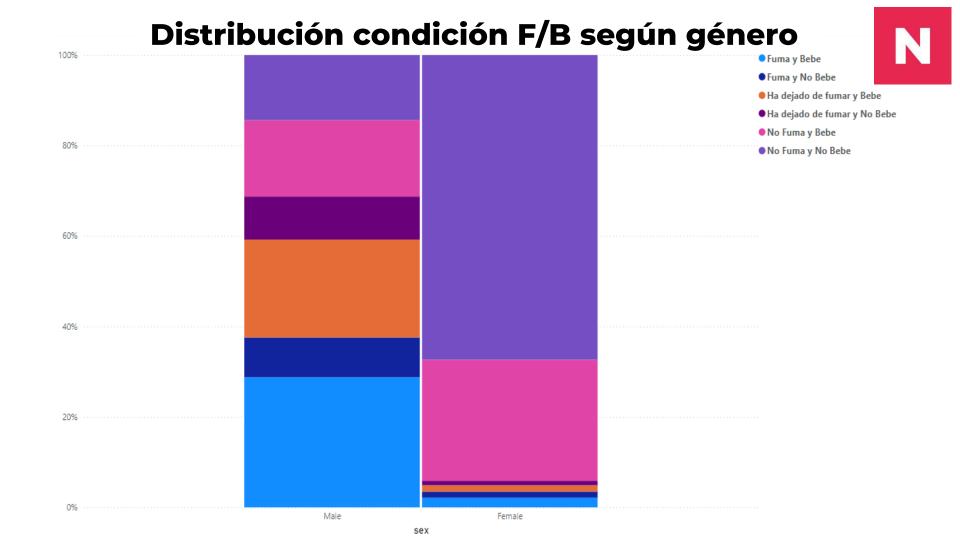
 Comprobación de valores nulos, vacíos o sin sentido dentro de la base de datos original

Creación de nuevas columnas como IMC o Smoking_Drinking_Status

EDA

- Modelos predictivos
 - Regresión logística: binarios y dummies
 - Árbol: binarios
 - Modelo XGBoost: LabelEncoder()

