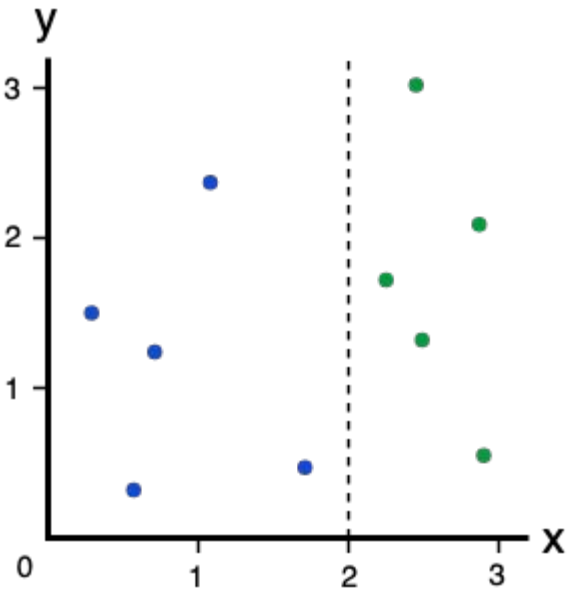
Aprendizaje Automático

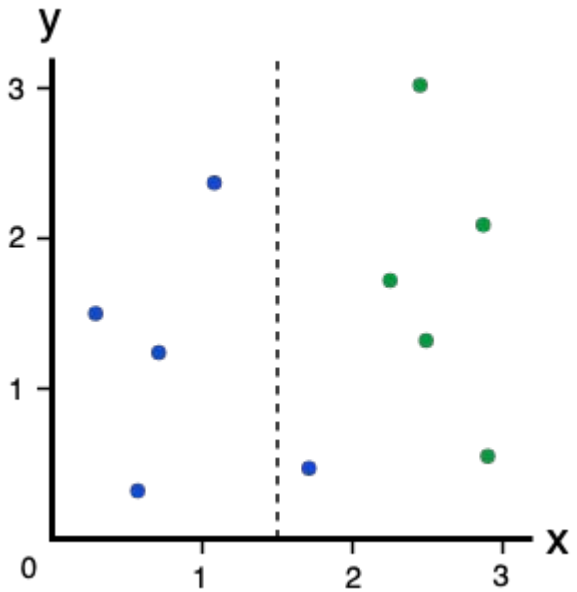
Árboles de Decisión

Algoritmo: CART (Classification and Regression Trees)

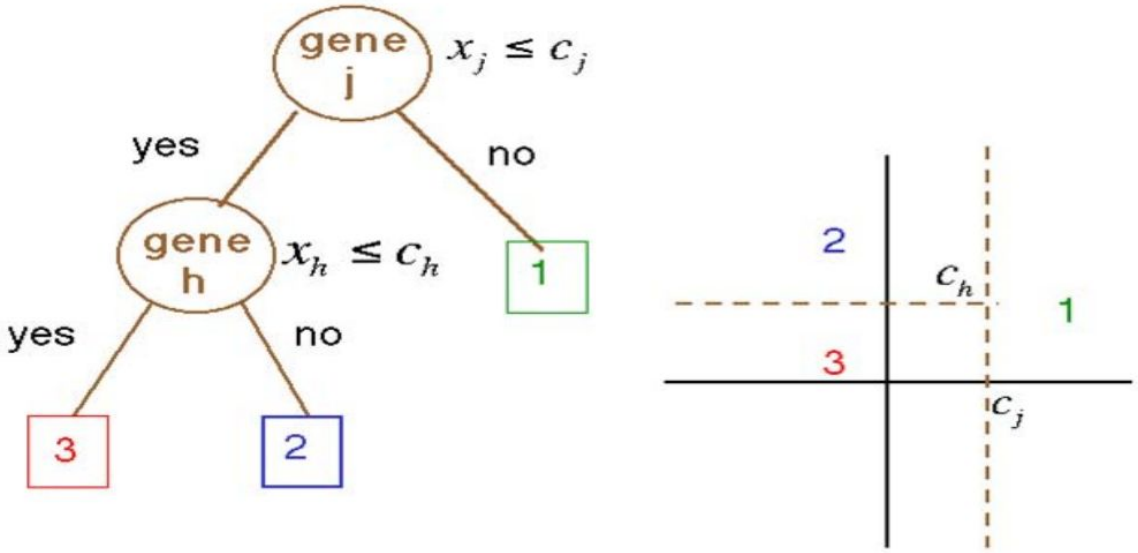
- Árboles de clasificación: predicen categorías de objetos.
- Árboles de regresión: predicen valores continuos.
- Partición binaria recursiva.
- En cada iteración se selecciona la variable predictiva y el punto de separación que mejor reduzcan la **'impureza'**.

