Proyecto Final

Luis Daniel Casais Mezquida

Análisis Inteligente de Datos 24/25

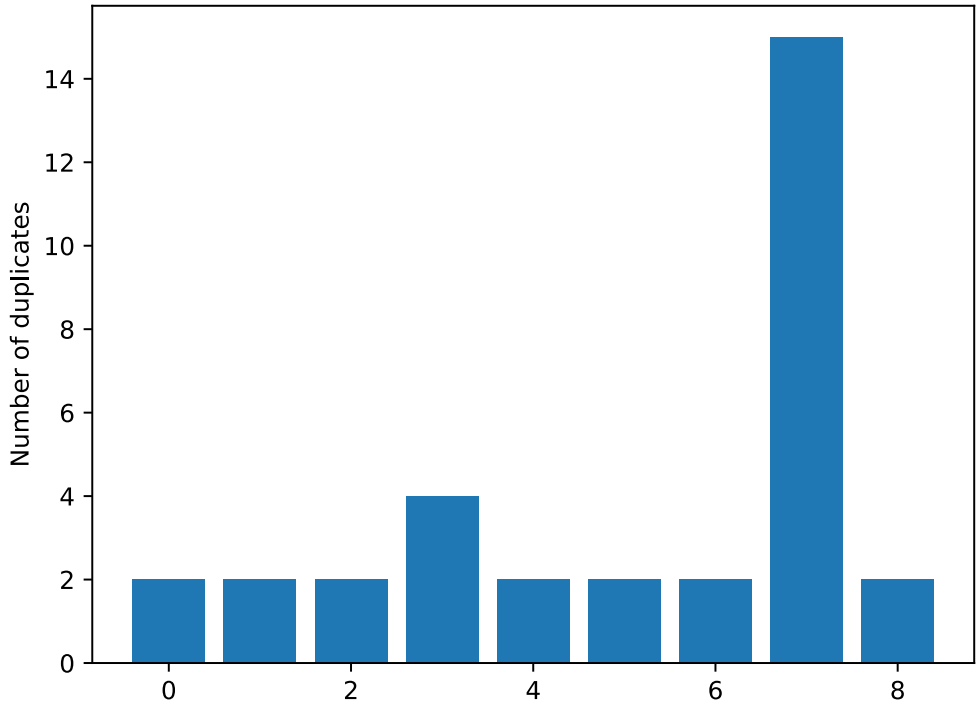
Universidad Carlos III de Madrid

Dataset

"Obesity Prediction Dataset", por Stephen Adeniran (2025)

- Estimar nivel de obesidad
- Hábitos alimenticios, físicos, y médicos
- México, Perú, y Colombia

- 2112 observaciones
- 17 características
  - categóricas (5), numéricas (8), booleanas (4)
- 0 celdas vacías, 9 (33) duplicados (1.56%)

