Big Data, Machine Learning & Business Intelligence

Por: Carlos Carreño

ccarreno@cienciadedatos.es

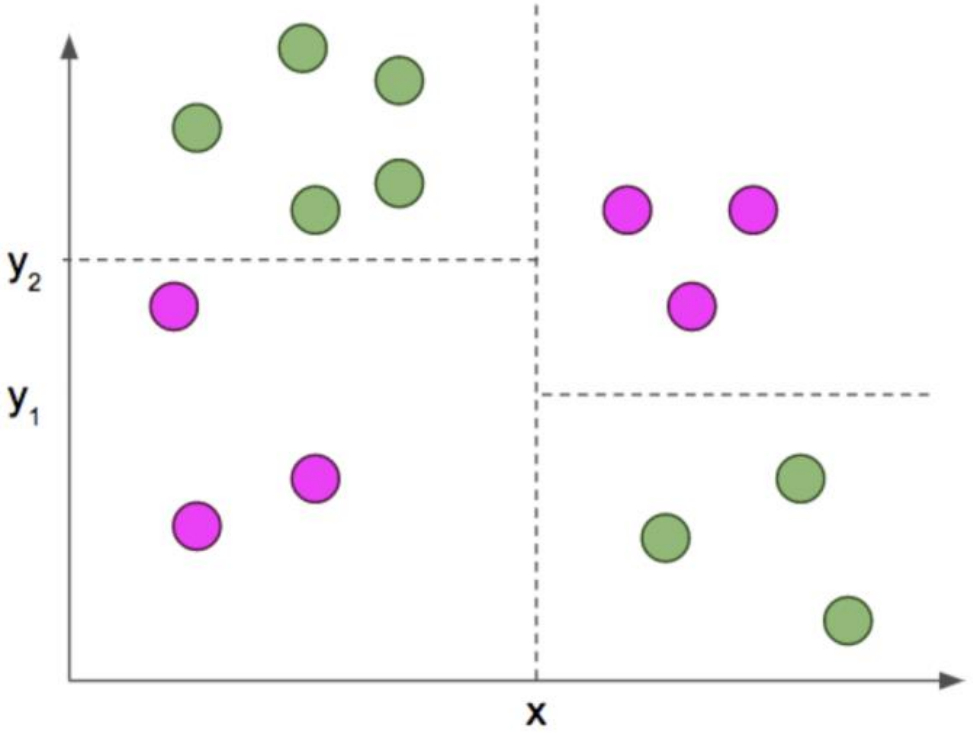
Arboles de Decision

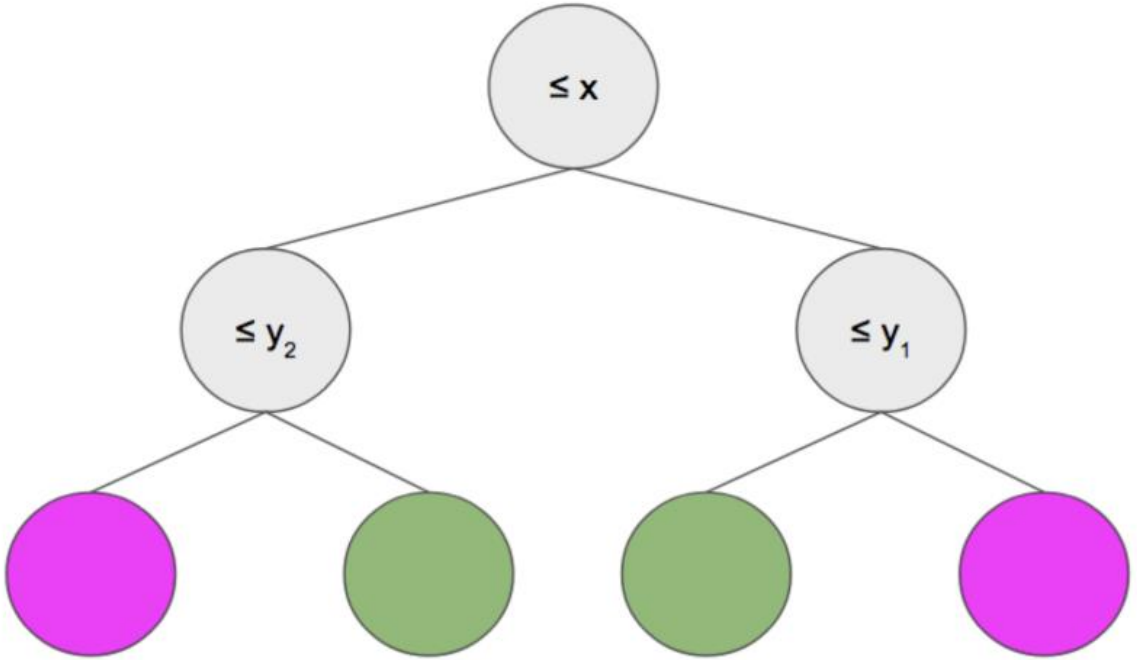
Introducción

Arboles de Decision

- El enfoque *classification and regression tree (CART)* fue desarrollado por Breiman et al. (1984).
- Son un tipo de algoritmos de aprendizaje supervisado (i.e., existe una variable objetivo predefinida).
- Principalmente usados en problemas de clasificación.
- Las variables de entrada y salida pueden ser categóricas o continuas.
- Divide el espacio de predictores (variables independientes) en regiones distintas y no sobrepuestas

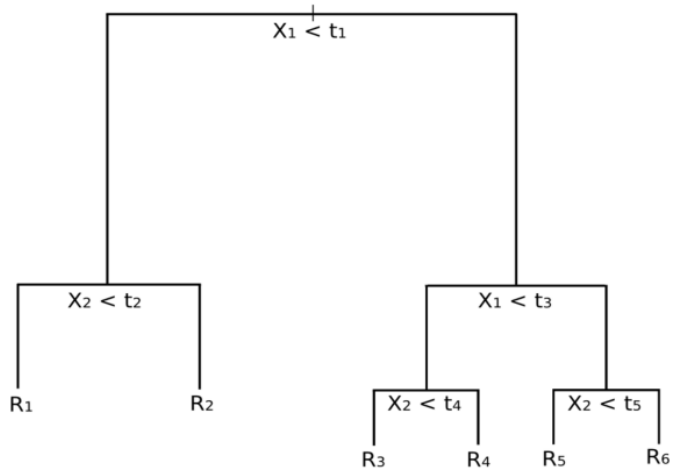
… continua



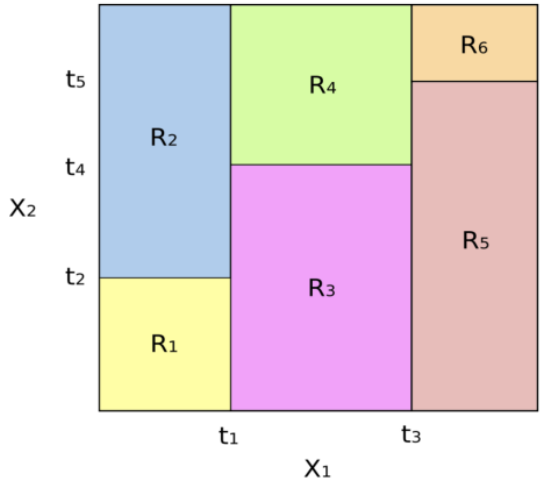


...

