

HOMEWORK4

DECISION TREE CLASSIFIER

GROUP5 : MIN_WEIGHT_FRACTION_LEAF

กลุ่มได้หมด & กลุ่มจะดีหรอ

ສມາຜິກ

1. ນາຍປີຍພັກ ປະລິກ 643020507-4
2. ນາງສາວພື້ນຄົມ ຖອນບ່ອ 643020508-2
3. ນາງສາວພິມຈະນກ ວັດທະຍາເຊື່ອ 643020510-5
4. ນາຍກູງວິດ ເຄື່ອງຈາກ 643020514-7
5. ນາງສາວວິກາດາ ໜ່ວງສູງເນື່ອນ 643020520-2
6. ນາງສາວສິຮກັກ ໄຊຍມາຕົມ 643020523-6
7. ນາງສາວຫ້ຍຈະນກ ສຽງແຈ້ງກູມ 643020525-2
8. ນາຍອາຄຸນສົ່ງ ຈຽວຮັກສົ່ງ 643020528-6
9. ນາຍຮນພຣ ກໍານກິ່ງ 643021264-9
10. ນາງສາວຈິນດາພຣ ໂພຮົ້ງກູມ 643021262-3

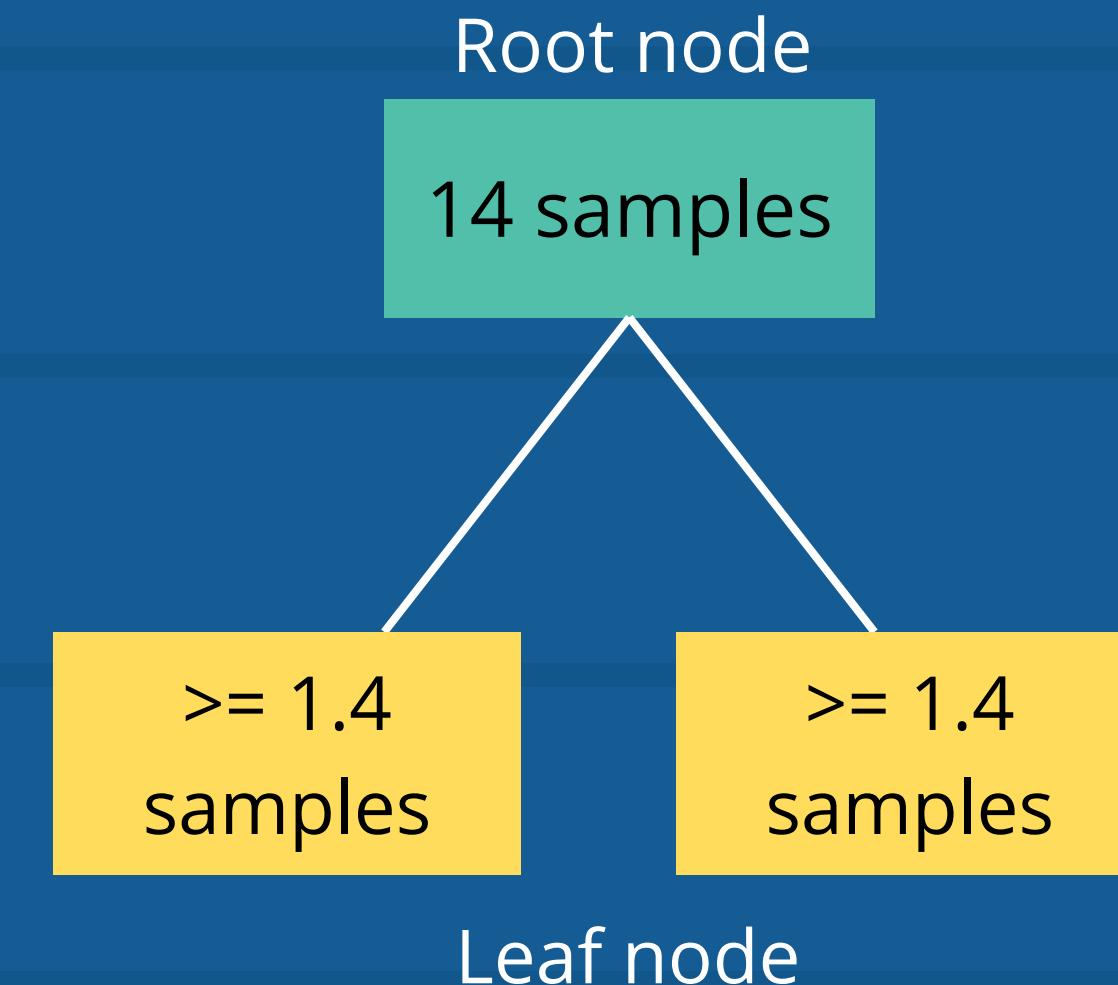
MIN WEIGHT FRACTION LEAF

เป็นพารามิเตอร์ของ `DecisionTreeClassifier` ใน scikit-learn ที่ระบุเศษส่วนก่วงน้ำหนัก

ขั้นต่ำของผลรวมของน้ำหนัก (ของตัวอย่างอันพุตทึ้งหมด) ที่โอนดีพ

พารามิเตอร์นี้เป็นวิธีหนึ่งในการควบคุมการโอเวอร์ฟิต โดยมีค่าอยู่ระหว่าง "0" ถึง "0.5"

