

Homework 5

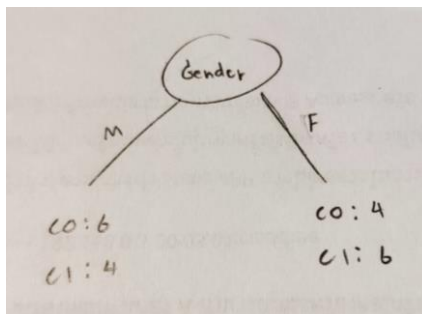
Exercise 1 (Feature selection):

(a) Compute the Gini index for the overall collection of training examples.

c0	10
c1	10

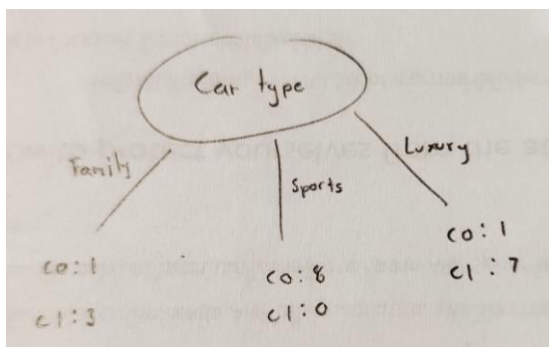
$$\begin{aligned}
 \text{Gini index} &= 1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2 \\
 &= 1 - \frac{1}{4} - \frac{1}{4} \\
 &= 0.5
 \end{aligned}$$

(b) Compute the Gini index for the Gender attribute.



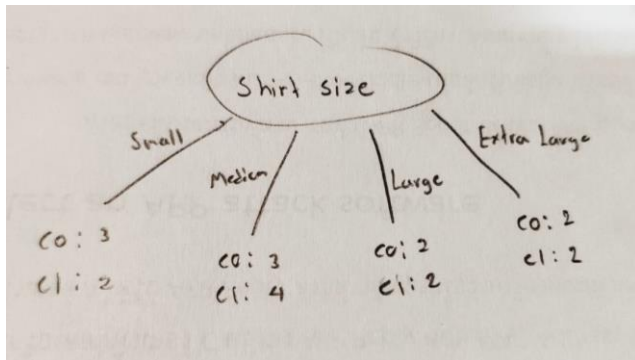
$$\begin{aligned}
 \text{Gini}(M) &= 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0.48 \\
 \text{Gini}(F) &= 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 0.48 \\
 \text{Gini}(\text{Gender}) &= \left(\frac{10}{20}\right)(0.48) + \left(\frac{10}{20}\right)(0.48) = 0.48
 \end{aligned}$$

(d) Compute the Gini index for the Car Type attribute using multi-way split.



$$\begin{aligned}
 \text{Gini}(\text{Family}) &= 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375 \\
 \text{Gini}(\text{Sports}) &= 1 - \left(\frac{8}{8}\right)^2 = 0 \\
 \text{Gini}(\text{Luxury}) &= 1 - \left(\frac{1}{8}\right)^2 - \left(\frac{7}{8}\right)^2 = 0.21875 \\
 \text{Gini}(\text{Car type}) &= \left(\frac{4}{20}\right)(0.375) + \left(\frac{8}{20}\right)(0) + \left(\frac{8}{20}\right)(0.21875) = 0.1625
 \end{aligned}$$

(e) Compute the Gini index for the Shirt Size attribute using multi-way split.



$$\begin{aligned}
 \text{Gini (Small)} &= 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48 \\
 \text{Gini (Medium)} &= 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.49 \\
 \text{Gini (Large)} &= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5 \\
 \text{Gini (Extra Large)} &= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5 \\
 \text{Gini (Shirt size)} &= \left(\frac{5}{20}\right)(0.48) + \left(\frac{7}{20}\right)(0.49) + \left(\frac{4}{20}\right)(0.5) + \left(\frac{4}{20}\right)(0.5) \\
 &= 0.49
 \end{aligned}$$

(f) Which attribute is better, Gender, Car Type, or Shirt Size? Why?

