

- The algorithm goes on splitting the node with a feature k and a threshold t_k for which the following cost function is minimized

$$J(k, t_k) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}}$$

Where

m_{left} : number of observations in the left sub-setted node

G_{left} : Impurity of the left sub-setted node

m_{right} : number of observations in the right sub-setted node

G_{right} : Impurity of the right sub-setted node

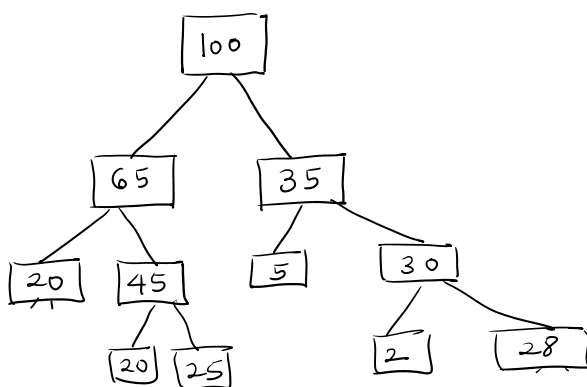
	$J()$
X_1, t_{11}	—
X_1, t_{12}	—
X_1, t_{13}	
\vdots	
X_1, t_{1n}	
X_2, t_{22}	← min
X_2, t_{23}	
\vdots	
X_2, t_{2m}	

Features
 $X_1 \ X_2 \ \dots \ X_n$
 —
 —
 —
 —

samples \equiv observations

min_samples_split : int or float, default=~~2~~/~~5~~/~~10~~/~~20~~ 30

The minimum number of samples required to split an internal node





min_samples_leaf : int or float, default=10

The minimum number of samples required to be at a leaf node.