ยกตัวอย่าง
จำนวนตัวอย่างทั้งหมด = 14 samples
'min_weight_fraction_leaf=0.1'
หมายความว่าจำนวนตัวอย่างแต่ละลีฟ
จะต้องมีอย่างน้อย 10% ของผลรวมน้ำหนัก
ของตัวอย่างอันพุตทึ้งหมด
หรือ $0.1 * 14 = 1.4$ samples



การคำนวณ INFORMATION GAIN

กำหนดพารามิเตอร์ **MIN_WEIGHT_FRACTION_LEAF = 0.1** หมายความว่าบ้าน้ำ SAMPLE ใน NODE สุดท้ายต้องมากกว่าหรือเท่ากับ 1.4

หา Root Node

$$Info(D) = I_{(9,5)} = -\frac{9}{14} \log_2 \left(\frac{9}{14}\right) - \frac{5}{14} \log_2 \left(\frac{5}{14}\right) = 0.940$$

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$$\begin{aligned} Info_{income}(D) &= \frac{4}{14} I(2,2) + \frac{6}{14} I(4,2) + \frac{4}{14} I(3,1) \\ &= \frac{4}{14} \left[-\frac{2}{4} \log_2 \left(\frac{2}{4}\right) - \frac{2}{4} \log_2 \left(\frac{2}{4}\right) \right] + \frac{6}{14} \left[-\frac{4}{6} \log_2 \left(\frac{4}{6}\right) - \frac{2}{6} \log_2 \left(\frac{2}{6}\right) \right] + \\ &\quad \frac{4}{14} \left[-\frac{3}{4} \log_2 \left(\frac{3}{4}\right) - \frac{1}{4} \log_2 \left(\frac{1}{4}\right) \right] = 0.911 \end{aligned}$$

$$\begin{aligned} Info_{student}(D) &= \frac{7}{14} I(3,4) + \frac{7}{14} I(6,1) \\ &= \frac{7}{14} \left[-\frac{3}{7} \log_2 \left(\frac{3}{7}\right) - \frac{4}{7} \log_2 \left(\frac{4}{7}\right) \right] + \frac{7}{14} \left[-\frac{6}{7} \log_2 \left(\frac{6}{7}\right) - \frac{1}{7} \log_2 \left(\frac{1}{7}\right) \right] \\ &= 0.788 \end{aligned}$$

$$\begin{aligned} Info_{credit_rating}(D) &= \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3) \\ &= \frac{8}{14} \left[-\frac{6}{8} \log_2 \left(\frac{6}{8}\right) - \frac{2}{8} \log_2 \left(\frac{2}{8}\right) \right] + \frac{6}{14} \left[-\frac{3}{6} \log_2 \left(\frac{3}{6}\right) - \frac{3}{6} \log_2 \left(\frac{4}{6}\right) \right] \\ &= 0.892 \end{aligned}$$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

