Estimación empírica del error de generalización

Uno de los enfoques para estimar el error de generalización es el método de retención (*holdout method*) en el que el conjunto de datos dado se divide aleatoriamente en dos conjuntos: Conjuntos de entrenamiento y prueba (Rokach, L. & Maimon, O, 2015) .

Por lo general, dos tercios de los datos se consideran para el conjunto de entrenamiento y los datos restantes se asignan al conjunto de prueba.

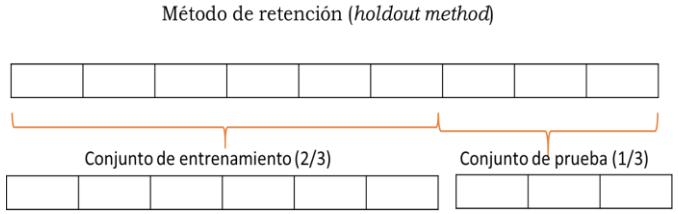
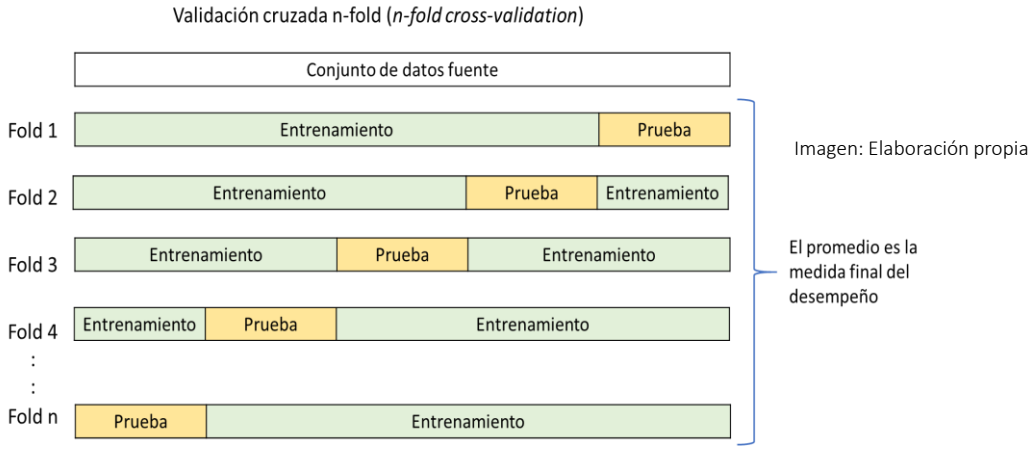


Imagen: Elaboración propia

El submuestreo aleatorio (*Random subsampling*) y la validación cruzada n-fold (*n-fold cross-validation*) son dos métodos comunes de remuestreo (Rokach, L. & Maimon, O, 2015):

- En el submuestreo aleatorio, los datos se dividen aleatoriamente varias veces en conjuntos de entrenamiento y pruebas disjuntos. Los errores obtenidos de cada partición se promedian.

- En la validación cruzada n-fold, los datos se dividen aleatoriamente en n subconjuntos mutuamente excluyentes de aproximadamente el mismo tamaño. Un inductor es entrenado y probado n veces; cada vez se prueba en uno de los k pliegues (fold) y se entrena utilizando los n-1 pliegues (fold) restantes.

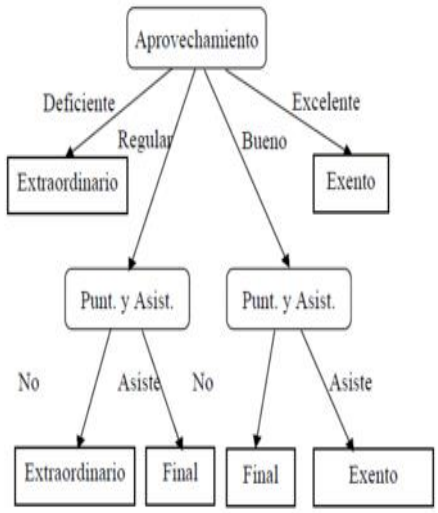


En este caso n vale 5, ya que se dividió el conjunto de datos en 5 partes (fold1, fold2, fold3, fold4, fold5), por citar:
Modelo 1: entrenado en fold1+fold2+fold3+fold4 y probado en fold5
Modelo 2: entrenado en fold1+fold2+fold3+fold5 y probado en fold4
Modelo 3: entrenado en fold1+fold2+fold4+fold5 y probado en fold3
Modelo 4: entrenado en fold1+fold3+fold4+fold5 y probado en fold2
Modelo 5: entrenado en fold2+fold3+fold4+fold5 y probado en fold1

Algoritmo ID3

Rokach & Maimon (2015) y Bhumika, Aditya, Akshay, Arpit & Naresh (2017) establecen que el ID3 tiene las siguientes características:

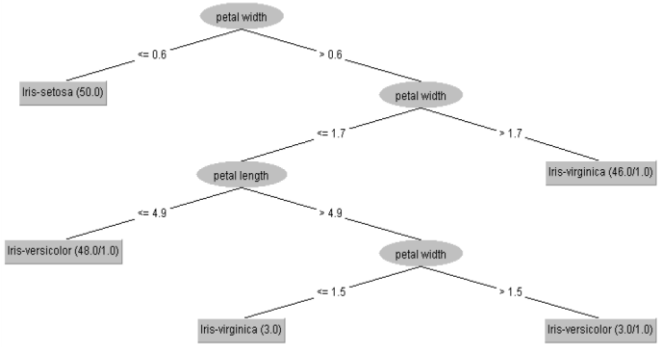
- Es un algoritmo desarrollado por Ross Quinlan.
- Sólo acepta atributos categóricos
- Usa la ganancia de información como criterio de división.
- Deja de crecer cuando:
  - Todas las instancias pertenecen a un solo valor de una característica objetivo o
  - Cuando la mejor ganancia de información no es mayor que cero.
- No aplica ningún procedimiento de poda.
- No maneja atributos numéricos o valores faltantes.



Algoritmo C4.5

Rokach & Maimon (2015) y Bhumika et al. (2017) establecen que el algoritmo C4.5 tiene las siguientes características:

- Es una evolución de ID3.
- Fue desarrollado por Ross Quinlan.
- Puede manejar atributos numéricos.
- Utiliza la relación de ganancia como criterio de división.
- Es n-ario con valores discretos y binario con datos continuos.
- La poda basada en errores se realiza después de la fase de crecimiento.



- La división termina cuando el número de instancias a dividir está por debajo de un cierto umbral.

Árbol CART

El árbol CART (Bhumika et al. (2017).

- Significa árboles de clasificación y regresión (*Classification And Regression Trees*).
- Fue presentado por Breiman en 1984.
- El algoritmo CART construye árboles de clasificación y regresión.
- CART construye el árbol de clasificación mediante la división **binaria** del atributo.
- El índice de Gini se usa para seleccionar el atributo de división
- Permite datos de atributos continuos y nominales.