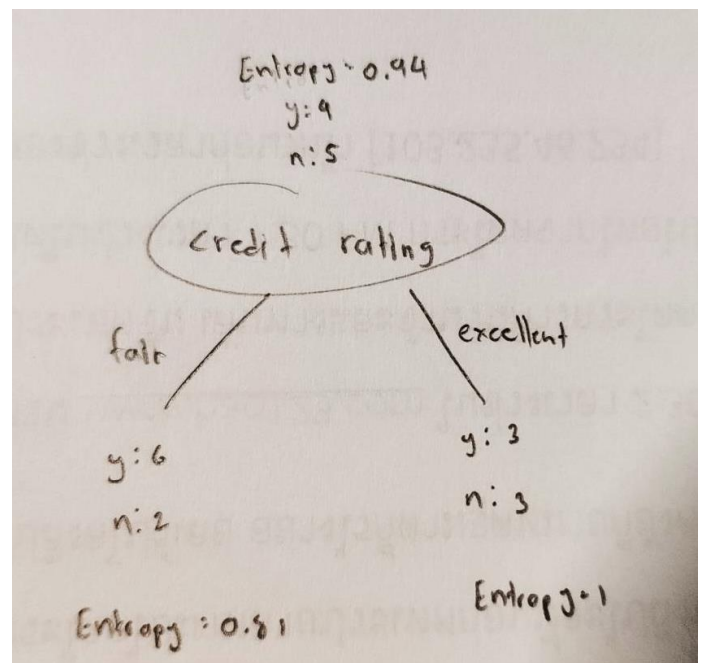
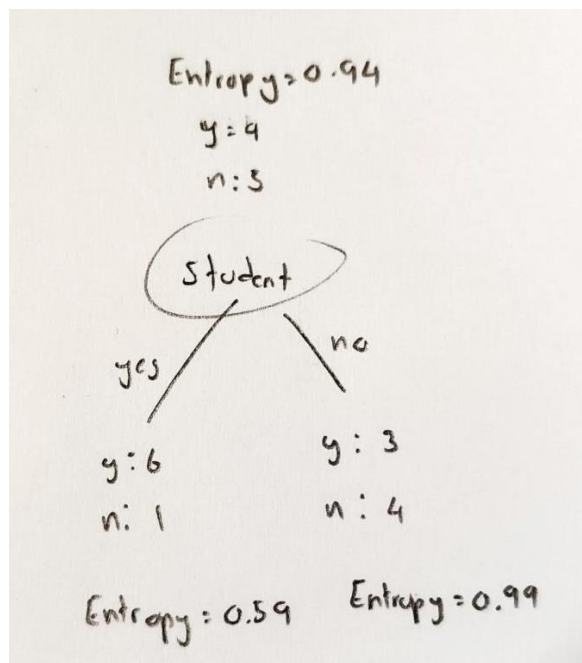
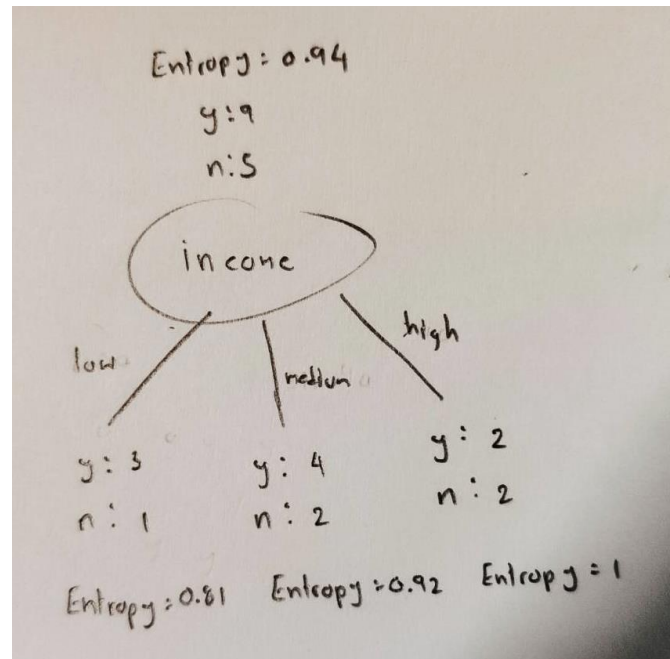
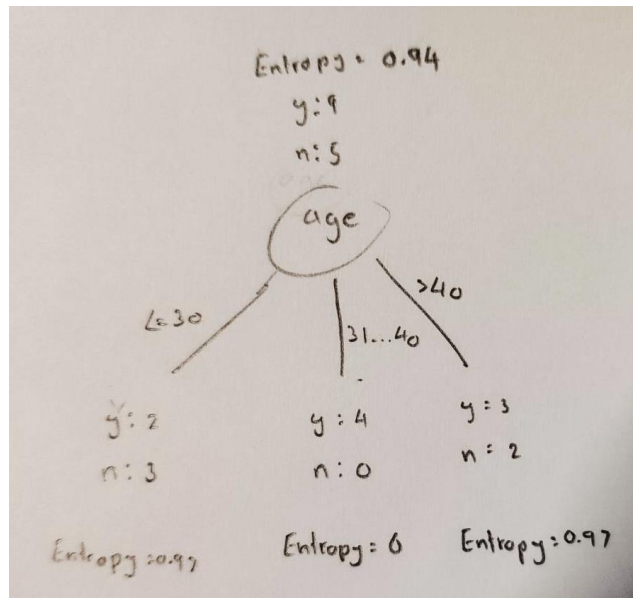
Attribute ที่ดีที่สุดคือ Car Type เนื่องจากมีค่า Gini Index ต่ำที่สุด (Gini Index = 0.1625)

