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

Objetivos

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

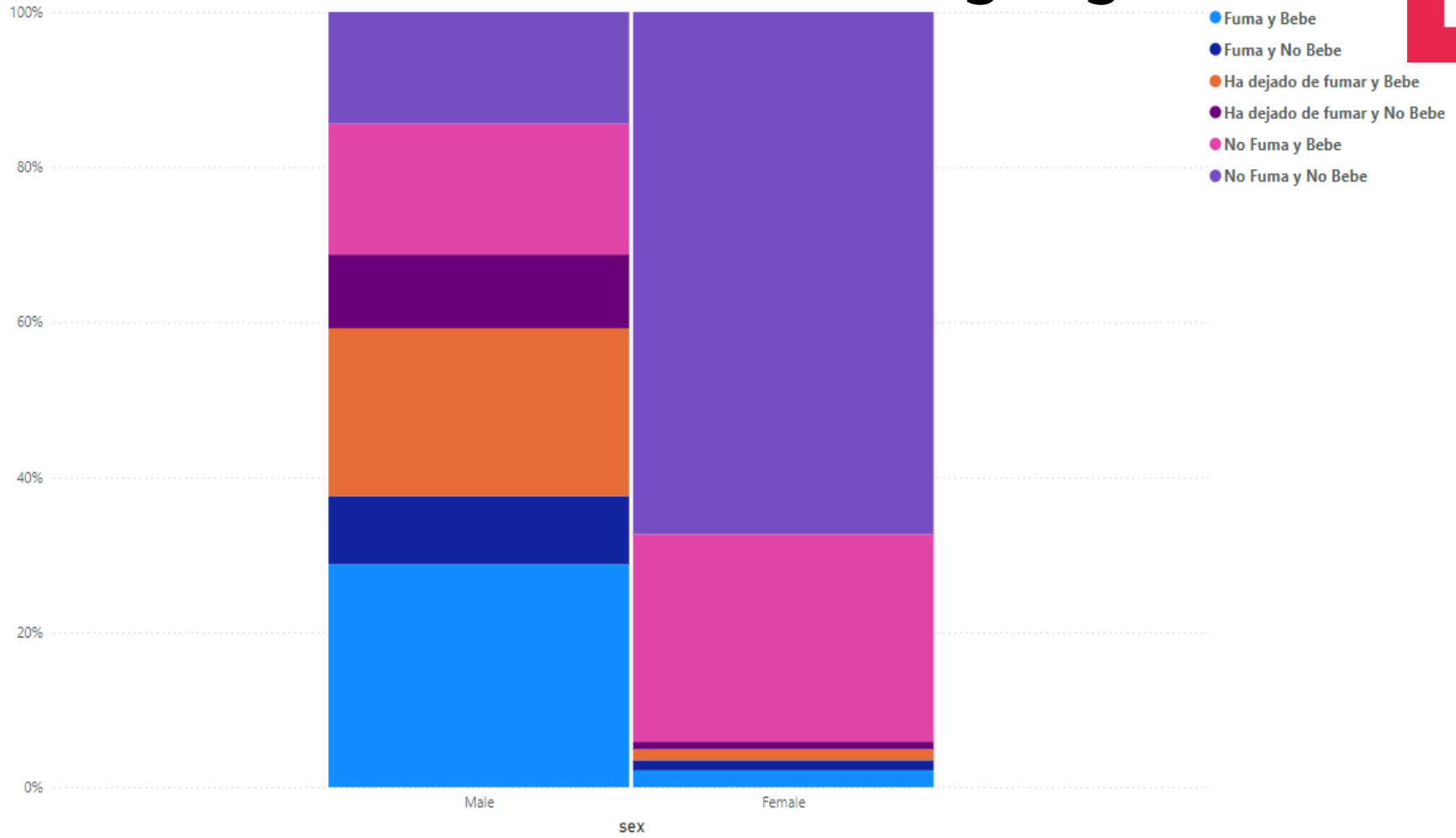
Limpieza de datos

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

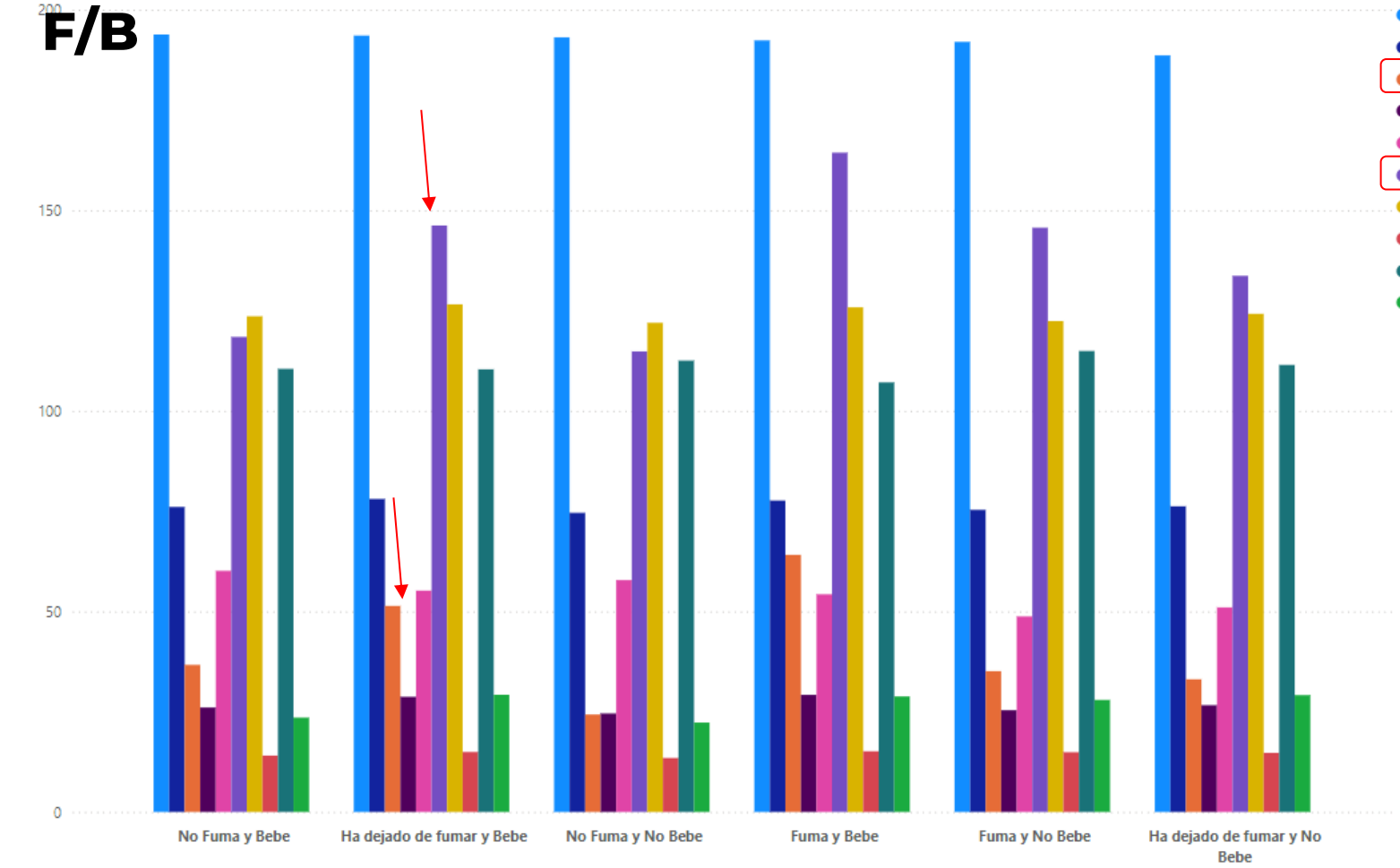
Distribución condición F/B según género



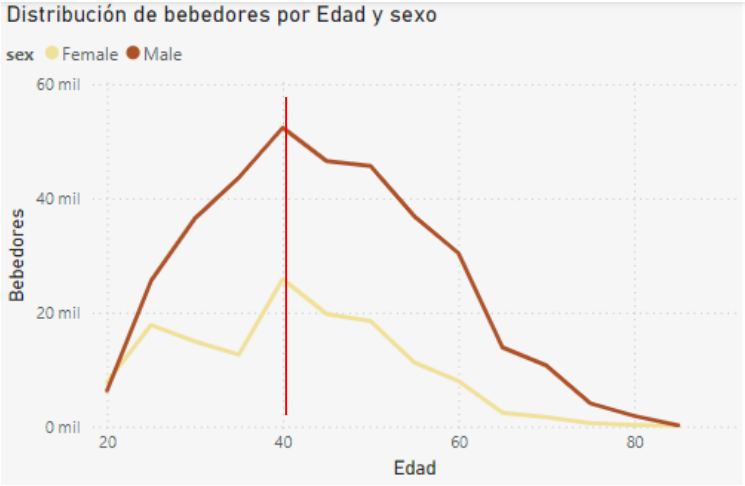
Datos Estadísticos más relevantes

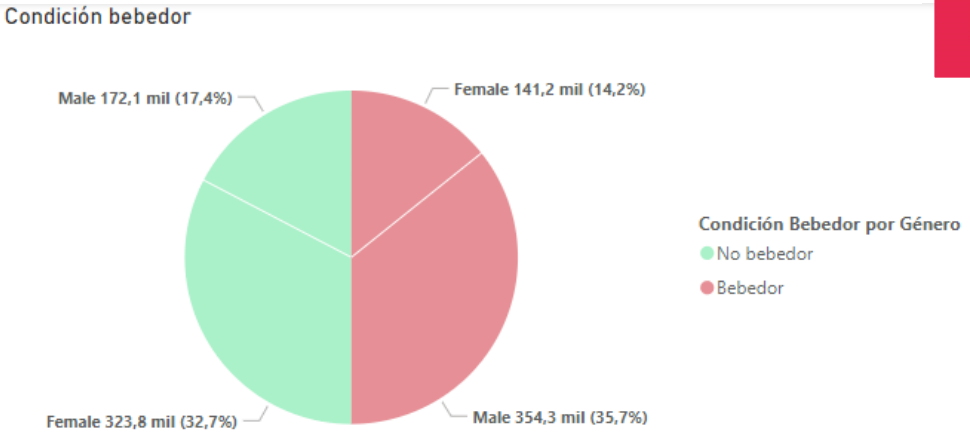
