





 $\times \times \times$

 $\times \times \times$



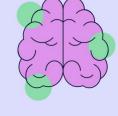


- Trastorno de la circulación cerebral que altera la función de una determinada región del cerebro
- 2^α causa de muerte en España, 1^ο en mujeres
- 120.000 casos al año













Proyecto de Machine Learning

**

Objetivo

 Predecir el riesgo de una persona de sufrir un ictus y prevenirlo

Dataset

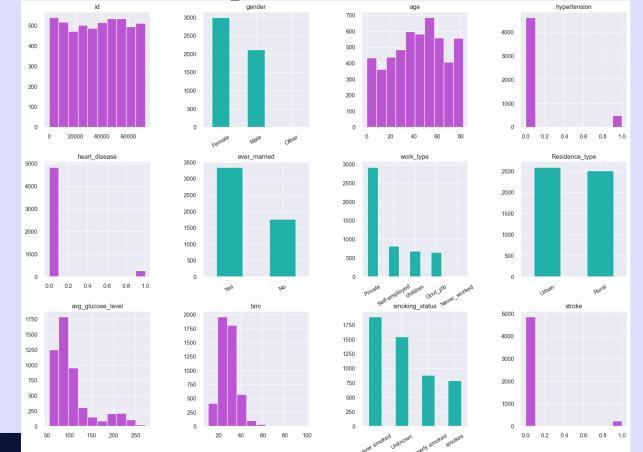
- Datos personales: edad, gérenero...
- Datos clínicos: hipertensión, niveles de glucosa, si fuma...

Tipo de problema

Clasificación binario



Análisis exploratorio de datos

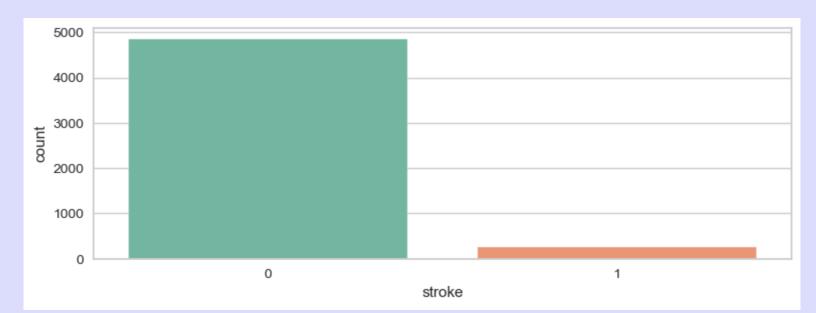






Análisis exploratorio de datos

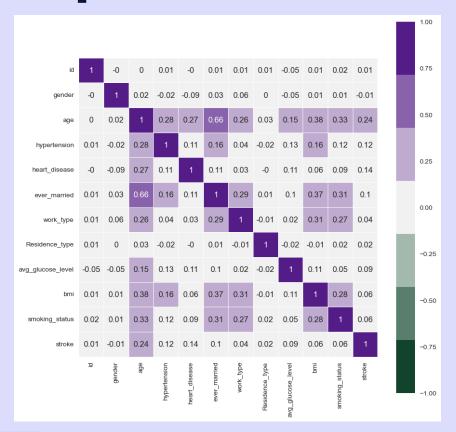
Target







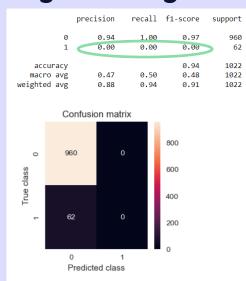
Análisis exploratorio de datos



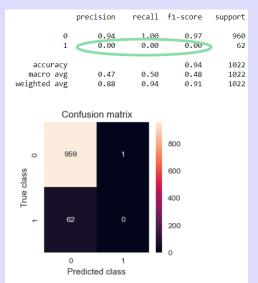




Regresión logística



Random Forest

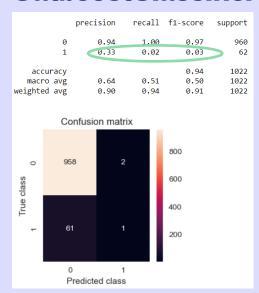




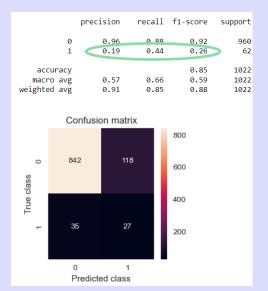




CathoostClassifier



XGBoostClassifier









Técnicas de balanceo de clases

Ol Utilizar parámetros, como weight='balanced'

