Predicción de cristalización de perovskitas

mediante aprendizaje automático

María Belén Ticona

Tesis de Licenciatura en Ciencias de la Computación

Director: Onna, Diego
Co-director: Turjanski, Pablo

DEPARTAMENTO DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Índice

01. Ciencia de los materiales

EXPERIMENTOS
(prueba y error)

✓ ÉXITO

✗ FRACASO

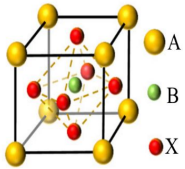
Ejemplo de datos experimentales

Concentraciones (M)					
Reactivo A	Reactivo B	Reactivo C	Temperatura (°C)	Solvente	Resultado
2,5	0,3	0,7	57	DMSO	✗
2,5	0,2	0,9	60	GBL	✗
2,1	0,2	0,5	55	DMSO	✓
…	…	…	…	…	…
2,1	0,1	0,7	55	DMSO	?

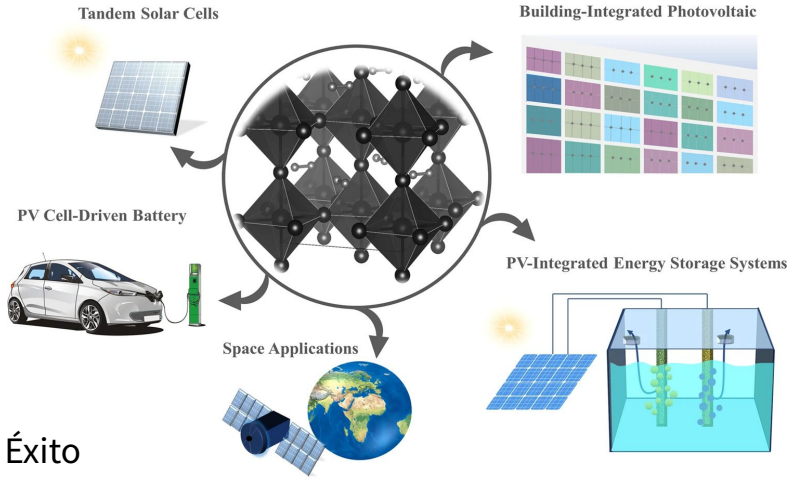
¿Caracterización de los datos?

¿Predicción?

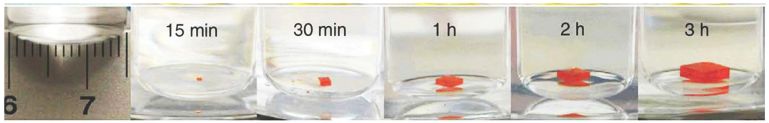
01. Caso de estudio: perovskitas



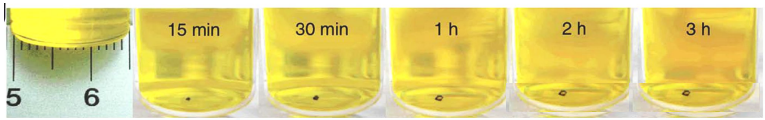
ESTRUCTURA CRISTALINA



FORMACIÓN DE MONOCRISTALES



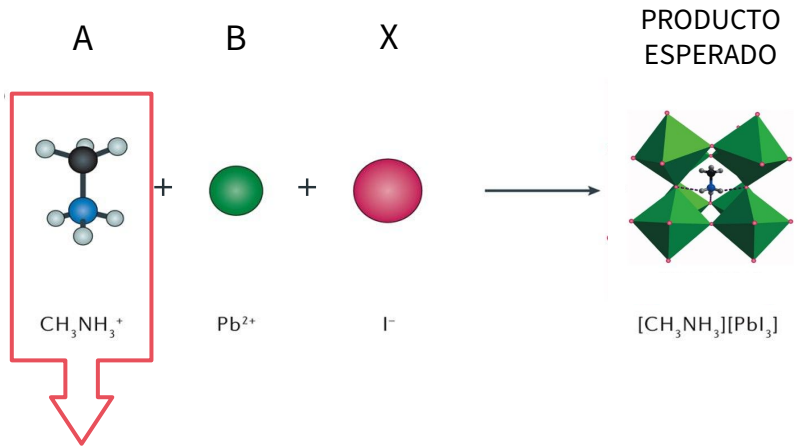
✓ Éxito



✗ Fracaso

02. Dataset de experimentos

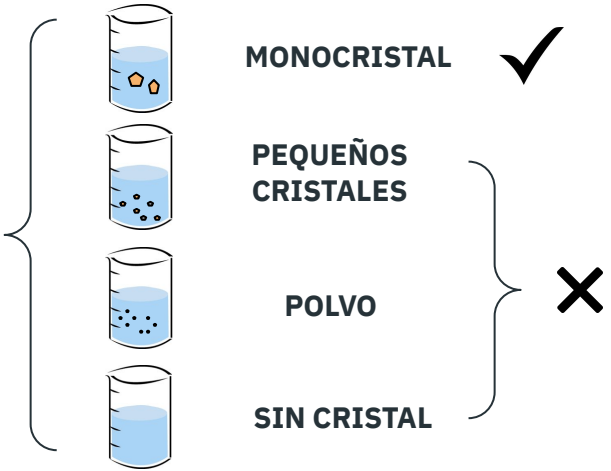
SÍNTESIS DE PEROVSKITA



¿Influencia de la organoamina?

