Question 1.

The dataset contain 4 Low and 4 High RiskLevel.

=> $I_E([4;4]) = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1.$

Split at ~~650, we~~ Credit Score = 650, we have: 4

- left Subset: (CreditScore ≤ 650) = ID: [2, 4, 6, 8] : 4 high

Information gain $I_E([0;4]) = 0$

- Right Subset (Credit Score > 650) = ID: [1, 3, 5, 7] : 4 low

$I_E([4;0]) = 0$

$I_E([0;4],[4;0]) = 0$

=> Information gain of the split:

$IG = I_E([4;4]) ~~*~~ - I_E([0;4],[4;0]) = 1 - 0 = 1.$

~~IG hi~~ Then split at CreditScore = 650 is good because we can clearly ~~split two~~ classify two type of RiskLevel

Question 2

x = Credit Score

mean of Credit Score = $\bar{x} = 685$

variance of dataset:

$$V(S) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$= \frac{1}{8}\sum_{i=1}^{8}(x_i - 685)^2 = 3575$$

=

Split dataset at age = 35

=> left subset (age ≤ 35): [1, 2, 4, 6, 8]

~~$V(left) = \frac{1}{N_{left}}\sum_{i=1}^{N_{left}}$~~

$\bar{x}_{left} = 648$

$$V(S_{left}) = \frac{1}{N_{left}}\sum_{i=1}^{N_{left}}(x_i - x_{left})^2$$

$$= \frac{1}{5}\sum_{i=1}^{5}(x_i - 648) = 1576$$

- Right subset: (age > 35) [3, 5, 7]

$\bar{x}_{Right} = 746{,}67$

$$V(S_{right}) = \frac{1}{N}\frac{1}{3}\sum_{i=3}^{3}(x_i - 746{,}67)^2 = 822{,}22$$

Weighted Variance after Split:

$$V_1(S) = \cancel{V}\frac{N_{left}}{N}V(S_{left}) + \frac{N_{right}}{N}V(S_{right})$$

$$= \frac{5}{8}\cdot 1576 + \frac{3}{8}\cdot 822{,}22 = 1293{,}33\cancel{7}$$

Variance Reduction = $V(S) - V_1(S) = 3575 - 1293{,}33$

$= \cancel{2281{,}67}\ 2281{,}67$