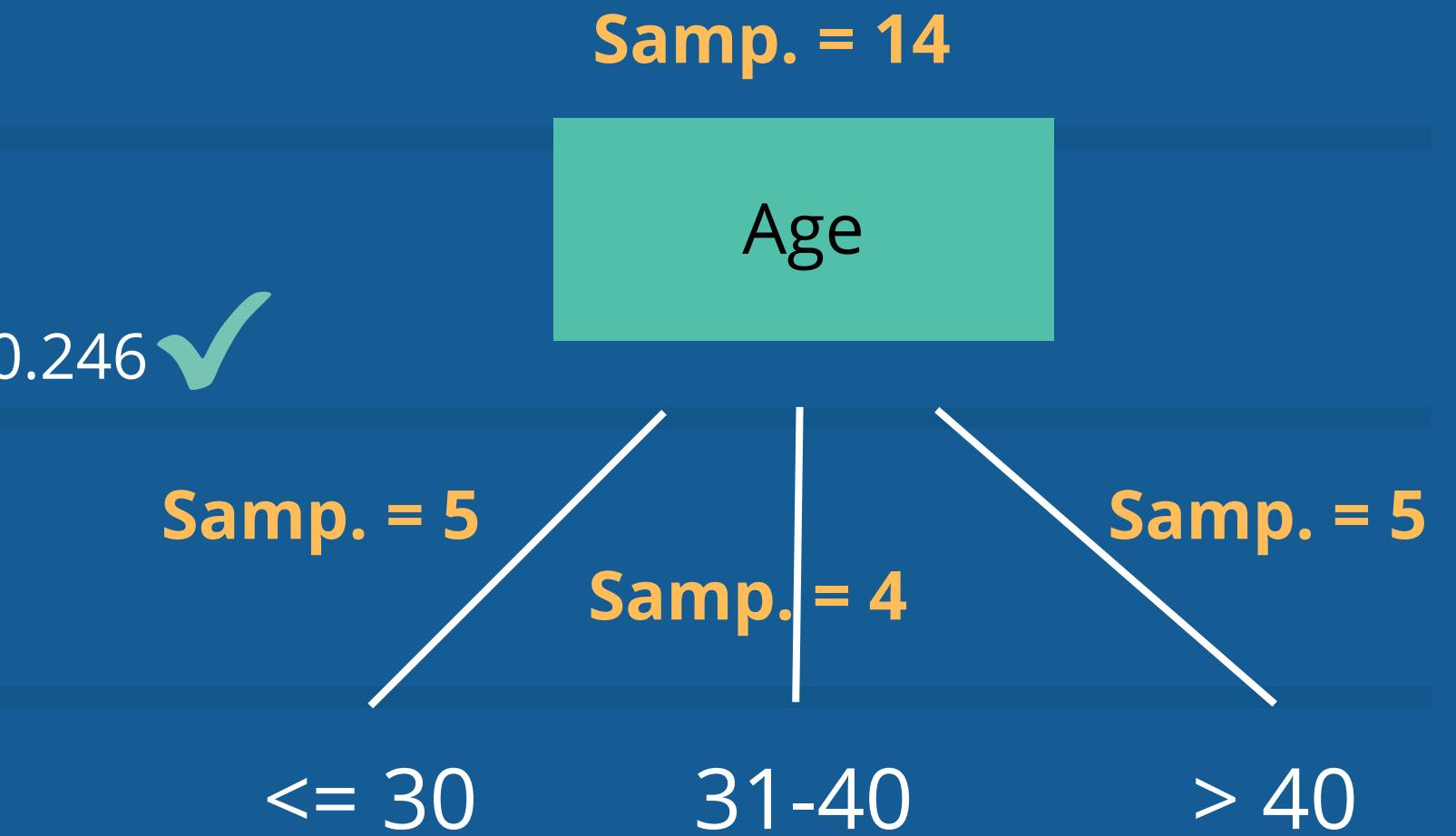
ROOT NODE

$$\text{Gain}(\text{Age}) = \text{Info}(D) - \text{Info}(D) = 0.940 - 0.694 = 0.246 \checkmark$$

$$\text{Gain}(\text{Income}) = 0.940 - 0.911 = 0.029$$

$$\text{Gain}(\text{Student}) = 0.940 - 0.788 = 0.152$$

$$\text{Gain}(\text{Credit_rating}) = 0.940 - 0.892 = 0.048$$



ดังนั้นจึงได้ Root Node = Age
 เพราะมีค่า Gain มากที่สุด
 และทำการแบ่งต่อไป เพราะ ตัวอย่างยังไม่เข้าเกล้า 1.4

አን Dicision Node እናሁ Age<= 30

$$\text{Info age } \leq 30 \text{ (D)} = I(2,3) = -\frac{2}{5} \log_2 \left(\frac{2}{5}\right) - \frac{3}{5} \log_2 \left(\frac{3}{5}\right) = 0.971$$

$$\begin{aligned}\text{Info income(D)} &= \frac{1}{5}I(1,0) + \frac{2}{5}I(1,1) + \frac{2}{5}I(0,2) \\ &= \frac{1}{5} \left[-\frac{1}{1} \log_2 \left(\frac{1}{1}\right) - 0 \right] + \frac{2}{5} \left[-\frac{1}{2} \log_2 \left(\frac{1}{2}\right) - \frac{1}{2} \log_2 \left(\frac{1}{2}\right) \right] + \frac{2}{5} \left[-\frac{2}{2} \log_2 \left(\frac{2}{2}\right) \right] \\ &= 0.4\end{aligned}$$

$$\text{Info student(D)} = \frac{3}{5}I(0,3) + \frac{2}{5}I(2,0) = \frac{3}{5} \left[-\frac{3}{3} \log_2(1) \right] + \frac{2}{5} \left[-\frac{2}{2} \log_2(1) \right] = 0$$

$$\begin{aligned}\text{Info credit(D)} &= \frac{3}{5}I(1,2) + \frac{2}{5}I(1,1) \\ &= \frac{3}{5} \left[-\frac{1}{3} \log_2 \left(\frac{1}{3}\right) - \frac{2}{3} \log_2 \left(\frac{2}{3}\right) \right] + \frac{2}{5} \left[-\frac{1}{2} \log_2 \left(\frac{1}{2}\right) - \frac{1}{2} \log_2 \left(\frac{1}{2}\right) \right] = 0.951\end{aligned}$$