- Generalización



A Decision Tree with six separate regions

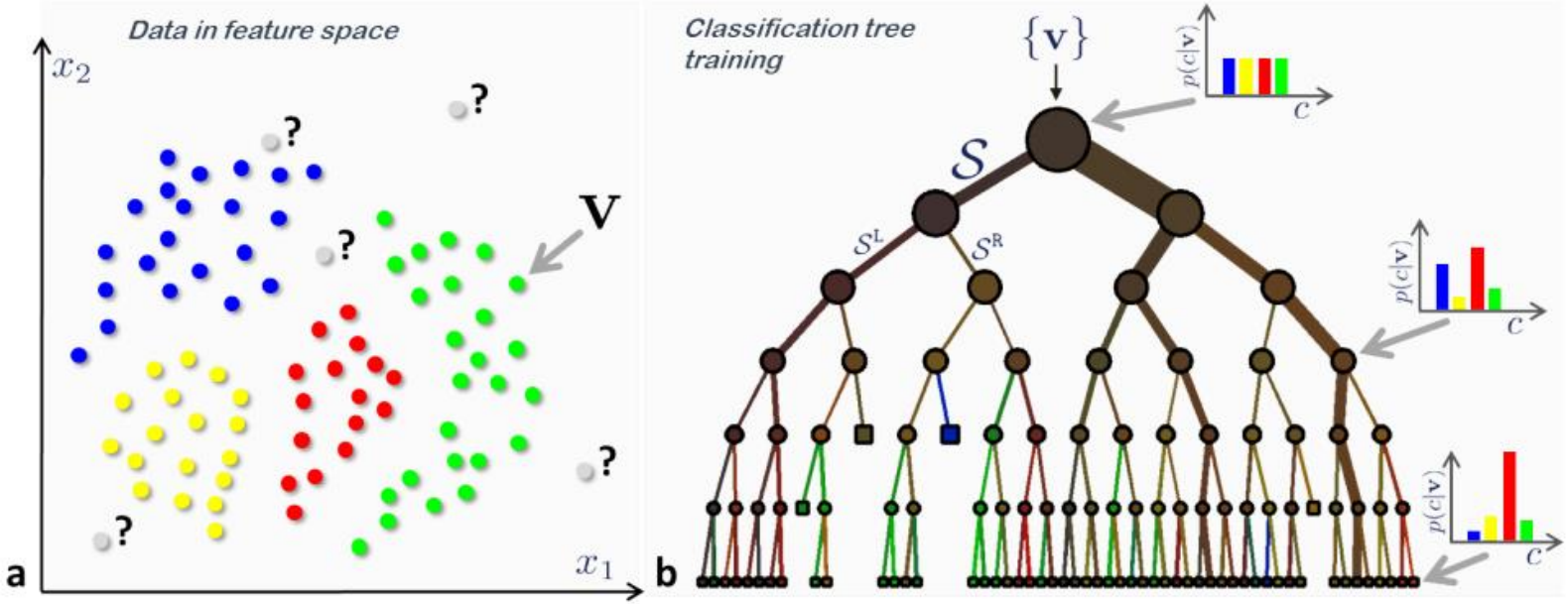


The resulting partition of the subset of $\mathbb{R}^2$ into six regional "blocks"

$$\text{RSS} = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y_{R_m}})^2$$

… continua

- Se divide la población o muestra en conjuntos homogéneos basados en la variable de entrada más significativa.
- La construcción del árbol sigue un enfoque de división binaria recursiva (top-down greddy approach). Greedy -> analiza la mejor variable para ramificación sólo en el proceso de división actual.



Como se ramifica un Arbol

- La decisión de hacer divisiones estratégicas afecta altamente la precisión del árbol.
- Los criterios de decisión son diferentes para árboles de clasificación y regresión.
- Existen varios algoritmos para decidir la ramificación. **Indice Gini**, Chi Cuadrado, Ganancia de la información y Reducción en la varianza
- La creación de subnodos incrementa la homogeneidad de los subnodos resultantes. Es decir, **la pureza del nodo se incrementa respecto a la variable objetivo.**
- Se prueba la división con todas las variables y se escoge la que produce subnodos más homogéneos.

Indice Gini

- “Si seleccionamos aletoriamente dos items de una población, entonces estos deben ser de la misma clase y la probabilidad de esto es 1 si la población es pura”.
- Variable objetivo categórica: “*Success*” o “*Failure*”
- Solo divisiones binarias
- A mayor valor de índice Gini, mayor la homogeneidad
- CART (Classification and Regression Tree) usa el método de Gini para la división binaria.