23 organoaminas

RESULTADOS REALES POSIBLES



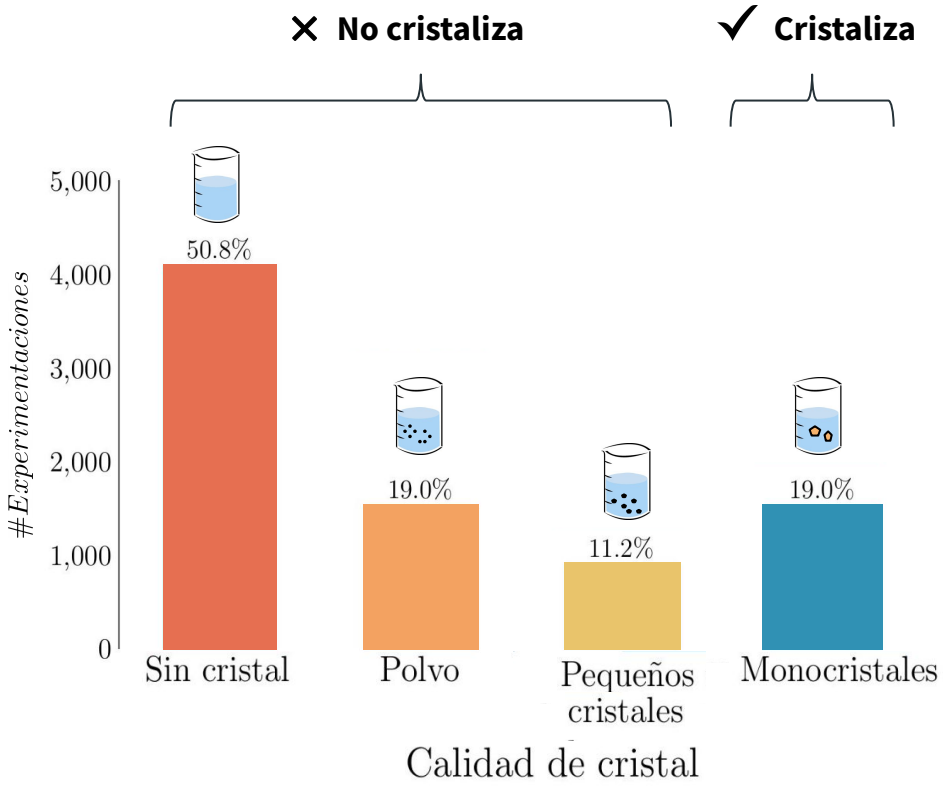
Calidad de Cristal

02. Variables predictoras del dataset

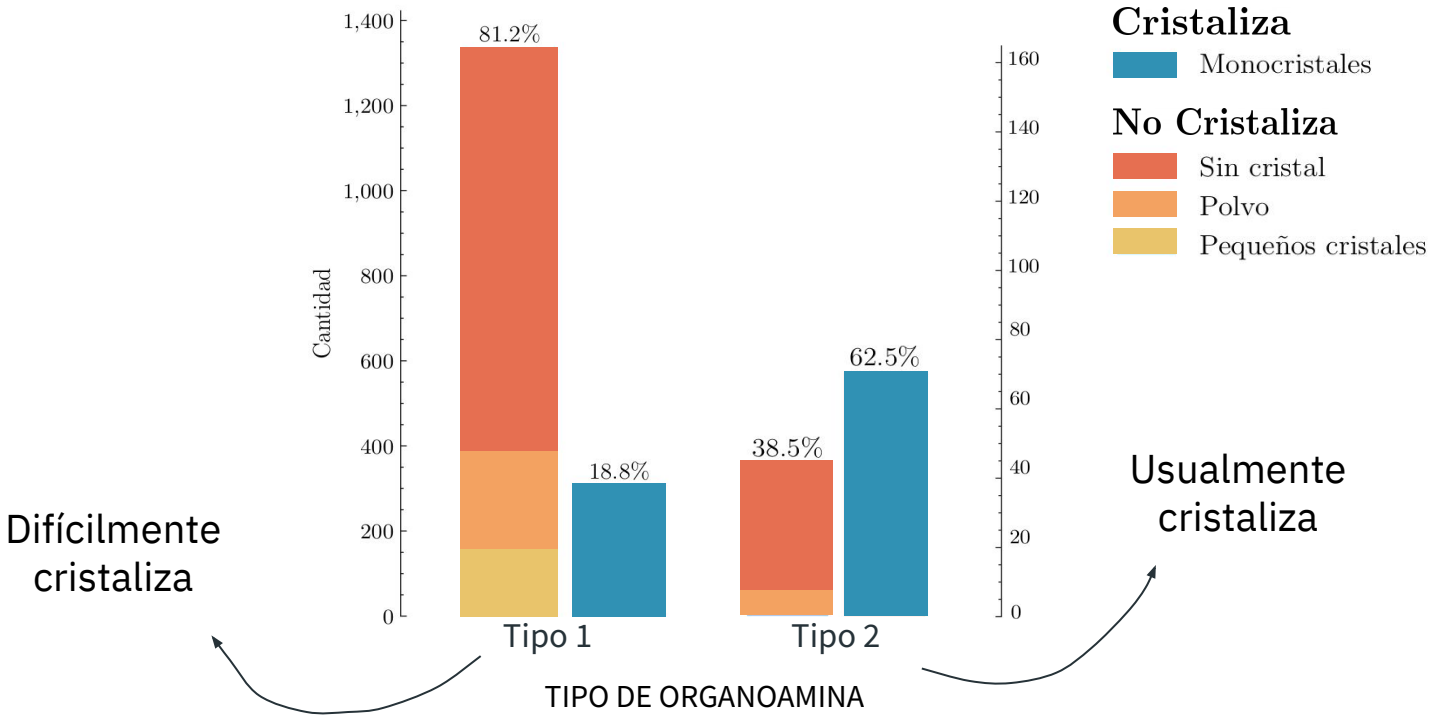
Condiciones de Reacción (FEAT_RXN)	Cantidad de Variables
concentración de reactivos (orgánico, inorgánico, ácido)	3
solvente (GBL, DMSO o DMF)	1
temperatura de reacción	1
tiempos de agitación	2
organoamina ID	1
Descriptor de Organoamina (FEAT_PHYCHEM)	
propiedades FQ[1]	61

