Detección de estímulos en tiempo real

Fundamentos de Aprendizaje Automático

Universidad Autónoma de Madrid

Alberto Altozano Fernández
Antonio Estebanez Yepes
Juan Antonio Martos Navarro
Marcos Martínez Jiménez

Índice

Introducción

- Objeto de estudio
- Objetivos
- Optimización

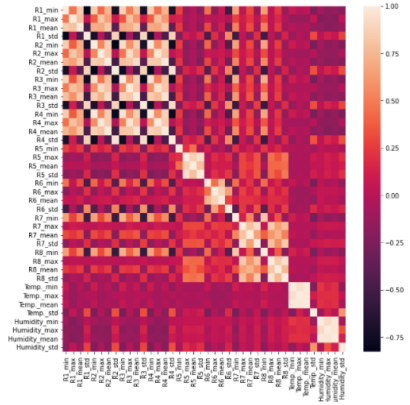
	id	time	R1	R2	R3	R4	R5	R6	R7	R8	Temp.	Humidity
0	0	-0.999750	0.779562	0.904831	0.853984	0.899352	-0.090317	-0.820466	0.917566	0.855313	-1.169299	0.307931
1	0	-0.999472	0.779101	0.904767	0.853527	0.899235	-0.090338	-0.820799	0.917442	0.855334	-1.163661	0.303182
2	0	-0.999194	0.777949	0.905024	0.853241	0.899177	-0.090355	-0.821011	0.917331	0.855355	-1.157359	0.298910
3	0	-0.998916	0.777373	0.905024	0.853241	0.899235	-0.090371	-0.821223	0.917279	0.855419	-1.151721	0.295011
4	0	-0.998627	0.776567	0.905152	0.853469	0.899352	-0.090382	-0.821465	0.917283	0.855570	-1.146636	0.291506
...	...	...	...	...	...	...	...	...	...	...	...	...
928986	99	1.675182	0.503447	1.041014	1.028350	1.195868	-0.136169	0.375476	0.281690	0.347078	0.491773	-0.708208
928987	99	1.675460	0.504829	1.040437	1.028579	1.196160	-0.136098	0.375234	0.281673	0.347121	0.495090	-0.710468
928988	99	1.675738	0.505751	1.040372	1.028636	1.196277	-0.136028	0.374901	0.281697	0.347081	0.498075	-0.712521
928989	99	1.676016	0.505635	1.040501	1.028807	1.196394	-0.135984	0.374810	0.281756	0.347087	0.500839	-0.714367
928990	99	1.676304	0.506211	1.040116	1.028807	1.196394	-0.135946	0.374537	0.281732	0.347012	0.503160	-0.716047

