ESTIMACIÓN DE LOS NIVELES DE OBESIDAD

MÍNERIA DE DATOS

ENTREGA 3

Juan David Castillo Garza
Daniela Alexandra Herrera Fandiño

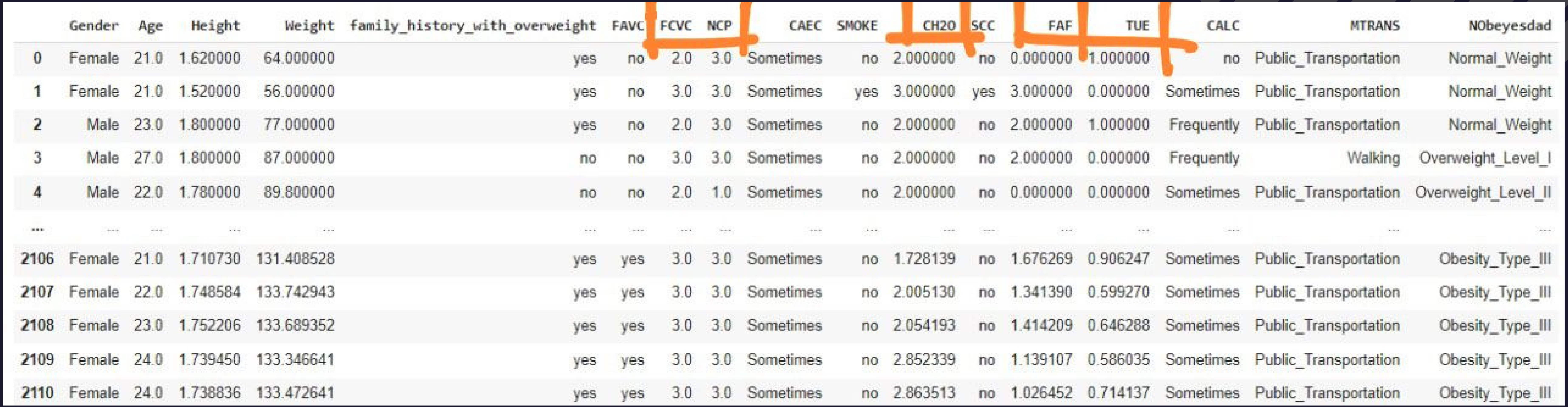
2022

PROCESAMIENTO

✗ NO DATOS FALTANTES

✓ DUPLICADOS (-24)

TRUNCAR DECIMALES



	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	Female	21.0	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	21.0	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	23.0	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	27.0	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	22.0	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	21.0	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	22.0	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	23.0	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	24.0	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	24.0	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

→ INCONSISTENTES

CONSUMO VERDURAS, CANTIDAD COMIDAS, CONSUMO AGUA, ACTIVIDAD FISICA, USO DISPOSITIVOS ELECTRONICOS

DUMMY

NIVEL DE PESO

BAJO PESO	**0**
NORMAL	0
EXCESO DE PESO	0
OBESIDAD I	**1**
OBESIDAD II	**1**
OBESIDAD III	**1**

EXPLICATIVAS

HISTORIAL FAM.