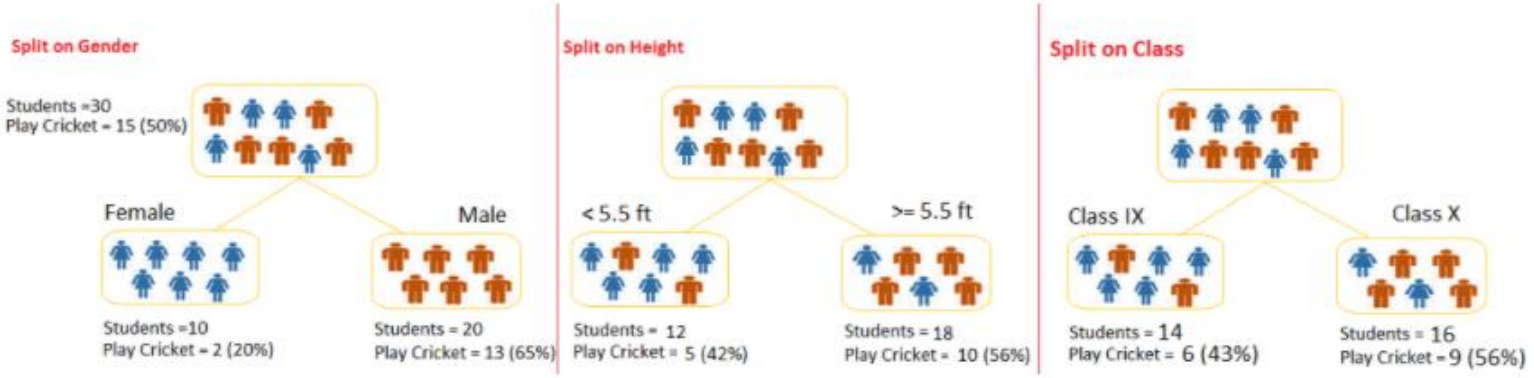
Calculo del Indice Gini

- Calcular Gini para los subnodos usando la fórmula de la suma de los cuadrados de probabilidad para *success* y *failure* **(p^2 + q^2)**.
- Calcular Gini para la división usando score Gini ponderado para cada nodo de la división.

Ejemplo: Calculo del Indice Gini

- 30 estudiantes
- 3 variables: Género (hombre/mujer), Clase (IX/X) y Altura (5 a 6 pies).
- 15 estudiantes juegan cricket en su tiempo libre
- Crear un modelo para predecir quien jugará cricket
- Segregar estudiantes basados en todos los valores de las 3 variables e identificar aquella variable que crea los conjuntos más homogéneos de estudiantes y que a su vez son heterogéneos entre ellos.