623569 rows × 12 columns

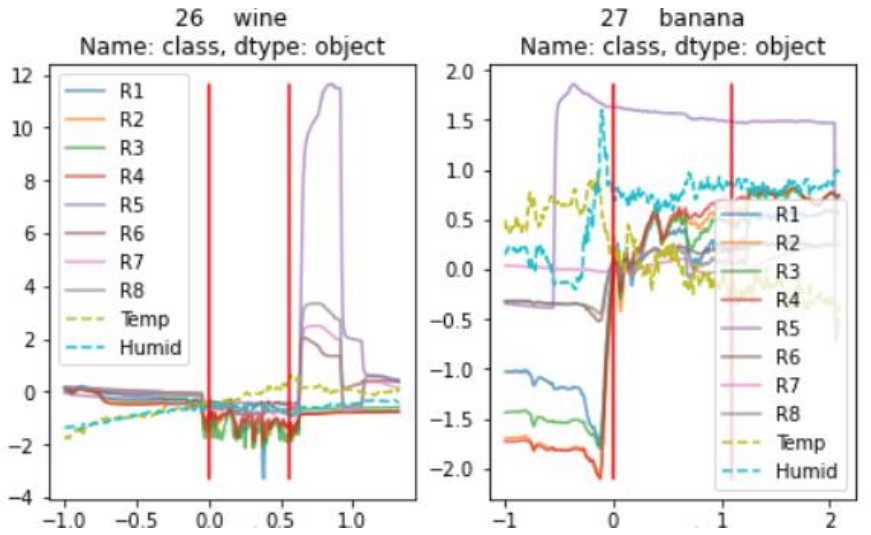
Análisis exploratorio de los datos

- Inspección de *Missing values*

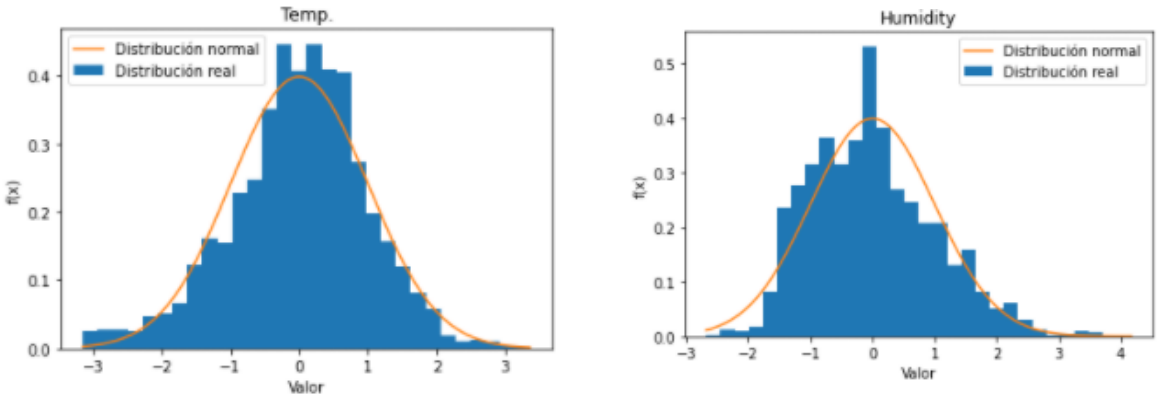
- Correlaciones entre atributos



- Detección de *outliers*

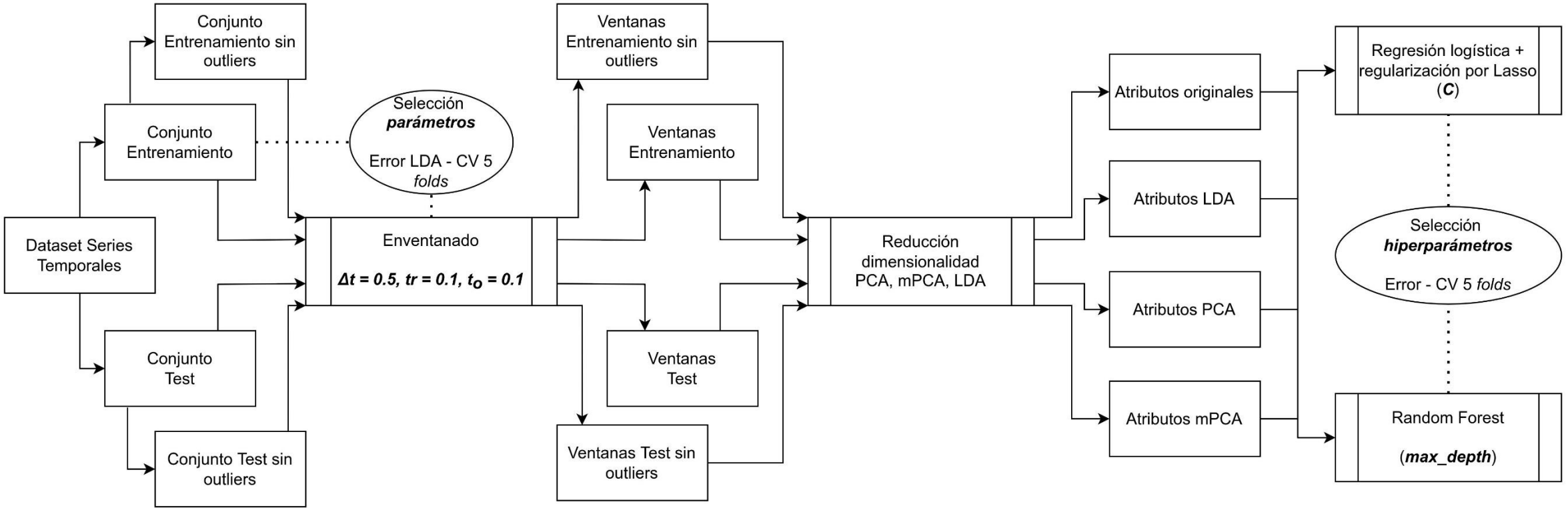


- Distribuciones de los atributos



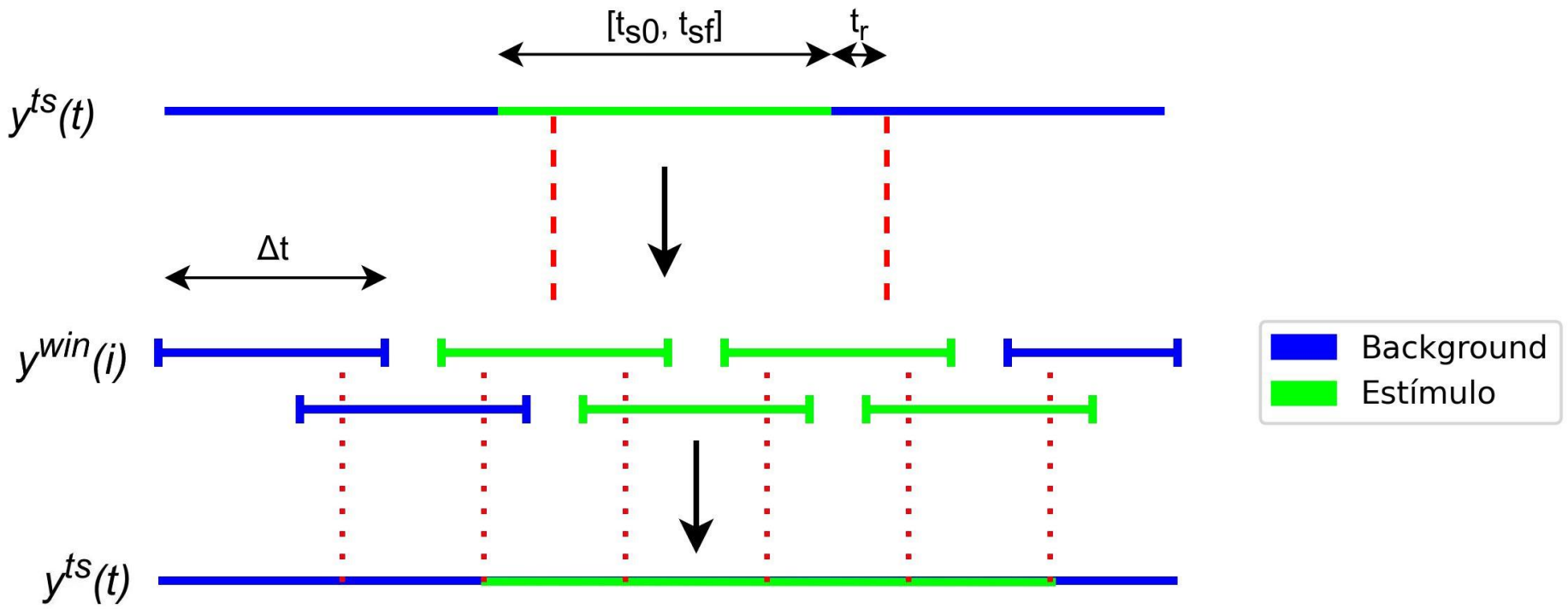
Metodología

Esquema del análisis



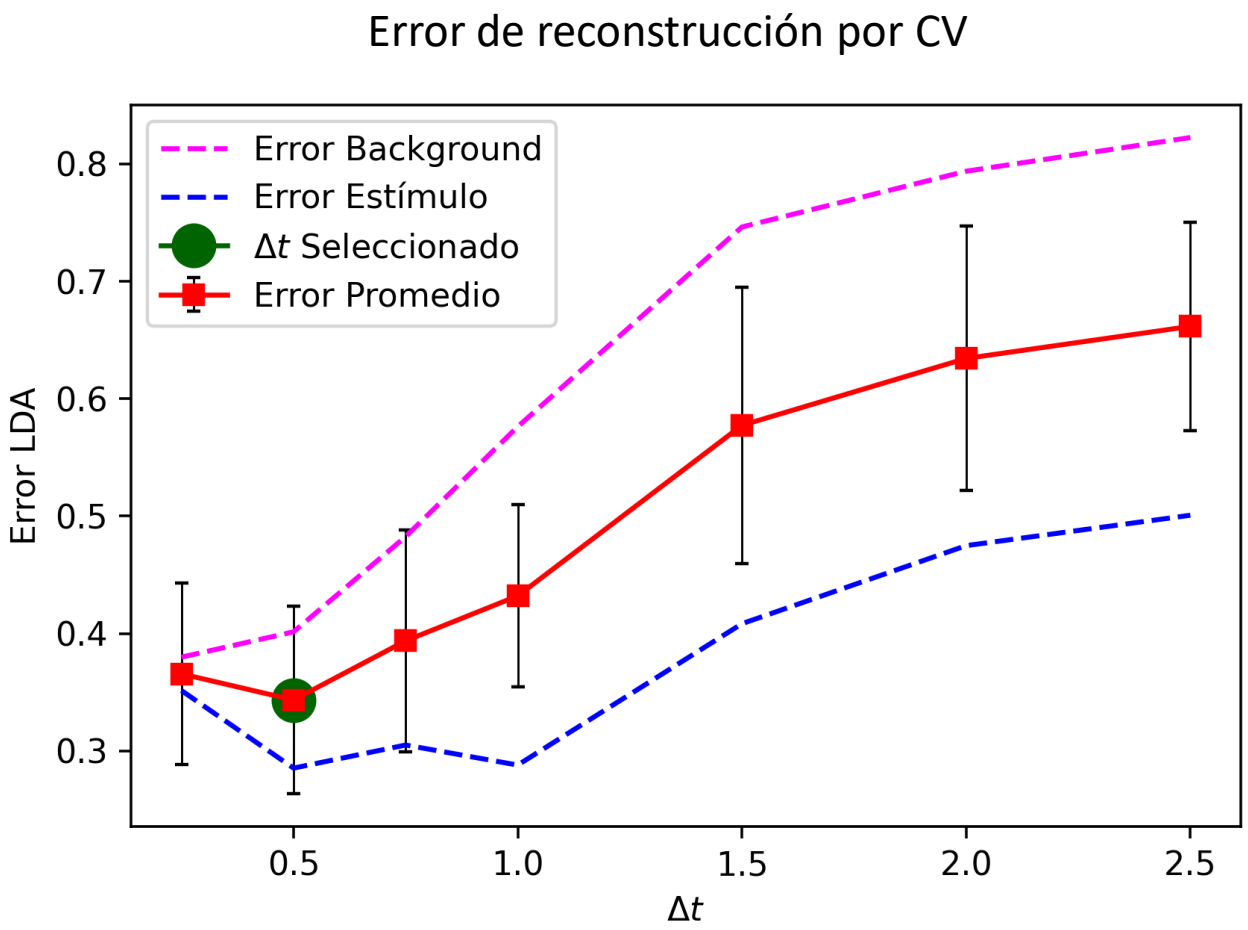
Metodología

Enventanado y reconstrucción



Resultados

Selección parámetros de enventanado



↑↑↑Δt:

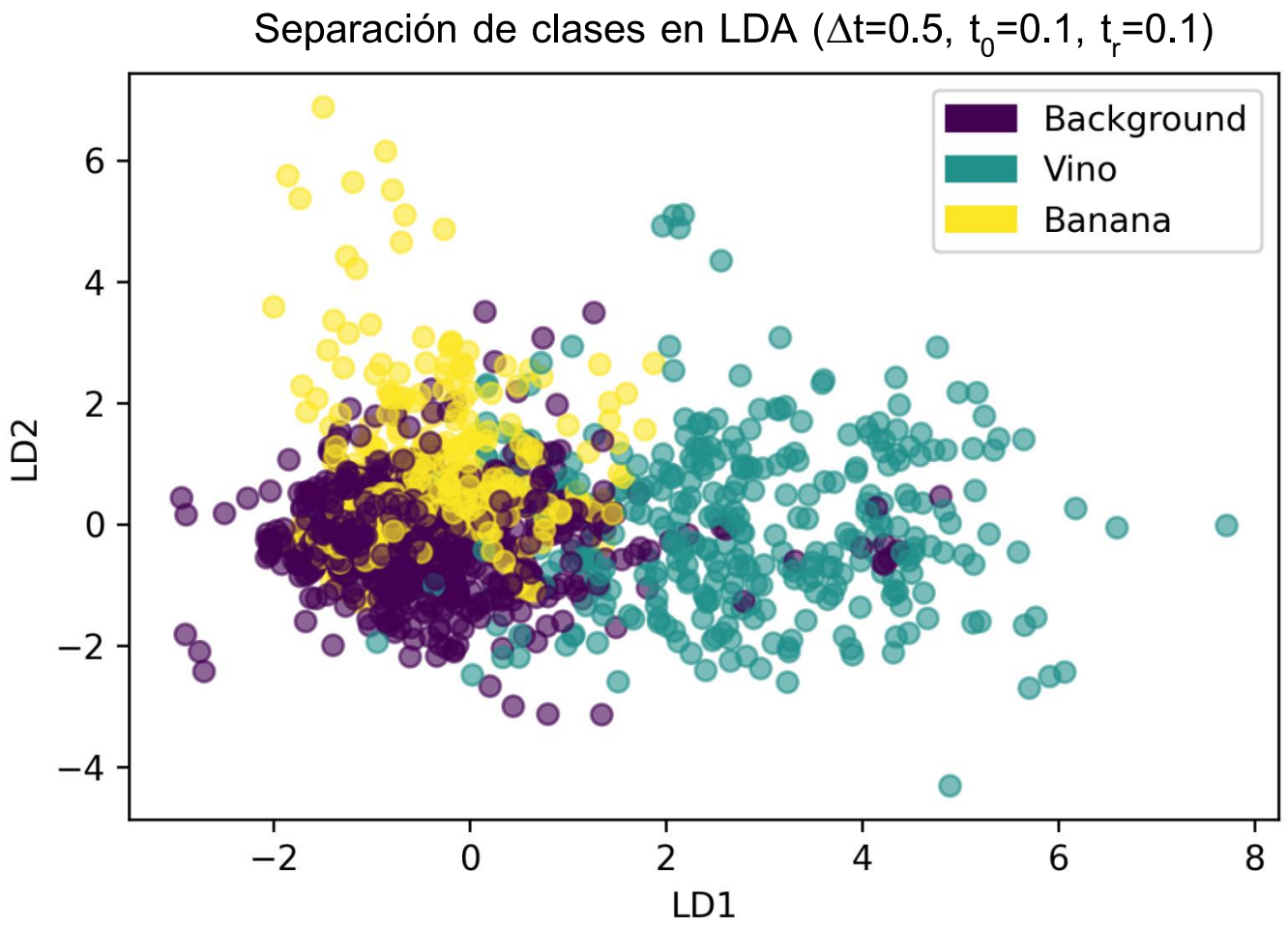
- ↑Separación clases
- ↑Error de asignación
- ↓Resolución

↓↓↓Δt:

- ↓Separación clases
- ↓Error de asignación
- ↑Resolución

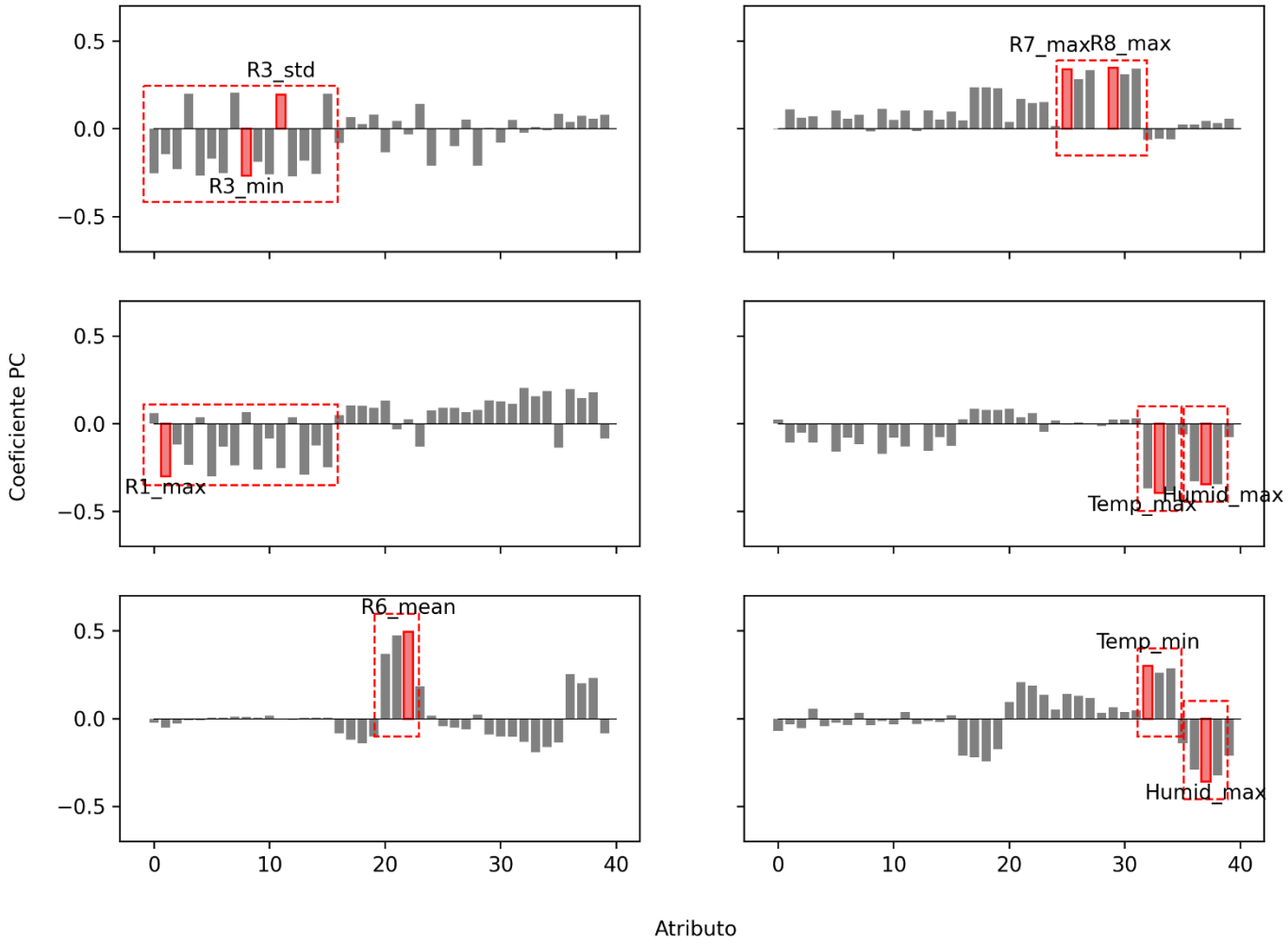
Resultados

Reducción de la dimensionalidad: LDA



Resultados

Reducción de la dimensionalidad: PCA y mPCA



PCA

6 componentes → 80% varianza

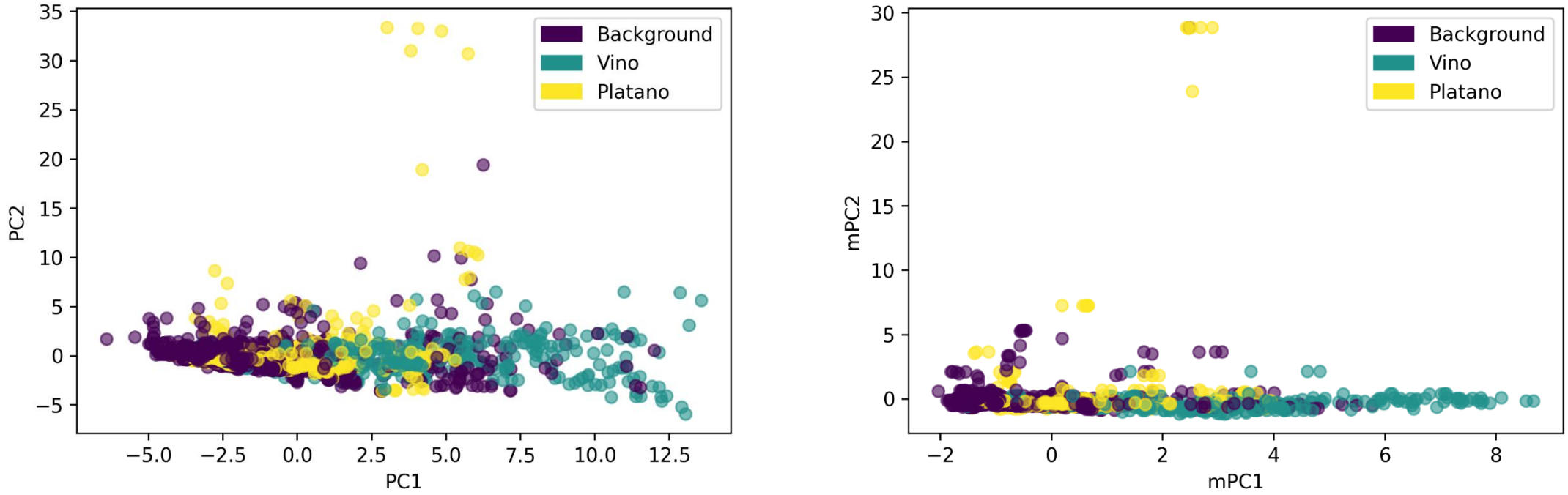
mPCA

- mPC1: R3_std – R3_min
- mPC2: R7_max + R8_max
- mPC3: R1_max
- mPC4: – Temp_max – Humid_max
- mPC5: R6_mean
- mPC6: Temp_min – Humid_max

Resultados

Reducción de la dimensionalidad: PCA y mPCA

Separación de clases en PCA y mPCA



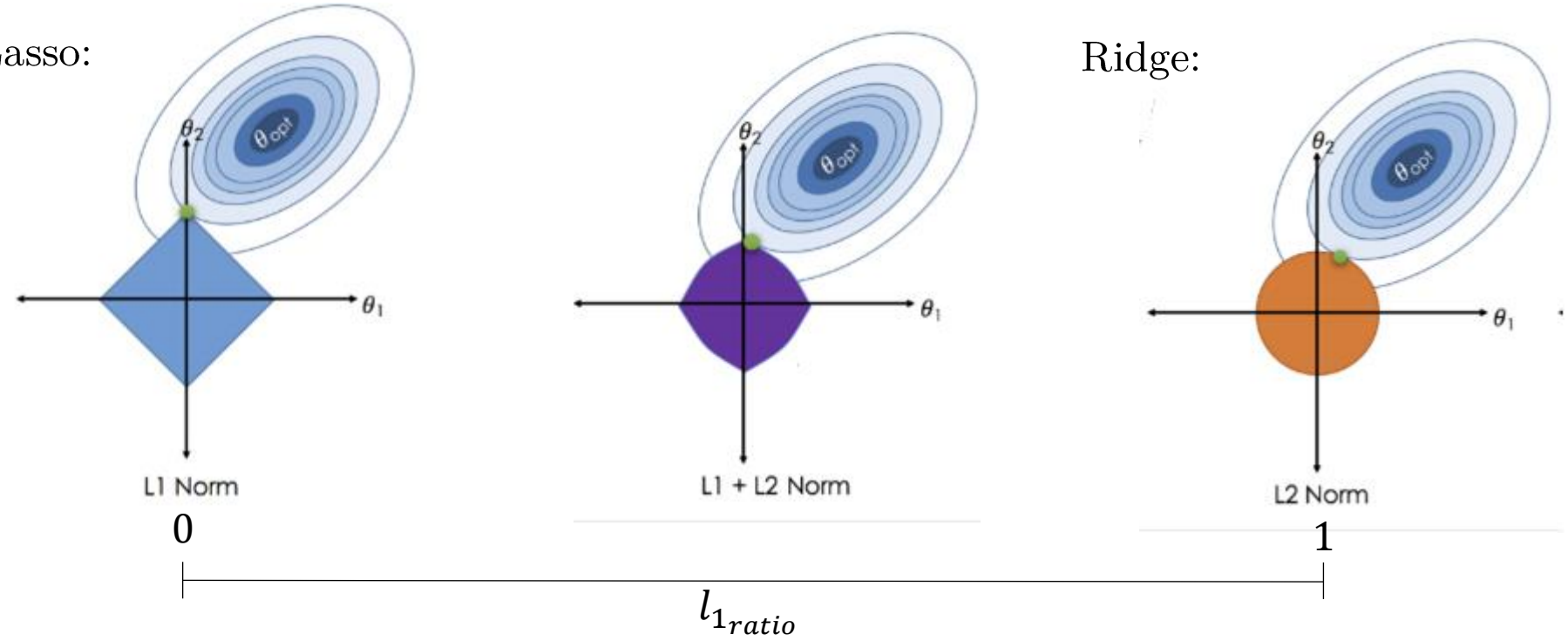
Resultados

Regresión logística: Tipo de Regularización

`sklearn.linear_model.LogisticRegression`

- $C \in (0, \infty)$ – Inversa de la fuerza de regularización
- $0 \leq l_{1_{ratio}} \leq 1$ – Ratio entre regularización lasso o ridge

Lasso:

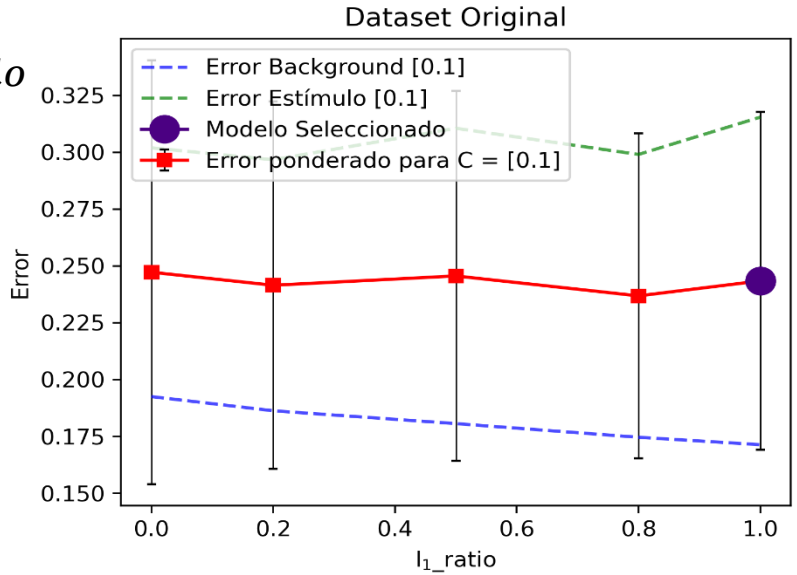


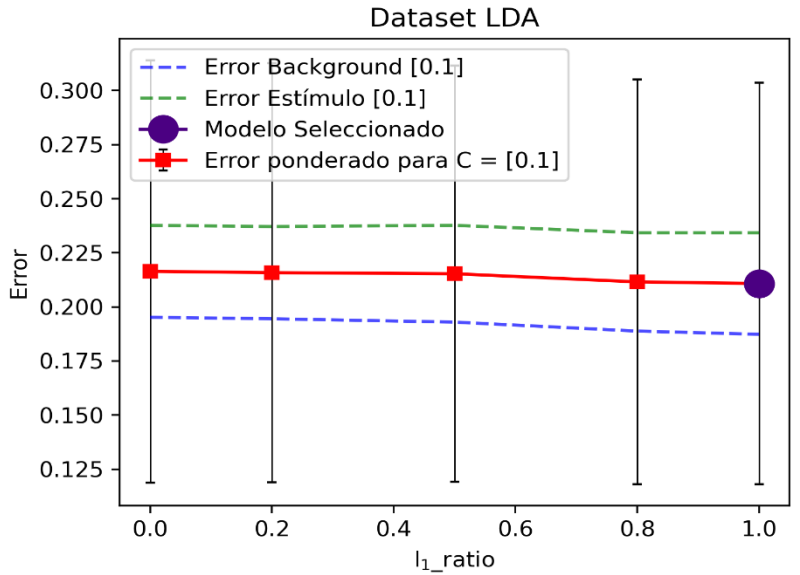
Resultados

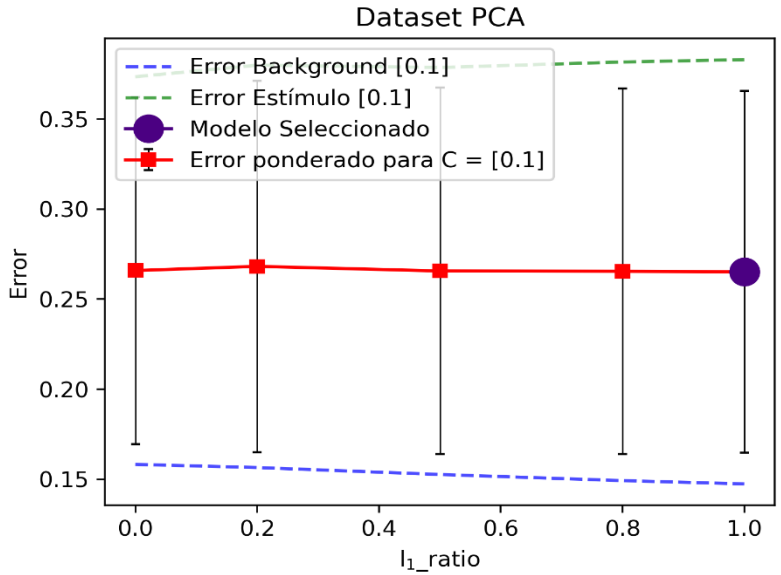
Regresión logística: $l_{1_{ratio}}$

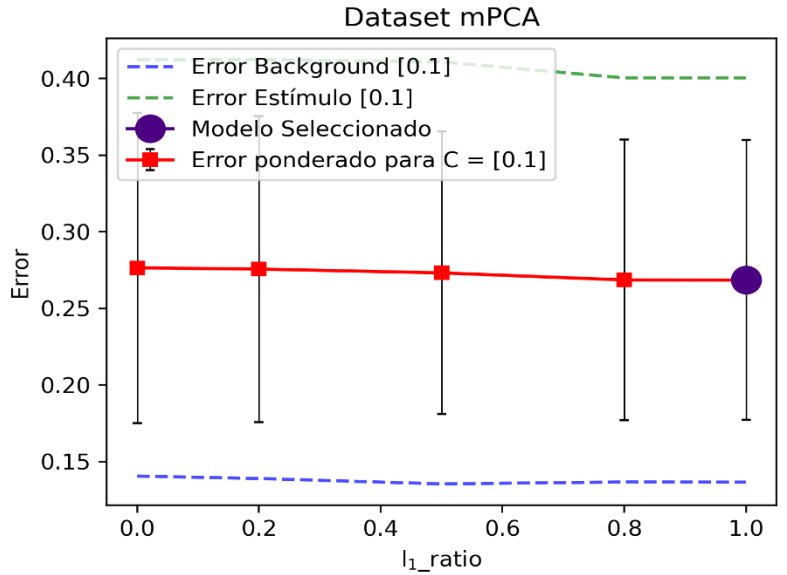
Lasso

$$l_{1_{ratio}} = 1$$







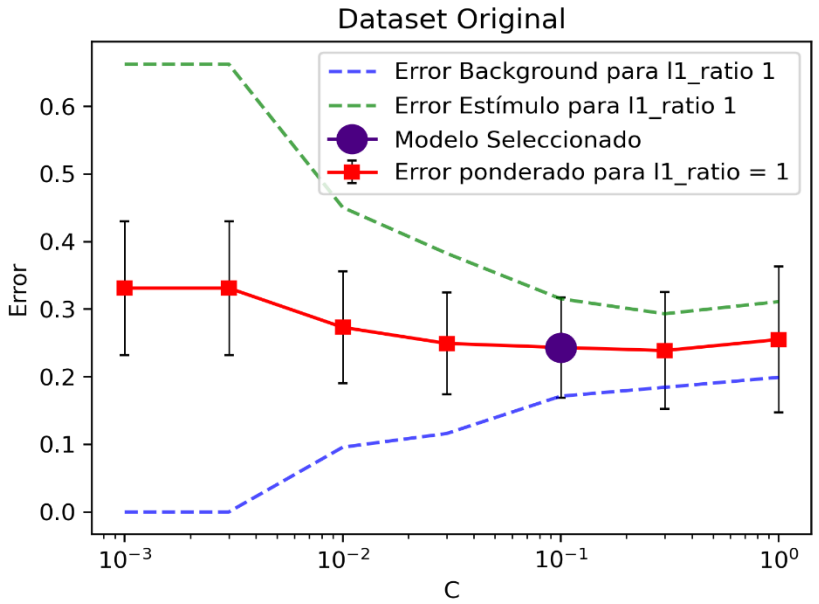


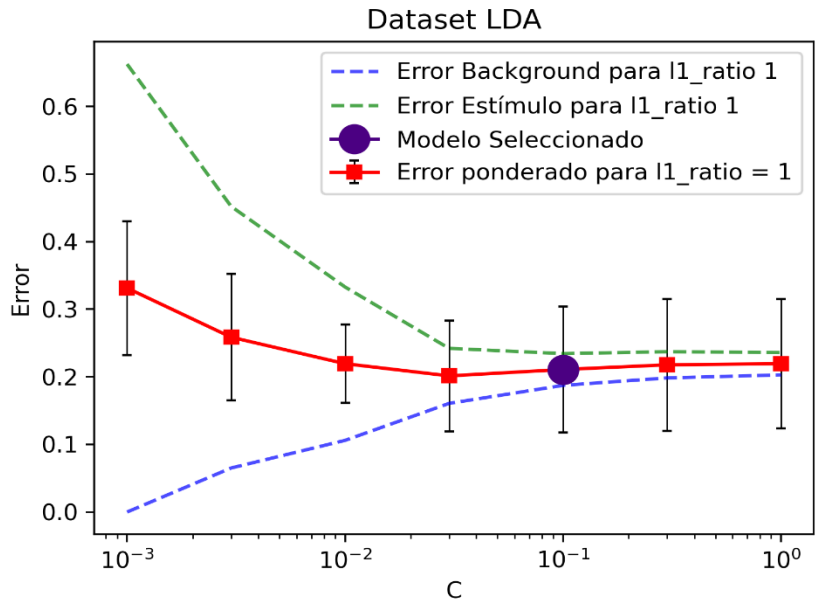
Resultados

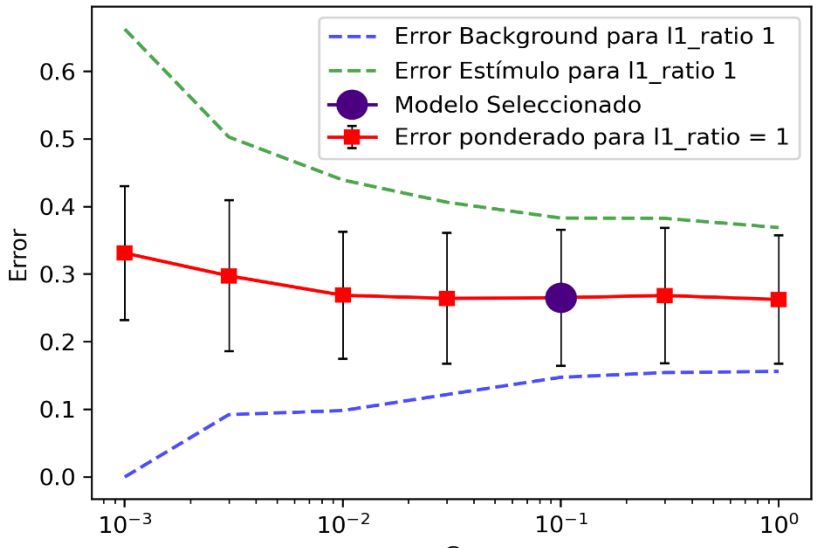
Regresión logística: C