SI	**1**
NO	0

CALORIAS

SI	**1**
NO	0

FUMA

SI	**1**
NO	0

GENERO

FEMENINO	**0**
MASCULINO	1

EDAD

Numerico

PESO

Numerico

VERDURAS

- NUNCA
- ALGUNAS VECES
- SIEMPRE

COMIDAS PRINCIPALES

- 1 - 2
- 3
- + 3

AGUA

- -1L
- 1L - 2L
- + 2L

ENTRE COMIDAS

- NO
- ALGUNAS VECES
- FRECUENTEMENTE
- SIEMPRE

ALCOHOL

- NO
- FRECUENTEMENTE
- SIEMPRE

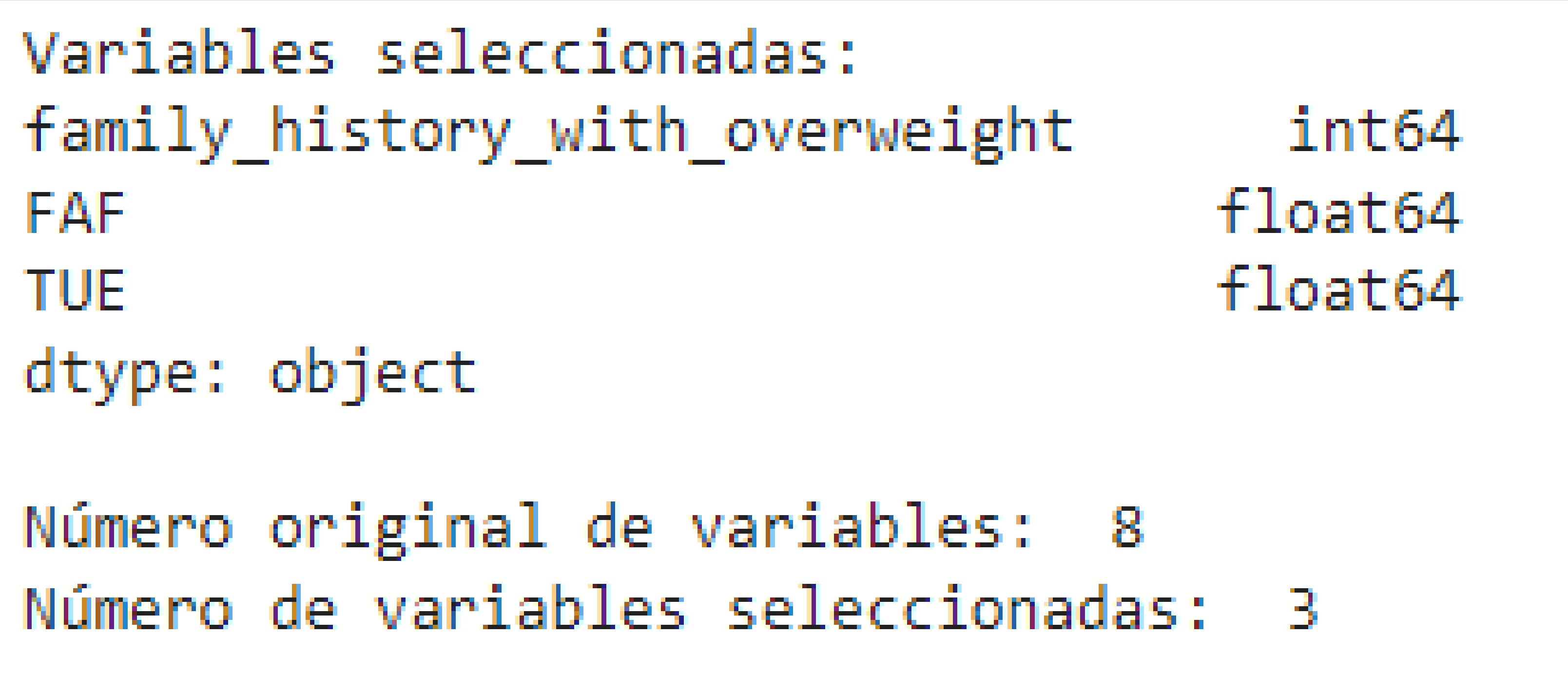
MEDIO TRANSPORTE

- AUTOMOVIL
- MOTO
- TRANSPORTE PUBLICO
- CAMINAR Y BICICLETA

SELECCION A PRIORI

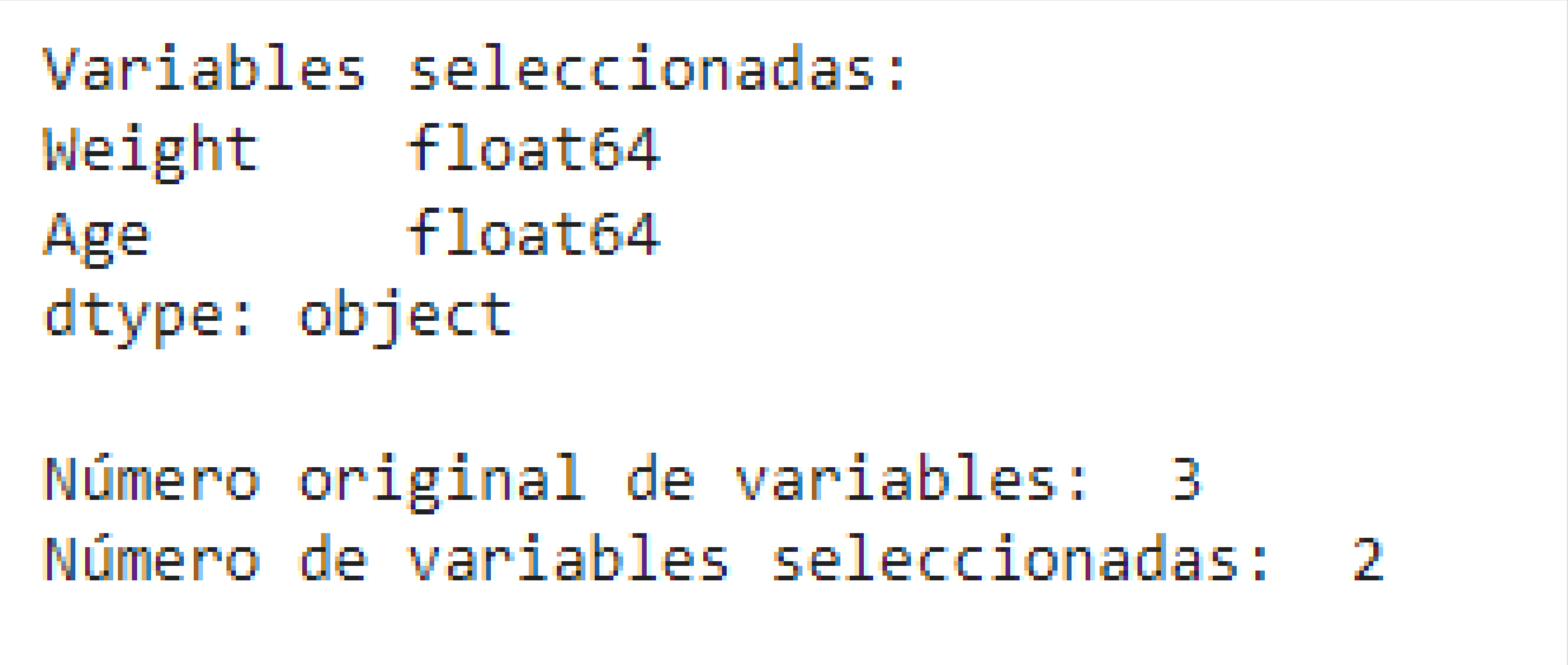
PRUEBA F Y CHI-CUADRADO

X = base[['family_history_with_overweight', 'FAVC', 'FCVC','NCP' ,'SMOKE' ,'CH2O' ,'FAF' ,'TUE']]



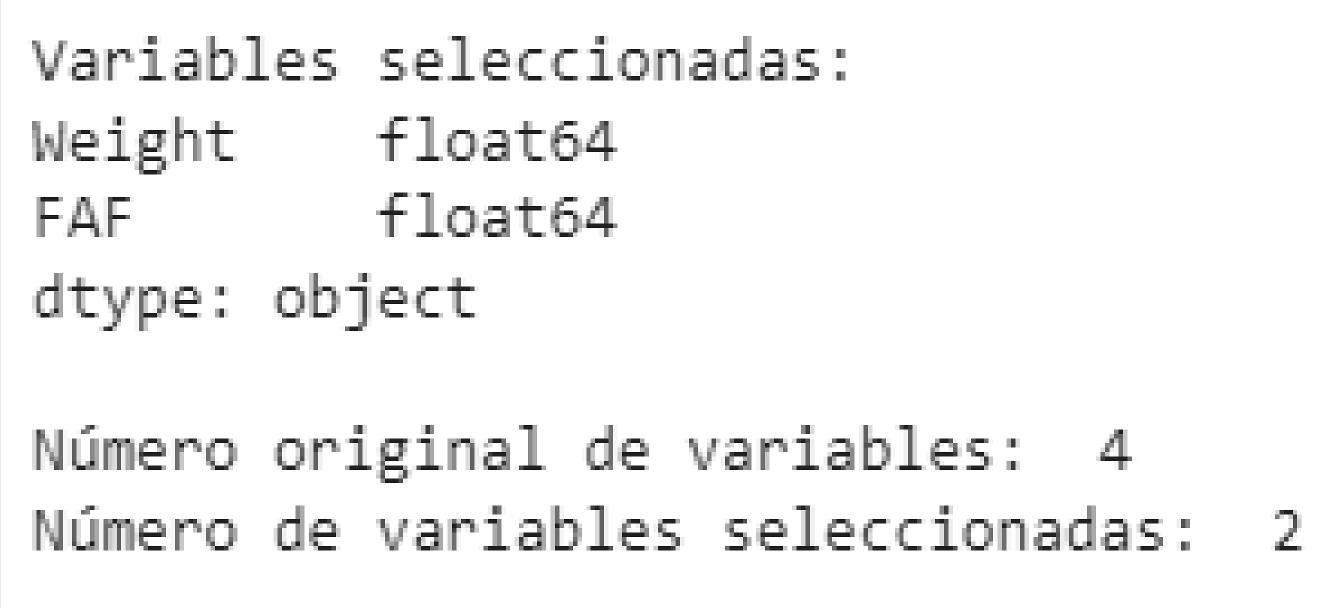
```
Variables seleccionadas:
family_history_with_overweight      int64
FAF                               float64
TUE                               float64
dtype: object

Número original de variables:  8
Número de variables seleccionadas:  3
```