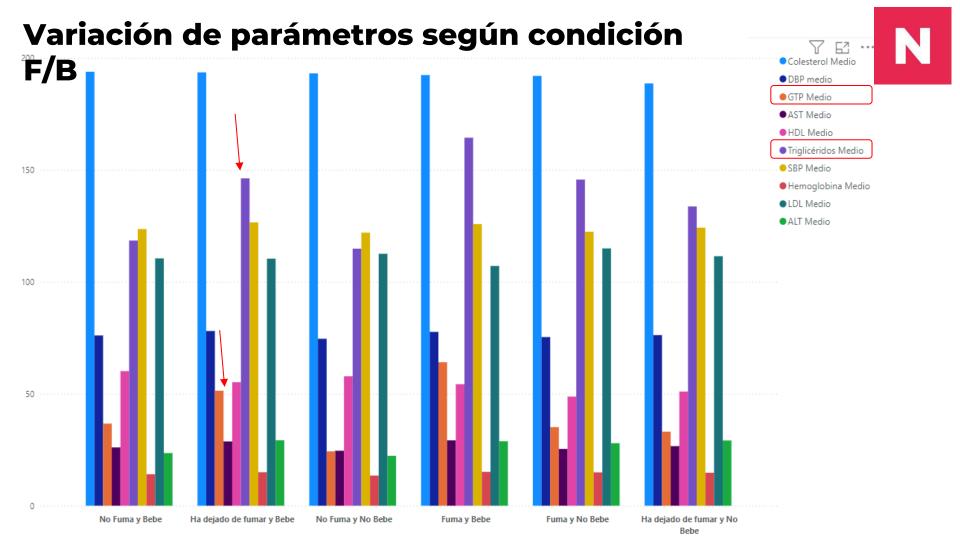
EDA



Datos Estadísticos más relevantes

_		Ŧ	
	ı		
	`		

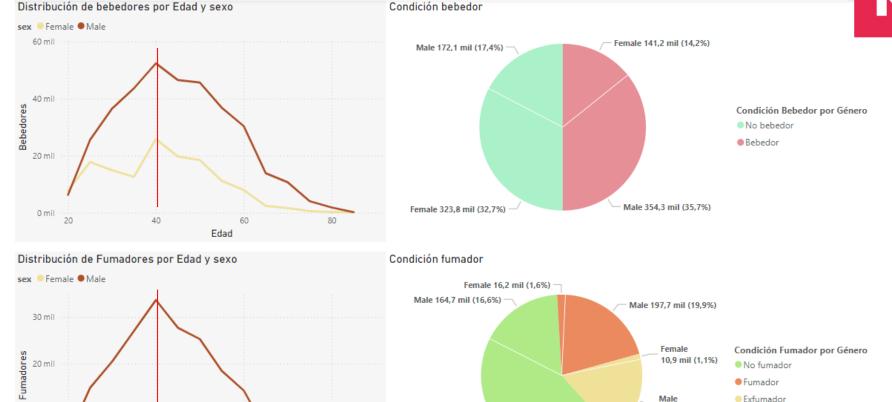
Variable	mean	std	min	max
sex	0.53101	0.499038	0.0	1.0
age	47.614941	14.181339	1.0	85.0
height	162.240625	9.282957	130.0	190.0
weight	61.883004	12.514241	30.0	99.9
waistline	81.233358	11.850323	74.0	99.0
sight_left	0.980834	0.060594	0.1	2.0
sight_right	0.978429	0.060774	0.1	2.0
hear_left	1.031495	0.112469	0.1	2.0
hear_right	1.030476	0.071842	0.1	2.0
SBP	122.432498	14.003444	67.0	185.0
DBP	76.052627	9.889365	32.0	135.0
BLDS	100.424747	24.17996	18.0	500.0
tot_chole	195.557402	36.660155	72.0	445.0
HDL_chole	56.9368	13.201235	18.0	150.0
LDL_chole	113.037692	35.842812	40.0	376.0
triglyceride	132.141577	102.196985	16.0	9490.0
hemoglobin	14.292824	1.554999	8.0	19.6
urine_protein	1.094024	0.437524	1.0	5.0
serum_creatinine	0.806047	0.161259	0.4	1.9
SGOT_AST	25.988803	23.493388	1.0	512.0
SGOT_ALT	25.751504	26.309596	1.0	453.0
gamma_GTP	37.136347	50.424153	4.0	968.0
SMK_stat_type_cd	0.608122	0.488128	0.0	2.0
DRK_YN	0.499811	0.500029	0.0	1.0
IMC	23.977179	3.150128	18.468993	45.124368
Smoking_Drinking_Status	3.283943	1.871912	1.0	5.0



Distribución de consumo de alcohol y tabaco según género



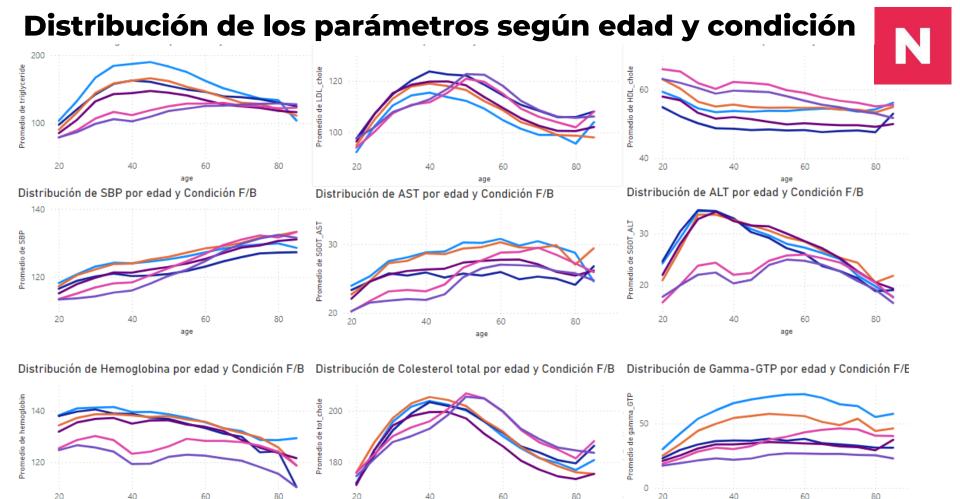
164,0 mil (16,5%)