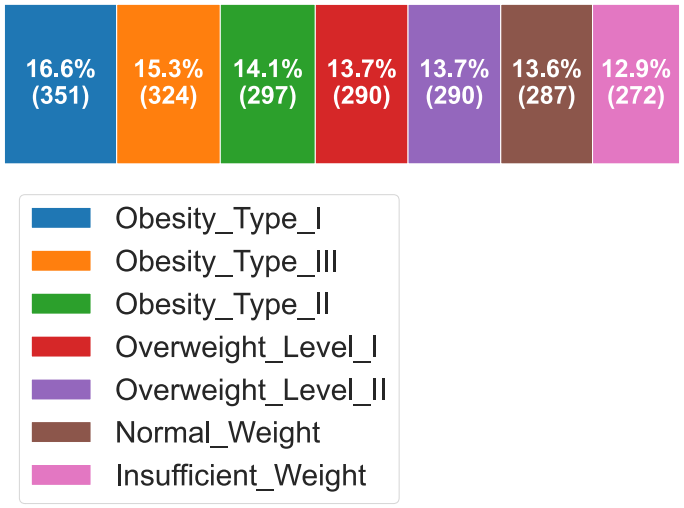


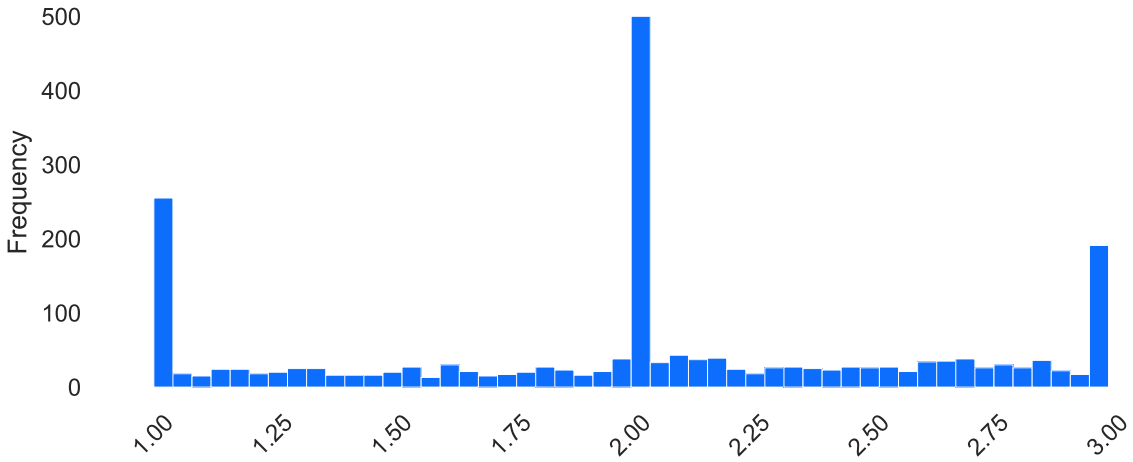
Gender	Male
Age	21
Height	1.62
Weight	70.0
family_history_with_overweight	no
FAVC	yes
FCVC	2.0
NCP	1.0
CAEC	no
SMOKE	no
CH2O	3.0
SCC	no
FAF	1.0
TUE	0.0
CALC	Sometimes
MTRANS	Public_Transportation
NObeyesdad	Overweight_Level_I
# duplicates	15

Análisis univariable

- Características muy desbalanceadas (>80%): *SMOKE*, *family_history_with_overweight*
- Característica objetivo balanceada
- Características de rangos numéricos con "ruido"