(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

ที่ไม่เลือก Customer ID มาเป็นหนึ่งใน attribute ทั้งที่มี Gini Index ต่ำเนื่องจาก Customer ID เป็น Primary key ในตารางนี้ ทำให้มีความแตกต่างกันในทุกๆแถวและเป็นค่า Unique ของแต่ละแถว

Exercise 2 (Building decision tree):

(a) Step by step, build a decision tree using Information Gain based on Entropy.



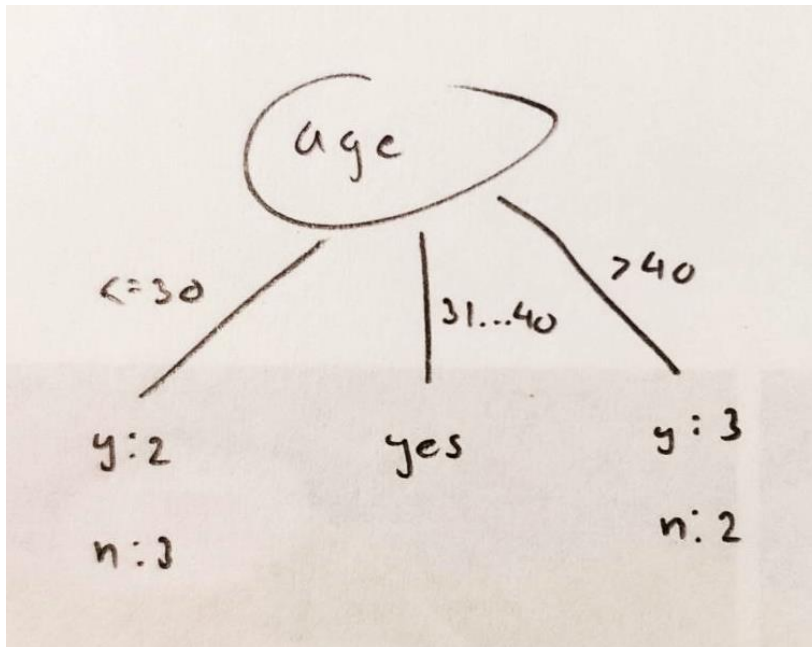
เลือก age เนื่องจากมีค่า Gain สูงที่สุดแล้วหา node ต่อไป