Cuál es el mejor corte?





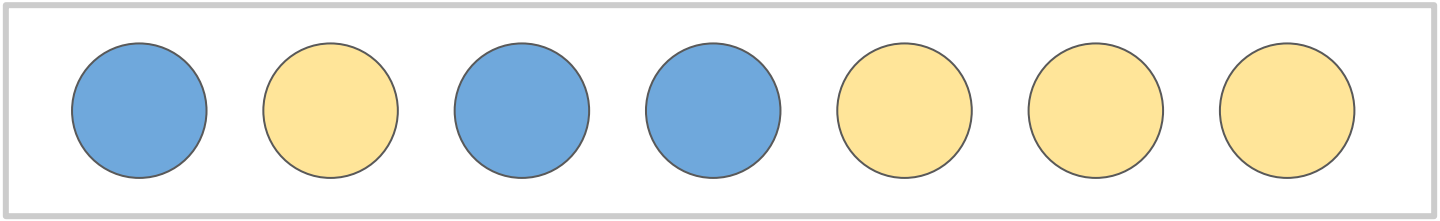
Clasificación CART ejemplo



Índice de impureza Gini

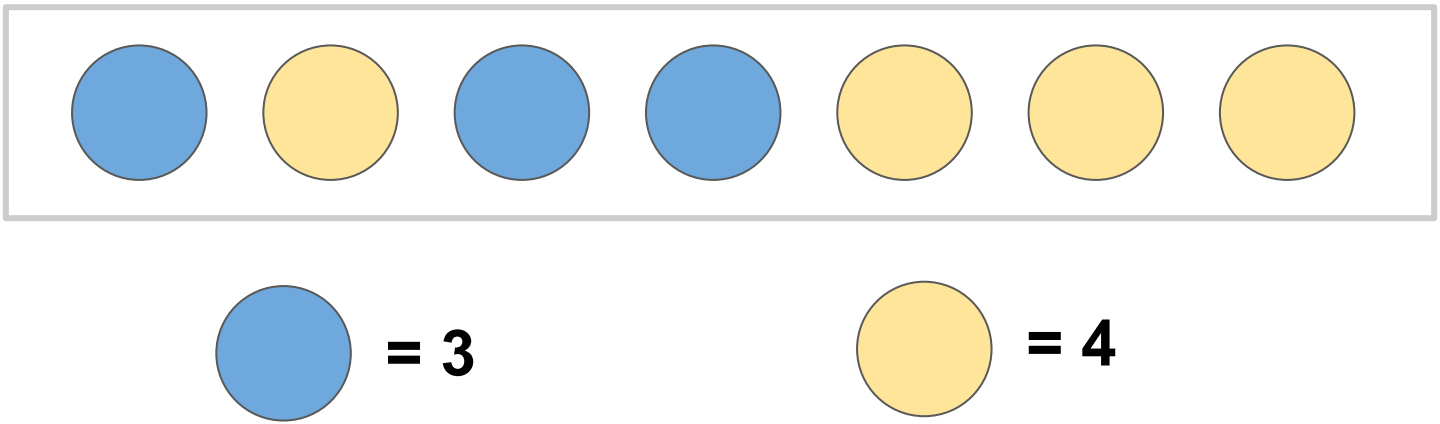
$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$