Dataset

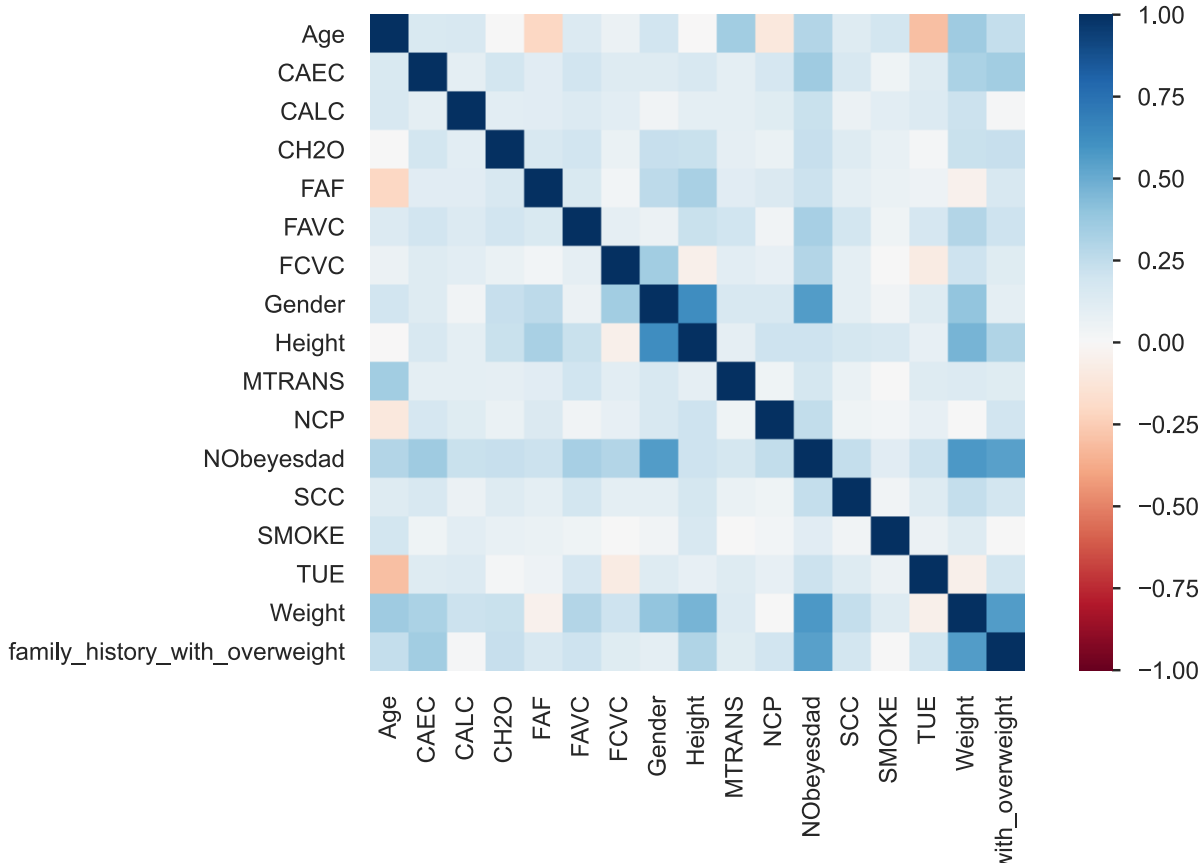

16.6%
(351)
15.3%
(324)
14.1%
(297)
13.7%
(290)
13.7%
(290)
13.6%
(287)
12.9%
(272)
Obesity_Type_I
Obesity_Type_III
Obesity_Type_II
Overweight_Level_I
Overweight_Level_II
Normal_Weight
Insufficient_Weight

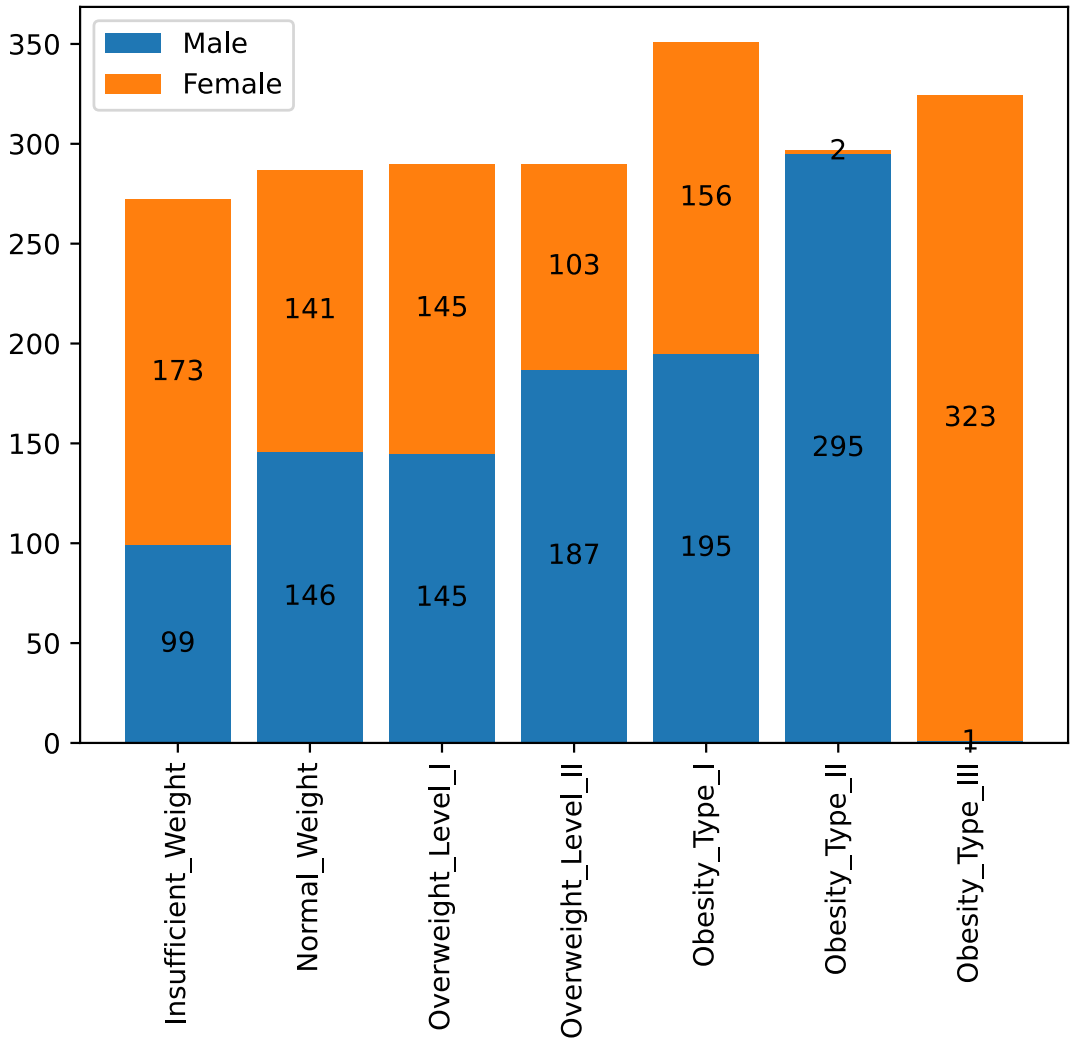

Frequency
500
400
300
200
100
0
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00