DICISION NODE

กรณี AGE <= 30

Age

$$\text{Gain}(A) = \text{Info age} \leq 30 (D) - \text{InfoA}(D)$$

$$\text{Gain(Income)} = 0.971 - 0.4 = 0.571$$

$$\text{Gain(Student)} = 0.971 - 0 = 0.971 \quad \checkmark$$

$$\text{Gain(Credit_rating)} = 0.971 - 0.951 = 0.020$$

<= 30

31-40

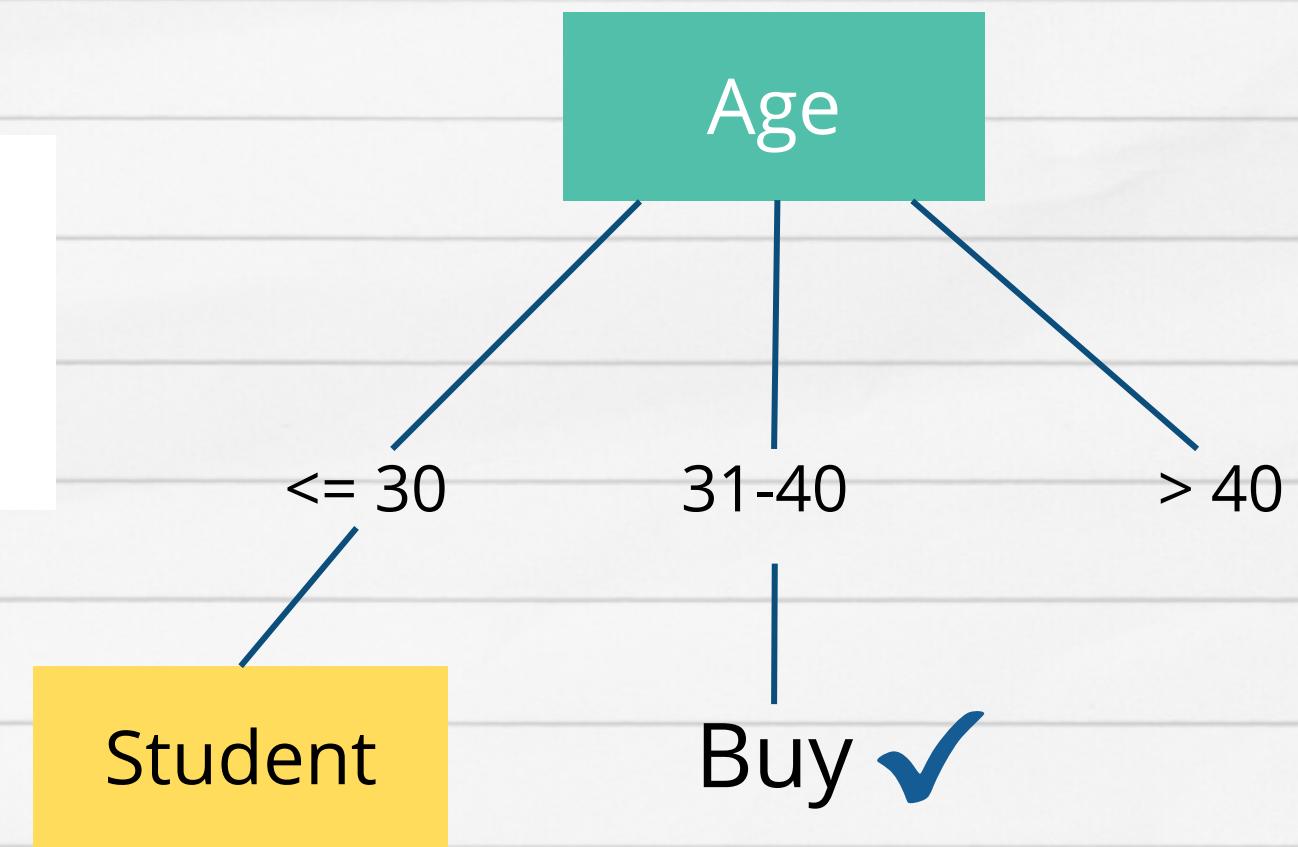
> 40

Student

ดังนั้นจะได้ Dicision Node ตัวที่ 2 คือ Student
 เพราะมีค่า Gain มากที่สุด

หัว Dicision Node กรณี Age = 31 - 40

$$\text{Info age 31-41 (D)} = I(4,0) = -\frac{4}{4} \log_2 \left(\frac{4}{4}\right) = 0$$



ดังนั้นกรณี Age = 31 - 40 จะเป็น Buy ทั้งหมด เพราะเป็น yes ทั้งหมด

ជាថាទី ក្រសុំ Age > 40

$$\text{Info age} > 40(D) = I(3,2) = -\frac{3}{5} \log_2 \left(\frac{3}{5}\right) - \frac{2}{5} \log_2 \left(\frac{2}{5}\right)$$
$$= 0.971$$

$$\text{Info income}(D) = \frac{2}{5} I(1,1) + \frac{3}{5} I(2,2) + 0$$
$$= \frac{2}{5} \left[-\frac{1}{2} \log_2 \left(\frac{1}{2}\right) - \frac{1}{2} \log_2 \left(\frac{1}{2}\right) \right] + \frac{3}{5} \left[-\frac{2}{3} \log_2 \left(\frac{2}{3}\right) - \frac{1}{3} \log_2 \left(\frac{1}{3}\right) \right]$$
$$= 0.951$$

$$\text{Info student}(D) = \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$$
$$= \frac{3}{5} \left[-\frac{2}{3} \log_2 \left(\frac{2}{3}\right) - \frac{1}{3} \log_2 \left(\frac{1}{3}\right) \right] + \frac{2}{5} \left[-\frac{1}{2} \log_2 \left(\frac{1}{2}\right) - \frac{1}{2} \log_2 \left(\frac{1}{2}\right) \right]$$
$$= 0.951$$

$$\text{Info credit}(D) = \frac{3}{5} I(3,0) + \frac{2}{5} I(0,2)$$
$$= \frac{3}{5} \left[-\frac{3}{3} \log_2 \left(\frac{3}{3}\right) \right] + \frac{2}{5} \left[-\frac{2}{2} \log_2 \left(\frac{2}{2}\right) \right]$$
$$= 0$$

