Validación Cruzada

Obtener mejores modelos con pocos datos

Tradicionalmente

Training set

Tradicionalmente

Training set	Test set

Tradicionalmente

Training set

Test set

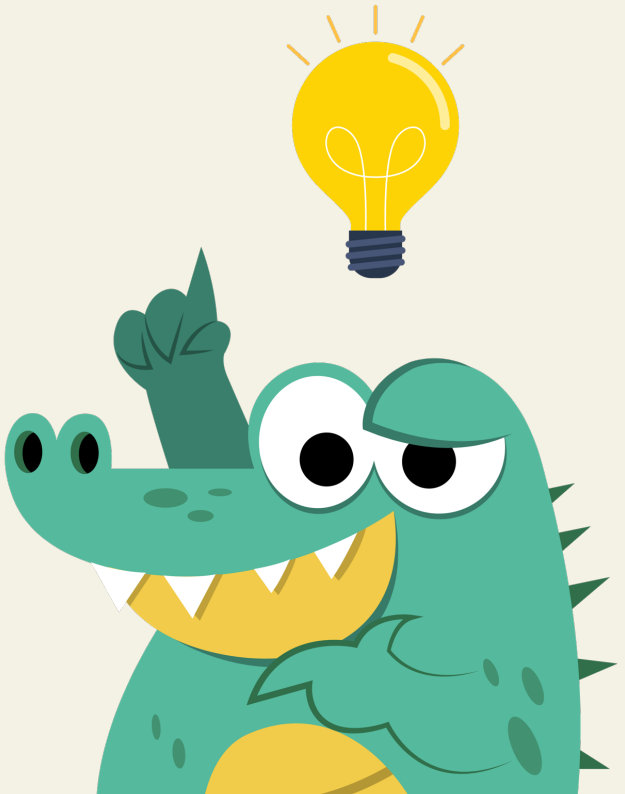
Accuracy: 0.91

Validación cruzada

Training set

Validación cruzada

Training set	Training set	Test set

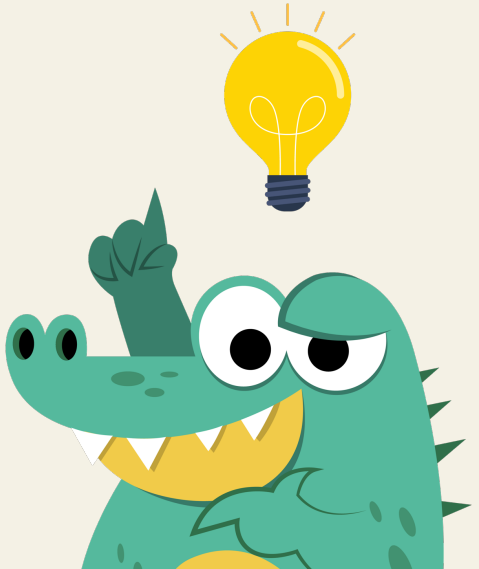


Validación cruzada

1st iteration

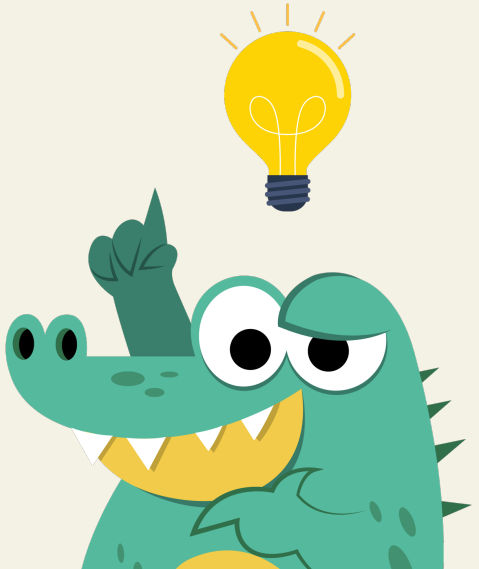
Training set	Training set	Test set

Acc: **0.98**



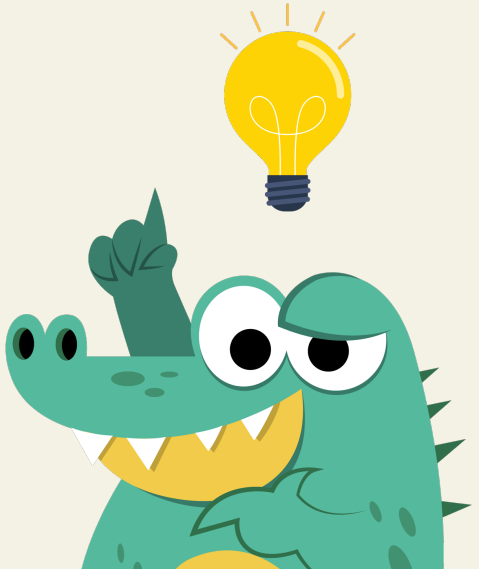
Validación cruzada

1st iteration	Trainng set	Training set	Test set	Acc: **0.98**
2nd iteration	Training set	Test set	Training set	Acc: **0.92**



Validación cruzada

1st iteration	Training set	Training set	Test set	Acc: **0.98**
2nd iteration	Training set	Test set	Training set	Acc: **0.92**
3rd iteration	Test set	Training set	Training set	Acc: **0.96**



Validación cruzada

k-fold x Validation

1st iteration	Training set	Training set	Test set
2nd iteration	Training set	Test set	Training set
3rd iteration	Test set	Training set	Training set