Análisis multivariable

- Alta correlación entre género, peso, e historial familiar con obesidad
- Obesidad por género:
  - Tipo II: apenas mujeres
  - Tipo III: apenas hombres

Dataset


Age
CAEC
CALC
CH2O
FAF
FAVC
FCVC
Gender
Height
MTRANS
NCP
NObeyesdad
SCC
SMOKE
TUE
Weight
family_history_with_overweight
Age
CAEC
CALC
CH2O
FAF
FAVC
FCVC
Gender
Height
MTRANS
NCP
NObeyesdad
SCC
SMOKE
TUE
Weight
vith_overweight
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Male
Female
350
300
250
200
150
100
50
0
173
99
141
146
145
145
103
187
156
195
2
295
323
1
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III

Metodología

Métrica de evaluación: F1-score, macro

Preprocesado:

1. Redondear rangos numéricos a entero
2. Escalado para variables numéricas
3. Encoding para categóricas

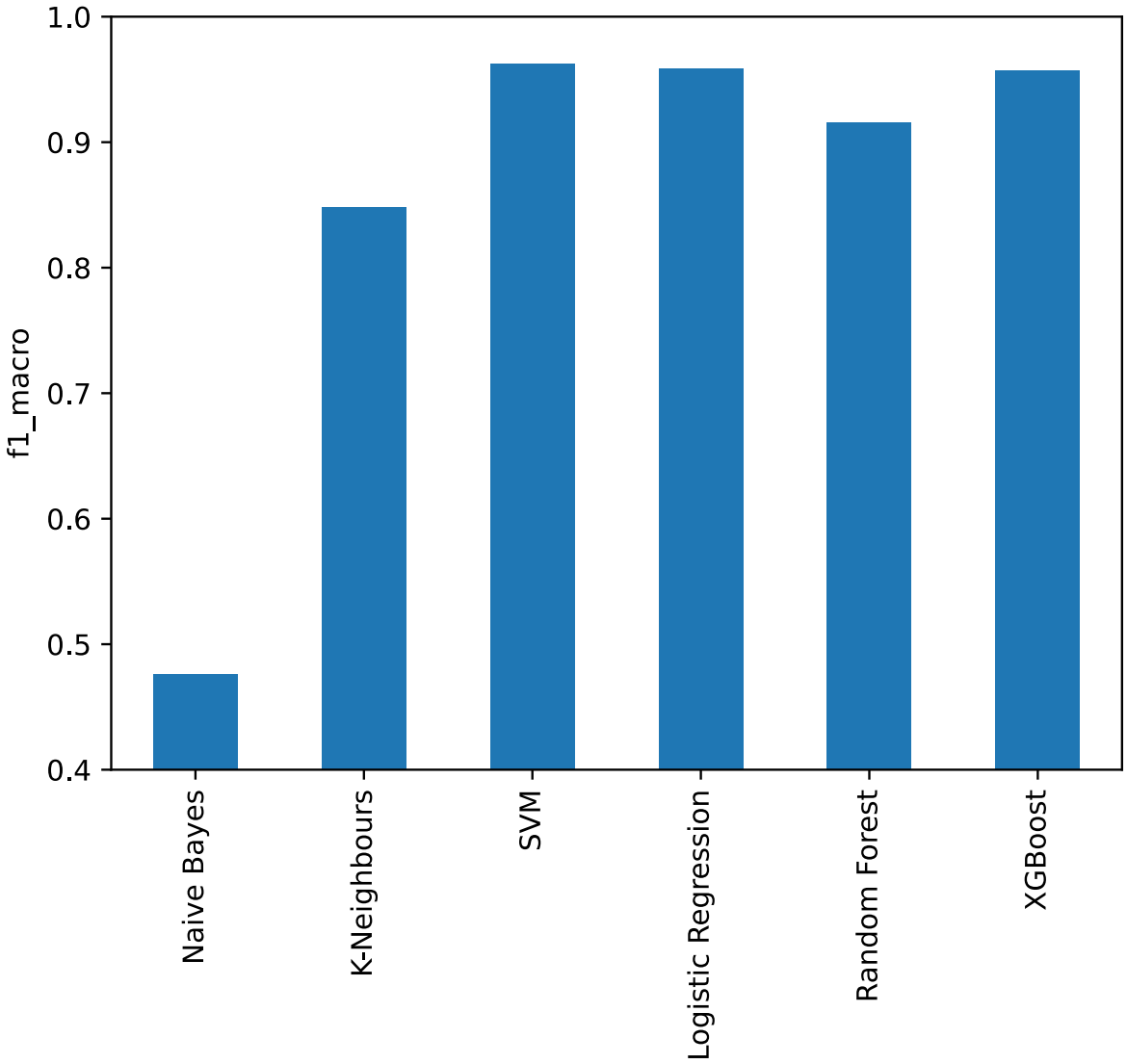
Experimentación

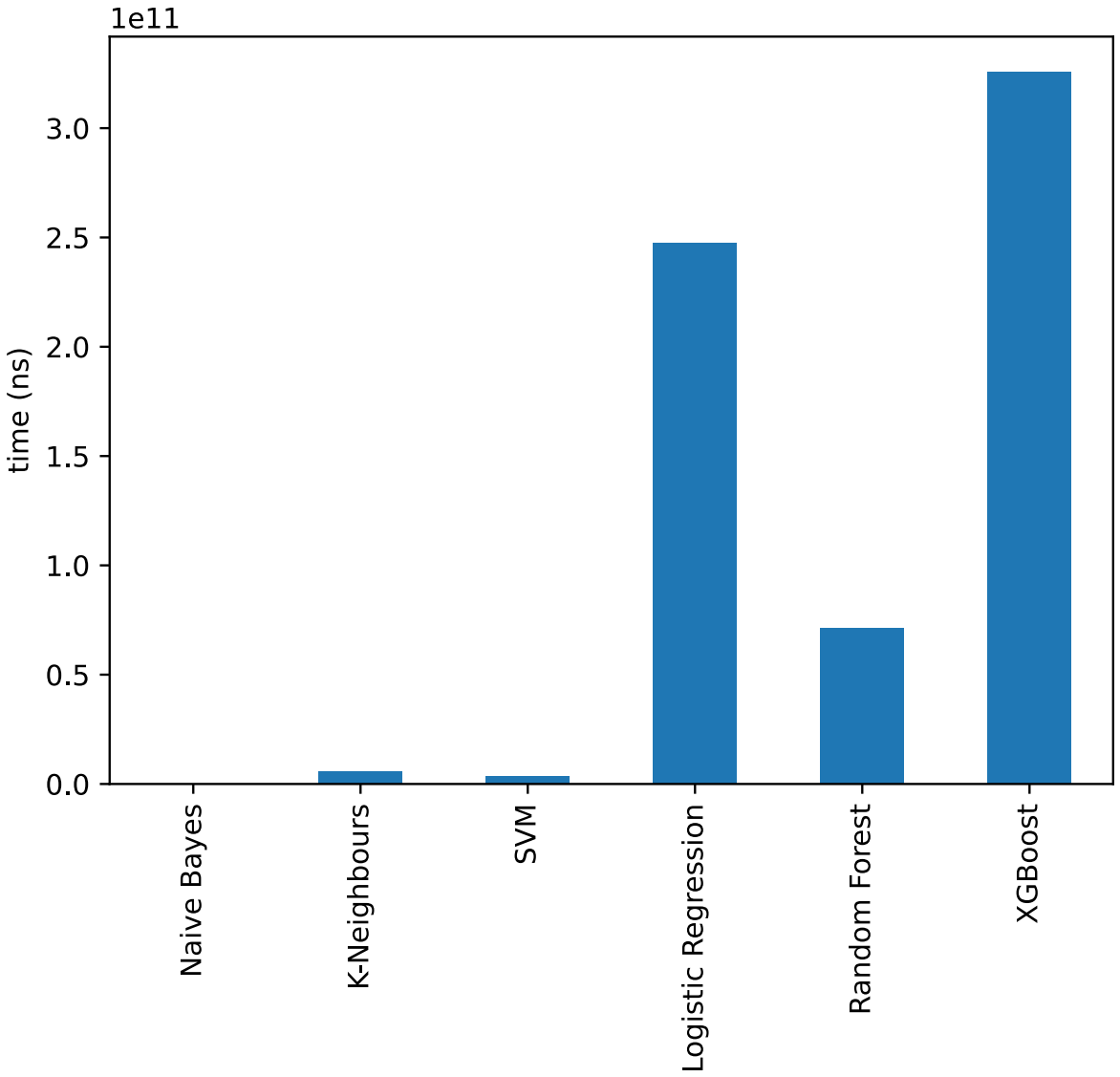
- 20% test
- *Cross Validation* de 5
- Ajuste hiperparámetros