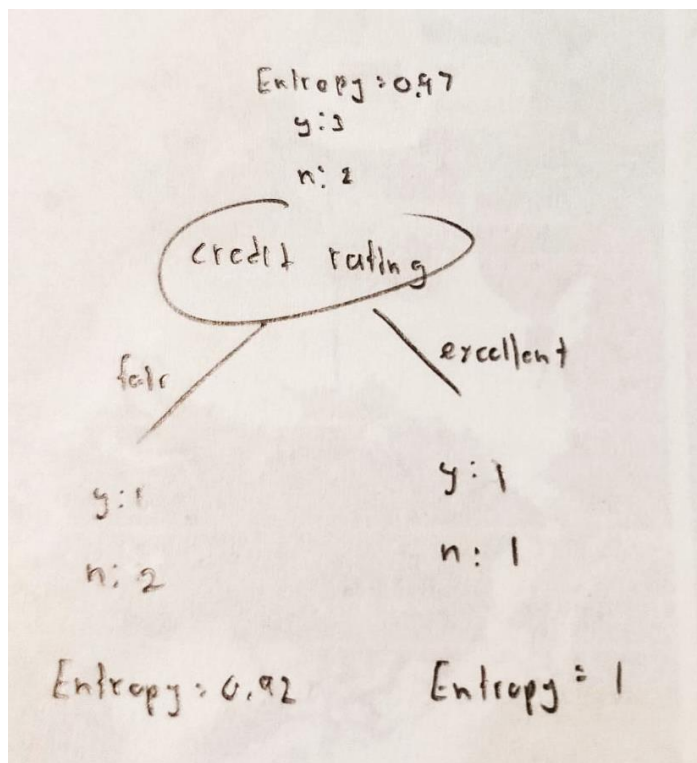
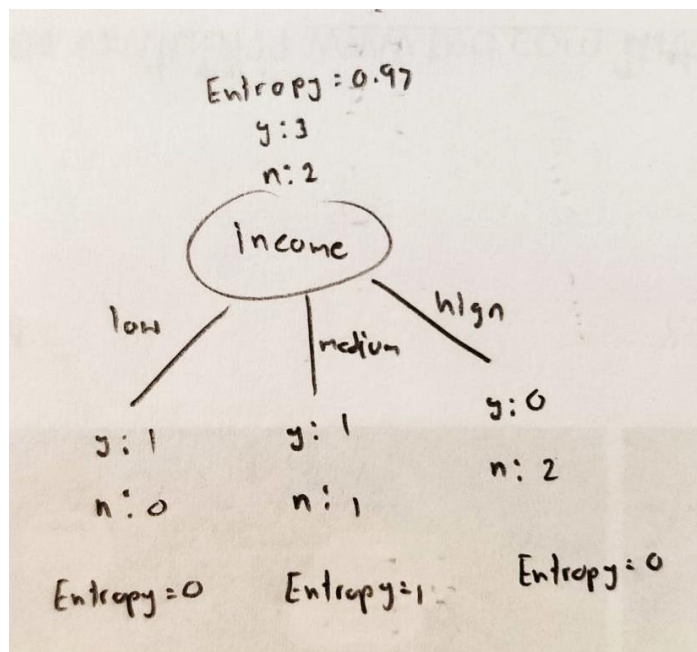
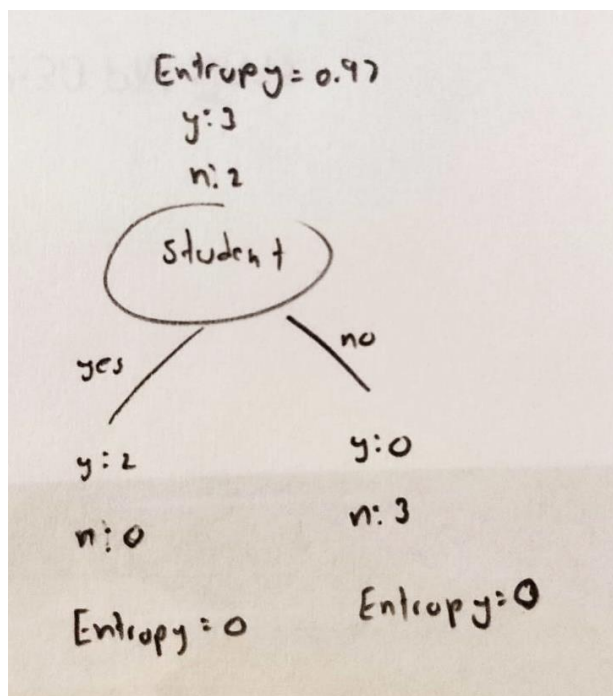
$$\text{Gain}(\text{age}) = 0.94 - \left(\frac{5}{14}\right)(0.97) - \left(\frac{4}{14}\right)(0) - \left(\frac{5}{14}\right)(0.97) = 0.246 \quad \checkmark$$

