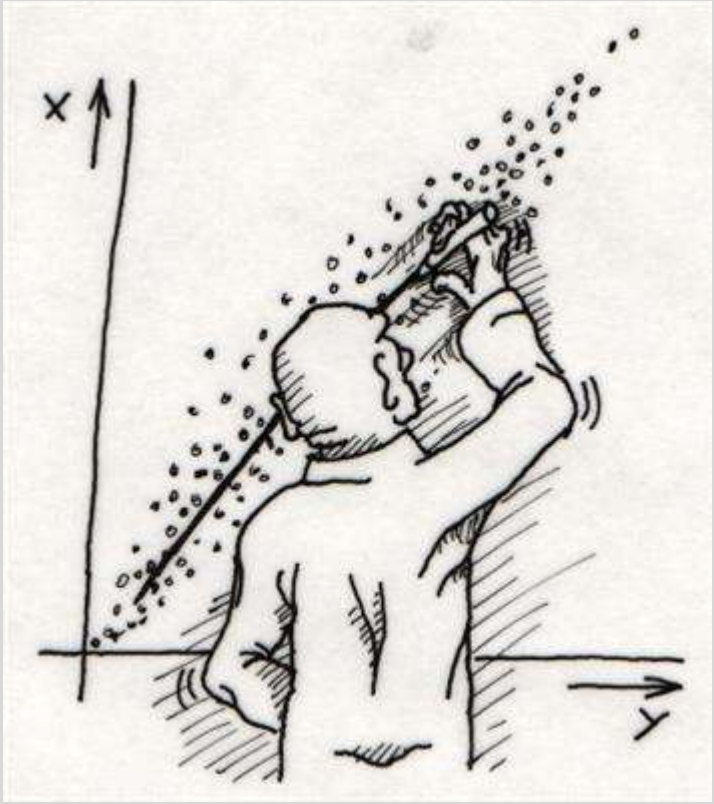


DigitalHouse > Coding School

DATA SCIENCE

UNIDAD 2
MÓDULO 3

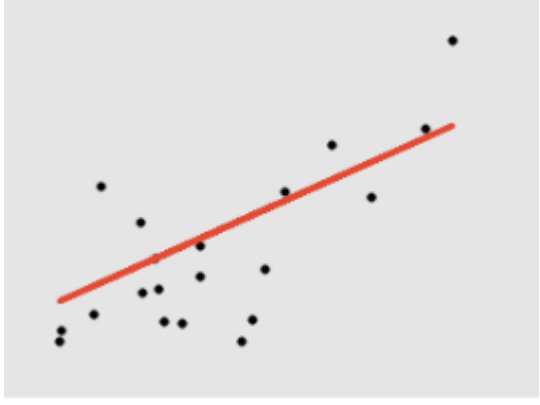
Split Train Test Workflow

Abril 2017

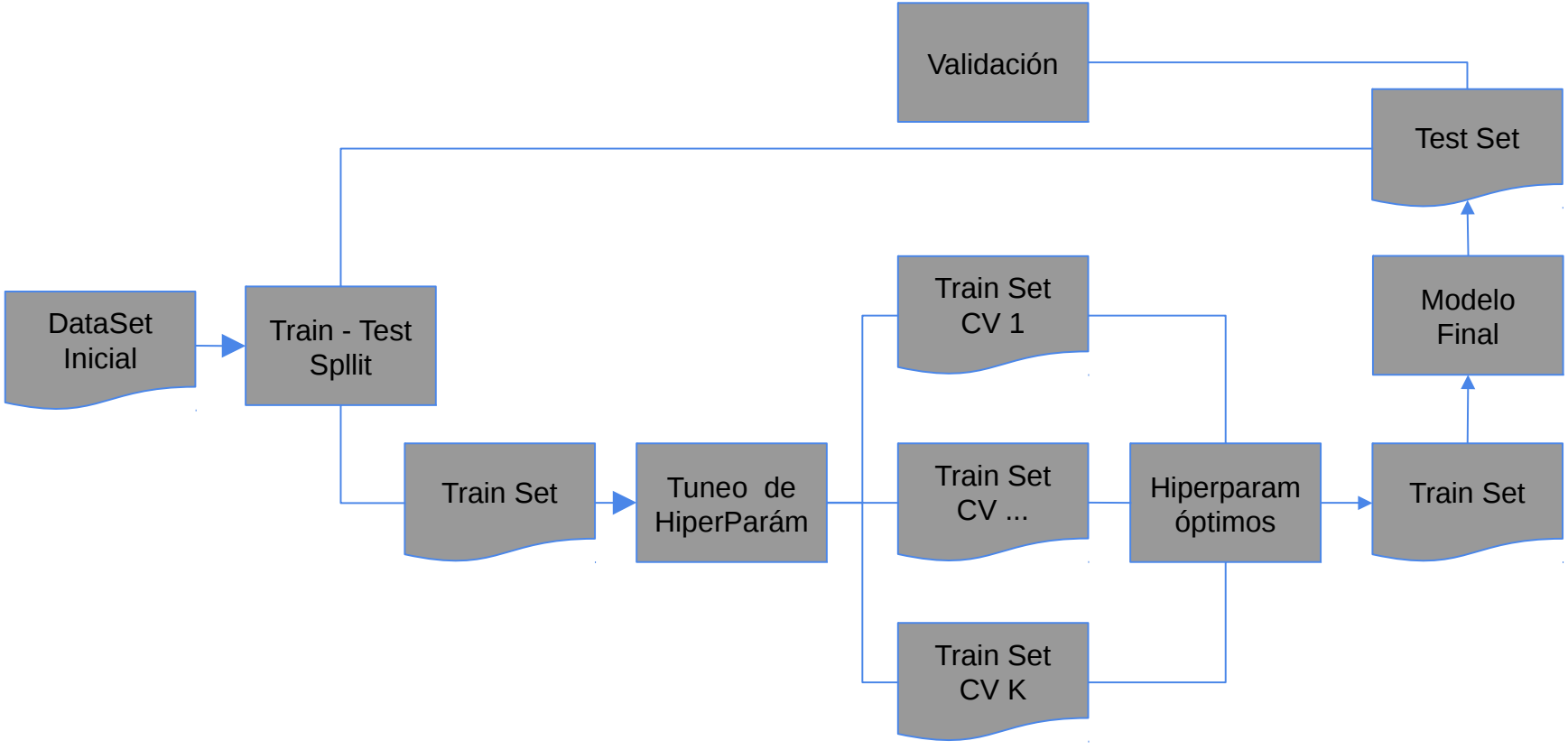
Split Train Test Workflow

OBJETIVOS DE LA CLASE

1. **Presentar las diferentes etapas de un proceso sencillo de entrenamiento**
2. **Comprender la lógica general del proceso de entrenamiento**
3. **Comprender las diferencias en el uso del Split Train Test y Cross Validation en este proceso**
4. **Evaluar tres modelos candidatos sobre un mismo dataset**



WorkFlow de Entrenamiento



Validación Cruzada

- Tenemos tres modelos candidatos (esto es generalizable sin demasiados problemas)
    - Regresión lineal
    - Regresión lineal regularizada con Ridge
    - Regresión lineal regularizada con LASSO
- Queremos elegir el mejor modelo. El que mejor performa. ¿En dónde?...
- Entonces lo primero que hacemos es dividir el dataset.
- Luego, tenemos que estimar los Hiperparámetros: en nuestro caso los alpha de Ridge y LASSO
    - Usamos Validación Cruzada dentro de Training Set
- Finalmente, estimamos las versiones finales de los modelos candidatos sobre TODO el Training Set

- Nos interesa conocer cómo funciona el modelo “en general”, no en el único dataset que tenemos o conocemos.
- Buscamos modelos que generen buenas predicciones sobre datos “nuevos”.
- En algunos casos, tendremos acceso a datos “nuevos” (por ejemplo, clientes nuevos). Pero en otros, no tendremos acceso inmediato.
- Diferentes estrategias para esto:
  - Train Test Split
  - Validación Cruzada

Práctica Guiada

WorkFlow Completo

Conclusión

- Entrenamos tres modelos diferentes.
- Usamos dos formas de split: split train-test y Cross Validation. Cada una tiene su lógica y en este workflow sirven para dos propósitos diferentes.
- Finalmente, evaluamos los tres modelos y seleccionamos el que mejor performa en test set.