Informe Regresión Logística

Santiago Arenas y Miguel Sánchez-Beato

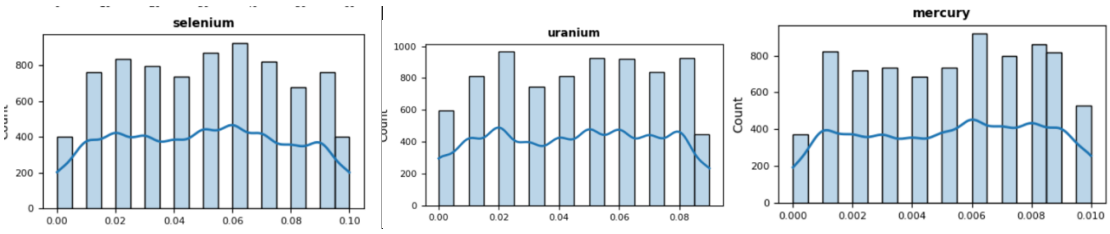
1. Preprocesado de datos

Se comprueba que todas las columnas del dataset tienen nombre, y que no existen valores nulos.

```
aluminium      0.0
ammonia        0.0
arsenic        0.0
barium         0.0
cadmium        0.0
chloramine     0.0
chromium       0.0
copper         0.0
flouride       0.0
bacteria       0.0
viruses        0.0
lead           0.0
nitrates       0.0
nitrites       0.0
mercury        0.0
perchlorate    0.0
radium         0.0
selenium       0.0
silver         0.0
uranium        0.0
is_safe        0.0
dtype: float64
```

2. Interpretación de distribuciones de las variables

Se observa que selenium, uranium y mercury tienen distribuciones muy similares, por lo que es posible que sean colineales.



3. Creación de variables dummies

No es necesario crear variables dummies en este caso, ya que todas las variables son cuantitativas.

4. Dividir el dataset en train y test

Hemos asignado un 70% de los datos al conjunto de train, y el 30% restante al de test.

5. Crear un primer modelo

Viendo los intervalos de confianza, se determina que las variables flouride y lead no explican la variable independiente, ya que éste pasa por 0.

Como LL-Null es significativamente menor que Log-Likelihood, podemos afirmar que como mínimo una de las variables independientes explica la variable dependiente.

Model:		Logit		Method:		MLE
Dependent Variable:		is_safe		Pseudo R-squared:		0.300
Date:		2023-10-04 17:38		AIC:		2841.0138
No. Observations:		5597		BIC:		2973.6135
Df Model:		19		Log-Likelihood:		-1400.5
Df Residuals:		5577		LL-Null:		-2000.0
Converged:		1.0000		LLR p-value:		1.5382e-242
No. Iterations:		8.0000		Scale:		1.0000
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	0.1687	0.2601	0.6484	0.5167	-0.3412	0.6785
aluminium	0.7132	0.0384	18.5683	0.0000	0.6380	0.7885
ammonia	-0.0224	0.0058	-3.8525	0.0001	-0.0338	-0.0110
arsenic	-3.1097	0.3716	-8.3679	0.0000	-3.8380	-2.3813
barium	0.1318	0.0473	2.7848	0.0054	0.0390	0.2246
cadmium	-19.8536	2.0837	-9.5281	0.0000	-23.9375	-15.7696
chloramine	0.1889	0.0243	7.7693	0.0000	0.1413	0.2366
chromium	1.1519	0.2131	5.4044	0.0000	0.7341	1.5696
copper	-0.3061	0.0838	-3.6541	0.0003	-0.4703	-0.1419
flouride	0.1230	0.1148	1.0714	0.2840	-0.1020	0.3480
bacteria	0.7627	0.2564	2.9745	0.0029	0.2601	1.2653
viruses	-1.3038	0.2199	-5.9285	0.0000	-1.7348	-0.8727
lead	-1.4550	0.8850	-1.6441	0.1002	-3.1896	0.2796
nitrates	-0.0423	0.0091	-4.6642	0.0000	-0.0601	-0.0245
nitrites	-0.3231	0.1152	-2.8041	0.0050	-0.5489	-0.0973
mercury	-52.6317	16.6924	-3.1530	0.0016	-85.3482	-19.9152
perchlorate	-0.0266	0.0037	-7.2205	0.0000	-0.0338	-0.0194
radium	-0.0480	0.0234	-2.0496	0.0404	-0.0938	-0.0021
selenium	-6.8061	1.7532	-3.8822	0.0001	-10.2422	-3.3700
silver	-1.6057	0.4147	-3.8719	0.0001	-2.4185	-0.7929

6. Quitar las variables flouride y lead y repetir el modelo

Como se puede observar, en el segundo modelo, el AIC es ligeramente más bajo, mientras que el Pseudo-R^2 ajustado es el mismo, lo cual tiene sentido, ya que hemos quitado 2 variables que apenas explicaban la variable dependiente is_safe.