$$\text{Gain}(\text{income}) = 0.94 - \left(\frac{4}{14}\right)(0.81) - \left(\frac{6}{14}\right)(0.92) - \left(\frac{4}{14}\right)(1) = 0.044 \quad \times$$

$$\text{Gain}(\text{student}) = 0.94 - \left(\frac{7}{14}\right)(0.59) - \left(\frac{7}{14}\right)(0.99) = 0.15 \quad \times$$

$$\text{Gain}(\text{credit}) = 0.94 - \left(\frac{8}{14}\right)(0.81) - \left(\frac{6}{14}\right)(1) = 0.049 \quad \times$$



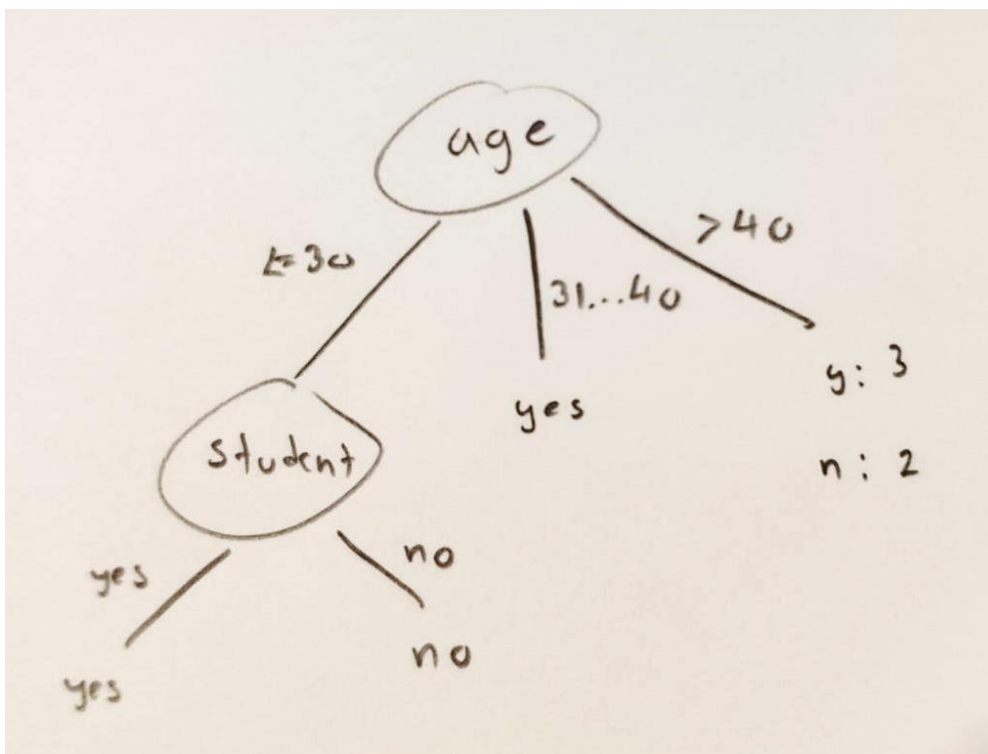


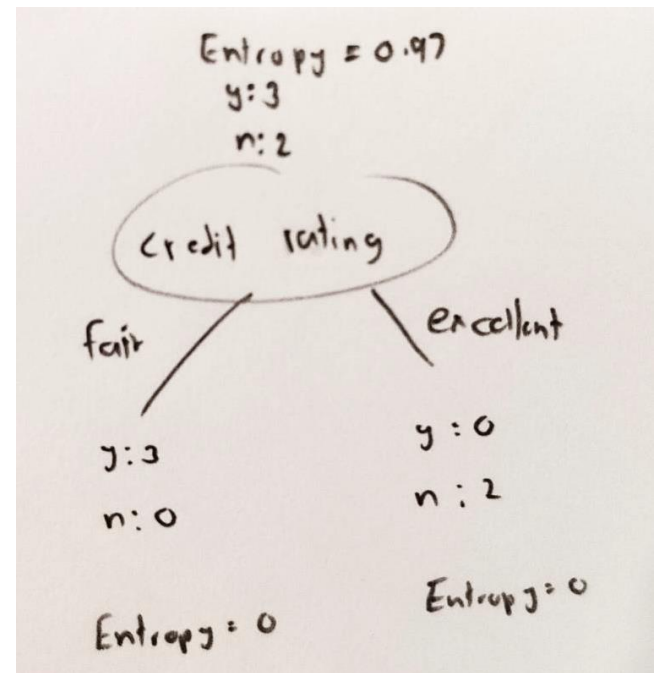