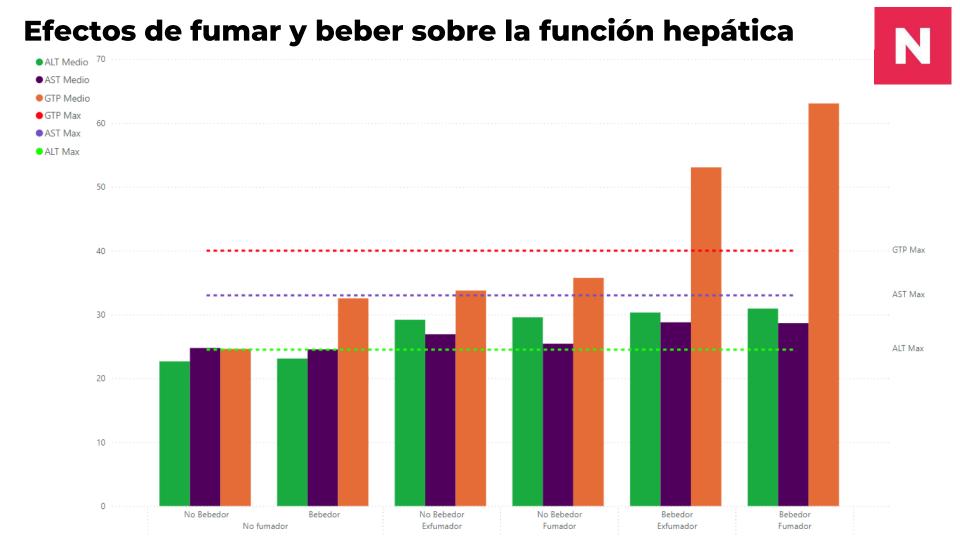
Female 437,8 mil (44,2%)

60 Fdad

10 mil

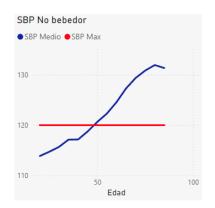


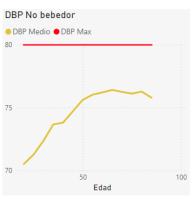
Smoking_Drinking_Status Fuma y Bebe Fuma y No Bebe Ha dejado de fumar y Bebe Ha dejado de fumar y No Bebe

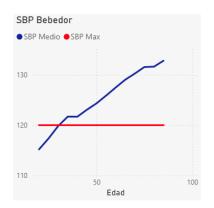


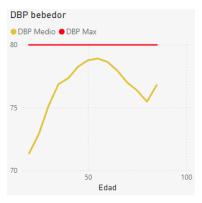












SBP: Presión sistólica, máxima presión arterial durante latidos cardíacos.

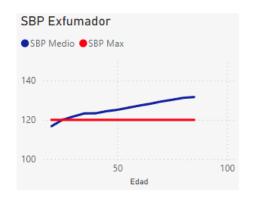
DBP: Presión diastólica, mínima presión arterial entre latidos cardíacos.

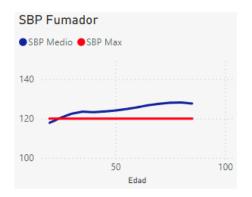
La línea roja establece el límite aconsejable de los valores SBP y DBP.

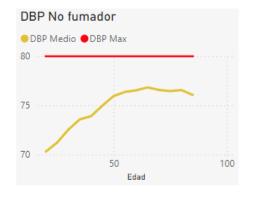
N

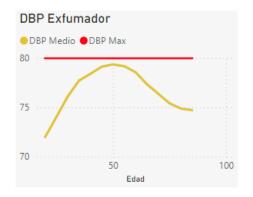
Los efectos de fumar sobre la presión arterial

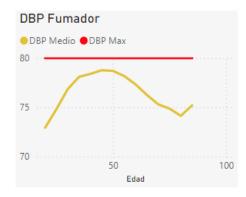












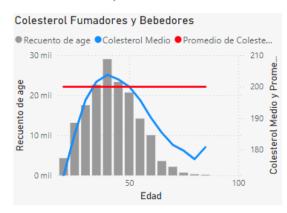
Evolución del colesterol según condición F/B