X = base[['Weight','Height','Age']]



```
Variables seleccionadas:
Weight    float64
Age       float64
dtype: object

Número original de variables:  3
Número de variables seleccionadas:  2
```

X=base[['Weight','FCVC', 'NCP', 'FAF']]



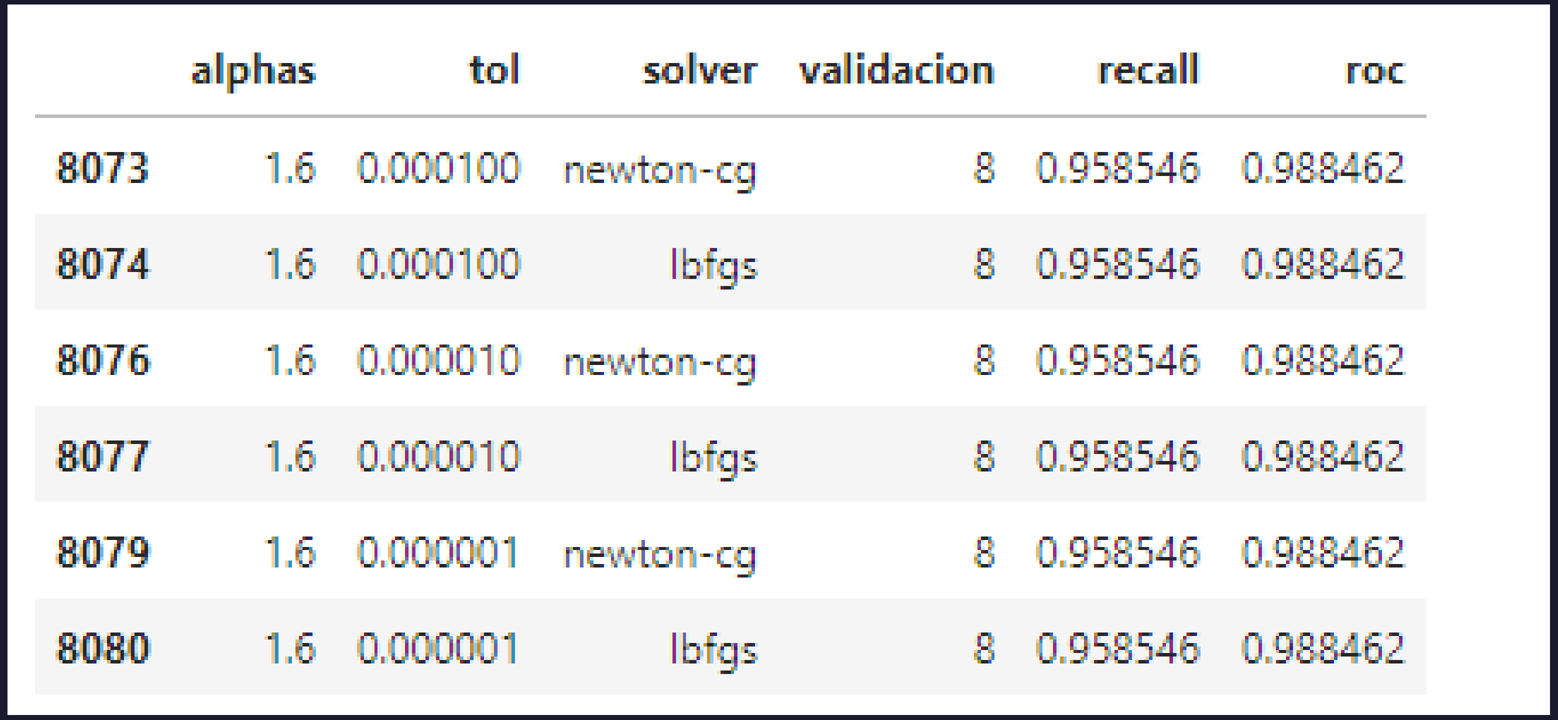
```
Variables seleccionadas:
Weight    float64
FAF       float64
dtype: object

Número original de variables:  4
Número de variables seleccionadas:  2
```

ALGORITMOS

HIPERPARAMETROS

LASSO

- ALPHAS : 115 ***
- CROOS VALIDATION : 9
- TOLERANCIAS : 3 ***
- SOLVER : newton-cg', 'lbfgs', 'liblinear ***
- CANTIDAD DE MODELOS : 9315



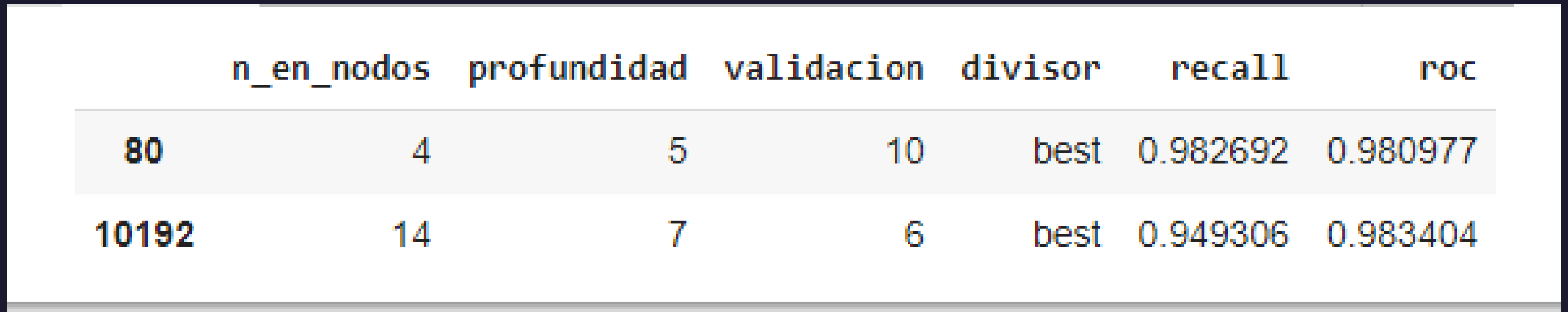
	alphas	tol	solver	validacion	recall	roc
8073	1.6	0.000100	newton-cg	8	0.958546	0.988462
8074	1.6	0.000100	lbfgs	8	0.958546	0.988462
8076	1.6	0.000010	newton-cg	8	0.958546	0.988462
8077	1.6	0.000010	lbfgs	8	0.958546	0.988462
8079	1.6	0.000001	newton-cg	8	0.958546	0.988462
8080	1.6	0.000001	lbfgs	8	0.958546	0.988462