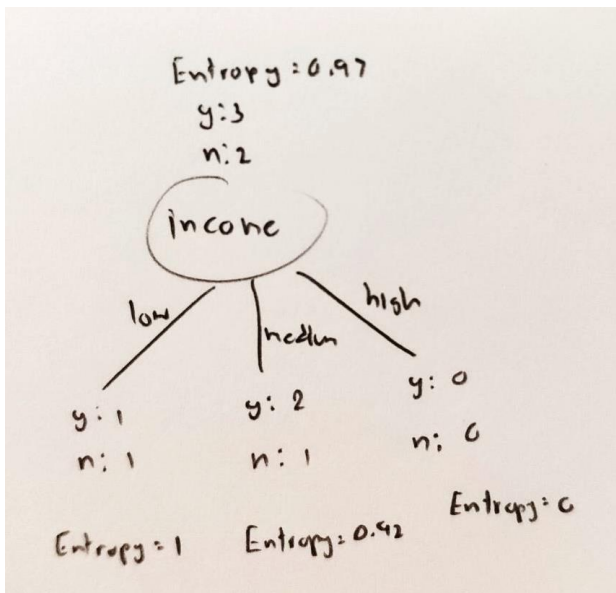
เลือก student เนื่องจากมีค่า Gain สูงที่สุดแล้วหา node ต่อไป

$$\text{Gain}(\text{income}) = 0.97 - \left(\frac{1}{5}\right)(0) - \left(\frac{2}{5}\right)(1) - \left(\frac{2}{5}\right)(0) = 0.57$$

$$\text{Gain}(\text{student}) = 0.97 - \left(\frac{2}{5}\right)(0) - \left(\frac{3}{5}\right)(1) = 0.97 \quad \checkmark$$

$$\text{Gain}(\text{credit}) = 0.97 - \left(\frac{3}{5}\right)(0.92) - \left(\frac{2}{5}\right)(1) = 0.018$$