Modelos:

1. Base
    - *Naive Bayes*
    - *K-Neighbours*
    - *SVM*
    - *Logistic Regression*
    - *Random Forest*
    - *XGBoost*
2. *Stacking*

Modelos base

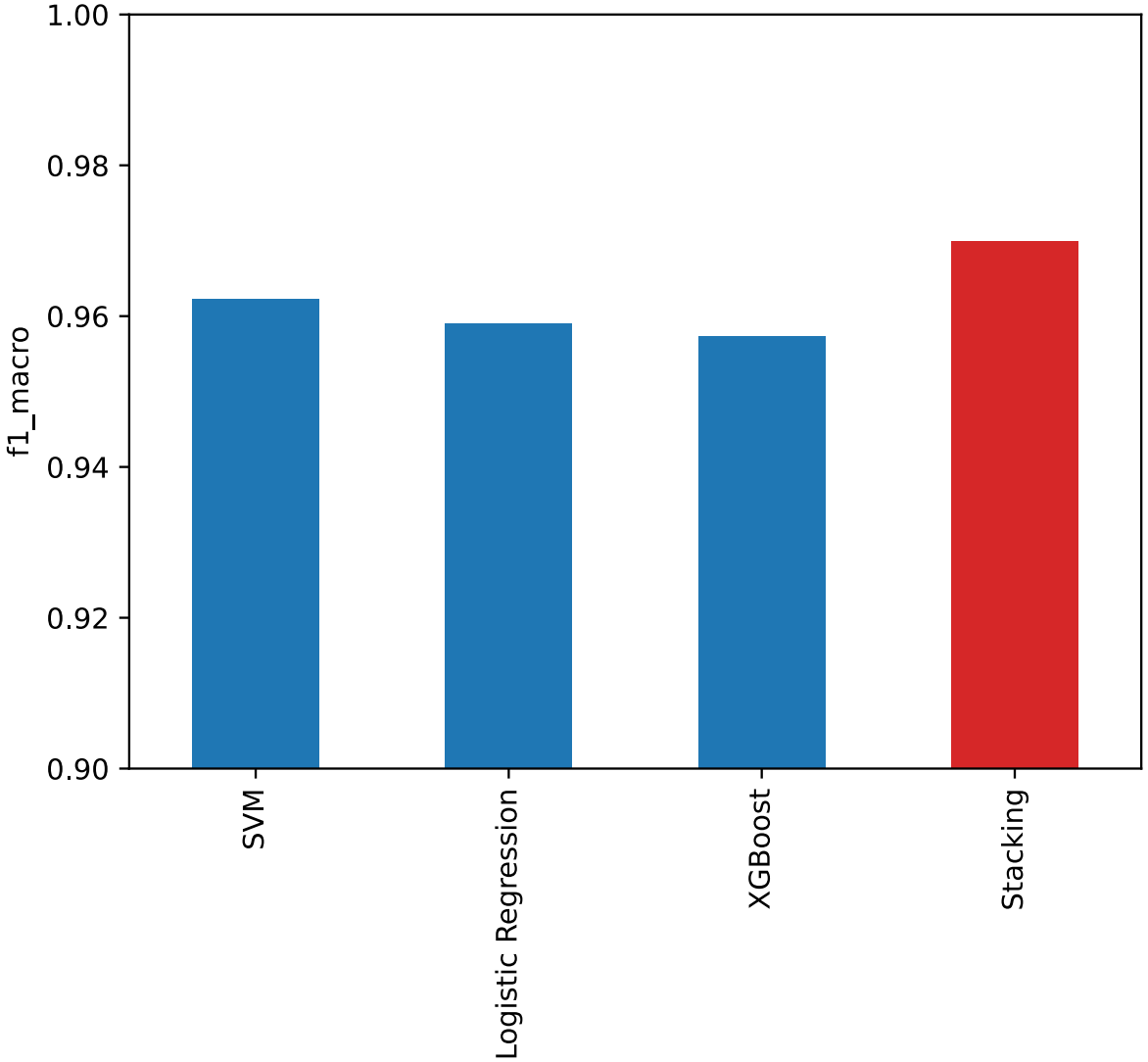


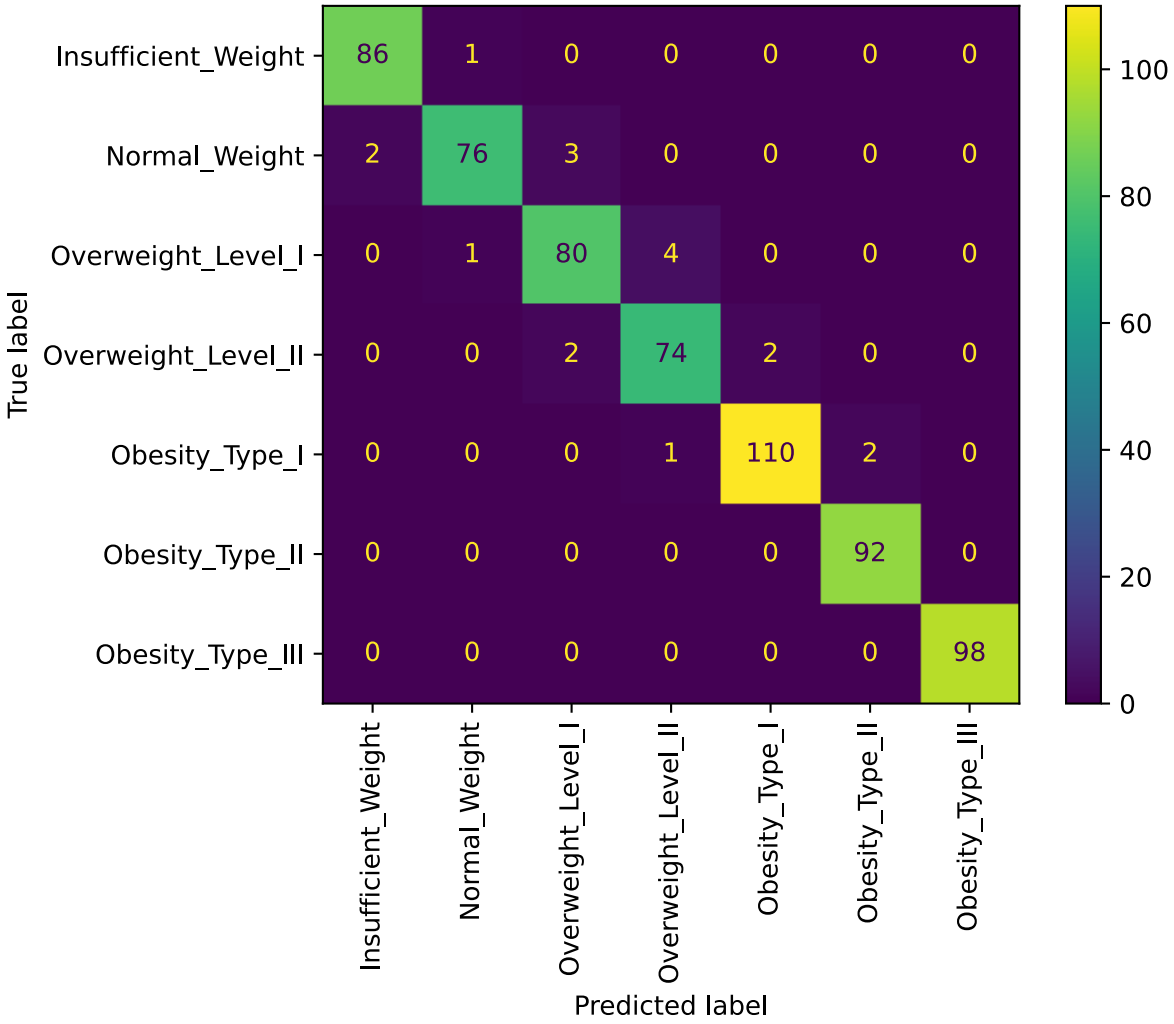

1e11
3.0
2.5
2.0
1.5
1.0
0.5
0.0
time (ns)
Naive Bayes
K-Neighbours
SVM
Logistic Regression
Random Forest
XGBoost