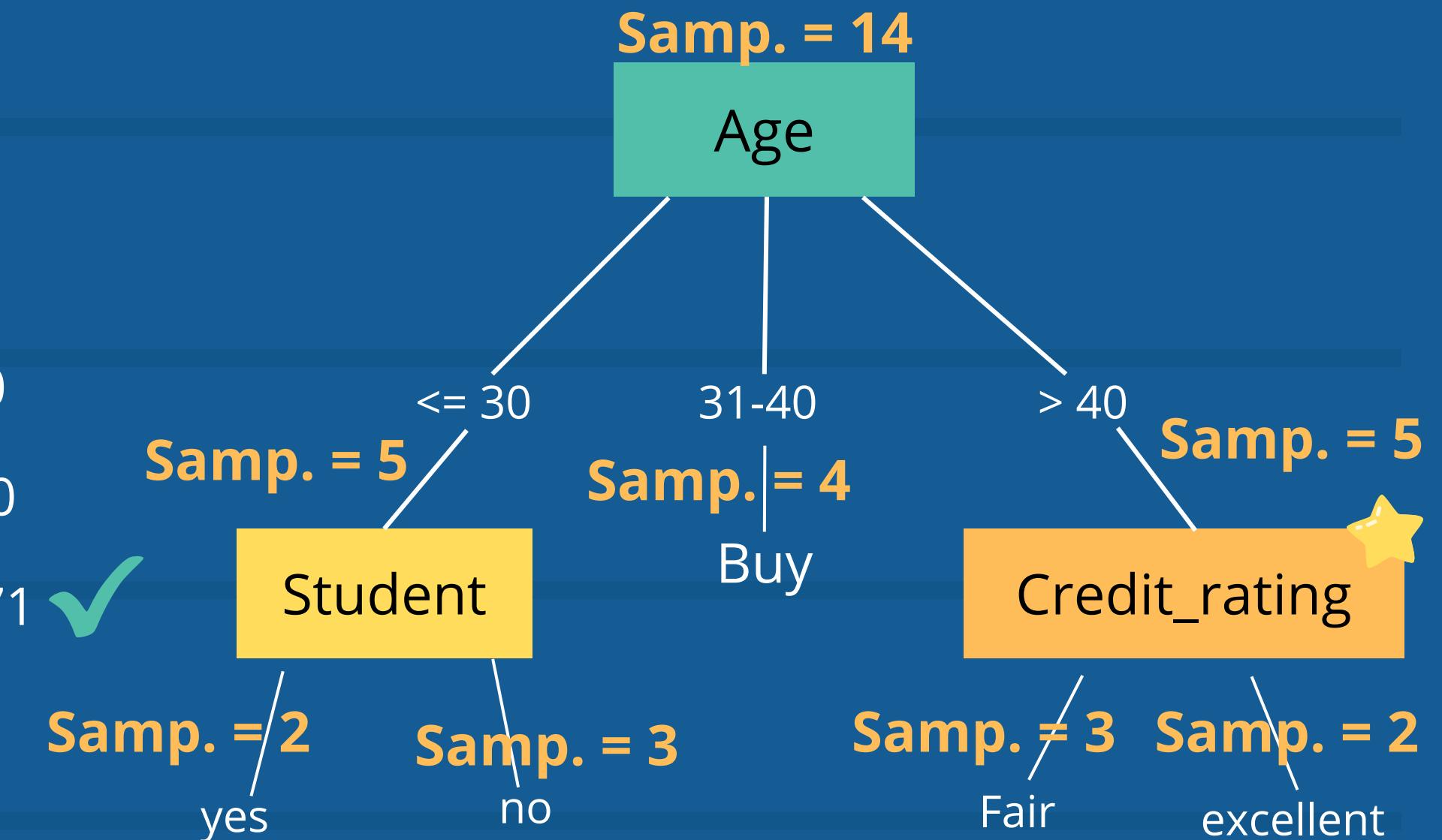
DICISION NODE กรณี AGE > 40

$$\text{Gain}(A) = \text{Info age}>40(D) - \text{InfoA}(D)$$

$$\text{Gain}(Income) = 0.971 - 0.951 = 0.020$$

$$\text{Gain}(Student) = 0.971 - 0.951 = 0.020$$

$$\text{Gain}(Credit_rating) = 0.971 - 0 = 0.971 \quad \checkmark$$



ดังนั้นจะได้ Decision Node ตัวที่ 3 คือ Credit_rating เพราะมีค่า Gain มากที่สุด และทำการแบ่งต่อไป เพราะ ตัวอย่าง ใน Leaf node ยังไม่เข้าใกล้ 1.4

ກາ Node ທີ່ຈະແປ່ງ Student = yes

$$Info_{student_{yes}}(D) = I(6,1) = -\frac{6}{7} \log_2 \left(\frac{6}{7}\right) - \frac{1}{7} \log_2 \left(\frac{1}{7}\right) = 0.591$$

$$\begin{aligned} Info_{income}(D) &= \frac{4}{7} I(3,1) + \frac{2}{7} I(2,0) + \frac{1}{7} I(1,0) \\ &= \frac{4}{7} \left[-\frac{3}{4} \log_2 \left(\frac{3}{4}\right) - \frac{1}{4} \log_2 \left(\frac{1}{4}\right) \right] + \frac{2}{7} \left[-\frac{2}{2} \log_2 \left(\frac{2}{2}\right) - 0 \right] + \frac{1}{7} \left[-\frac{1}{1} \log_2 \left(\frac{1}{1}\right) \right] \\ &= 0.464 \end{aligned}$$