Gini impurity



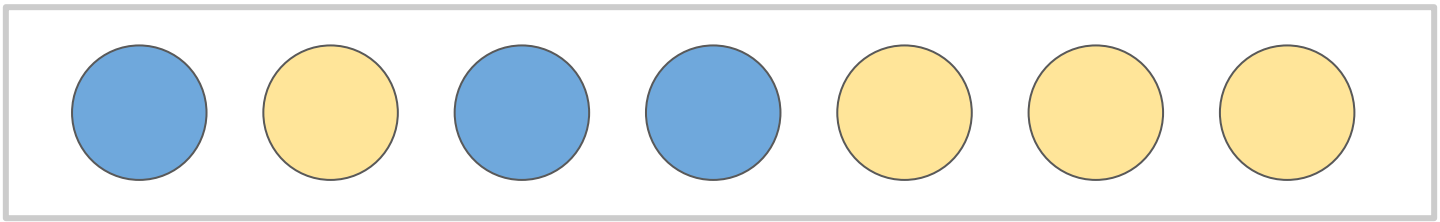
$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$

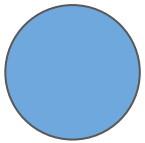
Gini impurity



$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$

Gini impurity



 = 3

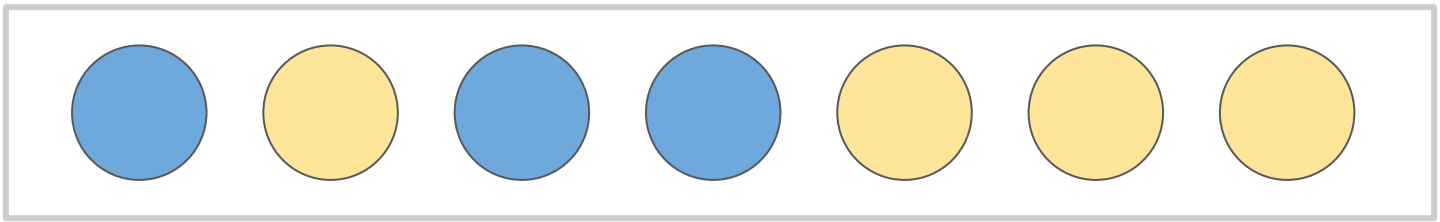
 = 4

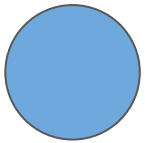
p() = 3/7 = 0.43

p() = 4/7 = 0.57

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$

Gini impurity



 = 3

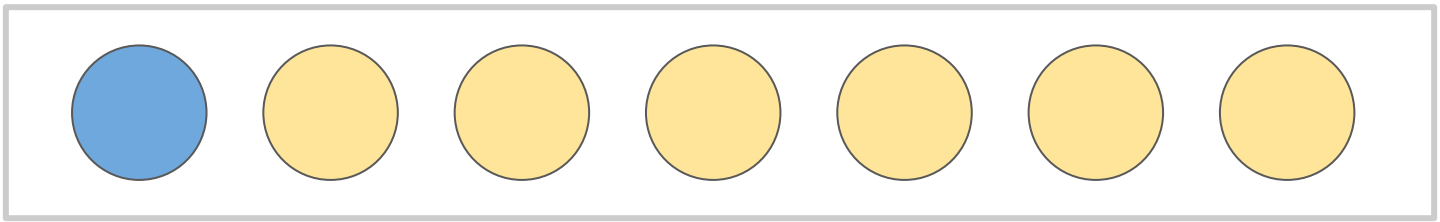
 = 4

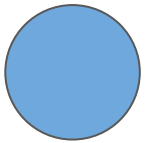
p() = 3/7 = 0.43

p() = 4/7 = 0.57

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$ = 1 - (0.43^2 + 0.57^2) = 0.49