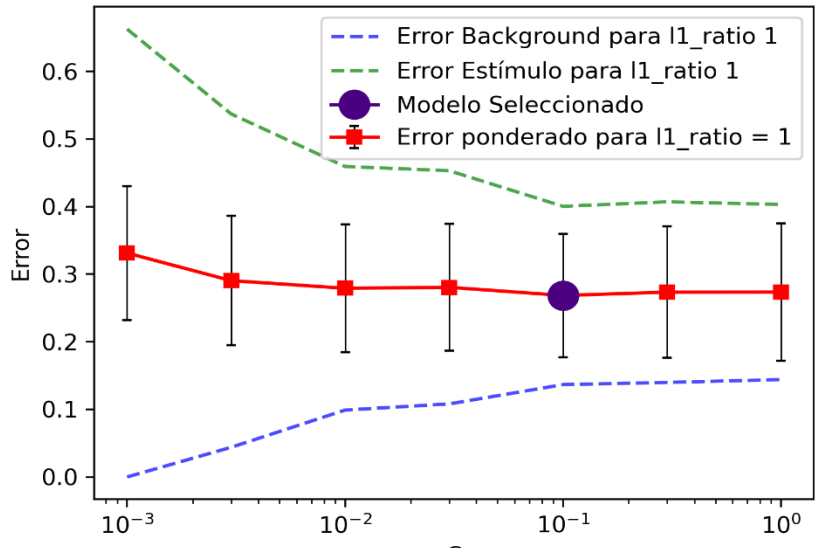
Fuerza de regularización

$C = 0.1$









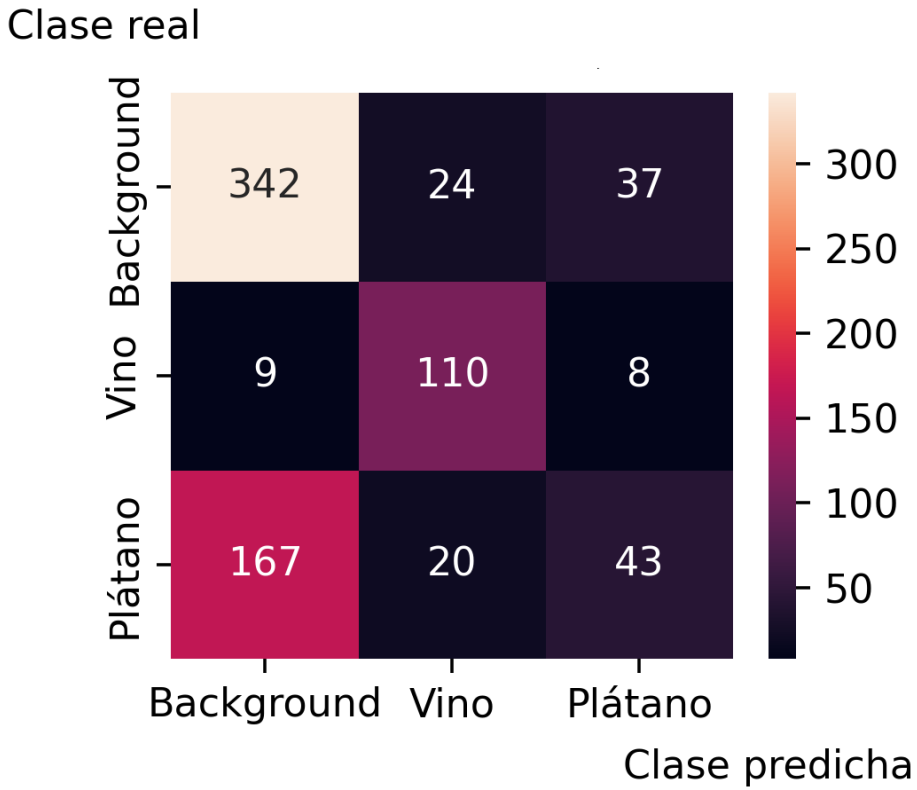
Resultados

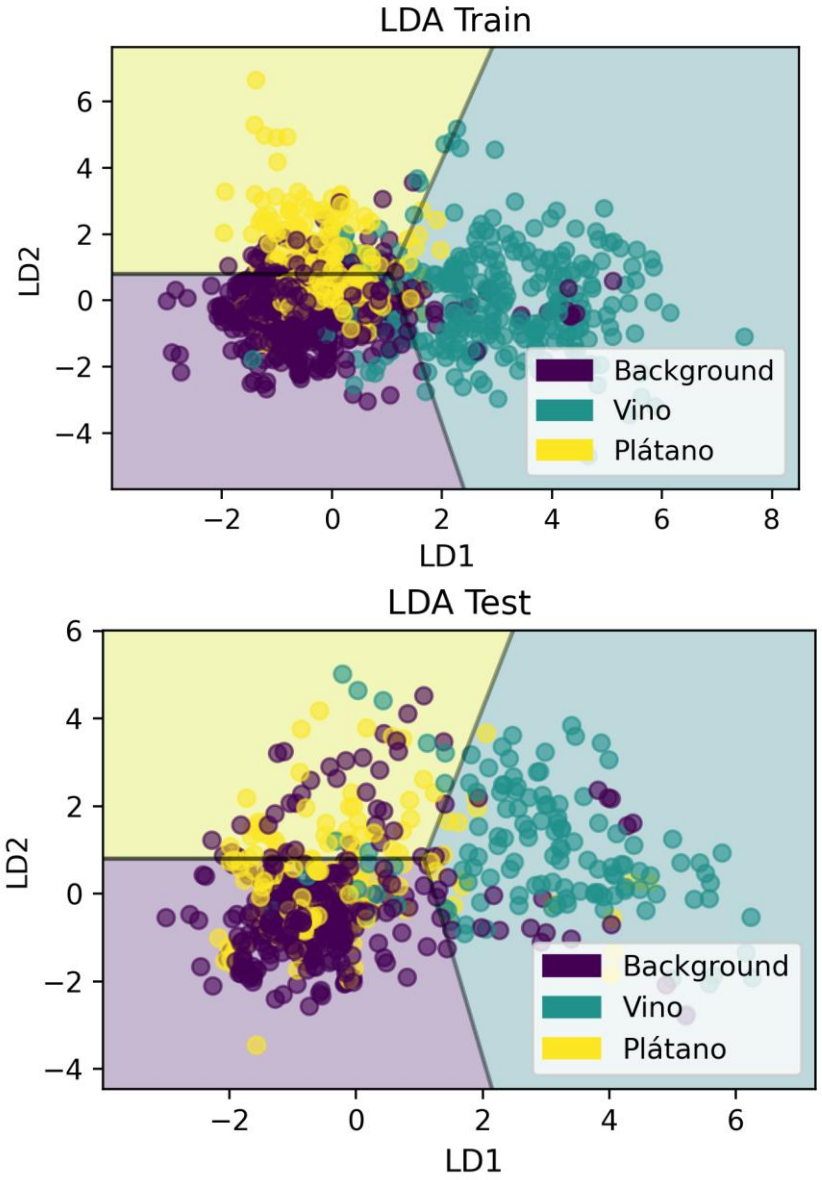
Regresión logística: Entrenamiento y Clasificación

Dataset	Rendimiento Background(%)	Rendimiento Estímulo (%)	Rendimiento Ponderado (%)	Rendimiento global (%)
Modelo Regresión Logística				
Original	78.19	64.32	71.26	65.26
PCA	75.74	57.53	66.63	58.94
