Ciencia de Datos

- **Módulo 4**

Cross-validation

UTNFra

GyEE Gestión y Evaluación Educativa

<aprendé programando Virtual>

SECUNDARIA del futuro

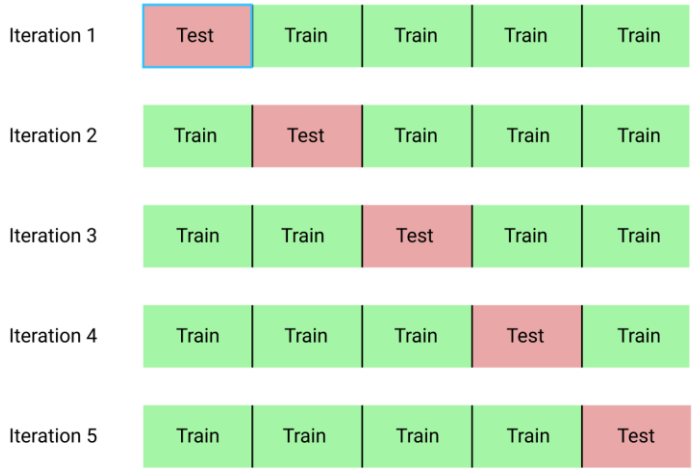
Cross -Validation

Para entrenar y probar los modelos, hasta ahora hemos visto que dividimos los datos en entrenamiento y testo, usamos los datos de entrenamiento para entrenar y guardar los datos de testo para evaluar el modelo. Pero es posible utilizar los mismos datos de entrenamiento divididos para hacer pruebas. Este método se llama Validación cruzada (Cross-validation).

Este método se utiliza por ejemplo cuando no hay muchos datos y no es posible dividir en entrenamiento y testeo

Fuentehttps://www.mathworks.com/discovery/cross-validation.html

Cross-validation



Ejemplo con 5 Folds

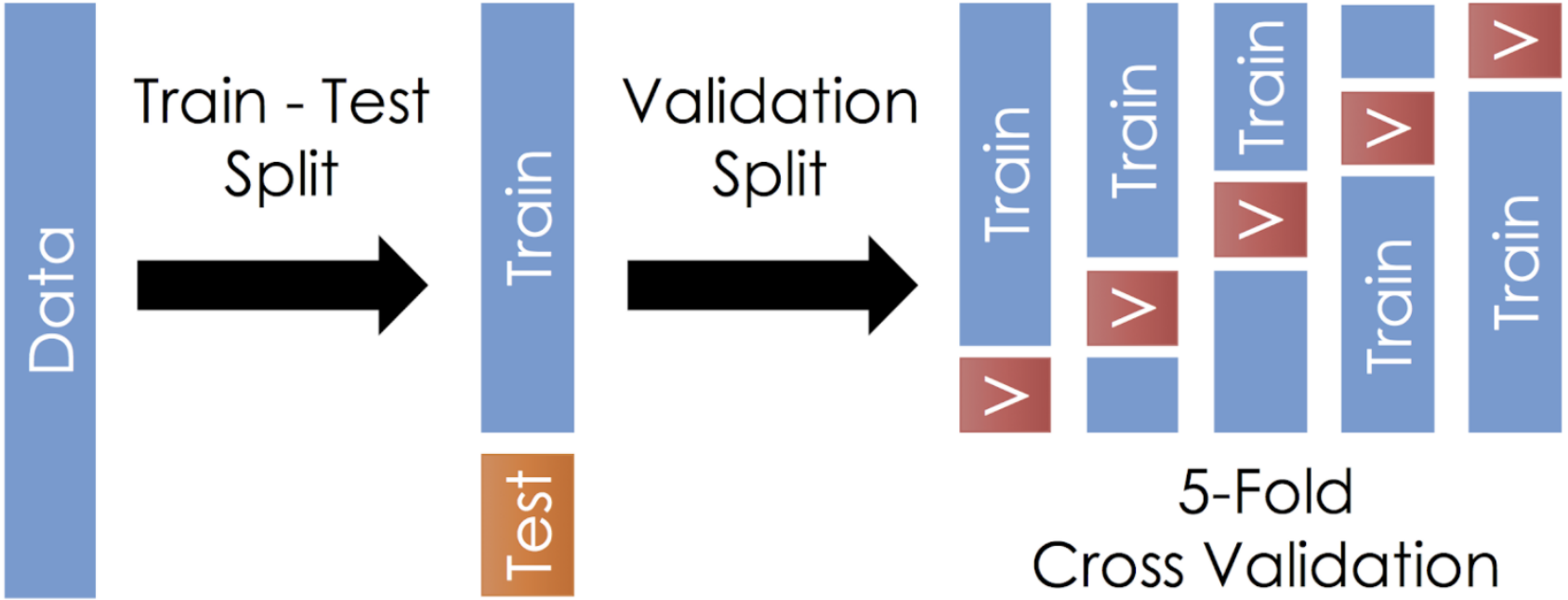
Fuente https://chinchurosajolly.medium.com/machine-learning-cross-validation-3d4a95452640

En primer lugar es necesario definir la cantidad de divisiones (folds) que se deseen utilizar (en general 5 o 10)

Solo sobre los datos de entrenamiento (los datos de testeo no se utilizan), se dividen los datos de entrenamiento de manera aleatoria, se utiliza una parte de los datos para entrenar, otra para evaluar, y luego se repite cambiando los datos.

De esta manera es posible asegurarse que no hay una distorsión en las evaluaciones por la división de los datos.

Cross-validation



Documentación Scikit-Learn: https://scikit-learn.org/stable/modules/cross_validation.html