Variable	mean	std	min	max
sex	0.53101	0.499038	0.0	1.0
age	47.614941	14.181339	1.0	85.0
height	162.240625	9.282957	130.0	190.0
weight	61.883004	12.514241	30.0	99.9
waistline	81.233358	11.850323	74.0	99.0
sight_left	0.980834	0.060594	0.1	2.0
sight_right	0.978429	0.060774	0.1	2.0
hear_left	1.031495	0.112469	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

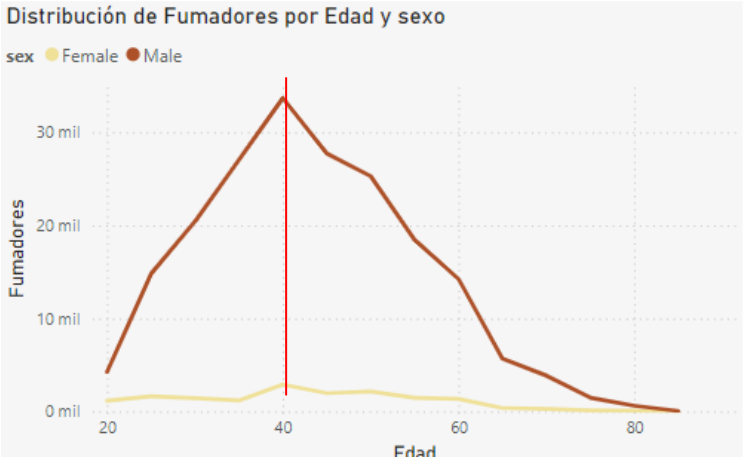
Variación de parámetros según condición F/B