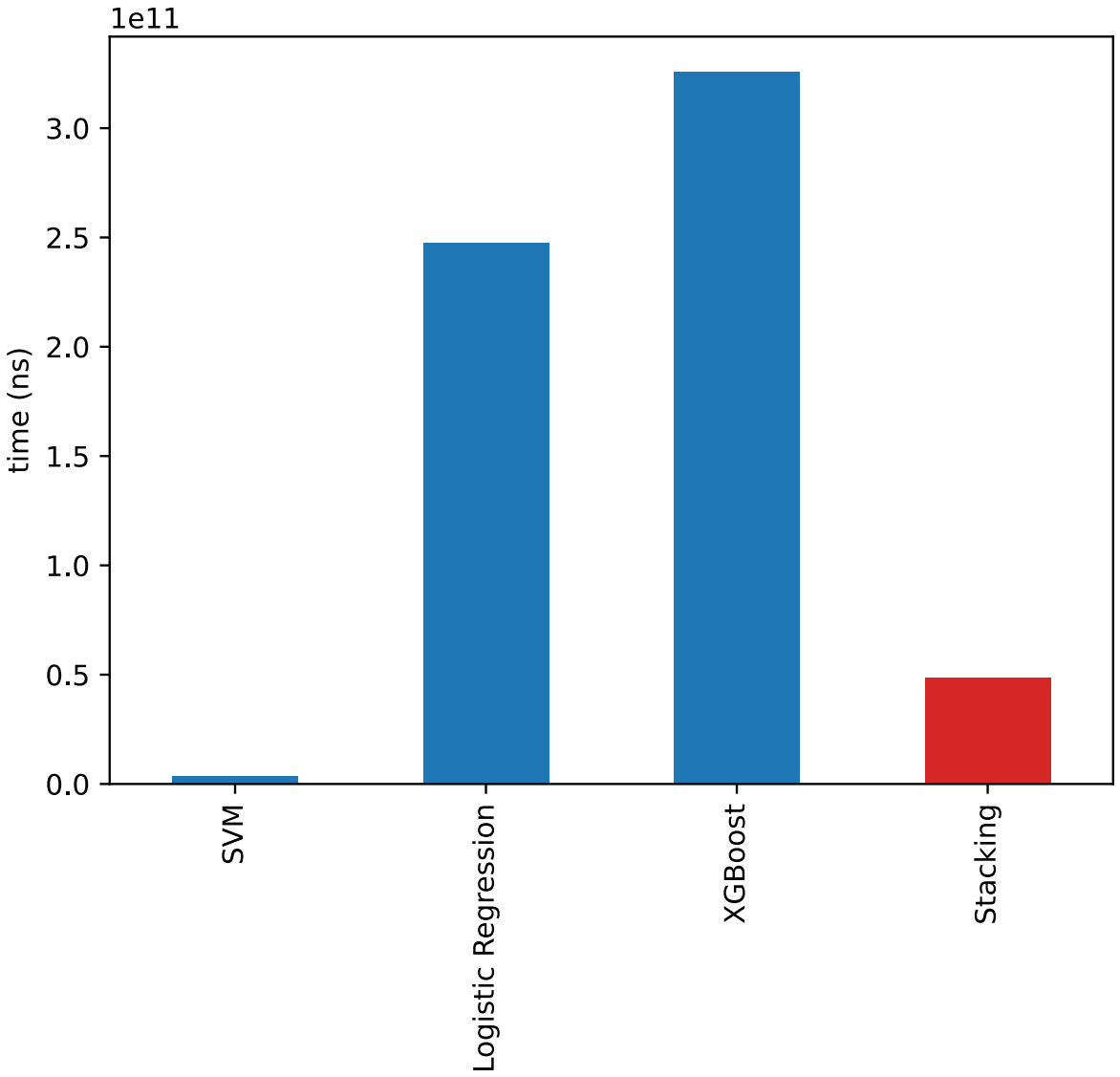
Stacking

- *SVM + Logistic Regression + XGBoost*
- Usando mejores hiperparámetros de búsqueda anterior
- Estimador final: *Logistic Regression*

Modelo elegido.


1.00
0.98
0.96
0.94
0.92
0.90
f1_macro
SVM
Logistic Regression
XGBoost
Stacking


Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
True label
86 1 0 0 0 0 0
2 76 3 0 0 0 0
0 1 80 4 0 0 0
0 0 2 74 2 0 0
0 0 0 1 110 2 0
0 0 0 0 0 92 0
0 0 0 0 0 0 98
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Predicted label
100
80
60
40
20
0


1e11
3.0
2.5
2.0
1.5
1.0
0.5
0.0
time (ns)
SVM
Logistic Regression
XGBoost
Stacking

Gracias por su atención