Gini impurity



 = _

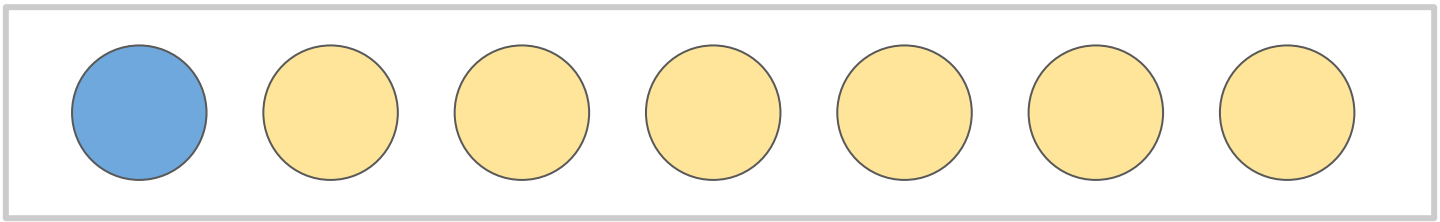
 = _

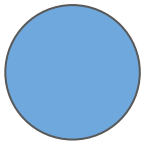
p() = _/7 = ?

p() = _/7 = ?

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$ = 1 - (_^{2} + _^{2}) = ??

Gini impurity



 = 1

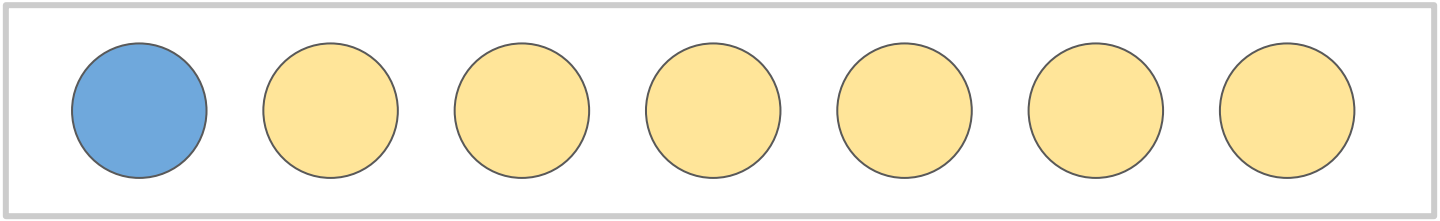
 = 6

p() = 1/7 = 0.14

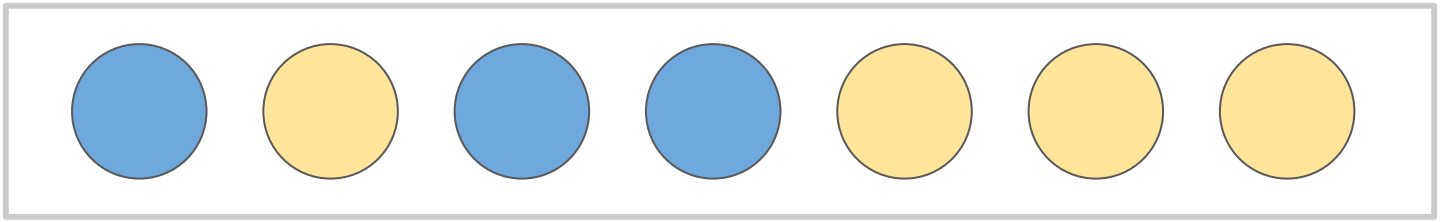
p() = 6/7 = 0.86

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$ = 1 - (0.14^2 + 0.86^2) = 0.24