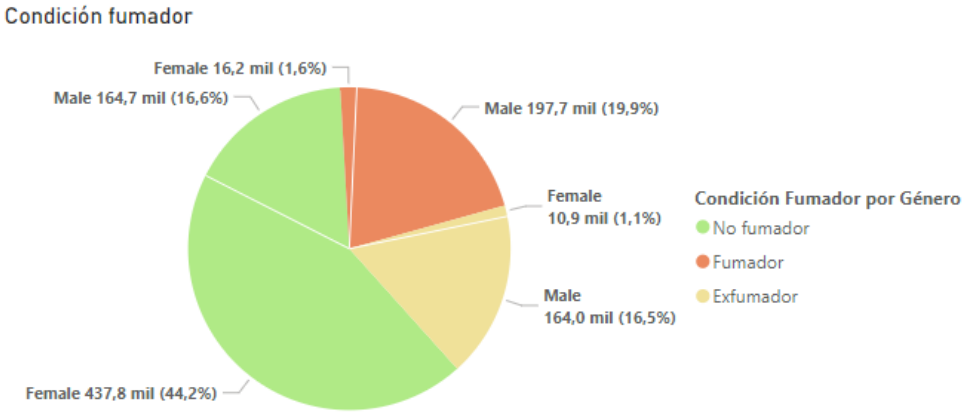


Distribución de consumo de alcohol y tabaco según género

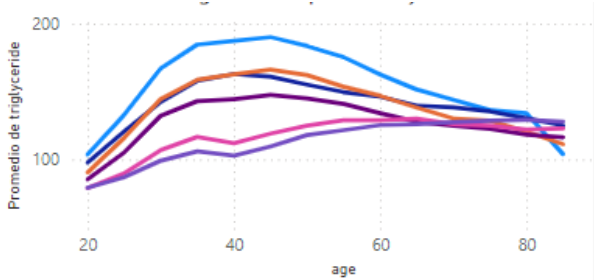


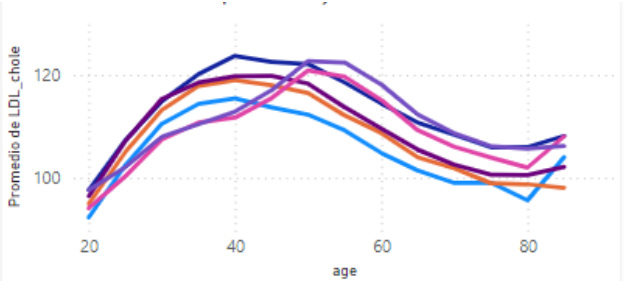


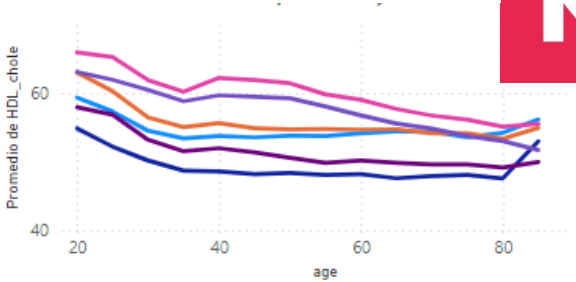




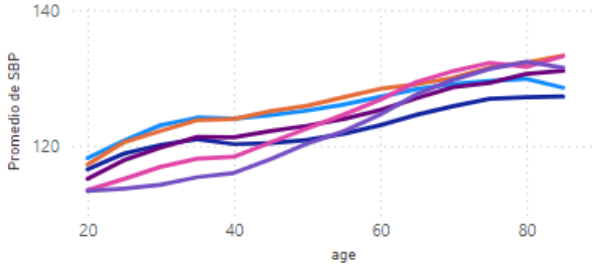
Distribución de los parámetros según edad y condición



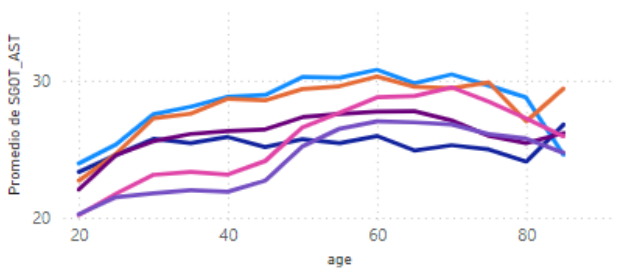




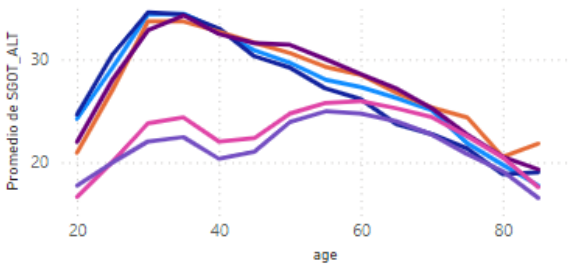
Distribución de SBP por edad y Condición F/B



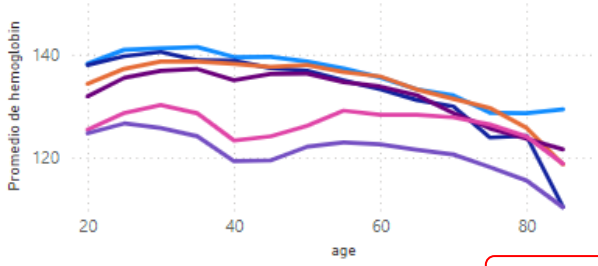
Distribución de AST por edad y Condición F/B



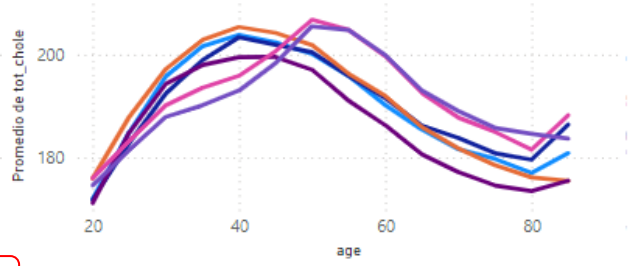
Distribución de ALT por edad y Condición F/B