SELECCION DE PREDICTORES
MENOS INFLUYENTE

ROC
ACCURACY

ARBOL DECISION

- PROFUNDIDAD : 80
- CROOS VALIDATION : 9
- SPLIT : BEST, RANDOM
- MINIMO DE MUESTRAS POR HOJA : 9
- CANTIDAD DE MODELOS : 12960



	n_en_nodos	profundidad	validacion	divisor	recall	roc
80	4	5	10	best	0.982692	0.980977
10192	14	7	6	best	0.949306	0.983404

Modelo

log(odds) = 2.07+6.75*Family +0.91*FAVC+ 1.83*Smoke

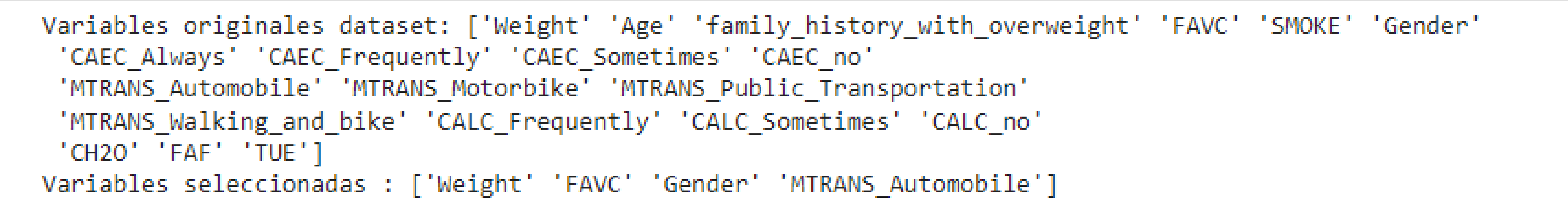
-2.18*Gender+ 0.40*CAEC_A,

-1.24*CAEC_f -0.75*CAEC_S, 0.25*CAEC_N ....

SELECCION

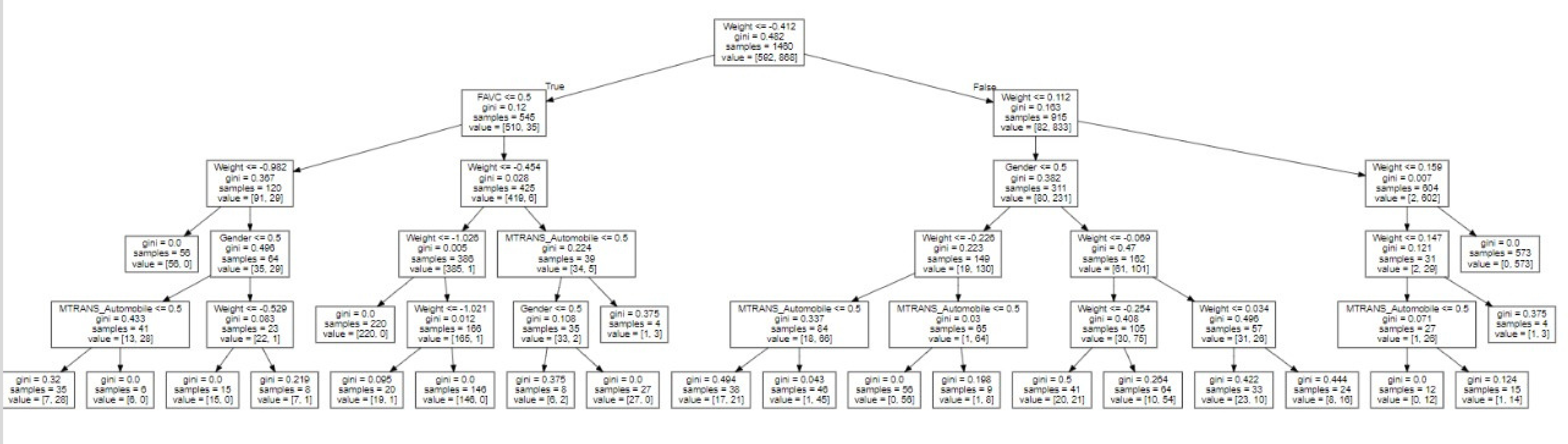
MEDIANTE ARBOL DE DECISION

SELECCION HACIA ADELANTE : 4 VARIABLES QUE MAXIMICEN ROC



```
Variables originales dataset: ['Weight' 'Age' 'family_history_with_overweight' 'FAVC' 'SMOKE' 'Gender'
 'CAEC_Always' 'CAEC_Frequently' 'CAEC_Sometimes' 'CAEC_no'
 'MTRANS_Automobile' 'MTRANS_Motorbike' 'MTRANS_Public_Transportation'
 'MTRANS_Walking_and_bike' 'CALC_Frequently' 'CALC_Sometimes' 'CALC_no'
 'CH2O' 'FAF' 'TUE']
Variables seleccionadas : ['Weight' 'FAVC' 'Gender' 'MTRANS_Automobile']
```