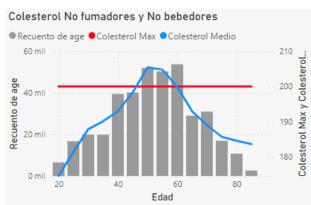


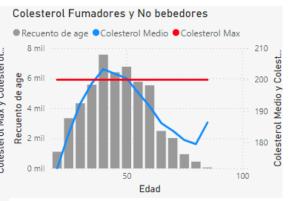
16,30 %

39,24 %

5,28 %





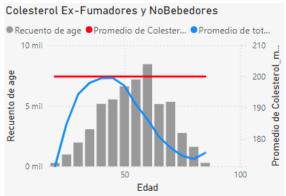


5,49 %

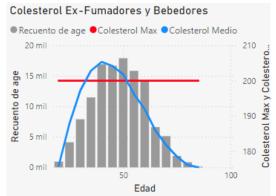
21.53 %



12,15 %





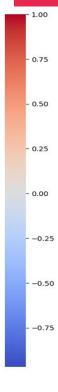


Análisis Predictivos

Correlaciones

											C	orrela	cion c	on He	eatma	р										
sex -	1	-0.083			0.34	0.099	0.1	0.0029	0.0007	5 0.18	0.2	0.12	-0.022	-0.26	0.0052	0.22		0.02	0.25	0.093	0.2	0.26		0.37	0.22	-0.62
age -	-0.083	1	-0.4	-0.2	0.13	-0.17	-0.17	0.2	0.2	0.27	0.11	0.2	0.011	-0.1	0.029	0.044	-0.17	0.041	0.023	0.059	-0.021	0.017	-0.13	-0.28	0.057	0.19
height -	0.72	-0.4	1	0.67	0.26	0.14	0.14	-0.081	-0.082	0.035	0.11	0.021	-0.023	-0.15	-0.015	0.14	0.53	0.0023	0.17	0.039	0.14	0.16	0.47	0.37	0.11	-0.51
weight -	0.58	-0.2	0.67	1	0.64	0.089	0.089	-0.044	-0.046	0.25	0.28	0.14	0.063	-0.29	0.068	0.28	0.5	0.029	0.15	0.1	0.28	0.22	0.37	0.26	0.81	-0.39
waistline -	0.34	0.13	0.26	0.64	1	0.0045	0.0062	0.033	0.031	0.27	0.24	0.18	0.063	-0.25	0.063	0.25	0.29	0.043	0.098	0.097	0.21	0.19	0.21	0.093	0.65	-0.2
sight_left -	0.099	-0.17	0.14	0.089	0.0045	1	0.31	-0.039	-0.041	-0.036	0.0012	-0.035	0.0044	0.0042	2 0.003	0.011	0.086	-0.01	0.02	-0.0054	0.019	0.014	0.064	0.077	0.0093	-0.077
sight_right -	0.1	-0.17	0.14	0.089	0.0062	0.31	1	-0.041	-0.039	-0.0340	0.0005	70.037	0.0034	0.0063	0.0022	0.012	0.087	-0.011	0.021	-0.0046	0.019	0.016	0.065	0.075	0.0095	-0.077
hear_left -	0.0029	0.2	-0.081	-0.044	0.033	-0.039	-0.041	1	0.54	0.056	0.0069	0.04	-0.024	-0.029	-0.015	0.0021	-0.036	0.02	0.015	0.011	0.0079	90.003	-0.021	-0.058	0.004	0.034
hear_right -C	0.00075	0.2	-0.082	-0.046	0.031	-0.041	-0.039	0.54	1	0.057	0.0075	0.041	-0.023	-0.028	-0.013	0.0022	-0.037	0.018	0.013	0.011	0.0087	0.0035	-0.022	-0.058	0.0034	0.035
SBP -	0.18	0.27	0.035	0.25	0.27	-0.036	-0.034	0.056	0.057	1	0.74	0.18	0.069	-0.11	0.034	0.19	0.17	0.064	0.063	0.081	0.12	0.16	0.085	0.033	0.31	-0.083
DBP -	0.2	0.11	0.11	0.28	0.24 -	0.0012	0.0005	D.0069	0.0075	0.74	1	0.14	0.11	-0.094	0.067	0.2	0.24	0.046	0.057	0.079	0.13	0.18	0.13	0.1	0.29	-0.14
BLDS -	0.12	0.2	0.021	0.14	0.18	-0.035	-0.037	0.04	0.041	0.18	0.14	1	0.013	-0.11	-0.03	0.2	0.1	0.11	0.044	0.068	0.11	0.17	0.086	0.015	0.17	-0.08
tot_chole -	-0.022	0.011	-0.023	0.063	0.063	0.0044	0.0034	-0.024	-0.023	0.069	0.11	0.013	1	0.16						0.032						
HDL_chole -	-0.26	-0.1	-0.15	-0.29	-0.25 -	0.0042	0.0063	-0.029	-0.028	-0.11	-0.094	-0.11			0.022	-0.27	-0.18	-0.022	-0.084	-0.034	-0.12	-0.056	-0.16	0.043	-0.27	0.13
LDL_chole -													100000000000000000000000000000000000000	0.022		0.03				0.002		-0.0085	-0.012	-0.044		0.022
triglyceride -									0.0022		0.2	0.2	0.27		0.03	1		0.043			0.19	0.3	0.22	0.1		-0.22
hemoglobin -	-	and the same of							-0.037			0.1		-0.18		0.24		0.0038	F		0.23	0.23	0.45	0.3		-0.48
urine_protein -																				0.035						
serum_creatinine -																				0.026	The same of					
SGOT_AST -									0.011												0.64		0.062			-0.065
SGOT_ALT -									90.0087									0.031		The second	1	0.37				-0.13
gamma_GTP -				0.22					0.0035									0.049			0.37	1	0.24			-0.27
SMK_stat_type_cd -									-0.022											0.062		0.24	1	0.35	0.13	-0.97
DRK_YN -									-0.058											0.039		0.21	0.35	1	0.06	-0.57
IMC -		0.057	ANIONATOR	0.81					0.0034				0.098			0.27		0.037		0.1	0.25	0.17	0.13	0.06	1	-0.13
Smoking_Drinking_Status -	-	0.19	-	-0.39	- 1	1	- 1	-	0.035	1	- 1	- 1	- 1	- 1	0.022	- 1	-0.48	- 1	- 1	-0.065	- 1	-0.27	-0.97	- 1	-0.13	1
	sex	age	height	weight	waistline	sight_left	sight_right	hear_left	hear_right	SBP	DBP	BLDS	tot_chole	HDL_chole	LDL_chole	triglyceride	hemoglobin	urine_protein	serum_creatinine	SGOT_AST	SGOT_ALT	gamma_GTP	SMK_stat_type_cd	DRK_YN	IMC	Smoking_Drinking_Status