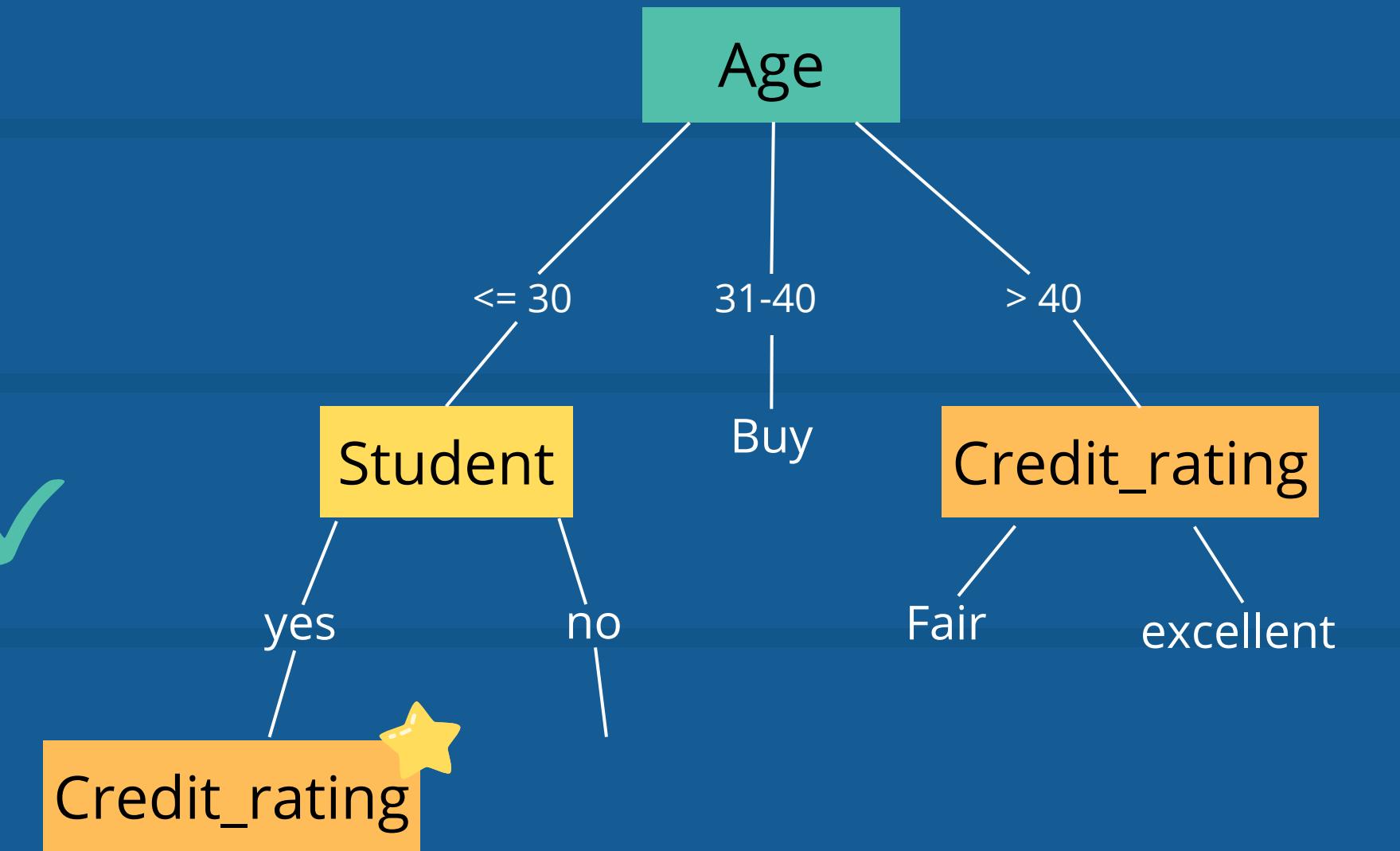
$$\begin{aligned} Info_{credit}(D) &= \frac{4}{7} I(4,0) + \frac{3}{7} I(2,1) \\ &= \frac{4}{7} \left[-\frac{4}{4} \log_2 \left(\frac{4}{4}\right) - 0 \right] + \frac{3}{7} \left[-\frac{2}{3} \log_2 \left(\frac{2}{3}\right) - \frac{1}{3} \log_2 \left(\frac{1}{3}\right) \right] = 0.394 \end{aligned}$$