Gini impurity

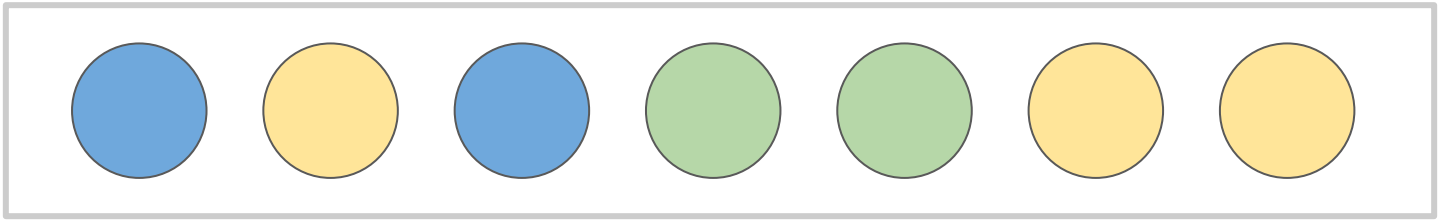


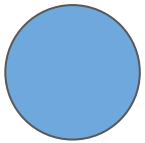
$$I_G(p) = \mathbf{0.24}$$



$$I_G(p) = \mathbf{0.49}$$

Gini impurity



 = _

 = _

 = _

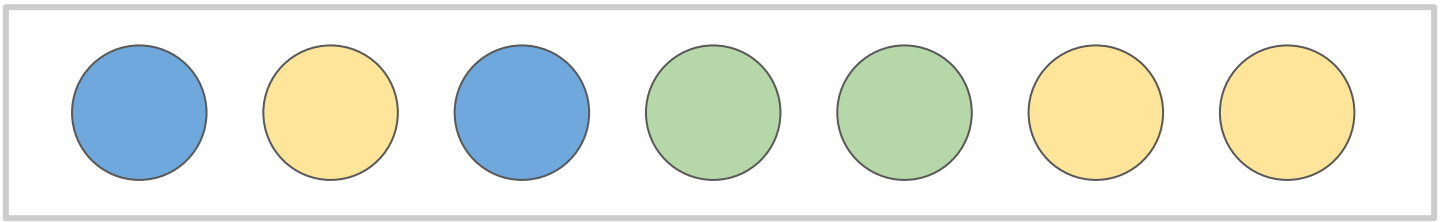
p() = _/7 = ?

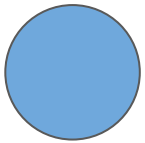
p() = _/7 = ?

p() = _/7 = ?

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$ = 1 - (__^{2} + __^{2} + __^{2}) = ??

Gini impurity



 = 2

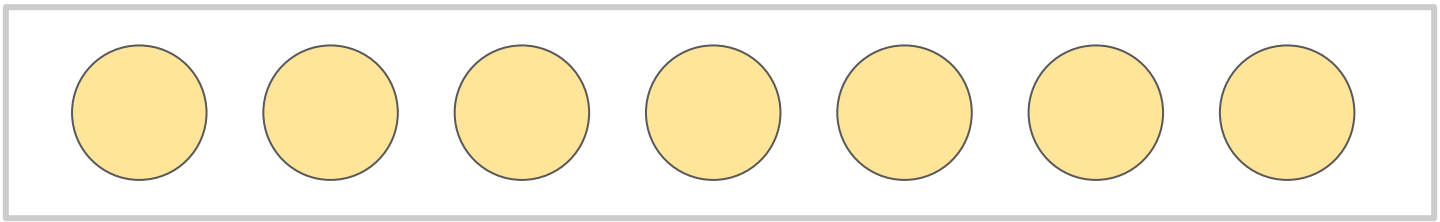
 = 3

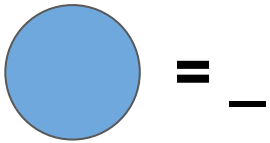
 = 2

p() = 2/7 = 0.29 p() = 3/7 = 0.42 p() = 2/7 = 0.29

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$ = 1 - (0.29^2 + 0.42^2 + 0.29^2) = 0.66