[1] E.g cantidad de grupos funcionales, donores o aceptores, cantidad de enlaces

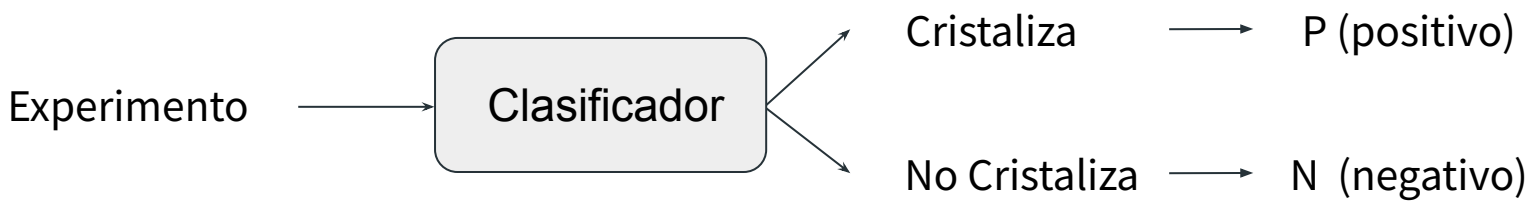
02. Desbalance en calidad de cristal



02. Organoaminas con distinto desbalance



03. Modelos de Clasificación Binaria



03. Evaluación de Clasificación

Matriz de confusión (CM)

	N_{PRED}	P_{PRED}
N_{REAL}	NN	NP
P_{REAL}	PN	PP

Ejemplo

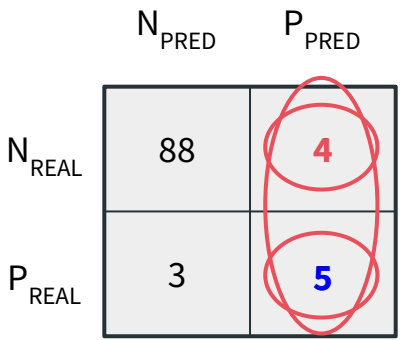
	N_{PRED}	P_{PRED}
N_{REAL}	88	4
P_{REAL}	3	5

Métricas

	Definición	*Ejemplo*
Recall	PP/(PP+PN)	0.63
MCC	$\frac{(PP*NN-PP*PN)}{\sqrt{(PP+NP)...(NN+PN)}}$	0.55