*

Regresión logística

				_	
	precis	ion	recall	f1-score	support
		.98 .16	0.73 0.76	0.84 0.26	960 62
accu macro weighted	avg 0	.57 .93	0.75 0.73	0.73 0.55 0.80	1022 1022 1022
	Confusio	on matri	ix	- 700	
dass 0	704	256	6	- 600 - 500 - 400	
True class	15	47		- 300 - 200 - 100	
	0 Predicte	1 ed class	3	_	

Random Forest

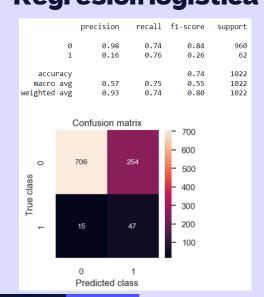
	preci	sion	recall	f1-score	support	
		0.94	1.00	0.97	960	
	1 (0.00	0.00	0.00	62	
accu	racy			0.94	1022	
macro	avg (2.47	0.50	0.48	1022	
weighted		3.88	0.94	0.91	1022	
U	0					
	Confusi	on mo	strise			
	Cornusi	OH HIS	aurix			
				000		
				- 800		
0	959		1			
Ø				- 600		
as				000		
ਹ						
True class				- 400		
F						
	62		0			
-	02		U	- 200		
				- 0		
	0		1			
	Predict	od cla	100			
	rieulo	ou olo	100			



Técnicas de balanceo de clases

02 Oversampling

Regresión logística



Random Forest

	pre	cision	recall	f1-score	support	
	0	0.95	0.95	0.95	960	
	1	0.16	0.15	0.15	62	
2001	nacu			0.90	1022	
accu						
macro		0.55	0.55	0.55	1022	
weighted	avg	0.90	0.90	0.90	1022	
	Confi	ısion ma	atriv			
	COINC	1310111116	u i x			
				- 800		
0	911		49			
10				- 600		
SS						
60						
Φ				- 400		
True class				400		
-						
←	53		9	- 000		
				- 200		
				_		
	0		1			
	Prod	icted cla	221			
	1 100	lotod old				





Técnicas de balanceo de clases

03 Undersampling

Regresión logística

	precis	ion	recall	f1-score	support	
	0 6	.98	0.74	0.84	960	
	1 6	.16	0.76	0.26	62	
accu	nacy			0.74	1022	
macro		.57	0.75	0.55	1022	
weighted		9.93	0.74	0.80	1022	
weighted	uv _B		0174	0.00	1022	
0	Confusion 706		trix 54	- 700 - 600		
(0				- 500		
class				- 400		
True class				- 300		
_	15	4	17	- 200		
				- 100		
	0		1	-		
	Predict	ed clas	22			
r redicted class						

Random Forest

	prec	ision	recall	f1-score	support	
	0	0.95	0.95	0.95	960	
	1	0.16	0.15	0.15	62	
26611	nacu			0.90	1022	
accu						
macro		0.55	0.55	0.55	1022	
weighted	avg	0.90	0.90	0.90	1022	
	Confus	sion ma	atrix			
				- 800		
_	911		49			
0	311		45			
S				- 600		
as						
True class						
P				- 400		
ᆮ						
~	53		9	- 200		
				200		
	0		1			
Predicted class						



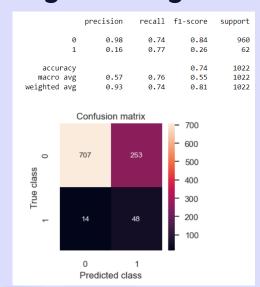


Técnicas de balanceo de clases

O4 Combinar Undersampling y OverSampling (SMOTETomek)