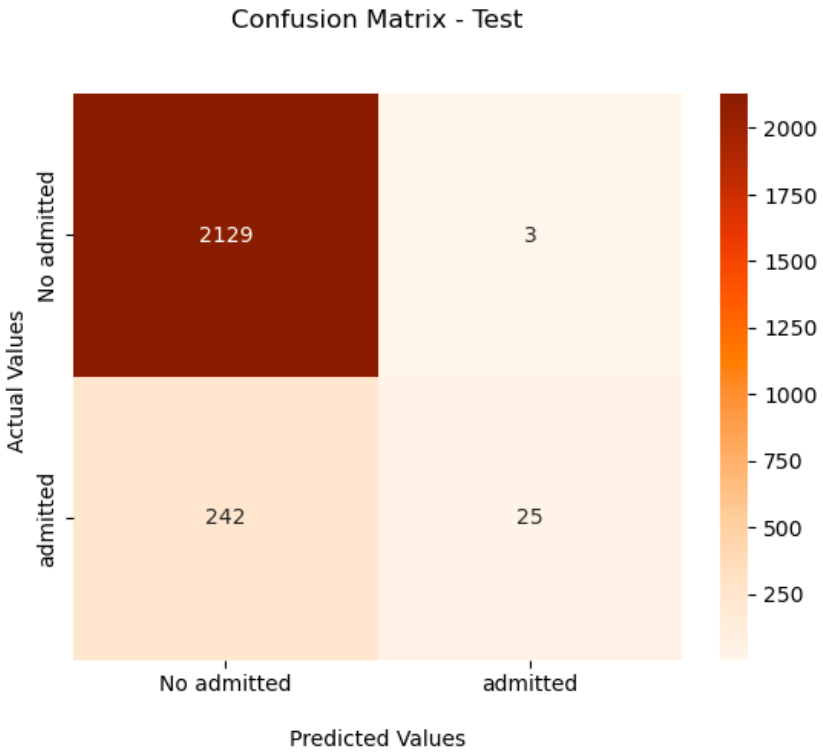
Model:		Logit		Method:		MLE
Dependent Variable:		is_safe		Pseudo R-squared:		0.288
Date:		2023-10-04 17:38		AIC:		2885.9006
No. Observations:		5597		BIC:		3005.2404
Df Model:		17		Log-Likelihood:		-1425.0
Df Residuals:		5579		LL-Null:		-2000.0
Converged:		1.0000		LLR p-value:		6.5793e-234
No. Iterations:		8.0000		Scale:		1.0000
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-0.1414	0.2549	-0.5546	0.5792	-0.6409	0.3582
aluminium	0.6735	0.0373	18.0368	0.0000	0.6003	0.7467
ammonia	-0.0250	0.0058	-4.3386	0.0000	-0.0362	-0.0137
arsenic	-3.1804	0.3721	-8.5464	0.0000	-3.9097	-2.4510
barium	0.1285	0.0470	2.7356	0.0062	0.0364	0.2206
cadmium	-20.5015	2.0784	-9.8641	0.0000	-24.5750	-16.4279
chloramine	0.1849	0.0240	7.6908	0.0000	0.1378	0.2320
chromium	1.1255	0.2108	5.3402	0.0000	0.7124	1.5386
nitrates	-0.0408	0.0090	-4.5377	0.0000	-0.0584	-0.0232
nitrites	-0.1760	0.1105	-1.5924	0.1113	-0.3927	0.0406
mercury	-58.4857	16.5763	-3.5283	0.0004	-90.9746	-25.9968
perchlorate	-0.0260	0.0036	-7.1287	0.0000	-0.0332	-0.0189
radium	-0.0384	0.0232	-1.6538	0.0982	-0.0839	0.0071
selenium	-6.3376	1.7311	-3.6610	0.0003	-9.7306	-2.9447
silver	-1.5800	0.4114	-3.8404	0.0001	-2.3863	-0.7736

Como se puede observar, el mercurio es la variable que más contribuye a explicar si una muestra de agua es segura o no, ya que su coeficiente es mucho mayor que los demás.