03. Ruido en los datos de evaluación

CM original



Recall: 0.63

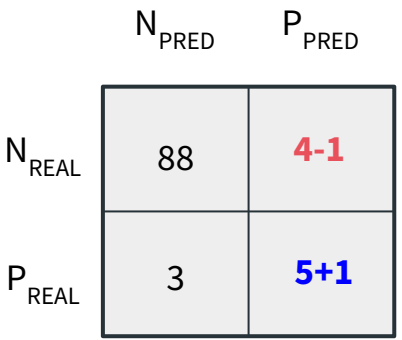
MCC: 0.55

Confusión de etiquetado real

1 NP → 1 PP

Gran variabilidad de las métricas

CM con ruido

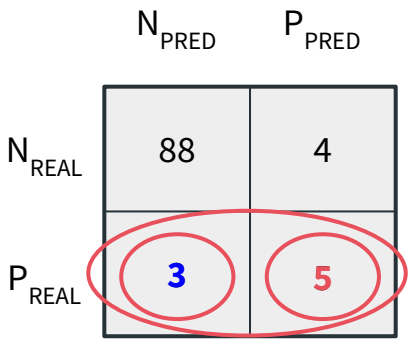


Recall: 0.66

MCC: 0.63

03. Ruido en la predicción del modelo

CM original



Recall: 0.63

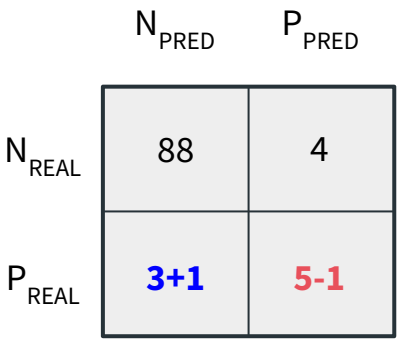
MCC: 0.55

Confusión de etiquetado predicho

1 PP → 1 PN

Gran variabilidad de las métricas

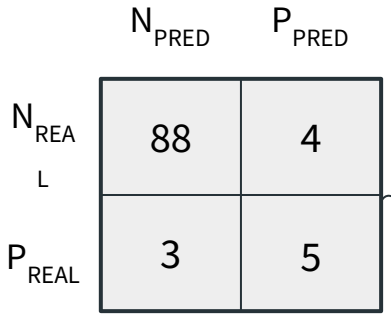
CM con ruido



Recall: 0.50

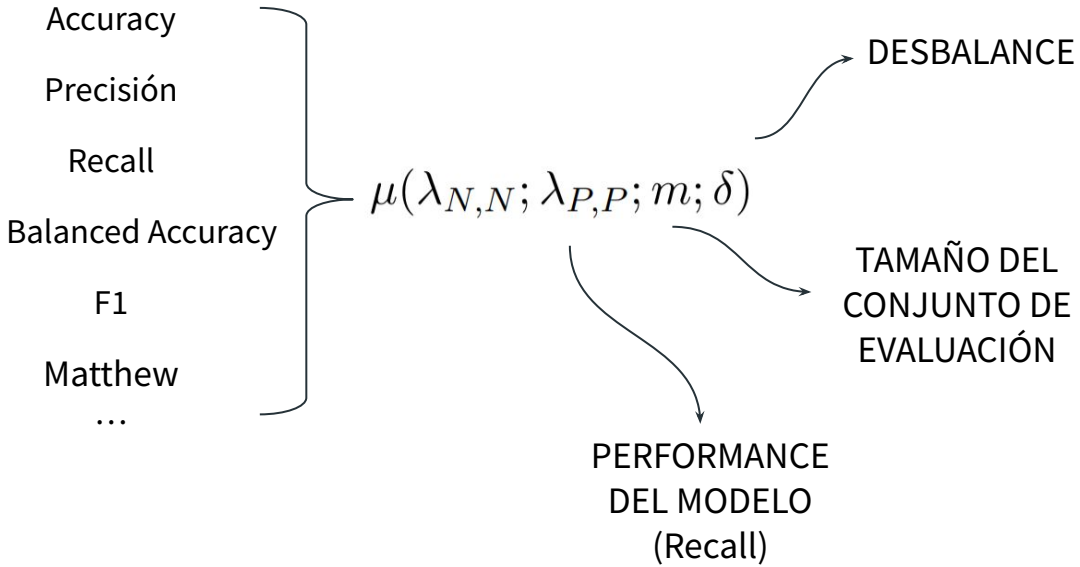
MCC: 0.46

03. Parametrización de la matriz de confusión

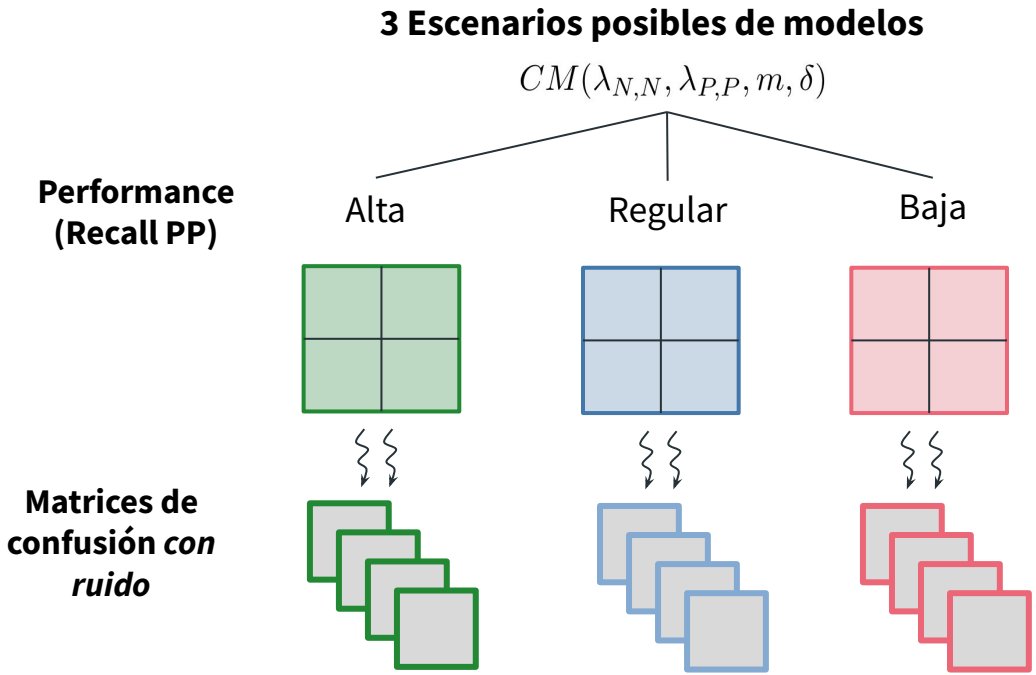


$$\begin{bmatrix} m_{N,N} & m_{N,P} \\ m_{P,N} & m_{P,P} \end{bmatrix} \rightsquigarrow CM(\lambda_{N,N}; \lambda_{P,P}; m; \delta)$$