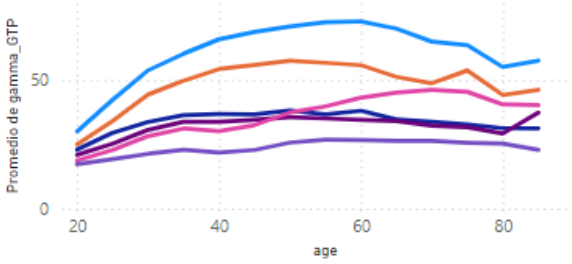
Distribución de Hemoglobina por edad y Condición F/B



Distribución de Colesterol total por edad y Condición F/B

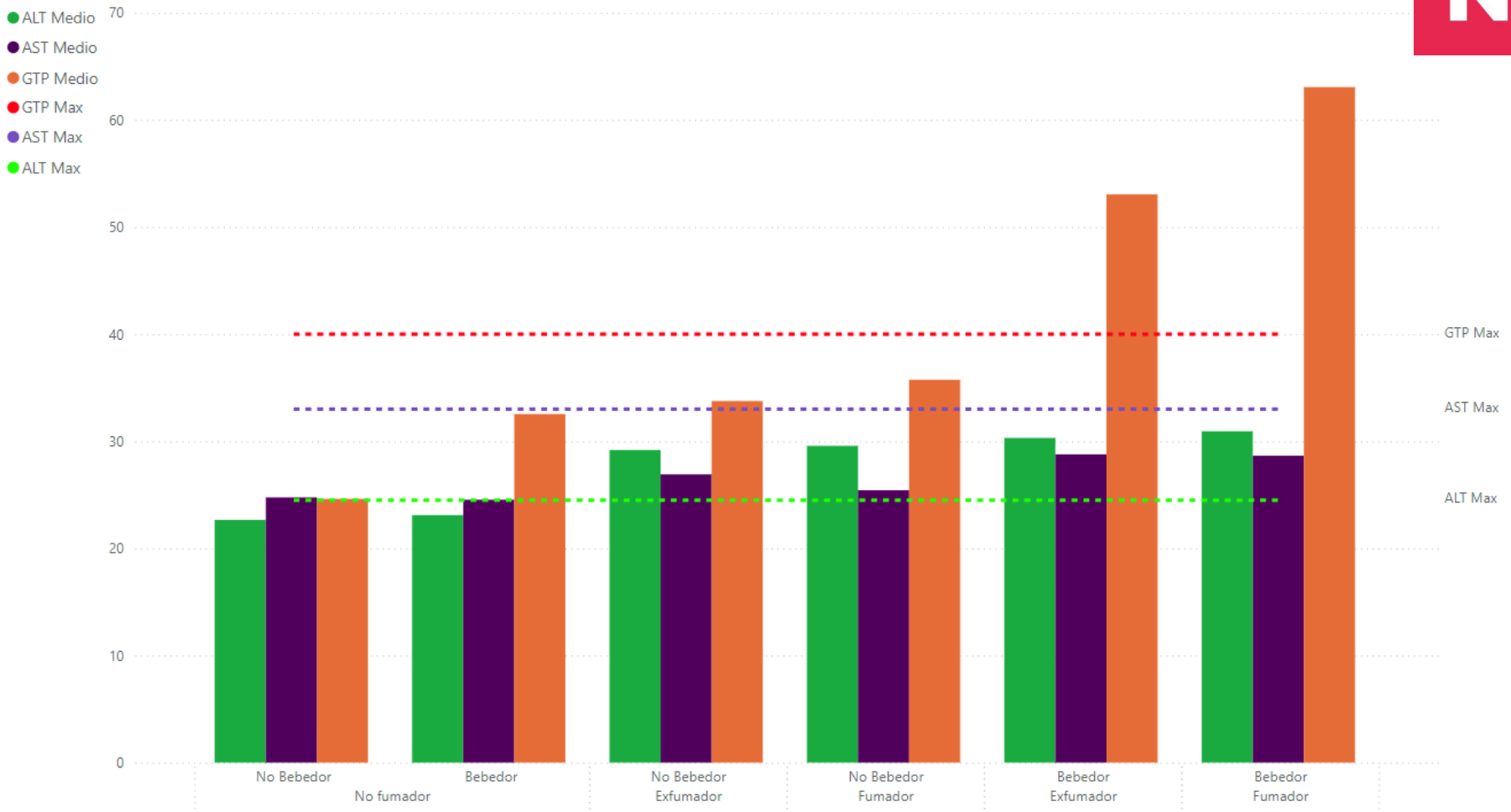


Distribución de Gamma-GTP por edad y Condición F/E

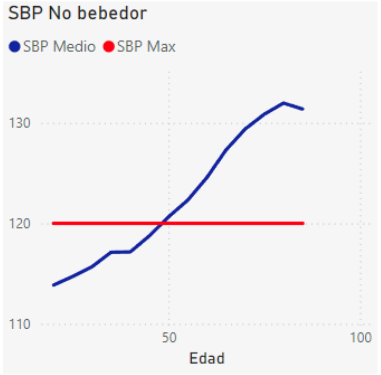


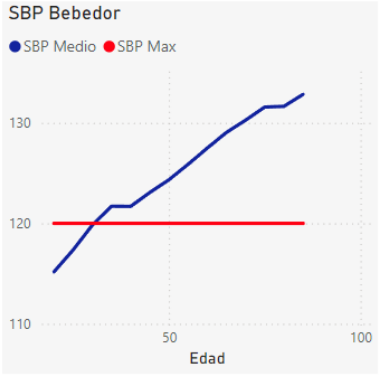
Smoking_Drinking_Status Fuma y Bebe Fuma y No Bebe Ha dejado de fumar y Bebe Ha dejado de fumar y No Bebe No Fuma y Bebe No Fuma y No Bebe

