CS685A

ADITYA RAGHUWANSHI 170052

D.T.

i) Dataset: A(60) B(30) C(10)

$p_A = \frac{60}{100} = 0.6$ $p_B = \frac{30}{60+30+10} = 0.3$ $p_C = \frac{10}{100} = 0.1$

$$gini(D) = 1 - \sum_{i=1}^{3} p_i^2 = 1 - (0.6^2 + 0.3^2 + 0.1^2)$$
$$= 1 - 0.46$$
$$= \boxed{0.54}$$ Ans

ii) D_1: A(45) B(25) C(5) D_2: A(15) B(5) C(5)

~~gini = 1 -~~ $p_1 = 0.6$ $p_2 = 0.33$ $p_3 = 0.06$ $p_1 = 0.6$ $p_2 = p_3 = 0.2$

$gini(D_1) = 1 - (0.6^2 + 0.33^2 + 0.06^2)$ $gini(D_2) = 1 - (0.6^2 + 0.2^2 + 0.2$

$info(D) = -0.6 \log(0.6) - 0.3 \log(0.3) - 0.1 \log(0.1)$

$Info_S(D) = \left(\frac{75}{100}\right) info(D_1) + \frac{25}{100} info(D_2)$

$info(D_1) = -0.6 \log 0.6 - 0.33 \log 0.33 - 0.06 \log 0.06$
$info(D_2) = -0.6 \log 0.6 - 0.2 \log(0.2) - 0.2 \log(0.2)$

$gain_S(D) = info(D) - info_S(D)$