N





Con **beber** como variable a predecir

VIF

	PYTHON	R
Variables X eliminadas	weight, tot_chole	weight, tot_chole
Accuracy	72,54%	49,86%



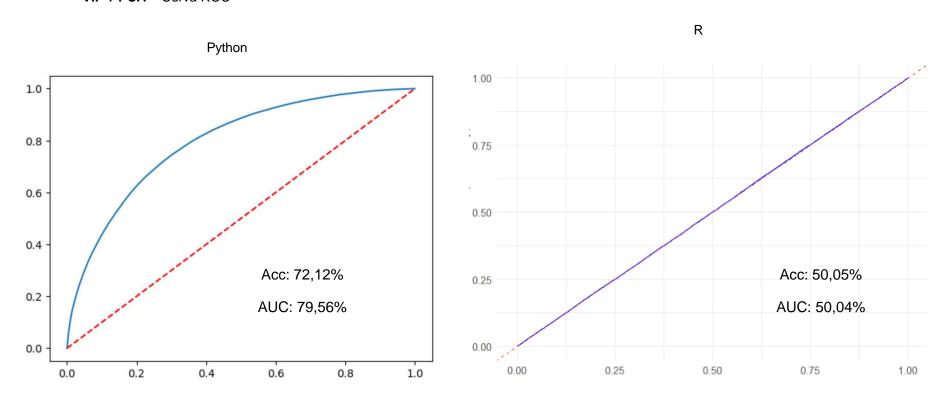
Con **beber** como variable a predecir

VIF + PCA

	PYTHON	R
N⁰ variables X	16 (91,88%)	17 (91,16%)
Accuracy	72,12%	50,05%



VIF + PCA ~ Curva ROC





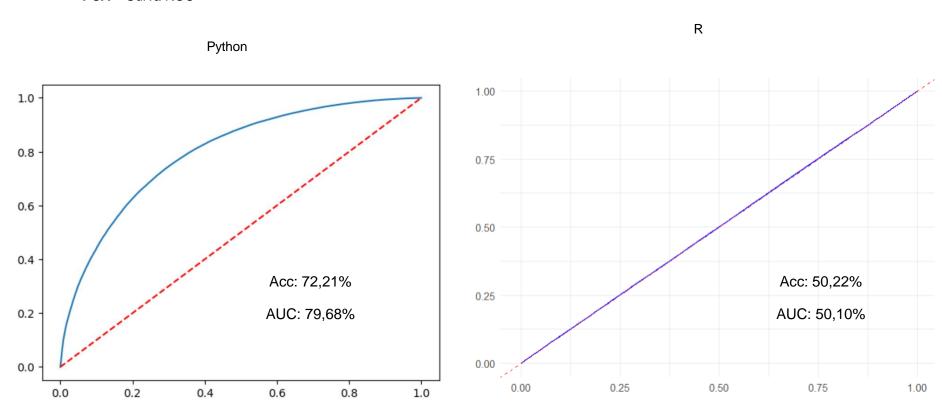
Con **beber** como variable a predecir

PCA

	PYTHON	R
Nº variables X	15 (90,06%)	16 (90,06%)
Accuracy	72,21%	50,22%



PCA ~ Curva ROC





Con **beber** como variable a predecir

Modelos de regresión logística	Python	R
VIFs	72,54%	49,86%
VIFs + PCA	72,12%	50,05%
PCA	72,21%	50,22%



