Formula for Gini index: $1 - \Sigma(p_i^2)$

923446482
Po-Han Chen

Node 1: Hight blood pressure

"Yes" branch

$5+2=7$

proportion of heart disease $(p_1) = \frac{5}{7}$

proportion of no heart disease $(p_2) = \frac{2}{7}$

$\text{Gini}_{yes} = 1 - \left(\left(\frac{5}{7}\right)^2 + \left(\frac{2}{7}\right)^2\right) = 1 - \left(\frac{25+4}{49}\right) = 1 - \frac{29}{49} = \frac{20}{49}$

"No" branch

$1+6=7$

proportion of heart disease $(p_1) = \frac{1}{7}$

proportion of no heart disease $(p_2) = \frac{6}{7}$

$\text{Gini}_{no} = 1 - \left(\left(\frac{1}{7}\right)^2 + \left(\frac{6}{7}\right)^2\right) = 1 - \left(\frac{1+36}{49}\right) = \frac{12}{49}$

Node two: Over 50 years old

"Yes" branch

$4+3=7$

proportion of heart disease $(p_1) = \frac{4}{7}$

proportion of no heart disease $(p_2) = \frac{3}{7}$

$\text{Gini}_{yes} = 1 - \left(\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right) = 1 - \left(\frac{16+9}{49}\right) = \frac{24}{49}$

"No" branch

$2+5=7$

proportion of heart disease $(p_1) = \frac{2}{7}$

proportion of no heart disease $(p_2) = \frac{5}{7}$

$\text{Gini}_{no} = 1 - \left(\left(\frac{2}{7}\right)^2 + \left(\frac{5}{7}\right)^2\right) = 1 - \left(\frac{4+25}{49}\right) = \frac{20}{49}$

High blood pressure split:

$Gini(Yes) = \frac{20}{49}$, $Gini(No) = \frac{12}{49}$

Over 50 years old split:

$Gini(Yes) = \frac{24}{49}$, $Gini(No) = \frac{20}{49}$

The Gini index for "High blood pressure" node is lower. Therefore, I'll pick "High blood pressure" as the better split for the decision tree.