Gini impurity



 = _

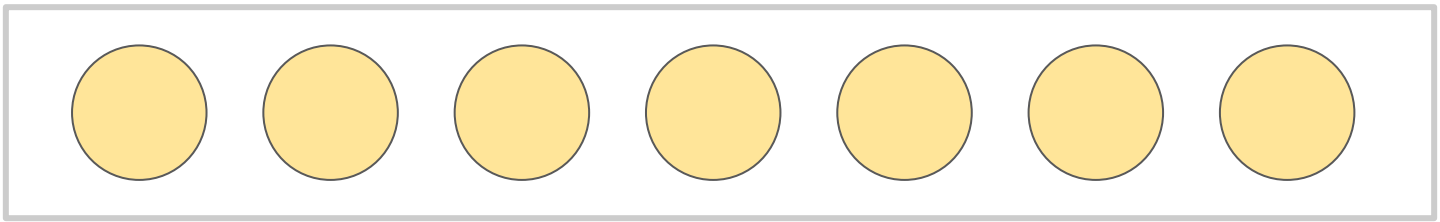
 = _

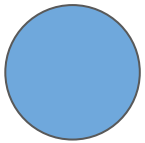
p() = _/7 = ?

p() = _/7 = ?

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$ = 1 - (_^{2} + _^{2}) = ??

Gini impurity



 = 0

 = 7

p() = 0/7 = 0

p() = 7/7 = 1

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$ = 1 - (0^2 + 1^2) = 0

Índice de impureza de Gini

$$G(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) G(C|A_{ij})$$

$$G(C|A_{ij}) = -\sum_{k=1}^{J} p(C_k|A_{ij}) p(\neg C_k|A_{ij}) =$$

$$= 1 - \sum_{k=1}^{J} p^2(C_k|A_{ij})$$

Índice de impureza de Gini

- A_i es el atributo para ramificar el árbol.
- M_i es el número de valores diferentes del atributo A_i.
- $p(A_{ij})$ es la probabilidad de que A_i tome su j-ésimo valor $(1 <= j <= M_i)$.

Índice de impureza de Gini

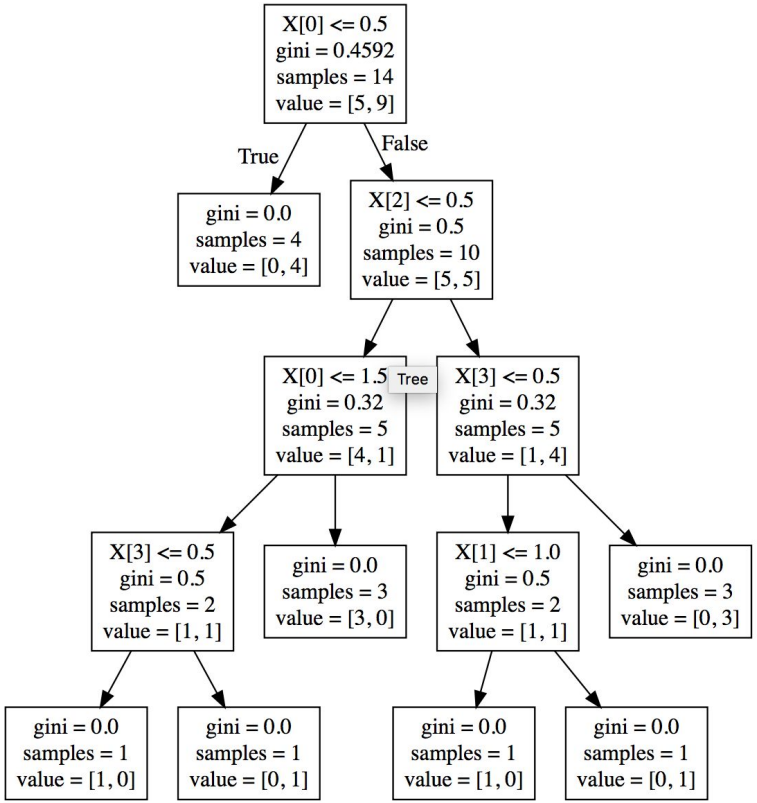
- $p(C_k|A_{ij})$ es la probabilidad de que un ejemplo pertenezca a la clase C_k cuando su atributo A_i toma su j-ésimo valor.
- $p(\neg C_k|A_{ij})$ es 1 - $p(C_k|A_{ij})$.
- Este índice es utilizado como una medida de impureza de la información al igual que la entropía.