NODE ที่จะแบ่ง STUDENT = YES

Gained by branching

$$\text{Gain}(Income) = 0.591 - 0.464 = 0.127$$

$$\text{Gain}(Credit_rating) = 0.591 - 0.394 = 0.197$$



ดังนั้นจะได้ Dicision Node ตัวที่4 คือ Credit_rating เพราะมีค่า Gain มากที่สุด

អាណ Node កំណត់ថា Student = no

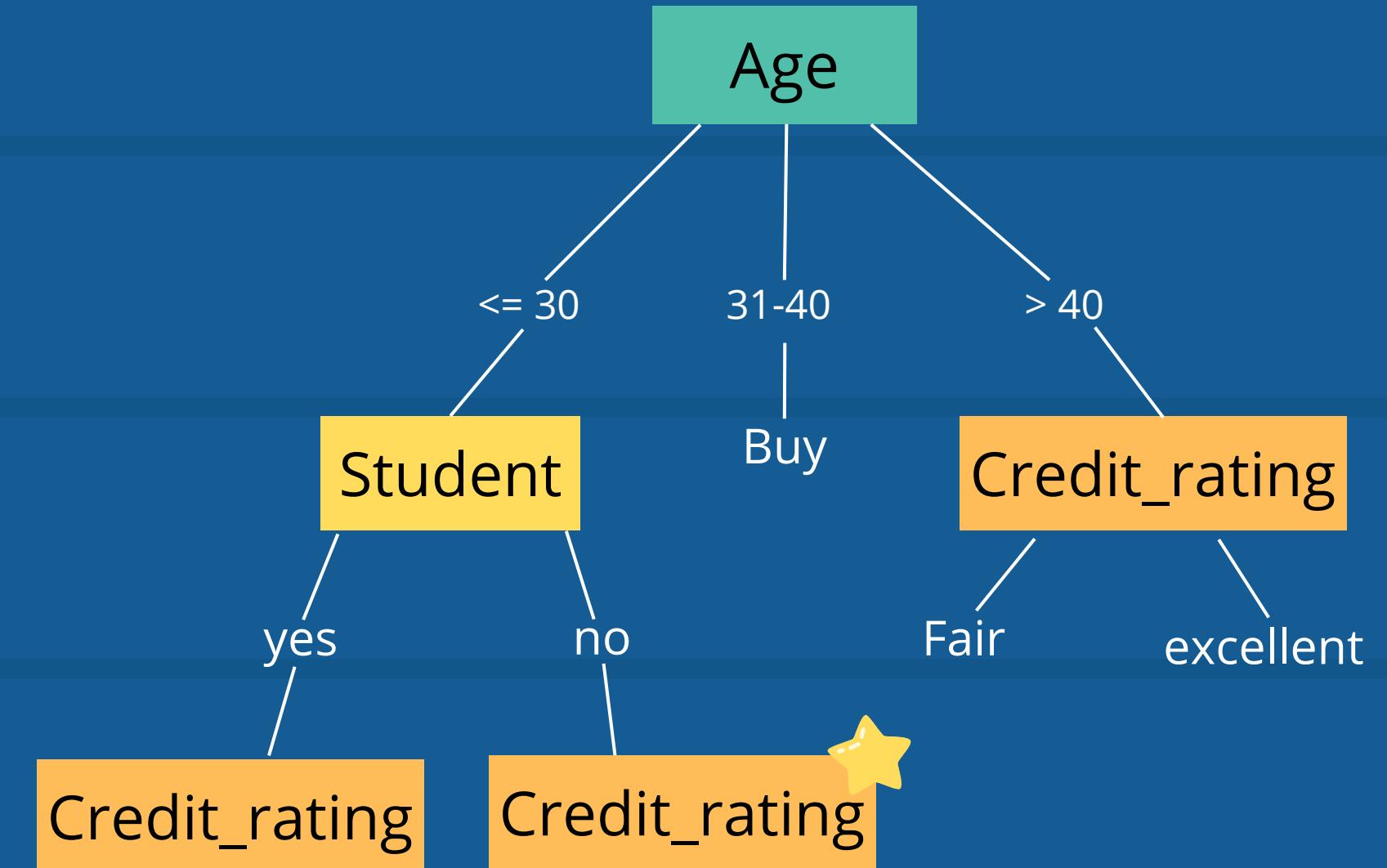
$$\text{Info}_{\text{stu_No}}(D) = I(3,4) = -\frac{3}{7} \log_2 \left(\frac{3}{7}\right) - \frac{4}{7} \log_2 \left(\frac{4}{7}\right) = 0.985$$

$$\begin{aligned}\text{Info}_{\text{income}}(D) &= 0 + \frac{4}{7} I(2,2) + \frac{3}{7} I(1,2) \\ &= \frac{4}{7} \left[-\frac{2}{4} \log_2 \left(\frac{2}{4}\right) - \frac{2}{4} \log_2 \left(\frac{2}{4}\right) \right] + \frac{3}{7} \left[-\frac{1}{3} \log_2 \left(\frac{1}{3}\right) - \frac{2}{3} \log_2 \left(\frac{2}{3}\right) \right] = 0.965\end{aligned}$$

$$\begin{aligned}\text{Info}_{\text{credit}}(D) &= \frac{4}{7} I(2,2) + \frac{3}{7} I(1,2) \\ &= \frac{4}{7} \left[-\frac{2}{4} \log_2 \left(\frac{2}{4}\right) - \frac{2}{4} \log_2 \left(\frac{2}{4}\right) \right] + \frac{3}{7} \left[-\frac{1}{3} \log_2 \left(\frac{1}{3}\right) - \frac{2}{3} \log_2 \left(\frac{2}{3}\right) \right] = 0.965\end{aligned}$$

NODE ที่จะแบ่ง STUDENT = NO

Gain(Income) = 0.985 - 0.965 = 0.02
Gain(Credit_rating) = 0.985 - 0.965 = 0.02



ดังนั้นจะได้ Decision Node ตัวที่ 5 คือ Credit_rating เพราะมีค่า Gain มากที่สุด

หา Node ที่จะแบ่งจาก Credit_rating = fair

$$\begin{aligned} Info_{credit_{fair}}(D) &= I(6,2) = -\frac{6}{8} \log_2 \left(\frac{6}{8}\right) - \frac{2}{8} \log_2 \left(\frac{2}{8}\right) \\ &= 0.811 \end{aligned}$$