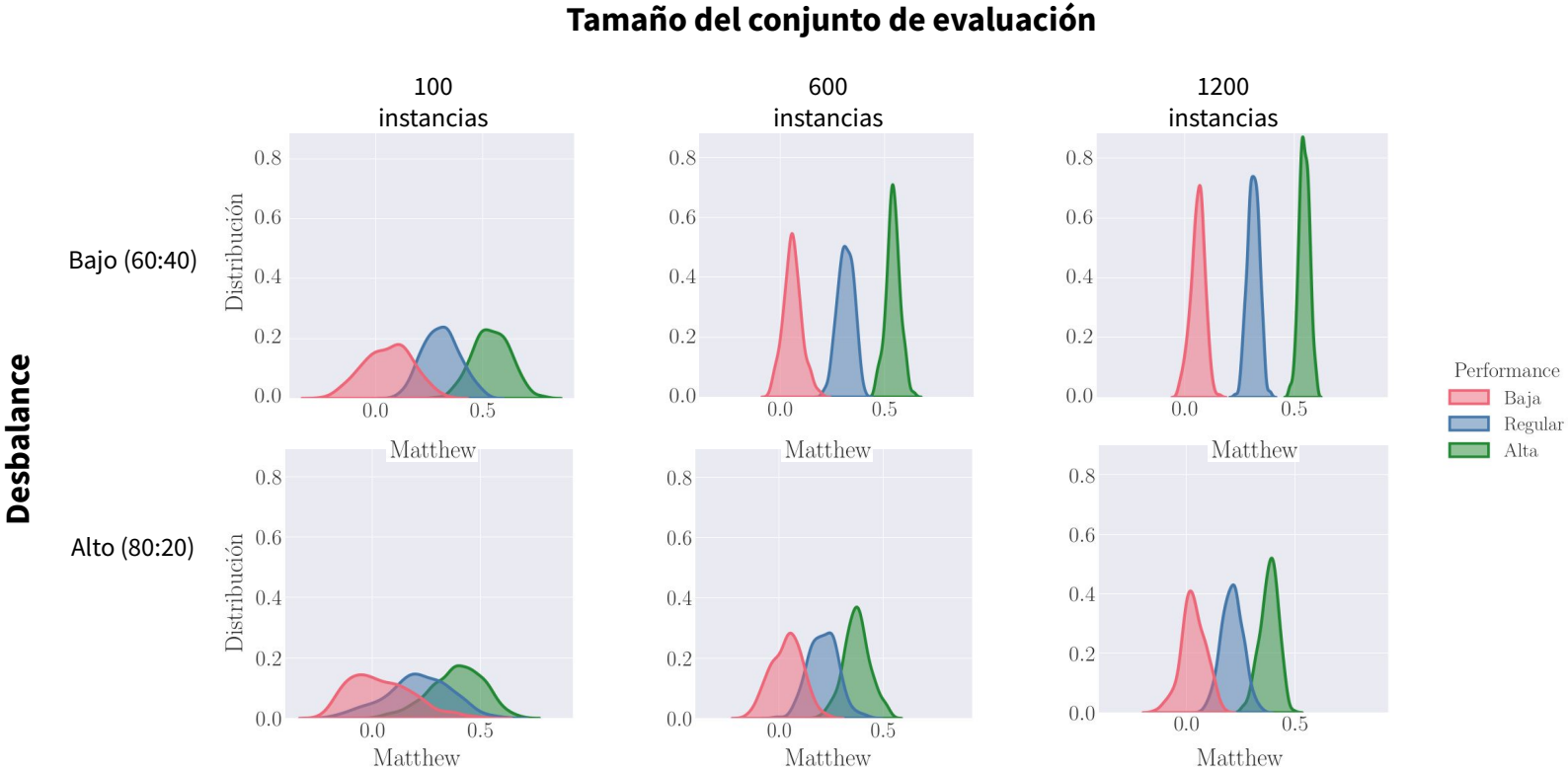
03. Parametrización de métricas de clasificación