mPCA	72.37	53.88	63.13	56.05
LDA	74.38	68.87	71.63	65.13
Modelo nulo: Predicción siempre background				
	100	24.86	62.43	53.02

Resultados

Regresión logística: Clasificación LDA





Resultados

Random Forest: Hiperparámetros

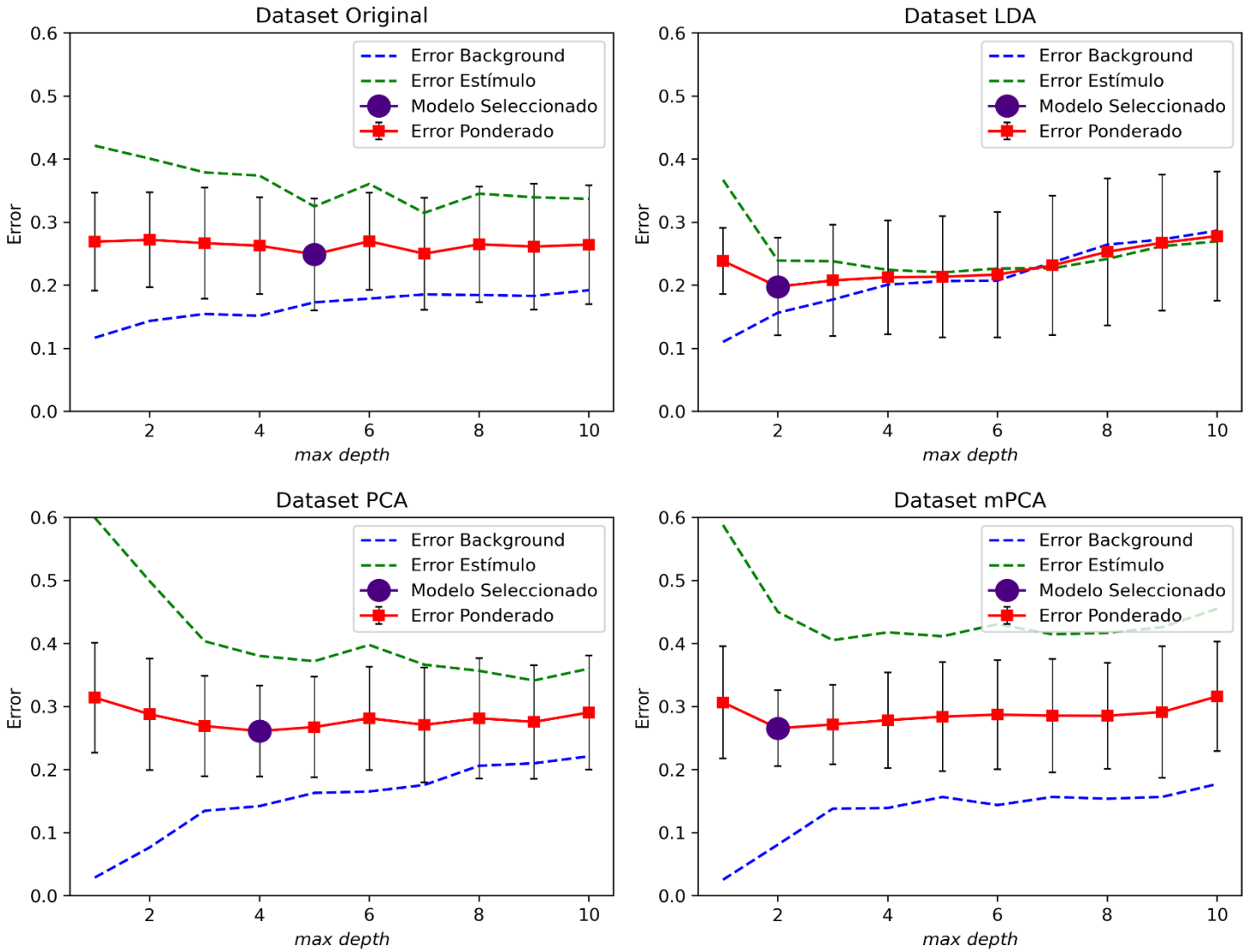
`sklearn.ensemble.RandomForestClassifier`

- $n_{estimators}$ – Número de árboles a usar
- $\max\_depth$ – Máxima profundidad de cada árbol.