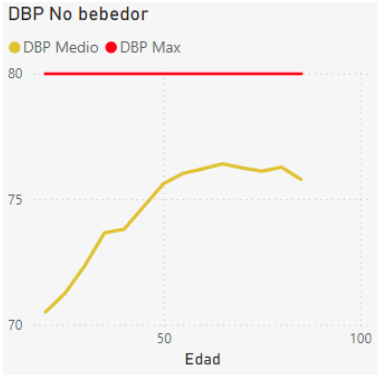
Efectos de fumar y beber sobre la función hepática

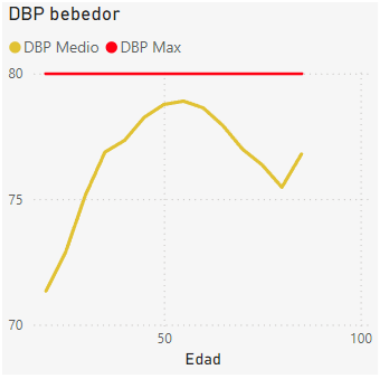


Los efectos de beber sobre la presión arterial







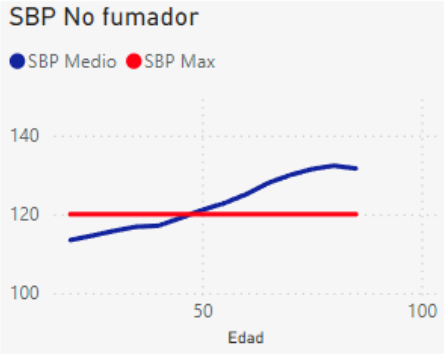


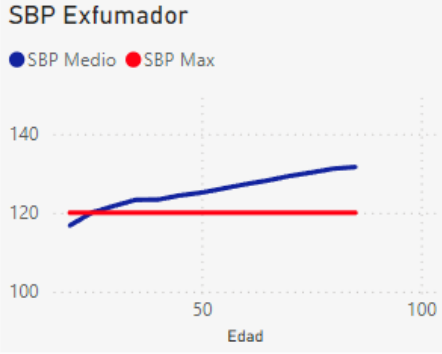
SBP: Presión sistólica, máxima presión arterial durante latidos cardíacos.

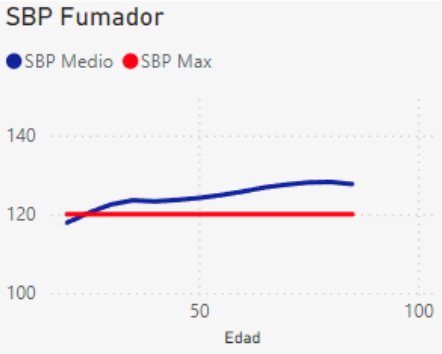
DBP: Presión diastólica, mínima presión arterial entre latidos cardíacos.

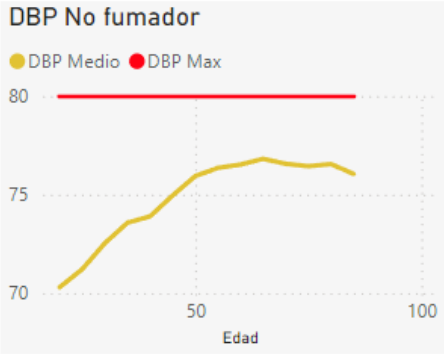
La línea roja establece el límite aconsejable de los valores SBP y DBP.

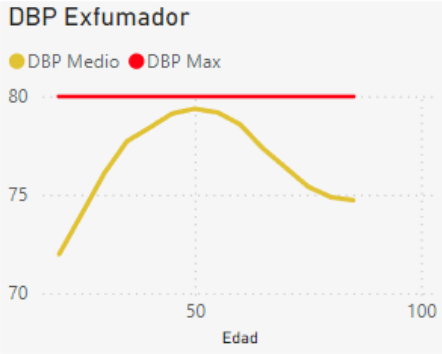
Los efectos de fumar sobre la presión arterial