04. Simulación de errores en CM

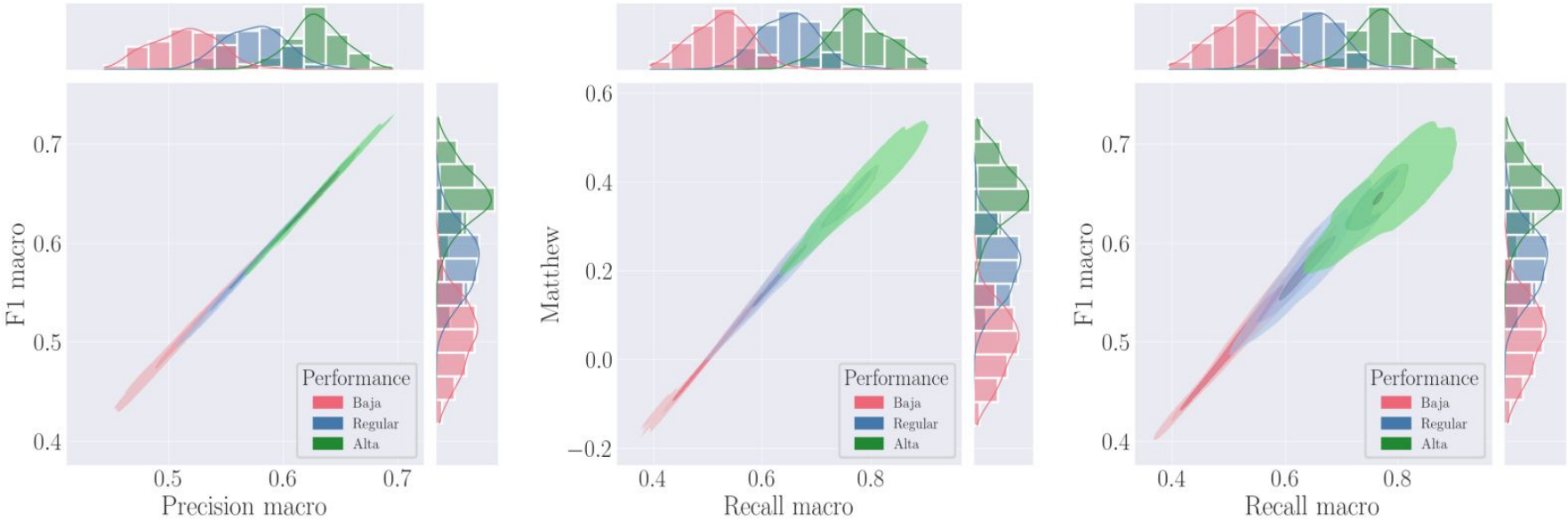


04. Influencia del desbalance y tamaño del dataset

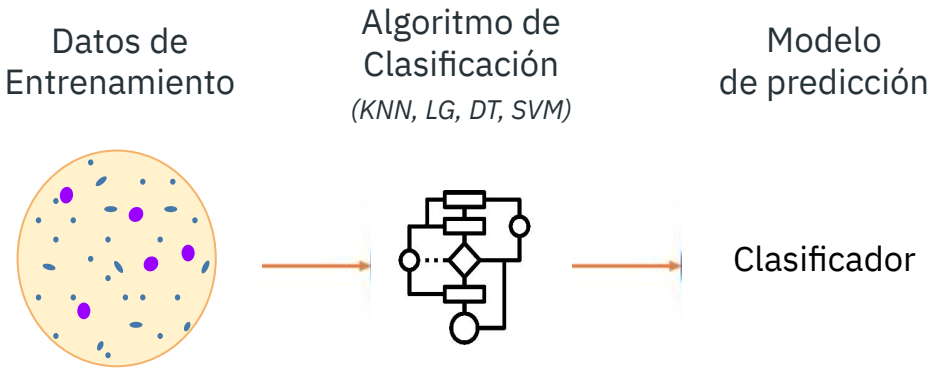


04. Métricas complementarias

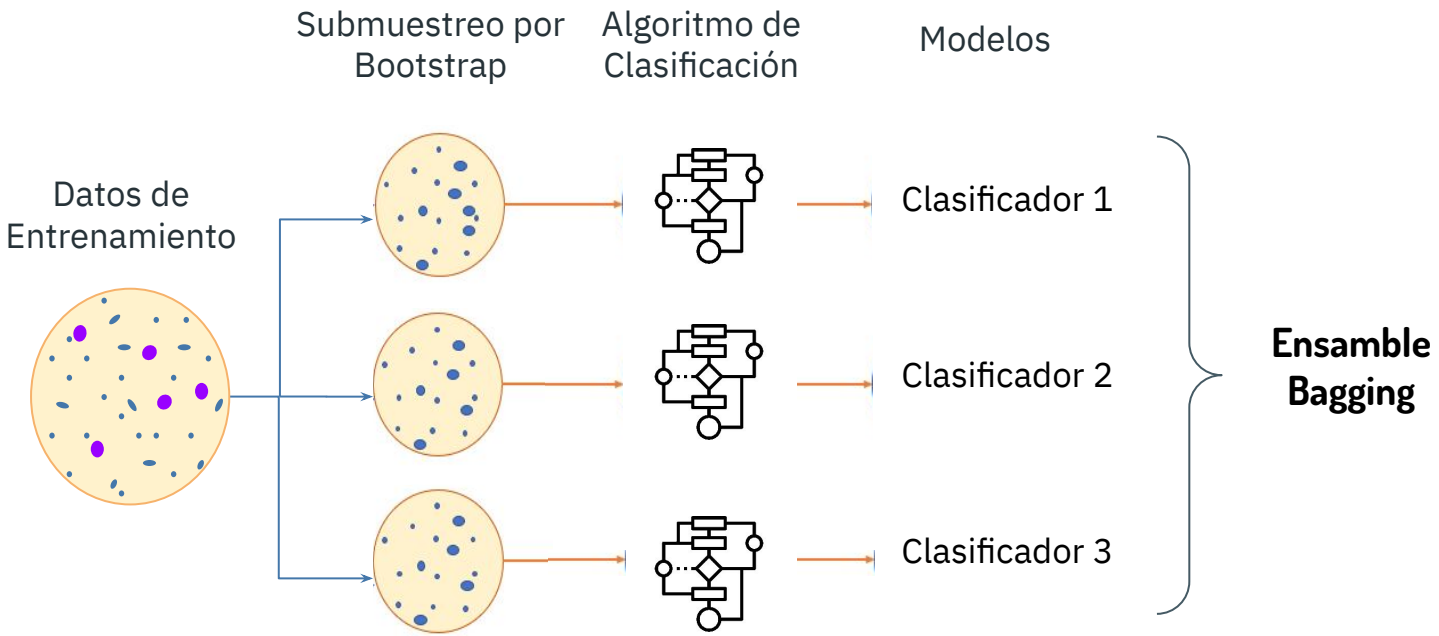
Complementariedad



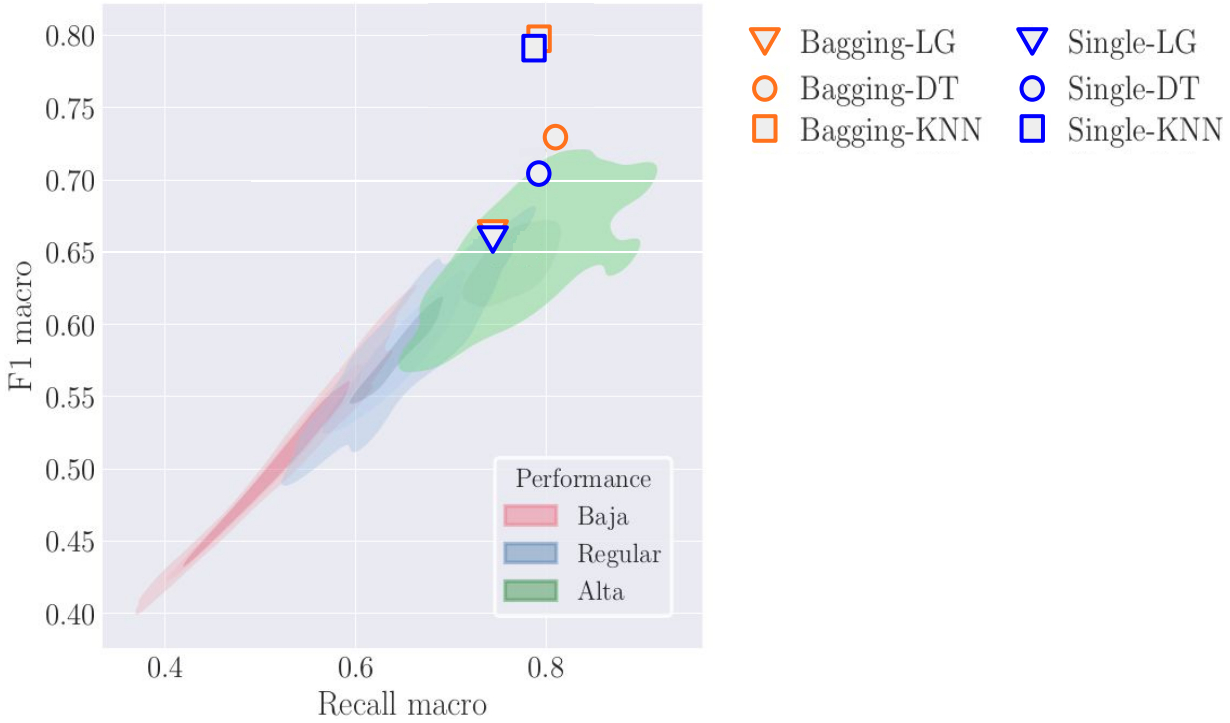
05. Modelos de clasificación *(single)*



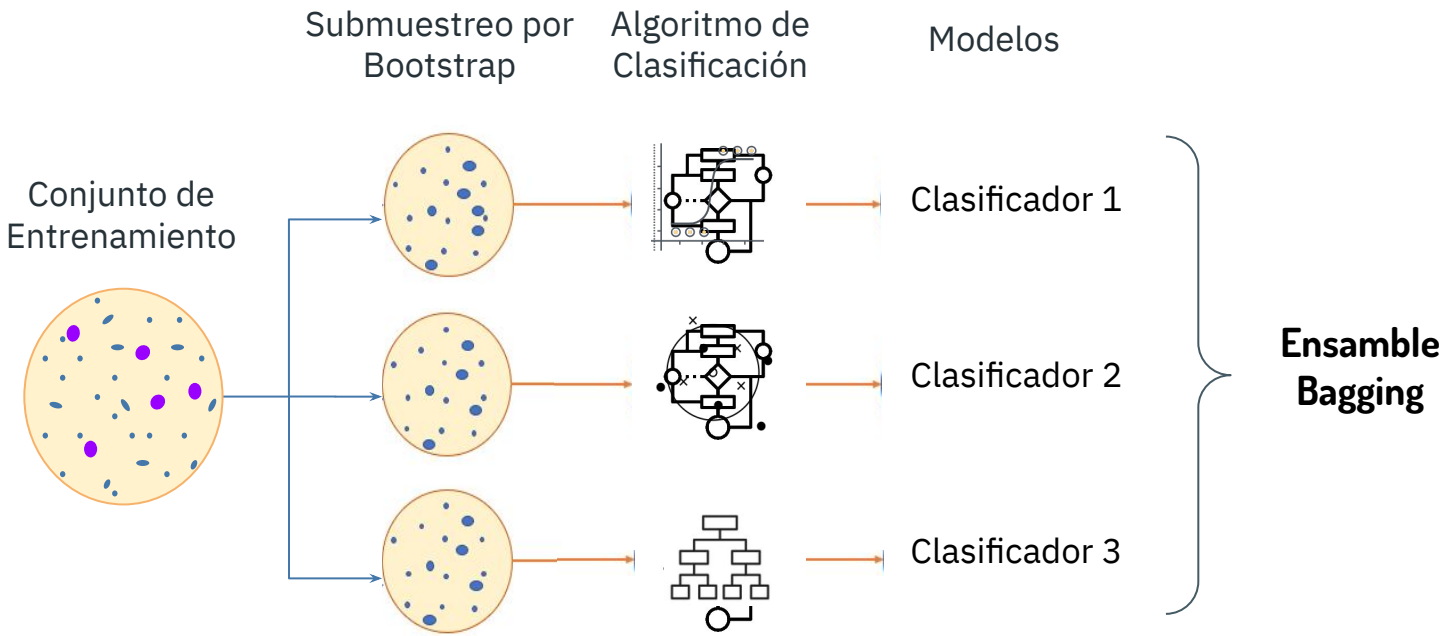
05. Ensamble de modelos *(bagging)*



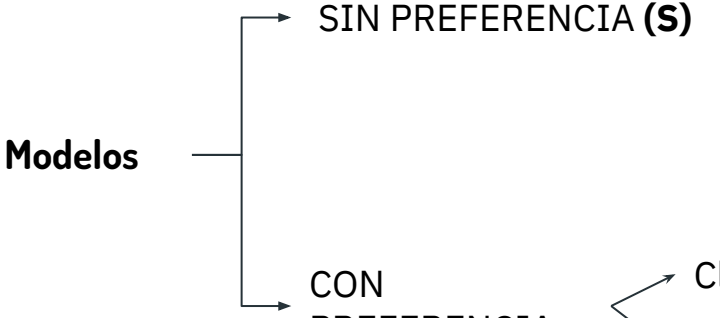