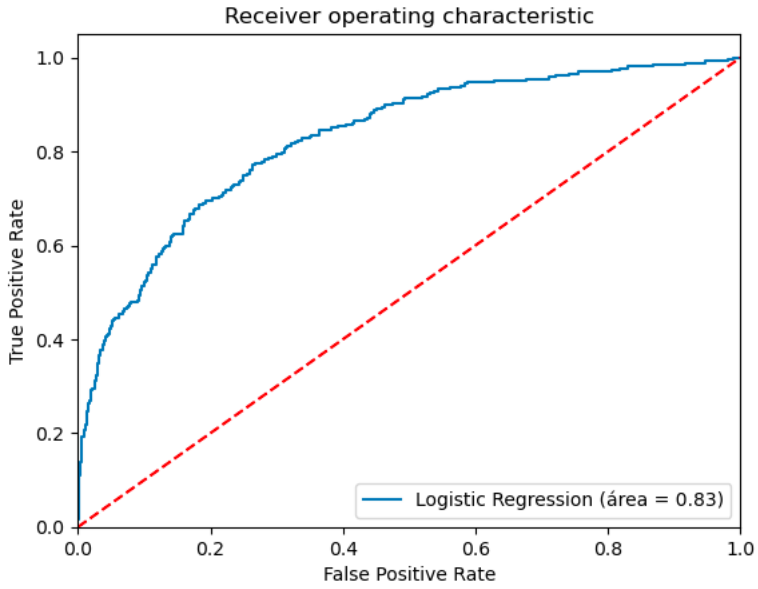
7. Matriz de confusión

En vista de la matriz de confusión del set de test:

- Precisión = 0.90
- Sensibilidad = 0.924
- Especificidad = 0.032

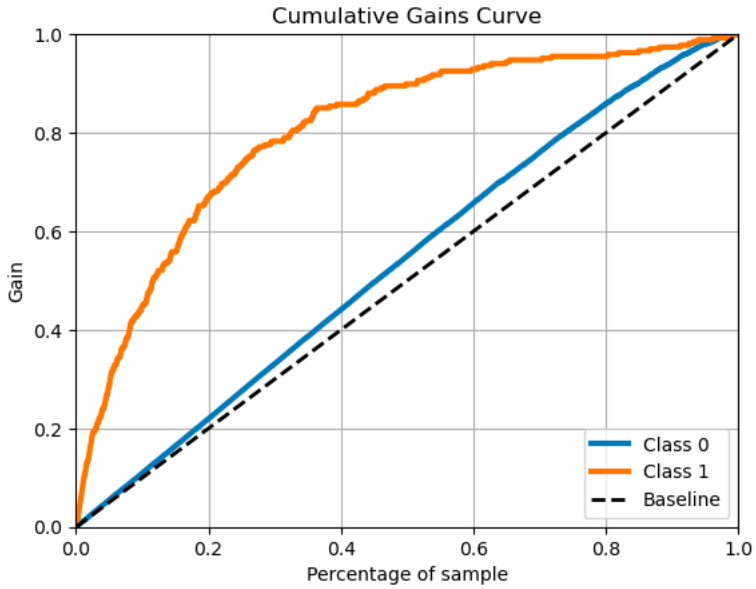


8. Curva ROC



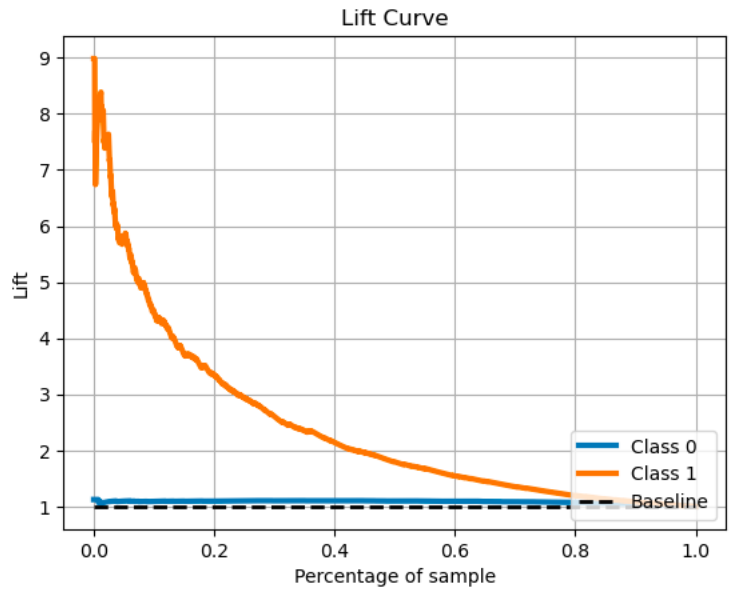
Con estos datos se puede una curva ROC, donde siempre se quiere estar por encima de la diagonal (asignado aleatorio). El área debajo de la curva, cuanto más alto mejor. Como en este caso el AUC es muy alto, se puede determinar que es un buen modelo.

9. Curva de ganancia acumulada



Proporción acumulativa de eventos positivos en función de los datos clasificados como positivos. Cuanto más cerca de la diagonal, mejor. Sirve para medir la efectividad del modelo ante eventos relevantes.

10. Curva Lift



La curva de lift compara el modelo frente a uno aleatorio, a mayor curva, más eventos positivos reconoce frente a la suposición aleatoria y mejor el modelo, por tanto.