Regresión logística



Random Forest

	preci	sion	recall	f1-score	support	
	0	0.95	0.95	0.95	960	
	1	0.17	0.18	0.18	62	
accu	racy			0.90	1022	
macro		0.56	0.56	0.56	1022	
weighted	avg	0.90	0.90	0.90	1022	
	Confus	ion ma	atrix			
				- 800		
				000		
0	908		52			
10				- 600		
38				000		
$\frac{6}{5}$						
e				- 400		
True class						
	51		11			
←	51		"	- 200		
				_		
	0		1			
	Predic	ted cla	ass			



Técnicas de balanceo de clases

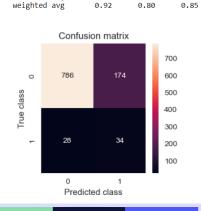
05

Ensamble de modelos con balanceo





support	f1-score	recall	precision		
960 62	0.89 0.25	0.82 0.55	0.97 0.16	0 1	
1022 1022	0.80 0.57	0.68	0.56	accuracy macro ave	



BalancedBaggingClassifier BalancedRandomForestClassifier

	preci	ision	reca	11	f1-score	support
	0 1	0.98 0.16		72 81	0.83 0.26	960 62
	1	0.10	٥.	01	0.20	02
	uracy				0.73	1022
macr weighte	o avg	0.57 0.93		76 73	0.55 0.80	1022 1022
weighte	u avg	0.93	٥.	/3	0.00	1022
	Confusio	on mate	riv			
	Cornusio	Jii iiiali	IX.			
					600	
0	692	26	8		500	
class					400	
True class					300	
	12	5	0		200	
					100	
	0	1				
Predicted class						

EasyEsembleClassifier

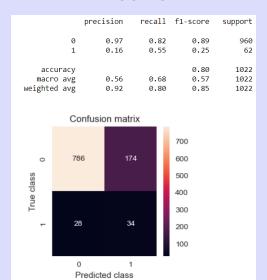
	preci	ision	recal	11	f1-score	support
	0	0.99	0.6	56	0.79	960
	1	0.14	0.8	35	0.24	62
aco	uracy				0.68	1022
macr	ro avg	0.56	0.7	76	0.52	1022
weighte	ed avg	0.93	0.6	58	0.76	1022
	Confusio	on matr	ix			
					600	
0	637	32:	3		500	
ass					400	
True class					300	
- t	9	53	;		200	
					100	
	0	1				
	Predicte	ed class				

Técnicas de balanceo de clases

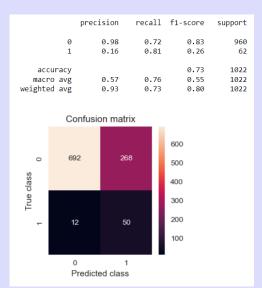
Ensamble de modelos con balanceo 05



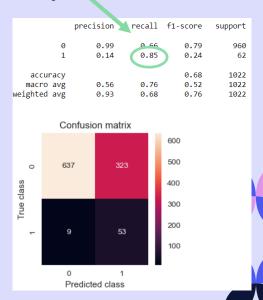




BalancedBaggingClassifier BalancedRandomForestClassifier



EasyExembleClassifier

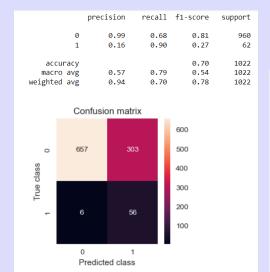


Mejor modelo y parámetros

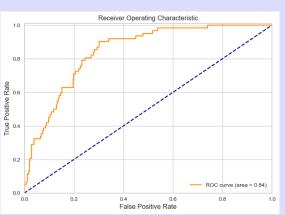
EasyEsembleClassifier

Con GridSearch y validación cruzada y los parámetros:

- base_estimator=CatboostClassifier (verbose=False, n_estimators=30)
- sampling_strategy=not minority
- replacement= True







Mejoras del modelo





Qué hacer



- Actualizar periodicamente los datos de los pacientes que hay en la base de datos
- Ampliar la información de la base de datos con otras bases de datos de otros hospitales
- Ampliar la información recogida con nuevas variables: como si hacen ejercicio físico, medicación, problemas de coagulación, etc....



 $\times \times \times$