$n_{estimators} = 100 \longrightarrow$ **No sobreajusta**

Resultados

Random Forest: max_*depth*



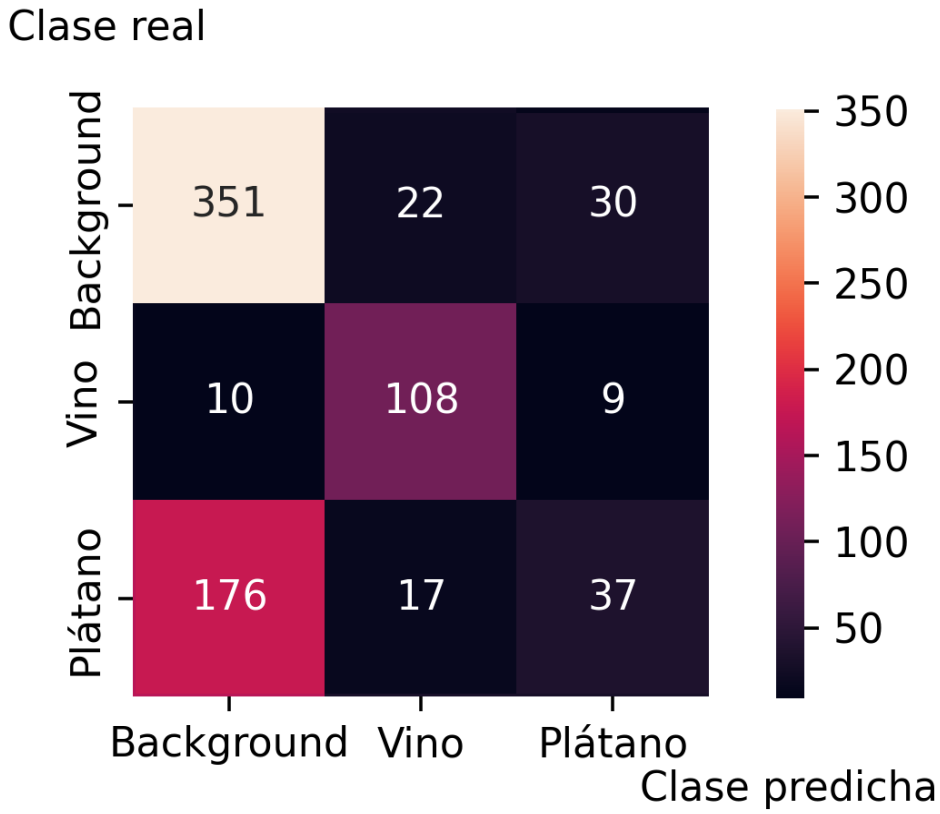
Resultados

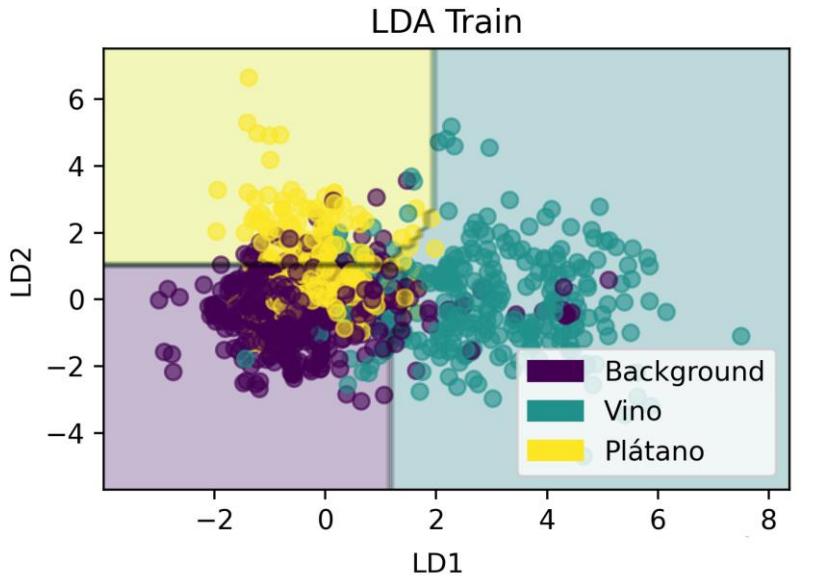
Random Forest: Entrenamiento y Clasificación

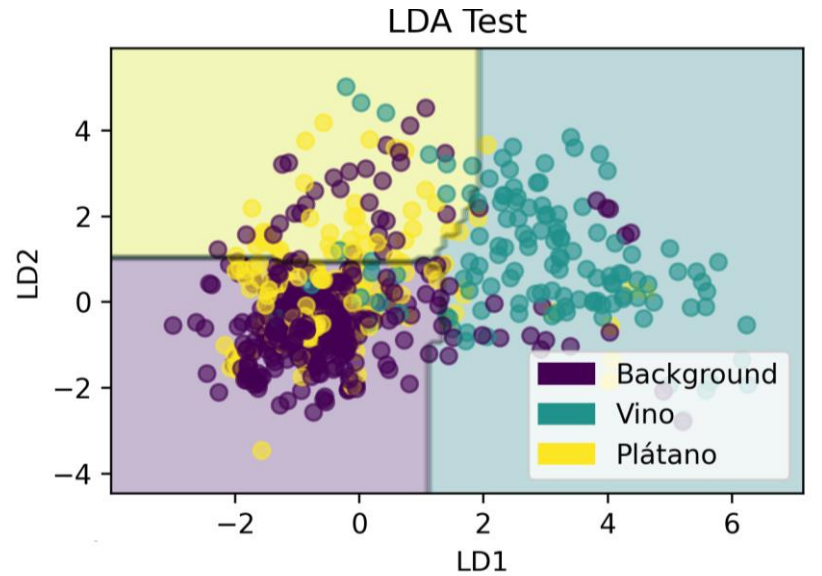
Dataset	Rendimiento Background(%)	Rendimiento Estímulo (%)	Rendimiento Ponderado (%)	Rendimiento global (%)
Modelo Regresión Logística				
Original	81.17	63.55	72.36	66,71
PCA	79.39	54.05	66.72	59,73
mPCA	77.20	56.05	66.63	56,18
LDA	79.03	66.04	72.53	65,26
Modelo nulo: Predicción siempre background				
	100	24.86	62.43	53.02

Resultados

Random Forest: Clasificación LDA

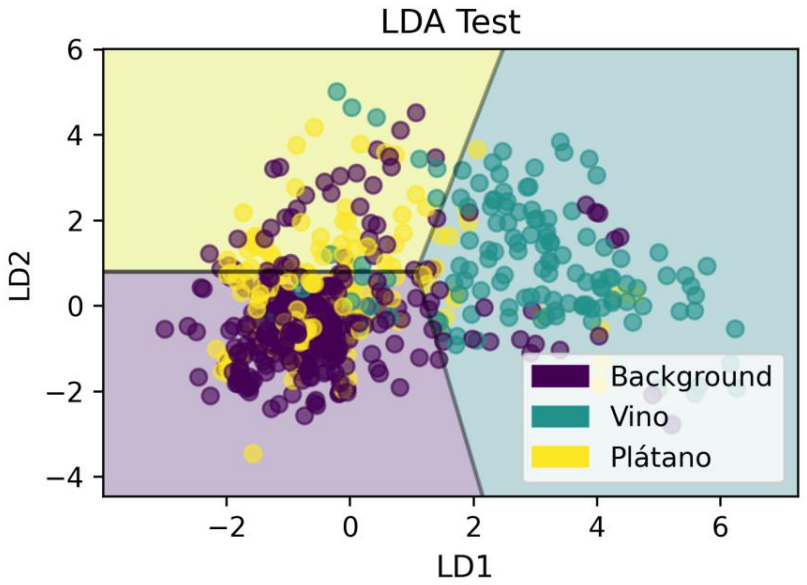






Discusión

- Rendimiento de los resultados
    - Confusión
    - Sparcity



Dataset	Rendimiento Background(%)	Rendimiento Estímulo (%)	Rendimiento Ponderado (%)	Rendimiento global (%)
mPCA	72.37	53.88	63.13	56.05
LDA	74.38	68.87	71.63	65.13

Referencias

[1] J.R. Berrendero, A. Justel, and M. Svarc. Principal components for multivariate functional data. *Computational Statistics Data Analysis*, 55(9):2619–2634, 2011.

[2] Ramon Huerta, Thiago Mosqueiro, Jordi Fonollosa, Nikolai F Rulkov, and Irene Rodriguez-Lujan. Online decorrelation of humidity and temperature in chemical sensors for continuous monitoring. *Chemometrics and Intelligent Laboratory Systems*, 157:169–176, 2016.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.