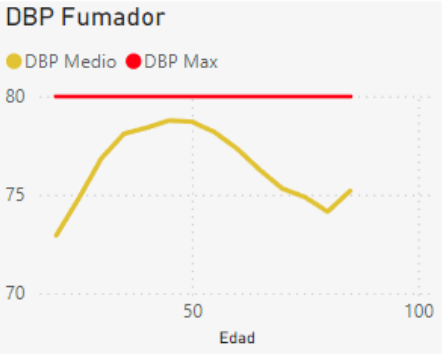






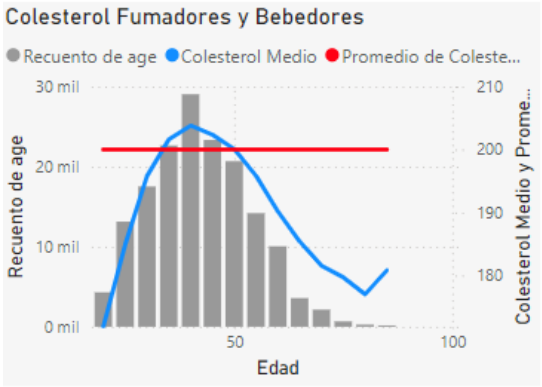




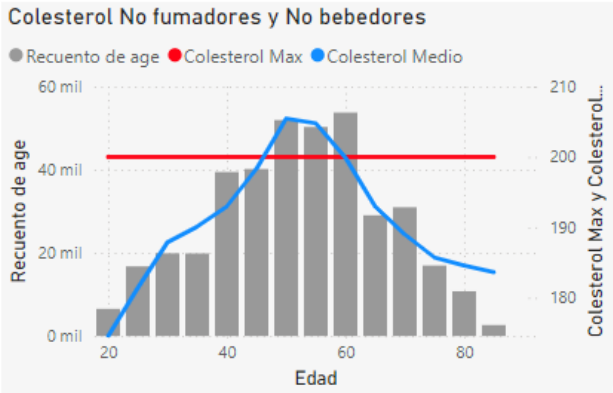


Evolución del colesterol según condición F/B

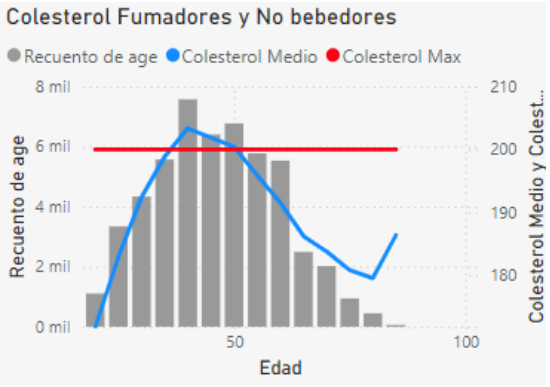
16,30 %



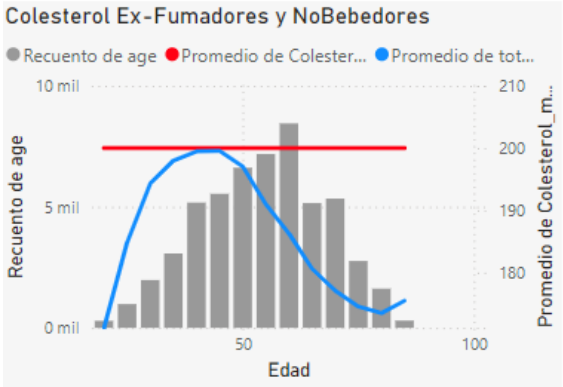
39,24 %



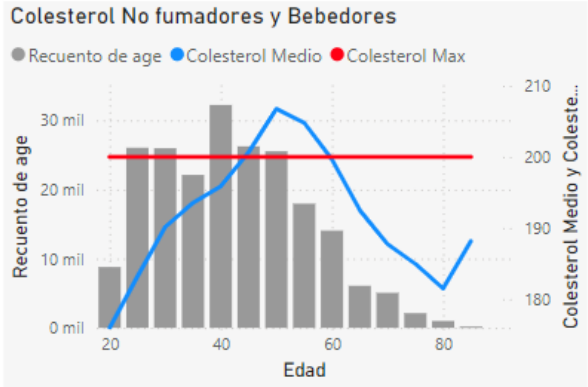
5,28 %



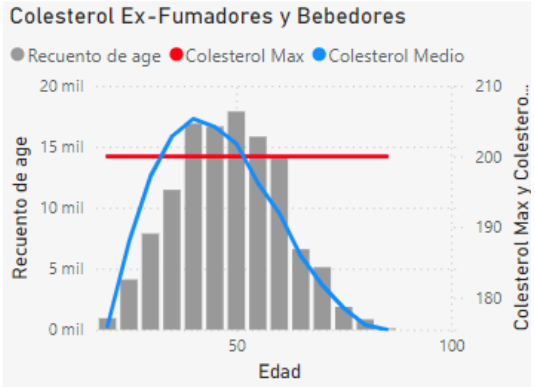
5,49 %



21,53 %

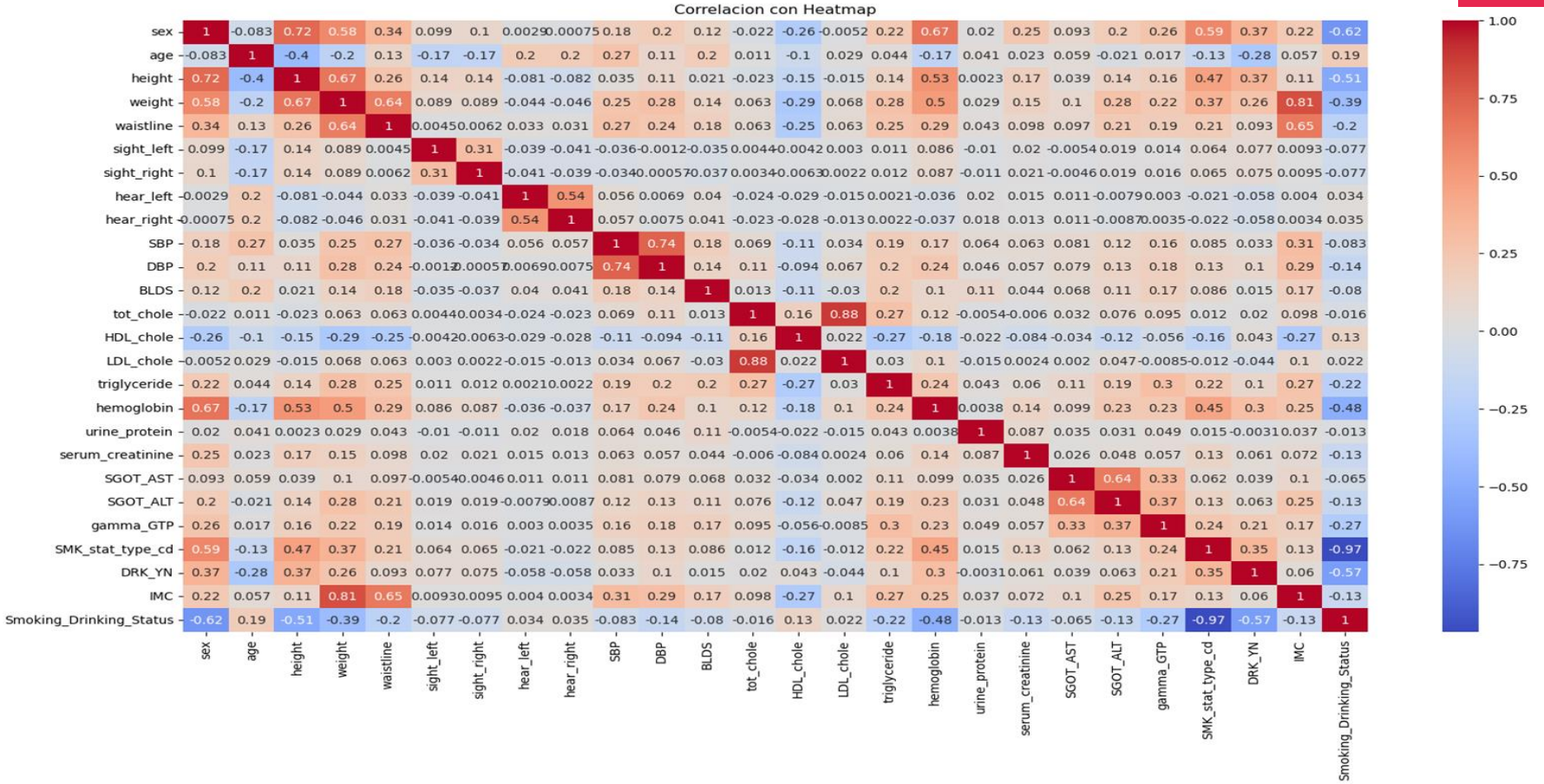


12,15 %



Análisis Predictivos

Correlaciones



Modelo regresión logística

Con **beber** como variable a predecir

VIF

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

Modelo regresión logística

Con **beber** como variable a predecir