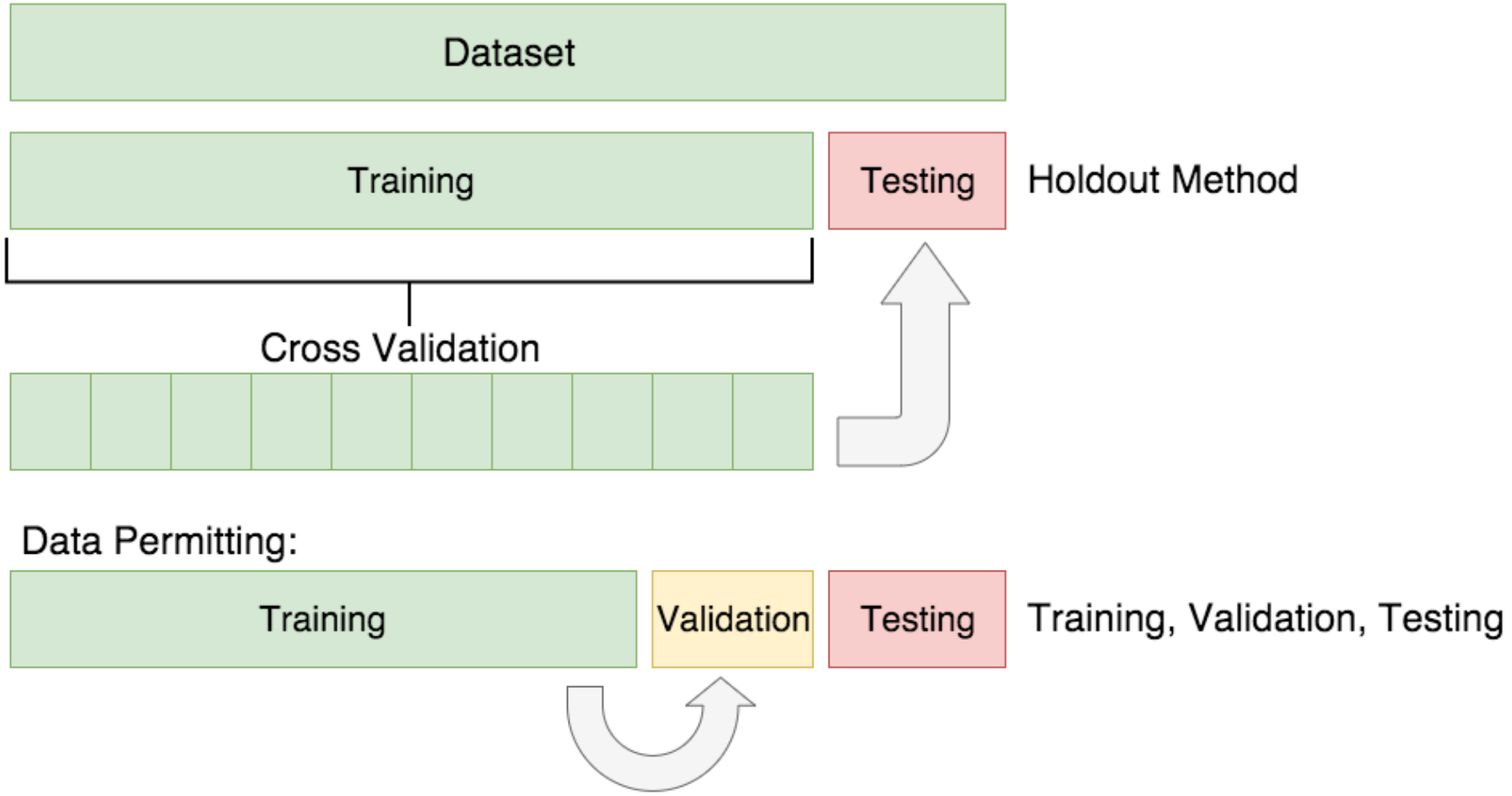
...continua



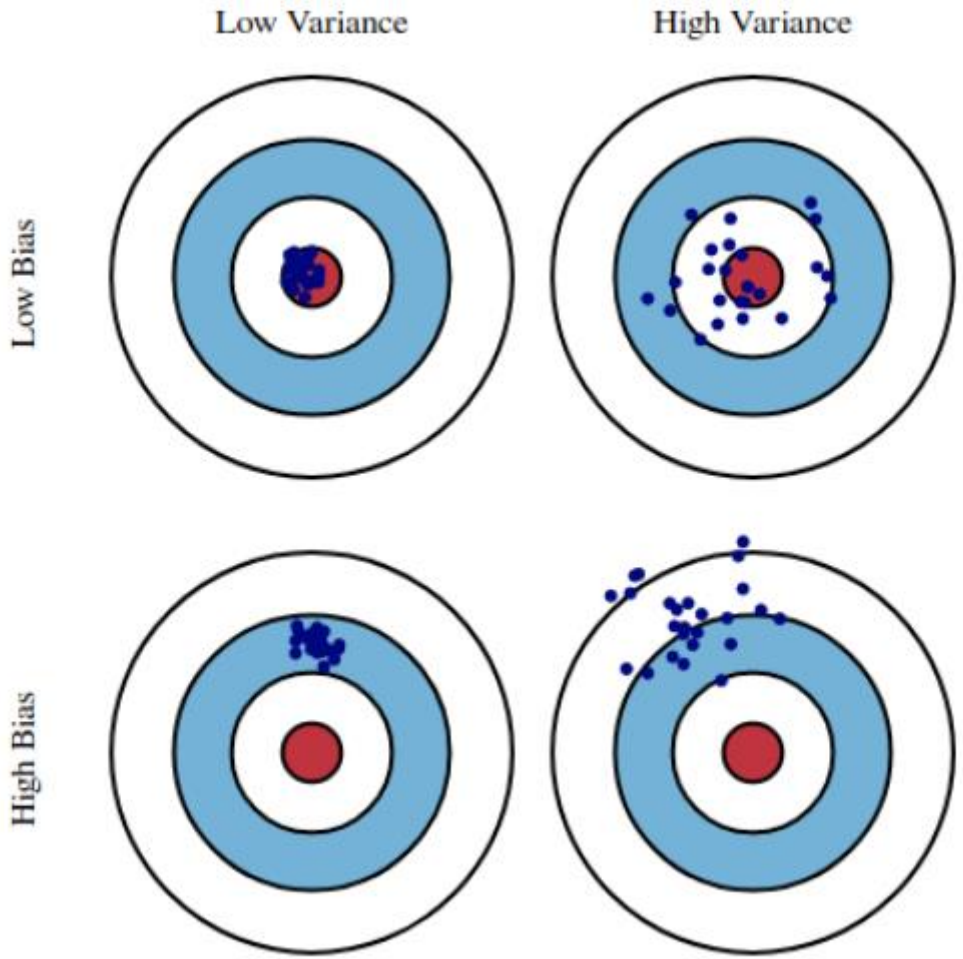
Género		**Clase**	
Mujer	(0.2)^2 + (0.8)^2 = 0.68	IX	(0.43)^2 +(0.57)^2 = 0.51
Hombre	(0.65)^2 + (0.35)^2 = 0.55	X	(0.56)^2 +(0.44)^2 = 0.51
Pond.	(10/30)*0.68* + *(20/30)*0.55 = **0.59**	Pond.	(14/30)*0.51* + *(16/30)*0.51 = **0.51**

- ¿Cuál es el resultado para la variable “Altura”
- ¿Cuál es el variable que se debe escoger para la primera división?