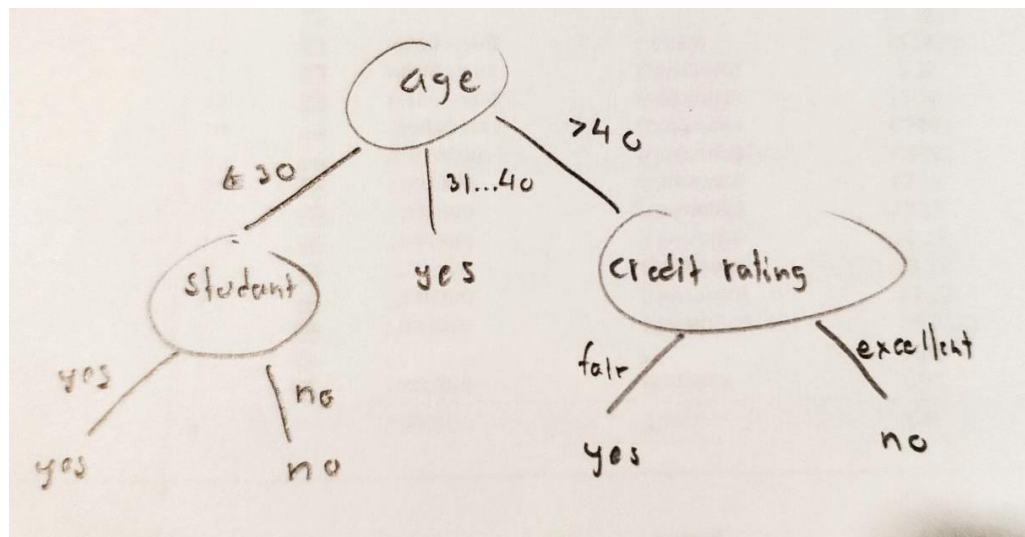




เลือก credit rating เนื่องจากมีค่า Gain สูงที่สุด หลังจากนั้นจะได้ decision tree ที่สมบูรณ์

$$\text{Gain}(\text{Income}) = 0.97 - \left(\frac{2}{5}\right)(1) - \left(\frac{3}{5}\right)(0.92) - \left(\frac{0}{5}\right)(0) = 0.018 *$$

$$\text{Gain}(\text{Credit}) = 0.97 - \left(\frac{3}{5}\right)(0) - \left(\frac{2}{5}\right)(0) = 0.97 \checkmark$$



- (b) Use the constructed tree to predict the class of the following new example: age \leq 30, income=medium, student=yes, credit-rating=fair.

Predict: yes

Exercise 3 (Building decision tree using software):

Parameters X

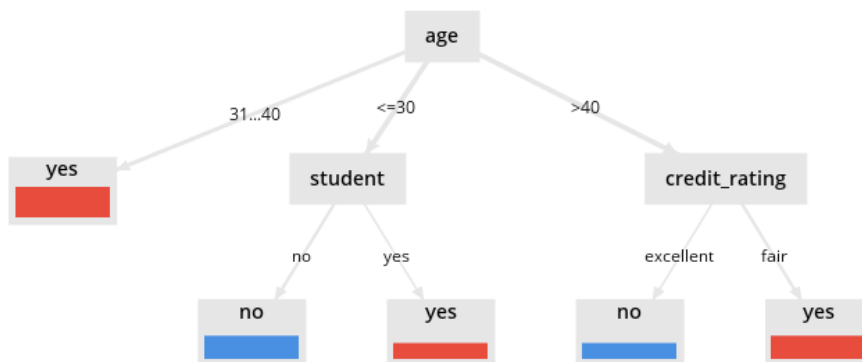
Decision Tree

criterion: information_gain

maximal depth: 10

☐ apply pruning

☐ apply prepruning



Decision tree ที่สร้างเองเหมือนกันกับที่สร้างจาก software (RapidMiner)