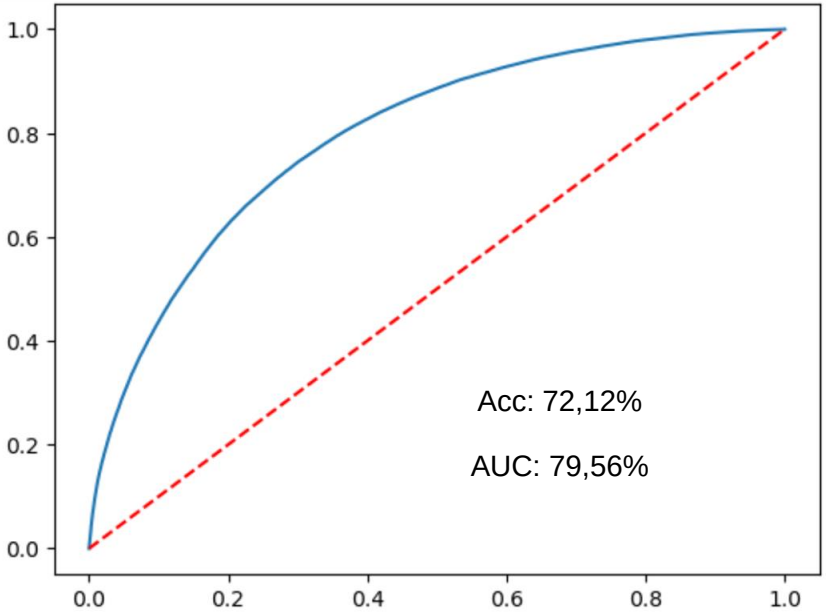
VIF + PCA

	PYTHON	R
Nº variables X	16 (91,88%)	17 (91,16%)
Accuracy	72,12%	50,05%

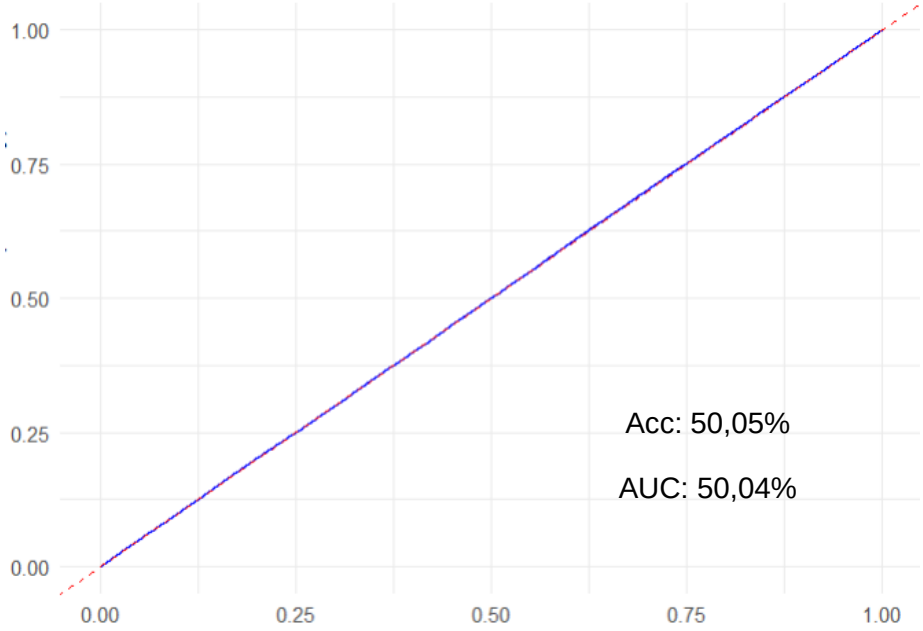
Modelo regresión logística

VIF + PCA ~ Curva ROC

Python



R



Modelo regresión logística

Con **beber** como variable a predecir

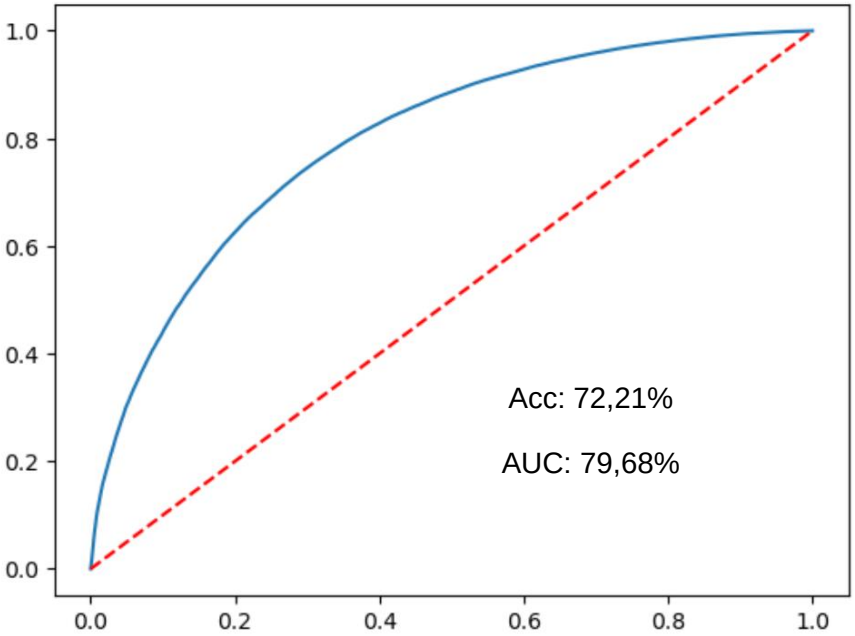
PCA

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

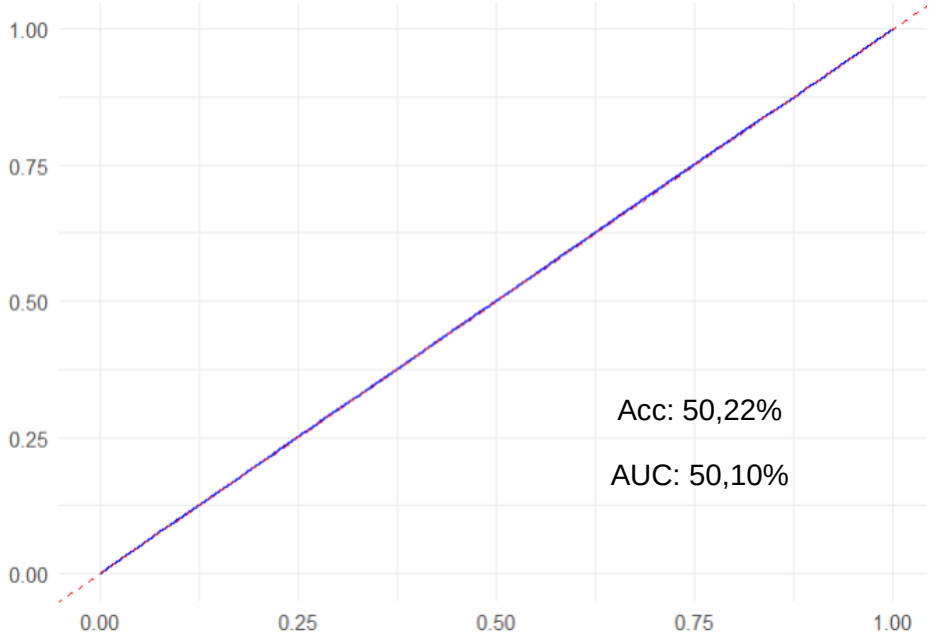
Modelo regresión logística

PCA ~ Curva ROC

Python



R



Modelo regresión logística

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%

Modelo regresión logística

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