05. Comparación Single vs Ensambles Bagging



05. Ensamble de modelos heterogeneos *(bagging)*



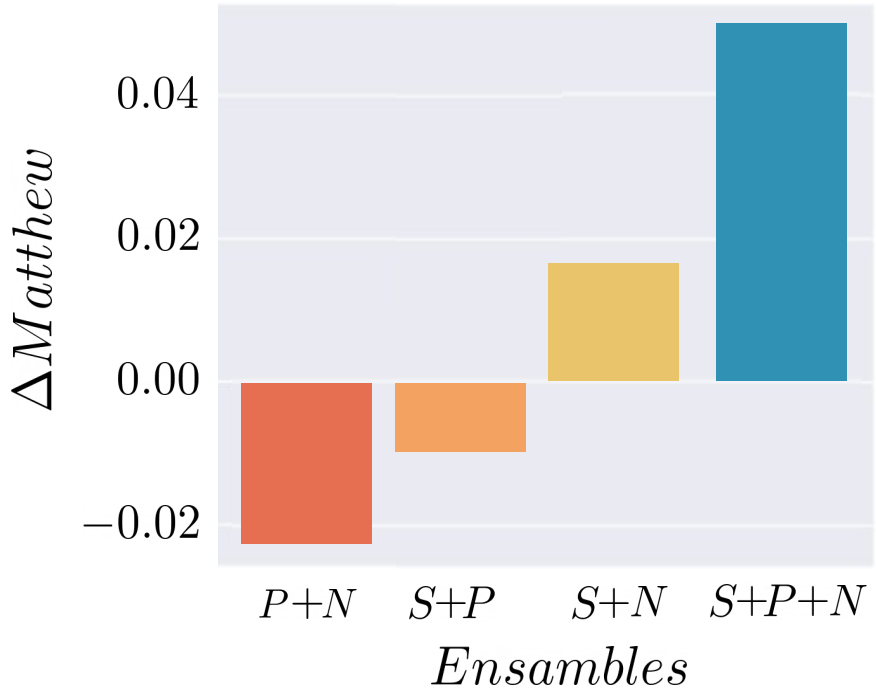
05. Ensambles según preferencia de clase



	Recall	
	Negativa	Positiva
DT	0.75	0.84
LG	0.69	0.79

		Recall	
		Negativa	Positiva
SVM		0.64	0.91
KNN		0.94	0.68

05. Ensambles según clase de preferencia



06. Conclusiones

- Caracterización de datos experimentales para evaluar sistemas predictivos
- Significancia del valor de una métrica de clasificación
- Ensambles de modelos complementarios para datasets chicos
- Contexto de recolección de datos vs aplicación de sistemas predictivos

06. Trabajo Futuro

- Dataset con ruido: variabilidad de evaluación de clasificadores
- Fusión de datos de síntesis de perovskita
- Enfoque multiclase de ensambles

¡Muchas gracias!

Al jurado:

Esteban Mocskos y Mario Tagliazucchi

A los directores:

Diego Onna y Pablo Turjanski

Programa de Becas de Iniciación a la Investigación en Cs. de la Computación