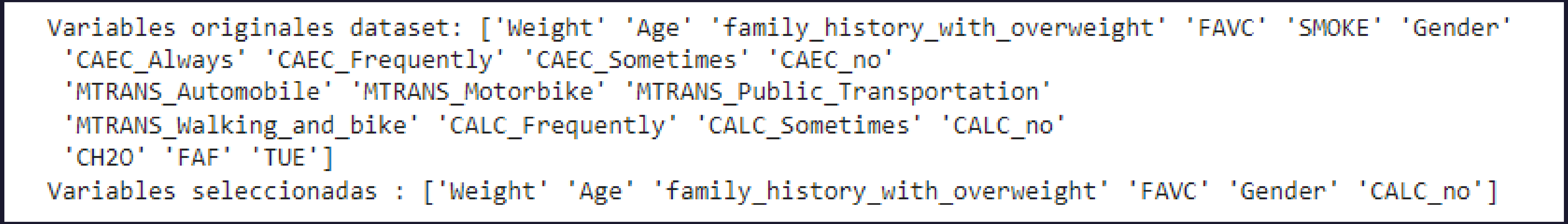
ANÁLISIS GRÁFICOS


Weight <= -0.412 gini = 0.482 samples = 1460 value = [592, 868]
True
False
FAVC <= 0.5 gini = 0.12 samples = 545 value = [510, 35]
Weight <= 0.112 gini = 0.163 samples = 915 value = [82, 833]
Weight <= -0.982 gini = 0.367 samples = 120 value = [91, 29]
Weight <= -0.454 gini = 0.028 samples = 425 value = [419, 6]
Gender <= 0.5 gini = 0.382 samples = 311 value = [80, 231]
Weight <= 0.159 gini = 0.007 samples = 604 value = [2, 602]
gini = 0.0 samples = 56 value = [56, 0]
Gender <= 0.5 gini = 0.496 samples = 64 value = [35, 29]
Weight <= -1.026 gini = 0.005 samples = 386 value = [385, 1]
MTRANS_Automobile <= 0.5 gini = 0.224 samples = 39 value = [34, 5]
Weight <= -0.226 gini = 0.223 samples = 149 value = [19, 130]
Weight <= -0.069 gini = 0.47 samples = 162 value = [61, 101]
Weight <= 0.147 gini = 0.121 samples = 31 value = [2, 29]
gini = 0.0 samples = 573 value = [0, 573]
MTRANS_Automobile <= 0.5 gini = 0.433 samples = 41 value = [13, 28]
Weight <= -0.529 gini = 0.083 samples = 23 value = [22, 1]
gini = 0.0 samples = 220 value = [220, 0]
Weight <= -1.021 gini = 0.012 samples = 166 value = [165, 1]
Gender <= 0.5 gini = 0.108 samples = 35 value = [33, 2]
gini = 0.375 samples = 4 value = [1, 3]
MTRANS_Automobile <= 0.5 gini = 0.337 samples = 84 value = [18, 66]
MTRANS_Automobile <= 0.5 gini = 0.03 samples = 65 value = [1, 64]
Weight <= -0.254 gini = 0.408 samples = 105 value = [30, 75]
Weight <= 0.034 gini = 0.496 samples = 57 value = [31, 26]
MTRANS_Automobile <= 0.5 gini = 0.071 samples = 27 value = [1, 26]
gini = 0.375 samples = 4 value = [1, 3]
gini = 0.32 samples = 35 value = [7, 28]
gini = 0.0 samples = 6 value = [6, 0]
gini = 0.0 samples = 15 value = [15, 0]
gini = 0.219 samples = 8 value = [7, 1]
gini = 0.095 samples = 20 value = [19, 1]
gini = 0.0 samples = 146 value = [146, 0]
gini = 0.375 samples = 8 value = [6, 2]
gini = 0.0 samples = 27 value = [27, 0]
gini = 0.494 samples = 38 value = [17, 21]
gini = 0.043 samples = 46 value = [1, 45]
gini = 0.0 samples = 56 value = [0, 56]
gini = 0.198 samples = 9 value = [1, 8]
gini = 0.5 samples = 41 value = [20, 21]
gini = 0.264 samples = 64 value = [10, 54]
gini = 0.422 samples = 33 value = [23, 10]
gini = 0.444 samples = 24 value = [8, 16]
gini = 0.0 samples = 12 value = [0, 12]
gini = 0.124 samples = 15 value = [1, 14]