Con **beber** como variable a predecir + 15 variables X

RFE

R

PYTHON

sex

age

height

weight

waistline

sight_left

sight_right

BLDS

tot_chole

HDL_chole

LDL_chole

hemoglobin

IMC

serum creatinine

SGOT_AST

hear_left

hear_right

urine_protein

SGOT_ALT

gamma_GTP

SMK_stat_type_cd_2.0

SMK_stat_type_cd_3.0



Modelo del árbol

Con la columna de **fumadores** como la variable a predecir

VIF

	PYTHON	R
Variables X eliminadas	weight (VIF=133), LDL_chole (VIF=7.27)	weight (VIF=133), LDL_chole (VIF=7.27)
Accuracy	69,61%	68,29%



Modelo del árbol

Con la columna de **fumadores** como la variable a predecir

VIF + PCA

	PYTHON	R
N⁰ variables X	15 (91,17%)	17 (91,16%)
Accuracy	68,22%	65,65%



Modelo del árbol

Con la columna de **fumadores** como la variable a predecir

PCA

	PYTHON	R
N⁰ variables X	15 (91,24%)	17 (91,63%)
Accuracy	68,04%	65,42%



Análisis predictivo

Con la columna de **fumadores** como la variable a predecir



Modelos con el método del árbol	Python	R
Modelo previo al PCA	69,61%	68,29%
Modelo PCA tras eliminar columnas VIF>5	68,22%	65,65%
Modelo PCA con todas las columnas	68,04%	65,42%

Análisis Predictivo

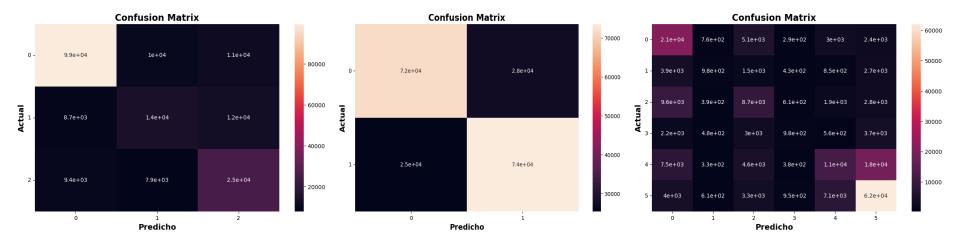


Con biblioteca **XGBoost** de Python:

Utilizamos como variables a predecir en un primer caso el estado de fumador, en un segundo caso el estado de bebedor, y en un tercer caso el estado de fumador y bebedor.

Utilizamos como variables explicativas todas las variables restantes del dataframe.

Modelo con XGBoost	Exactitud
Estado de fumador	70.11%
Estado de bebedor	73.26%
Estado fumador y bebedor	52.82%



Conclusiones

Conclusiones



 Edad, Altura y Peso consumo de alcohol es más frecuente entre los jóvenes y adultos de mediana edad

- Relación con Atributos Físicos — Altura : las personas más altas y pesadas consumir alcohol

Cintura: Personas con cinturas más grandes

- Perfiles Lipídicos
 - BLDS, tot_chole y HDL_chole bebedores = no bebedores.
 - Colesterol LDL: no bebedores > bebedores
 - Triglicéridos : bebedores > no bebedores.

Conclusiones



- Hemoglobina : niveles más altos en hombres
- Función Hepática : SGOT_ALT y gamma_GTP bebedores, indicando posibles efectos hepatotóxicos del alcohol
 - Capacidades Auditivas : Participantes con problemas auditivos consumen más alcohol
- Historial de Fumar : Personas que nunca han fumado tienen casi el doble de probabilidad de consumir alcohol
 - Género : Los hombres tienen más del doble de probabilidades de consumir alcohol que las mujeres
- Equilibrio de Género en el Estudio : distribución de participantes es casi igual entre hombres y mujeres