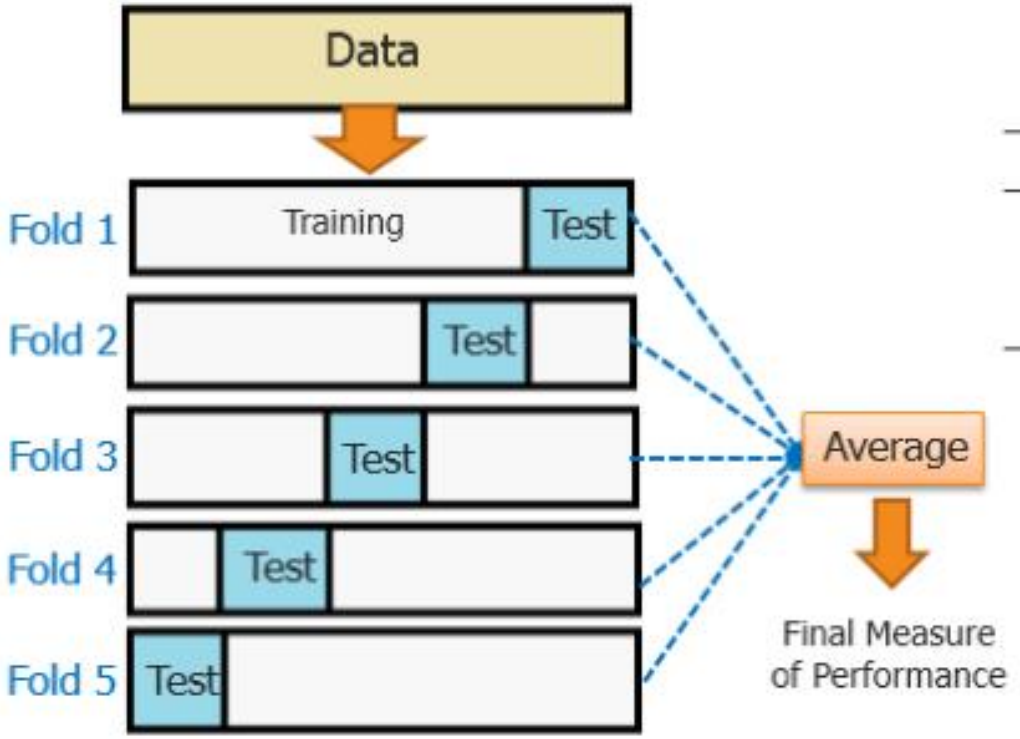
Conjunto de Datos de Entrenamiento y de Prueba



Sesgo y Varianza



Validación Cruzada (Cross-Validation)



→ Technique to validate models/classifiers

→ Method to estimate how accurately the model generalizes to unseen data i.e., how well it performs/predicts

→ K-fold CV

- » Most popular
- » k is typically set to 10
- » Every sample/record is used both in training and test sets

Matriz de Rendimiento (Matriz de Confusión)

- A partir de la matriz de confusión es posible calcular un medida de rendimiento del modelo.
- La matriz de confusión es utilizada en casos de clasificación.

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Accuracy test

- Accuracy test a partir de la matriz de confusión o tabla de contingencia
- Accuracy = TP + TN / (TP + TN + FP + FN)
- Proporción de las instancias predichas correctamente TP and TN sobre la suma total de elementos evaluados.

```
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for test', accuracy_Test))
```

```
## [1] "Accuracy for test 0.808612440191388"
```

Clasificación con Arboles de Decisión

- Paso 1: Importar los datos
- Paso 2: Limpiar los datos
- Paso 3: Crear los conjuntos de entrenamiento y test
- Paso 4: Construir el modelo
- Paso 5: Hacer la predicción
- Paso 6: Medir el rendimiento del modelo
- Paso 7: Ajustar los hyper-parámetros

Laboratorio

- Lab 6: Modelo de Clasificación con Arboles de Decisión para predecir la supervivencia de pasajeros del Titanic
  - ❑El propósito del laboratorio es a partir del conjunto de datos Titanic (Titanic.csv) es predecir que personas son más propensas a sobrevivir la colisión con el iceberg. El conjunto de datos contiene 13 variables y 1309 observaciones. Finalmente, este se encuentra ordenado por la variable X que es el índice.