SELECCION

MEDIANTE REGRESION LOGISTICA L2

SELECCION HACIA ATRAS

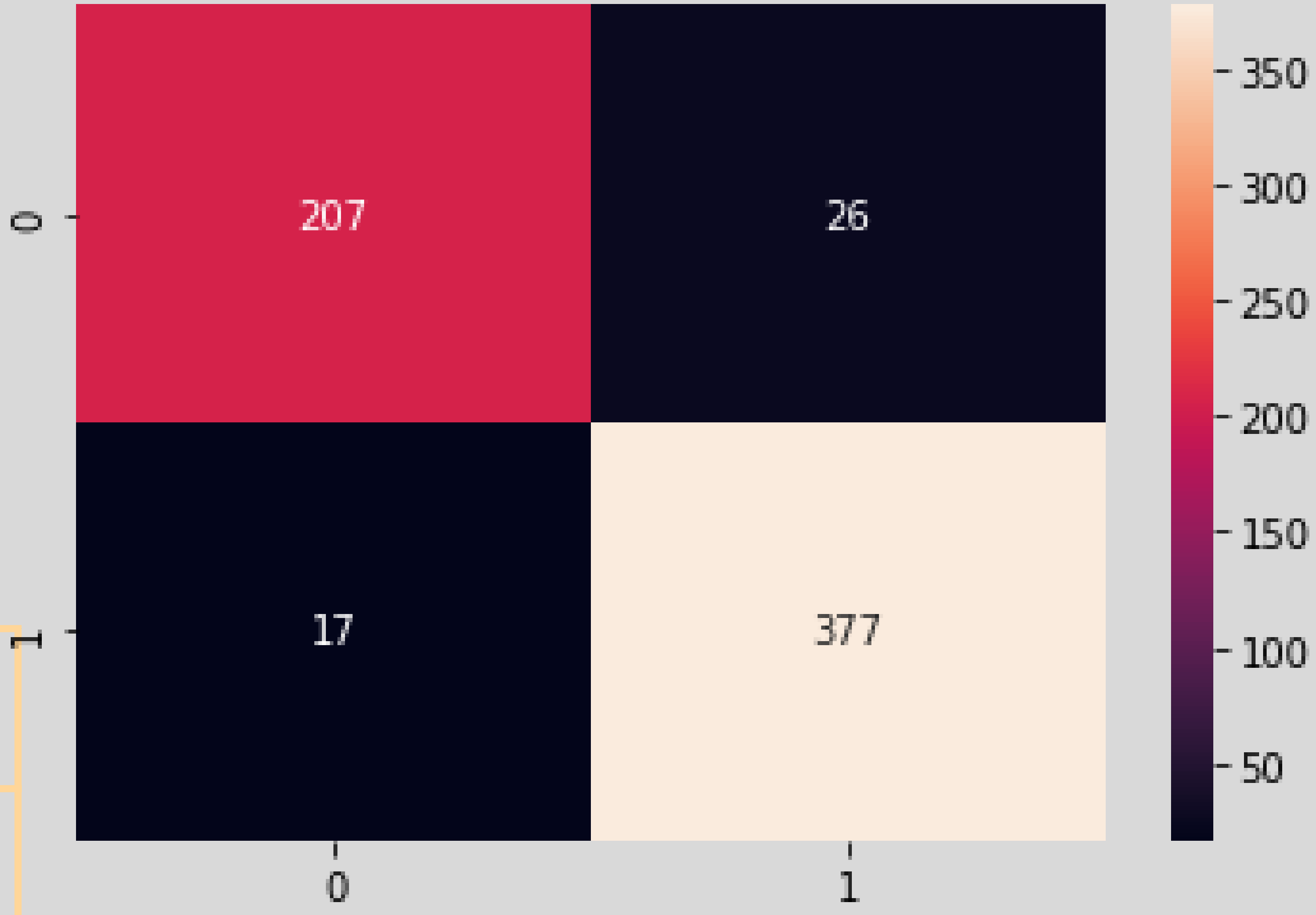


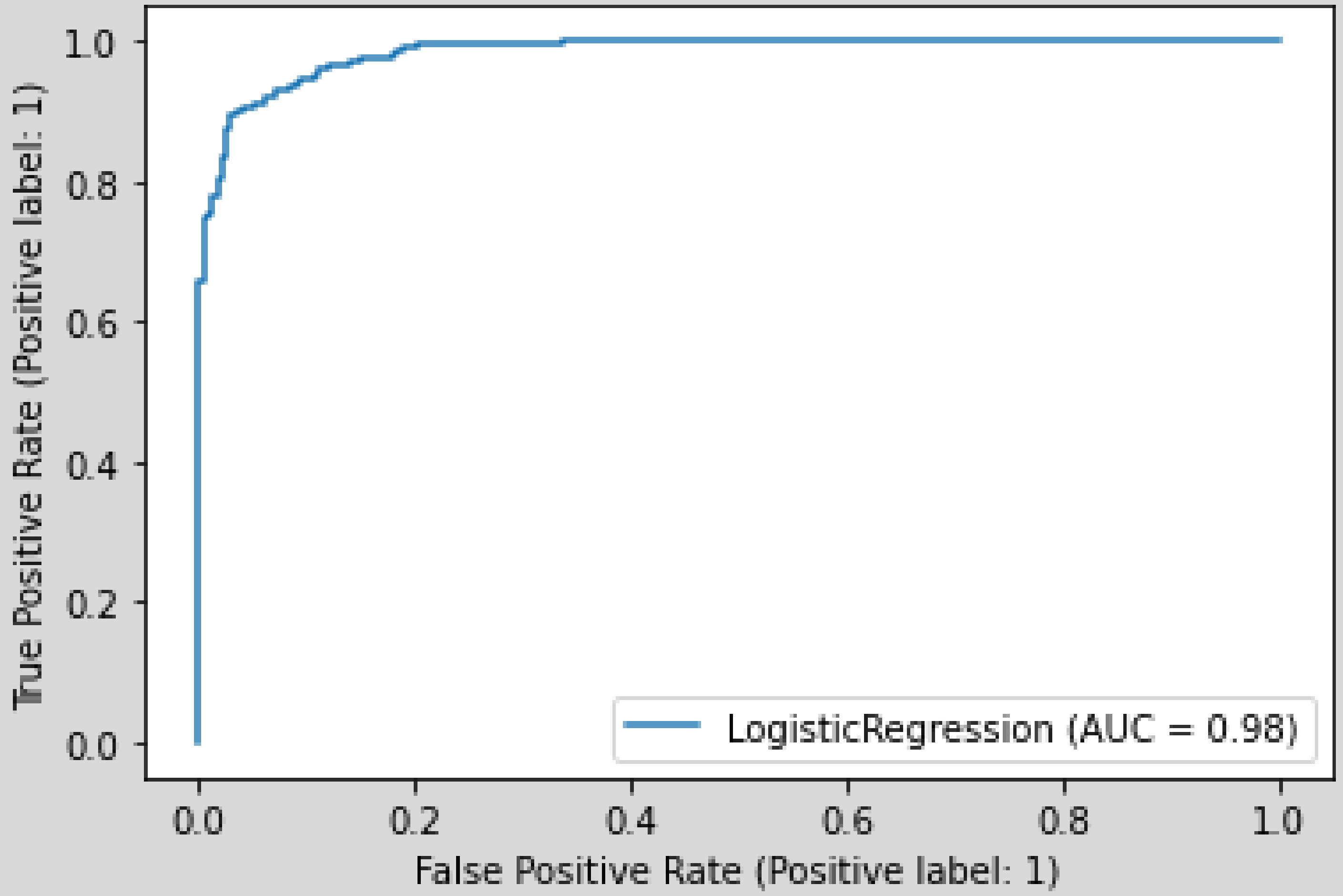
```
Variables originales dataset: ['Weight' 'Age' 'family_history_with_overweight' 'FAVC' 'SMOKE' 'Gender'
 'CAEC_Always' 'CAEC_Frequently' 'CAEC_Sometimes' 'CAEC_no'
 'MTRANS_Automobile' 'MTRANS_Motorbike' 'MTRANS_Public_Transportation'
 'MTRANS_Walking_and_bike' 'CALC_Frequently' 'CALC_Sometimes' 'CALC_no'
 'CH2O' 'FAF' 'TUE']
Variables seleccionadas : ['Weight' 'Age' 'family_history_with_overweight' 'FAVC' 'Gender' 'CALC_no']
```

