Quer. 01

The formula for entropy is:

$$H(S) = -P_1 \log_2 P_1 - P_2 \log_2 P_2$$

where P_1 & P_2 are the proportion of each class

Low Risk: 4 instances (ID: 1, 3, 5, 7)

High Risk: 4 instances (ID: 2, 4, 6, 8)

$$H(S) = -\left(\frac{4}{8} \log_2 \frac{4}{8} + \frac{4}{8} \log_2 \frac{4}{8}\right)$$

$$H(S) = 1.0$$

Credit Score > 690

Low Risk: 4

High Risk: 0

$$H(S_1) = -\left(\frac{4}{8} \log_2 \frac{4}{4} + \frac{0}{4} \log_2 \frac{0}{4}\right) = 0$$

Credit Score < 650

Low Risk: 0

High Risk: 4

$$H(S_2) = -\left(\frac{0}{4} \log_2 \frac{0}{4} + \frac{4}{4} \log_2 \frac{4}{4}\right) = 0$$

Gain: $IG = 1.0 - 0 = 1.0$

Ques: 2

Lg. Age ≤ 34   1D. {1, 2, 4}

Rg. Age > 34   1D3 {3,5,6,7,8}

Variance = $\frac{1}{n} \Sigma (x_i - \bar{x})^2$

$\bar{x} = \frac{720 + 650 + 750 + 600 + 780 + 630 + 710 + 600}{8} = \frac{5430}{8} = 655$

$\sigma^2 = (720 - 655)^2 + (650 - 655)^2 + (750 - 655)^2 + (600 - 655)^2 + (780 - 655)^2 + (630 - 655)^2$
$+ (710 - 655)^2 + (600 - 655)^2$

over 8

$\sigma^2 = \frac{28600}{8} = 3475$

Compute the variance
Lg Age ≤ 34

subset {1, 2, 4}

Credit Scores: {720, 650, 600}

$\bar{x}_L = \frac{720 + 650 + 600}{3} = \frac{1970}{3} = 656.67$

$\sigma_L^2 = \frac{(720 - 656.67)^2 + (650 - 656.67)^2 + (600 - 656.67)^2}{3}$

$= \frac{(63.33)^2 + (-6.67)^2 + (-56.67)^2}{3}$

$= \frac{4000.89 + 44.49 + 3211.69}{3} = \frac{7257.07}{3} = 2419.02$

Compute Variance

L_R Age > 35

Subset {3, 5, 6, 7, 8}

Credit Scores {750, 780, 630, 710, 640}

$\bar{a}_R = \dfrac{750 + 780 + 630 + 710 + 640}{5} = 702$

Variance

$6^2_R = \dfrac{(750-702)^2 + (780-702)^2 + (630-702)^2 + (710-702)^2 + (640-702)^2}{5}$

$= \dfrac{(48)^2 + (78)^2 + (-72)^2 + (8)^2 + (-62)^2}{5}$

$= 3496$

weighted Variance:

weighted Variance $= \dfrac{3}{8} \times 2419.02 + \dfrac{5}{8} \times 3496$

$= 0.375 \times 2419.02 + 0.625 \times 3496$

$= 3092.13$

Compute Variance Reduction:

Variance Reduction $= 3475 - 3092.13$

$= 182.87$

	CreditScore	Education	Risk Level
	600	14	High
31	640	12	High
29	630	14	High
33	610	12	High

Age Range: 28-33

CreditScore: Between 600-630

Education: Mostly 12 or 14

Age	credit-line	Education (missing)
30	615	missing

T_2 has a high Probability of being High Risk

Ques: 04

Acsong
using linear regression

$$\hat{y} = \theta_0 + \theta_1 x$$

where

- $\hat{y}$ is the predicted Credit-score
- x is the input feature (Age)
- $\theta_0 = 500$ (initial intercept)
- $\theta_1 = 5$ (initial slope)

Predicted, $\hat{y}_i = \theta_{00\theta} \times Age_i$

Cost of linear regression using MSE

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{M} (\hat{y}_i - Y_i)^2 \quad \left[\begin{array}{l} m = 8 \text{ (number of training examples)} \\ Y_i \text{ is actual credit score} \end{array} \right]$$

ID	Age	Y	$h_\theta(x)$
1	35	720	$500 + 5(35) = 675$
2	28	650	$500 + 5(28) = 640$
3	45	750	$500 + 5(45) = 725$
4	31	600	$500 + 5(31) = 655$
5	52	780	$500 + 5(52) = 760$
6	29	630	$500 + 5(29) = 645$
7	42	710	$500 + 5(42) = 710$
8	33	640	$500 + 5(33) = 665$