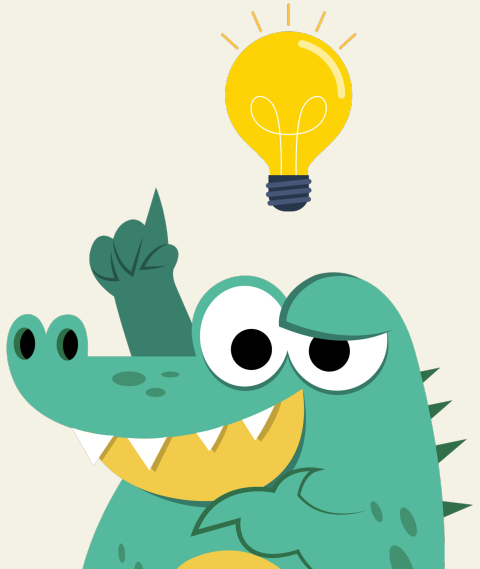
Acc:

Acc:

Acc:

$$K = 3$$

Cross-validated accuracy: (**0.98** + **0.92** + **0.96**) / 3 = **0.95**



Validación cruzada

Comúnmente usada cuando tenemos pocos datos, puesto que aprovecha al máximo lo poco que hay.

Aún recomendable con muchos datos si tu poder de cómputo lo permite.

Hay que tener especial cuidado cuando se trata de *forecasting*.

Búsqueda en *grid*

¿Cuáles son los mejores hiperparámetros?

Tradicionalmente

Los modelos tienen varios hiperparámetros

¿Cómo le hacemos para encontrar los mejores?

¿Ejecución manual de experimentos?

Tradicionalmente

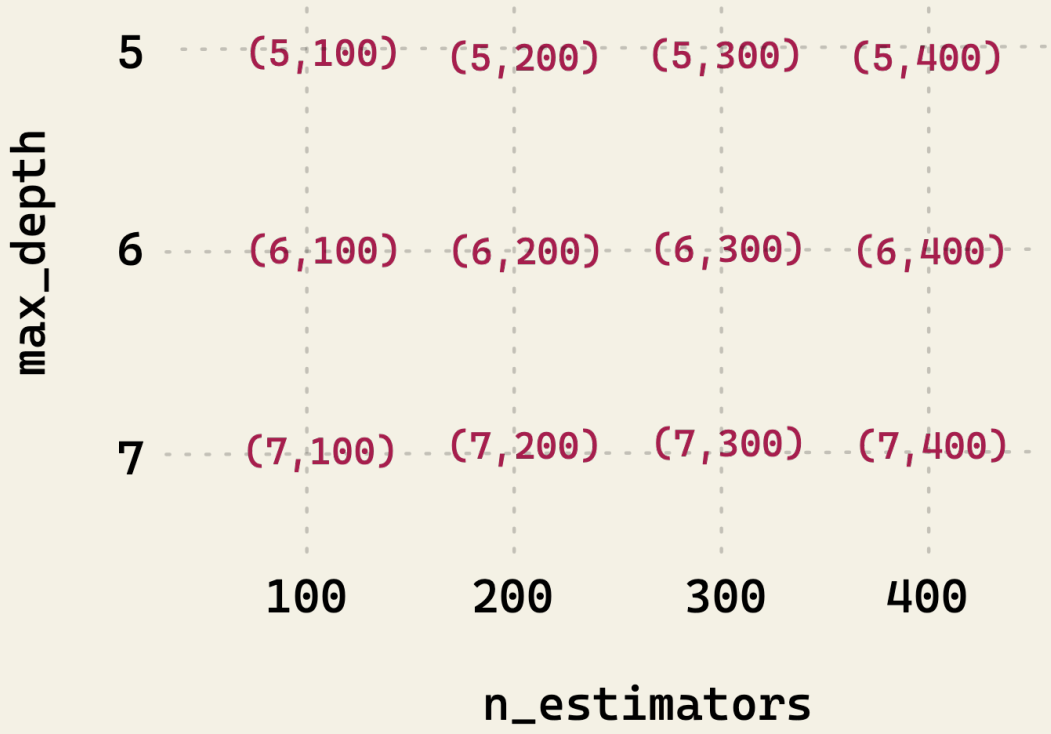
Los modelos tienen varios hiperparámetros

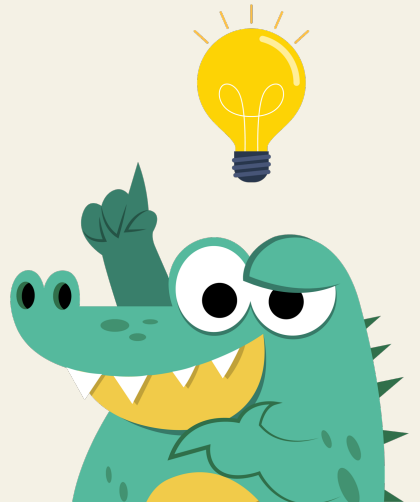
RandomForestClassifier

n_estimators: 100, 200, 300, 400
max_depth: 5, 6, 7

Grid search

`n_estimators: 100, 200, 300, 400 | max_depth: 5, 6, 7`





Grid search

Ventaja

- Espacio de búsqueda comprensivo – prueba todas las combinaciones que le demos.

Desventaja

- La complejidad crece con cada dimension – es exponencial
- Está limitada a las combinaciones que nosotros especificamos

Otras, mejores, opciones: Random search y optimización bayesiana