EVALUACION MODELOS INDIVIDUALES

LASSO

ACCURACY	0.931
ROC_AUC	0.929
RECALL	0.956
FBETA 2	0.952

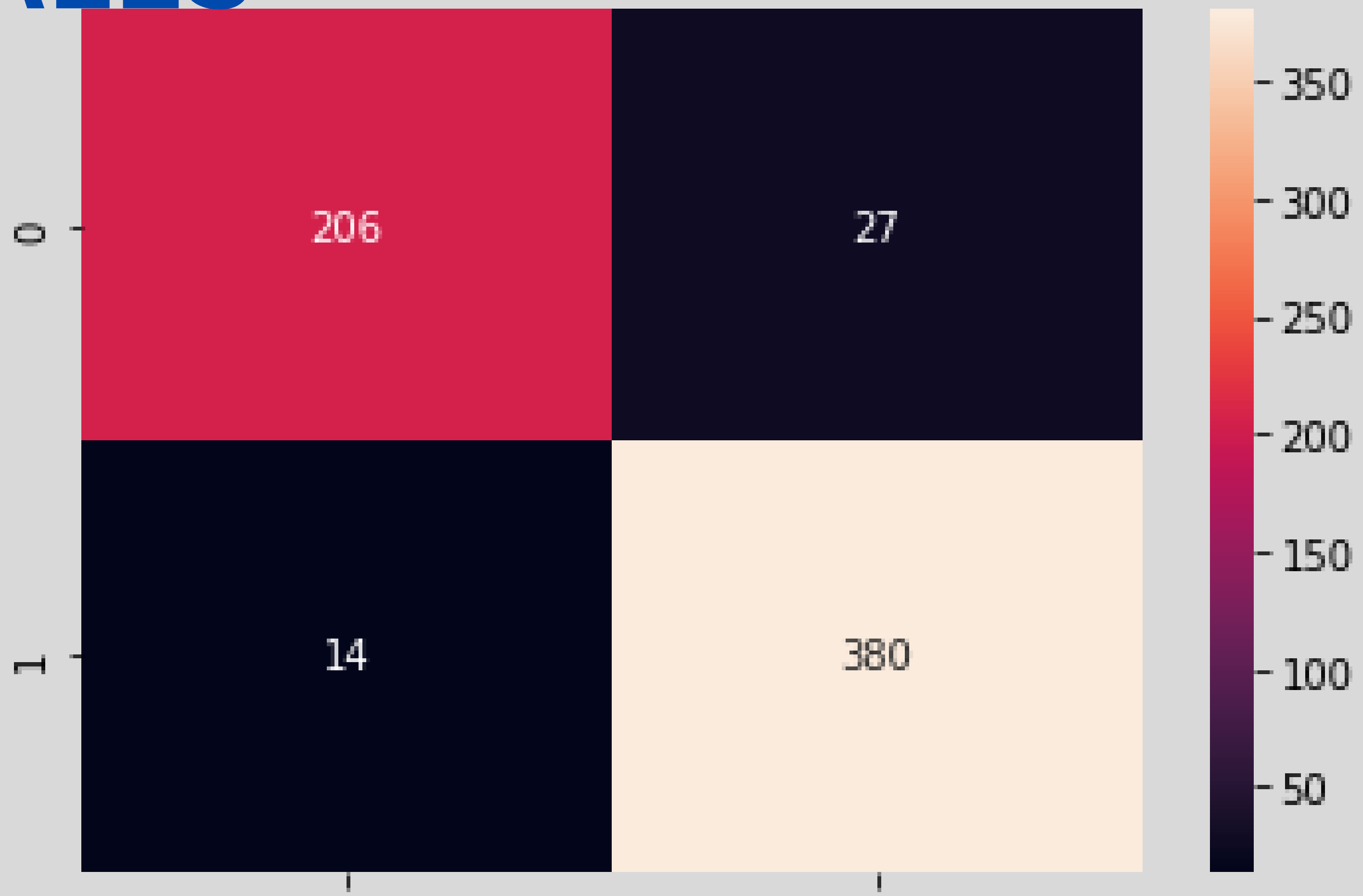


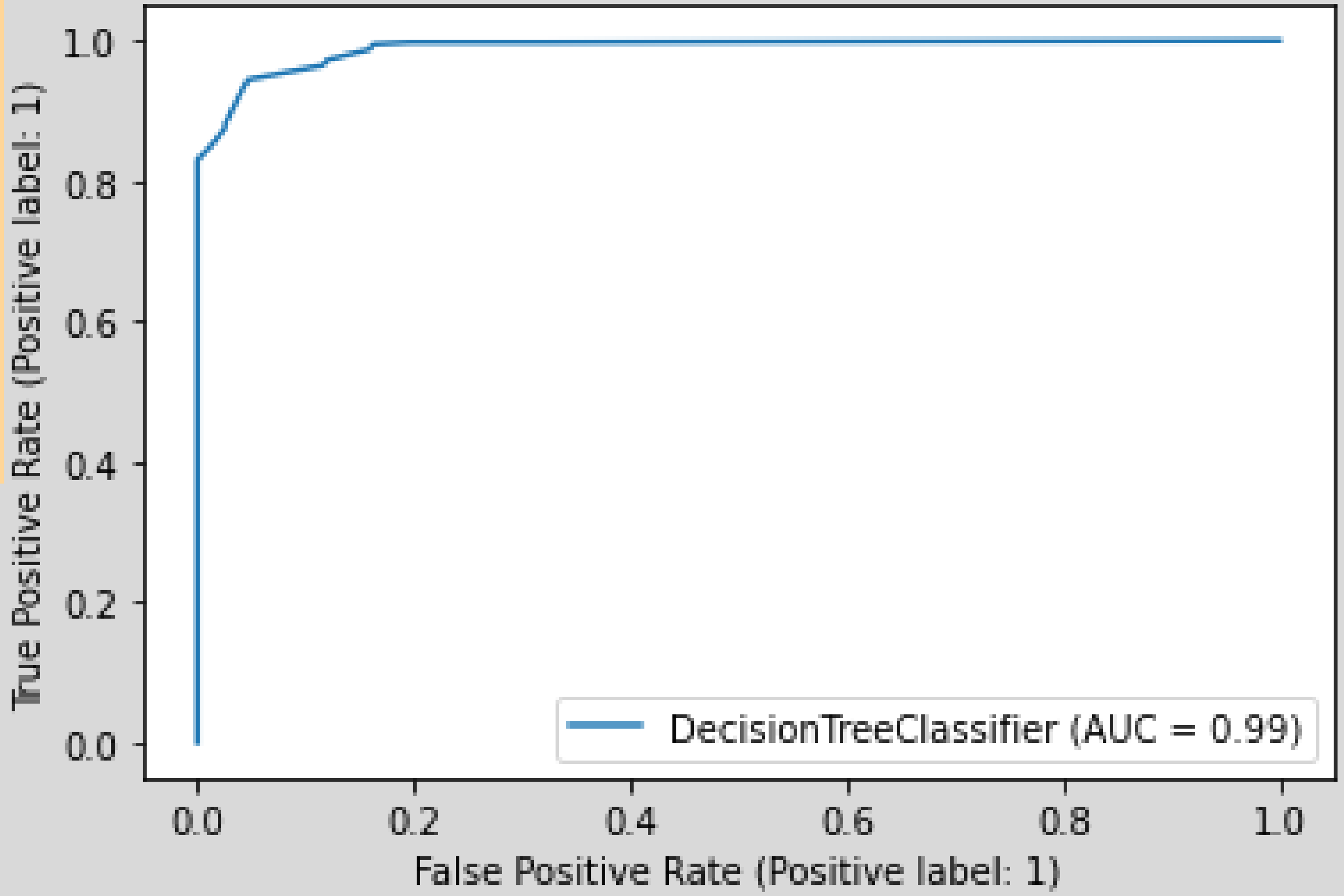


EVALUACION MODELOS INDIVIDUALES

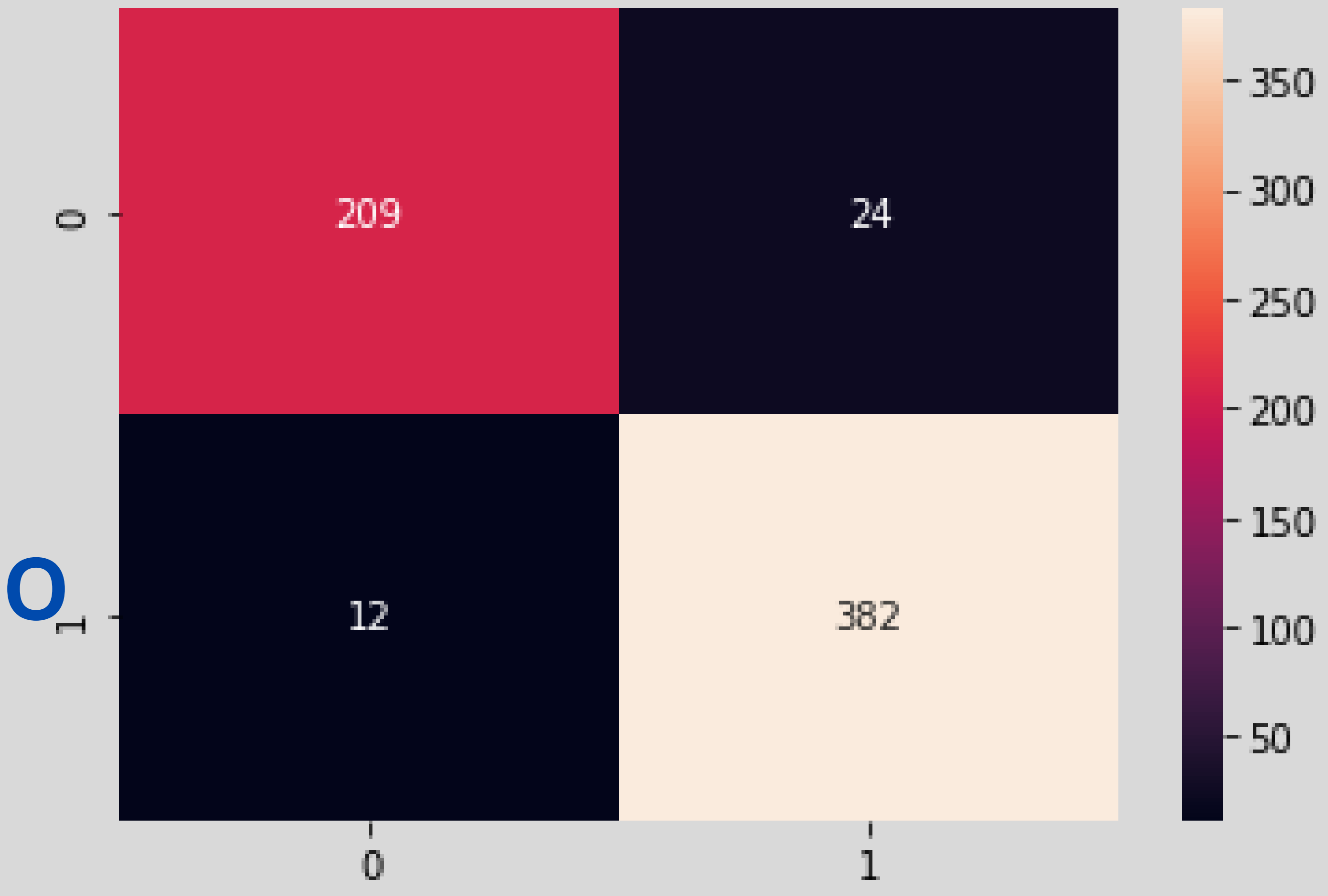
ARBOL DECISION

ACCURACY	0.934
ROC_AUC	0.935