Ejemplo, juego de Tenis

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

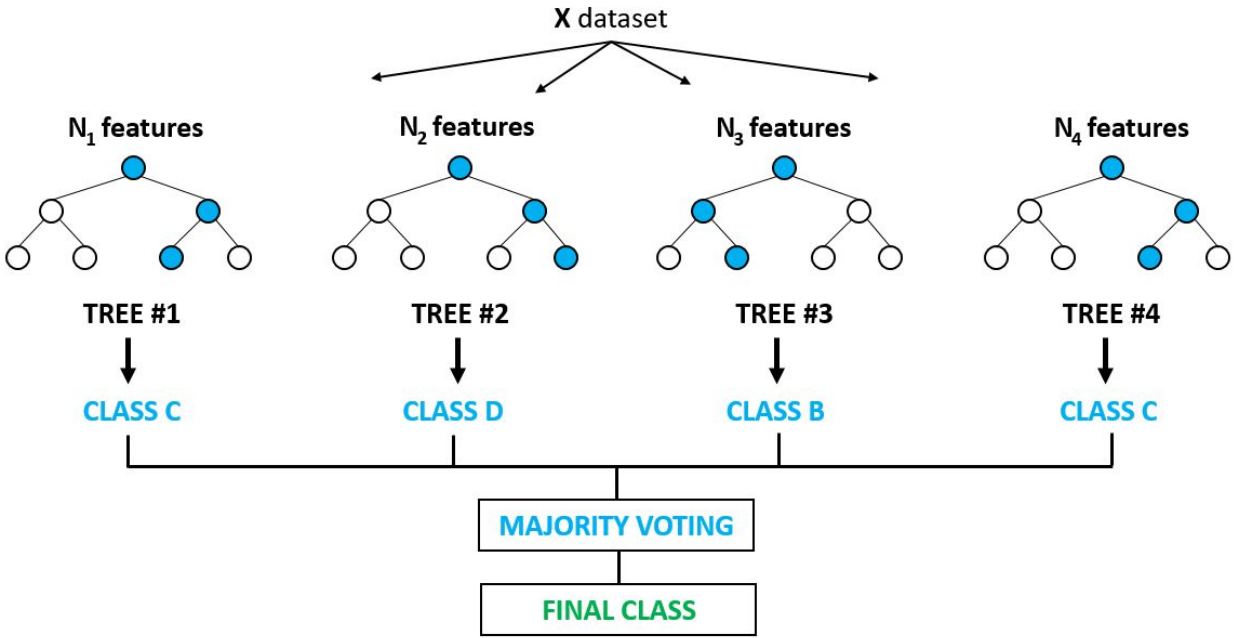
Day	Outlook	Temperature	Humidity	Wind	Play
D1	2	1	0	1	No
D2	2	1	0	0	No
D3	0	1	0	1	Yes
D4	1	2	0	1	Yes
D5	1	0	1	1	Yes
D6	1	0	1	0	No
D7	0	0	1	0	Yes
D8	2	2	0	1	No
D9	2	0	1	1	Yes
D10	1	2	1	1	Yes
D11	2	2	1	0	Yes
D12	0	2	0	0	Yes
D13	0	1	1	1	Yes
D14	1	2	0	0	No

Árbol de decisión - Clasificación Play tennis



Day	Outlook	Temperature	Humidity	Wind	Play
D1	2	1	0	1	No
D2	2	1	0	0	No
D3	0	1	0	1	Yes
D4	1	2	0	1	Yes
D5	1	0	1	1	Yes
D6	1	0	1	0	No
D7	0	0	1	0	Yes
D8	2	2	0	1	No
D9	2	0	1	1	Yes
D10	1	2	1	1	Yes
D11	2	2	1	0	Yes
D12	0	2	0	0	Yes
D13	0	1	1	1	Yes
D14	1	2	0	0	No

Random Forest



Sobreentrenamiento

- Se debe evitar el sobreentrenamiento
  - Parar de crecer el árbol temprano.
  - Postprocesamiento del árbol (poda)

  Cómo?
  - Usar un conjunto de ejemplos de validación
  - Usar estadísticas