

* Problem 2: Decision Trees based on GINI Index.

A	B	Class	Label
		+	-
T	T	0	20
T	F	20	10
F	T	15	0
F	F	0	35

①

$$* \text{GINI}(t) = 1 - \sum_j [p(j|t)]^2$$

$$* \text{GINI}_{\text{split}} = \sum_{i=1}^k \frac{n_i}{n} \text{GINI}(i)$$

$$\rightarrow P(+)=35/100 \text{ and } P(-)=65/100$$

→ Parent

$$\therefore \text{GINI}(P) = 1 - (35/100)^2 - (65/100)^2$$

$$= 1 - 0.1225 - 0.4225$$

$$= 0.455$$

* For attribute A, count matrix is,

A	+	-
T	20	30
F	15	35

$$\rightarrow \text{GINI}(T) = 1 - (20/50)^2 - (30/50)^2$$

$$= 1 - 0.16 - 0.36$$

$$= 0.48$$

$$\rightarrow \text{GINI}(F) = 1 - (15/50)^2 - (35/50)^2$$

$$= 1 - 0.09 - 0.49$$

$$= 0.42$$

$$\therefore \text{GINI}_{\text{split}} (=A) = \left(\frac{50}{100}\right)(0.48) + \left(\frac{50}{100}\right)(0.42)$$

$$= 0.24 + 0.21$$

$$= \boxed{0.45}$$

* for attribute B, count matrix is,

B	+	-
T	15	20
F	20	45

$$\rightarrow \text{GINI}(T) = 1 - \left(\frac{15}{35}\right)^2 - \left(\frac{20}{35}\right)^2$$

$$= 1 - 0.1837 - 0.3265$$

$$= 0.4898$$

$$\rightarrow \text{GINI}(F) = 1 - \left(\frac{20}{65}\right)^2 - \left(\frac{45}{65}\right)^2$$

$$= 1 - 0.0947 - 0.4793$$

$$= 0.426$$

$$\therefore \text{GINI}_{\text{split}} (=B) = \left(\frac{35}{100}\right)(0.4898) + \left(\frac{65}{100}\right)(0.426)$$

$$= 0.17143 + 0.2769$$

$$= \boxed{0.4483}$$

\Rightarrow So, attribute 'B' will be chosen as first splitting attribute as it has the least GINI index

② For attribute A, count and cost metric is

			Cost matrix	Attribute Value		
A	+	-		T	F	
T	20	30	Actual	+	-1	100
F	15	35	Class	-	0	-10

⇒ Cost of splitting A is,

$$\text{Cost}_{(\text{split}=A)} = (20)(-1) + (30)(0) + (15)(100) + (35)(-10)$$

$$= -20 + 0 + 1500 - 350$$

$$= 1130$$

* → For attribute B, count and cost metric is

			Cost	Attribute Value		
B	+	-	Matrix	T	F	
T	15	20	Actual	+	-1	100
F	20	45	Class	-	0	-10

⇒ Cost of splitting B is,

$$\text{Cost}_{(\text{split}=B)} = (15)(-1) + (20)(0) + (20)(100) + (45)(-10)$$

$$= -15 + 0 + 2000 - 450$$

$$= 1535$$

* ⇒ Hence, we will choose attribute 'A', as it has smaller cost of splitting.