Nº 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
Nº 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 → **Método del árbol**

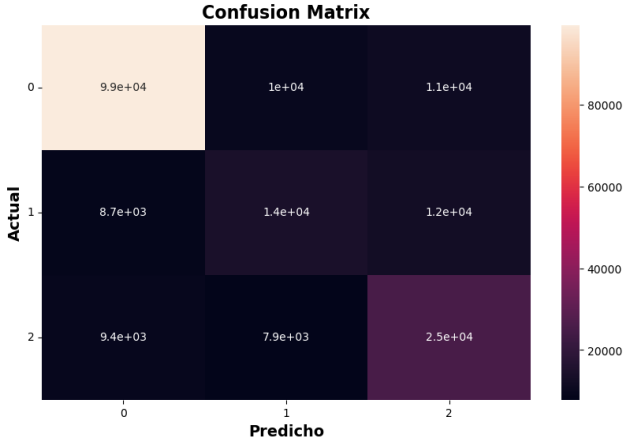
Modelos con el método del árbol	Python	R
Modelo previo al PCA	69,61%	68,29%
Modelo PCA tras eliminar columnas VIF>5	68,22%	65,65%
Modelo PCA con todas las columnas	68,04%	65,42%

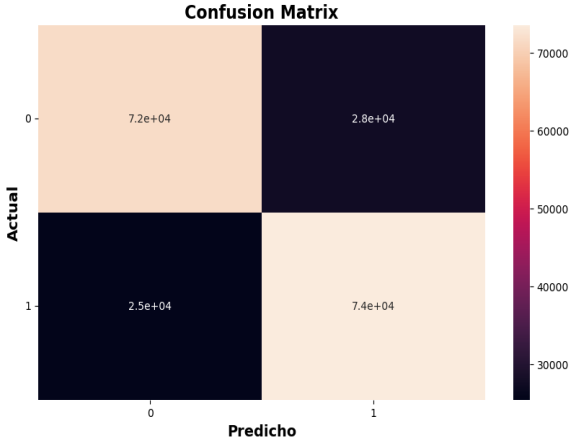
Análisis Predictivo

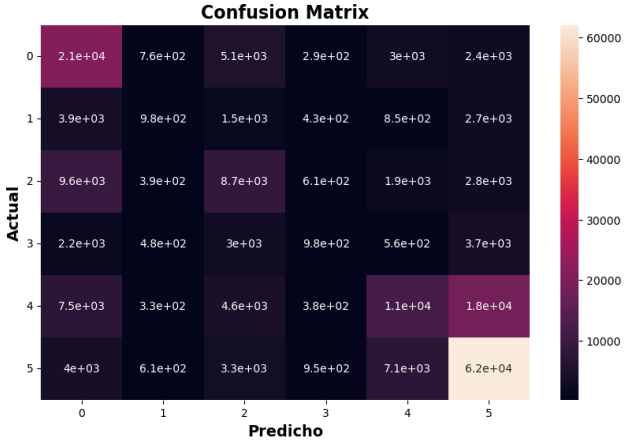
Con biblioteca **XGBoost** de Python:

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

  → **Cintura**: Personas con cinturas más grandes

  ← consumir alcohol

- **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