RECALL	0.964
FBETA 2	0.958



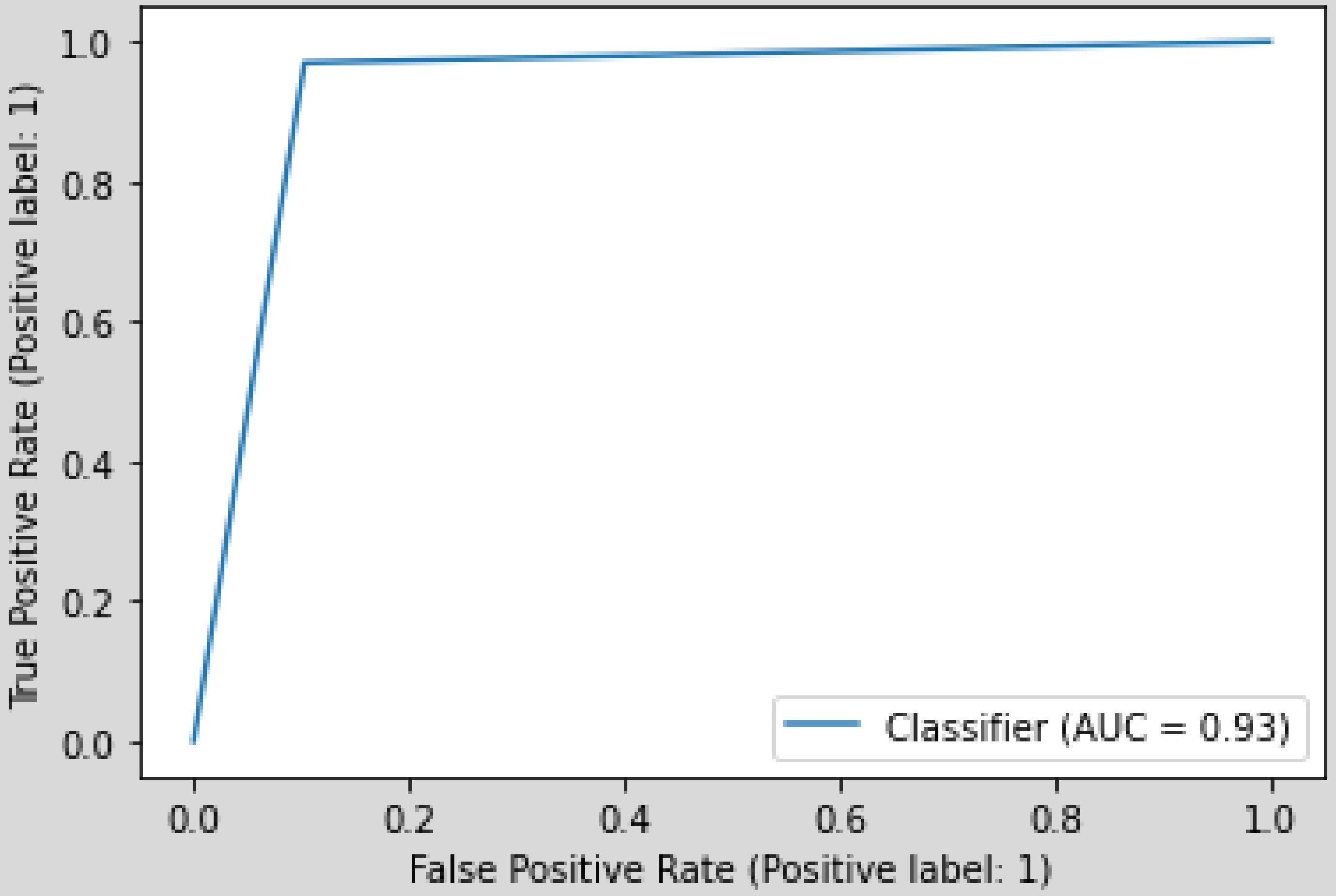


EVALUACION MODELOS INDIVIDUALES



MODELO ENSAMBLADO

ACCURACY	0.942
ROC_AUC	0.933
RECALL	0.969
FBETA 2	0.963



BIBLIOGRAFIA

- https://archive.ics.uci.edu/ml/index.php
- https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+