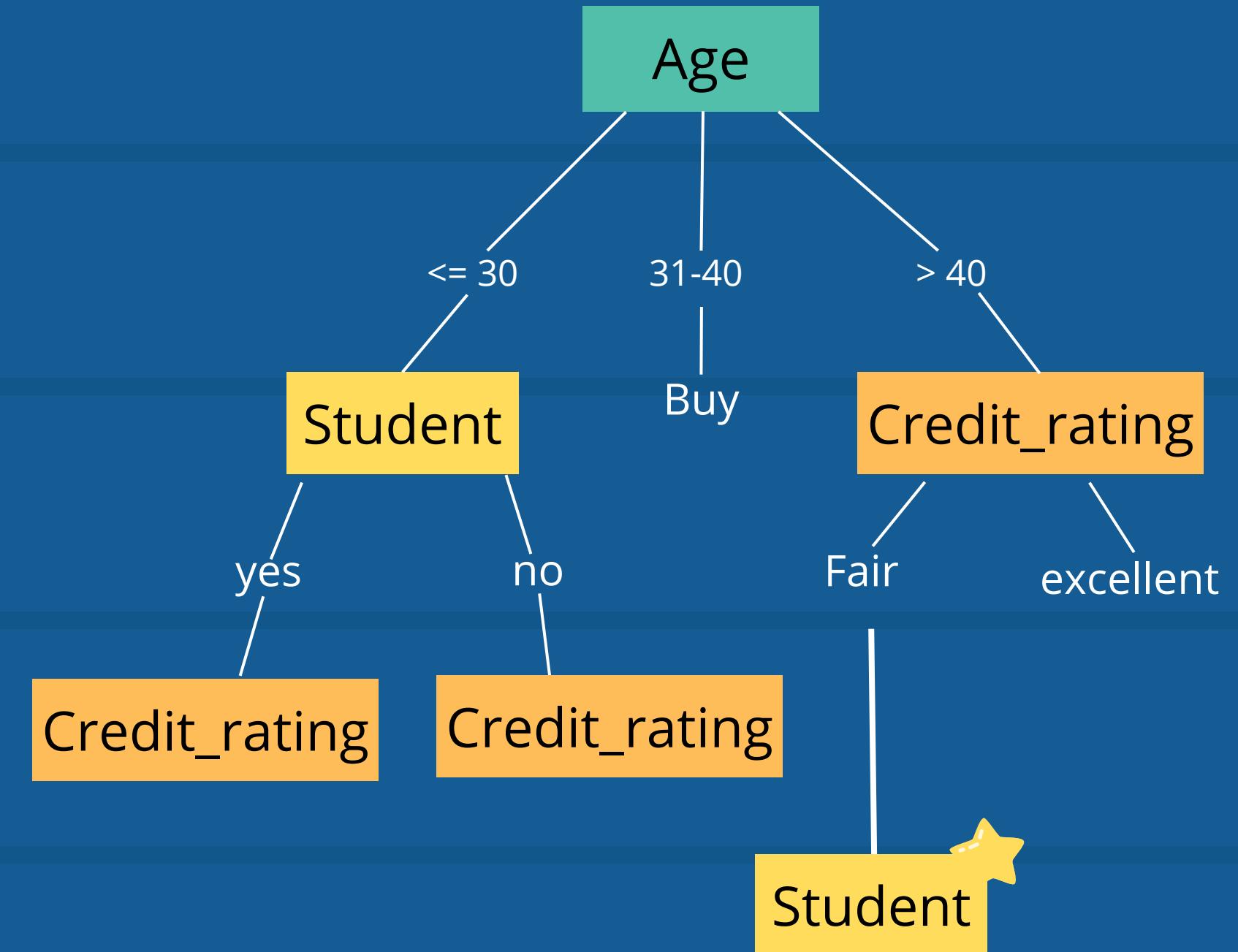
$$\begin{aligned} Info_{income}(D) &= \frac{2}{8} I(2,0) + \frac{3}{8} I(2,1) + \frac{3}{8} I(2,1) \\ &= \frac{2}{8} \left[-\frac{2}{2} \log_2 \left(\frac{2}{2}\right) - 0 \right] + \frac{3}{8} \left[-\frac{2}{3} \log_2 \left(\frac{2}{3}\right) - \frac{1}{3} \log_2 \left(\frac{1}{3}\right) \right] + \\ &\quad \frac{3}{8} \left[-\frac{2}{3} \log_2 \left(\frac{2}{3}\right) - \frac{1}{3} \log_2 \left(\frac{1}{3}\right) \right] \\ &= 0.564 \end{aligned}$$

$$\begin{aligned} Info_{student}(D) &= \frac{4}{8} I(4,0) + \frac{4}{8} I(3,1) \\ &= \frac{4}{8} \left[-\frac{4}{4} \log_2 \left(\frac{4}{4}\right) - 0 \right] + \frac{4}{8} \left[-\frac{3}{4} \log_2 \left(\frac{3}{4}\right) - \frac{1}{4} \log_2 \left(\frac{1}{4}\right) \right] \\ &= 0.405 \end{aligned}$$

NODE ที่จะแบ่งจาก CREDIT_RATING = FAIR

$$\text{Gain(Income)} = 0.811 - 0.564 = 0.247$$
$$\text{Gain(Student)} = 0.811 - 0.405 = 0.406 \quad \checkmark$$

ดังนั้นจะได้ Decision Node ตัวที่ 6 คือ Student
เพราะมีค่า Gain มากที่สุด



หา Node ที่จะแบ่งจาก Credit_rating = excellent

Info credit ex(D) = $I(3,3)$

$$= -\frac{3}{6} \log_2 \left(\frac{3}{6}\right) - \frac{3}{6} \log_2 \left(\frac{3}{6}\right)$$
$$= 1$$

Info income(D) = $\frac{2}{6}I(1,1) + \frac{3}{6}I(2,1) + \frac{1}{6}I(0,1)$

$$= \frac{2}{6} \left[-\frac{1}{2} \log_2 \left(\frac{1}{2}\right) - \frac{1}{2} \log_2 \left(\frac{1}{2}\right) \right] + \frac{3}{6} \left[-\frac{2}{3} \log_2 \left(\frac{2}{3}\right) - \frac{1}{3} \log_2 \left(\frac{1}{3}\right) \right] + \frac{1}{6} \left[0 - \frac{1}{1} \log_2 \left(\frac{1}{1}\right) \right]$$
$$= 0.79$$

Info student(D) = $\frac{3}{6}I(2,1) + \frac{4}{8}I(3,1)$

$$= \frac{3}{5} \left[-\frac{4}{4} \log_2 \left(\frac{4}{4}\right) - 0 \right] + \frac{4}{8} \left[-\frac{3}{4} \log_2 \left(\frac{3}{4}\right) - \frac{1}{4} \log_2 \left(\frac{1}{4}\right) \right]$$
$$= 0.405$$

Gain(income)

$$= 1 - 0.79$$

$$= 0.21$$

Gain(student)

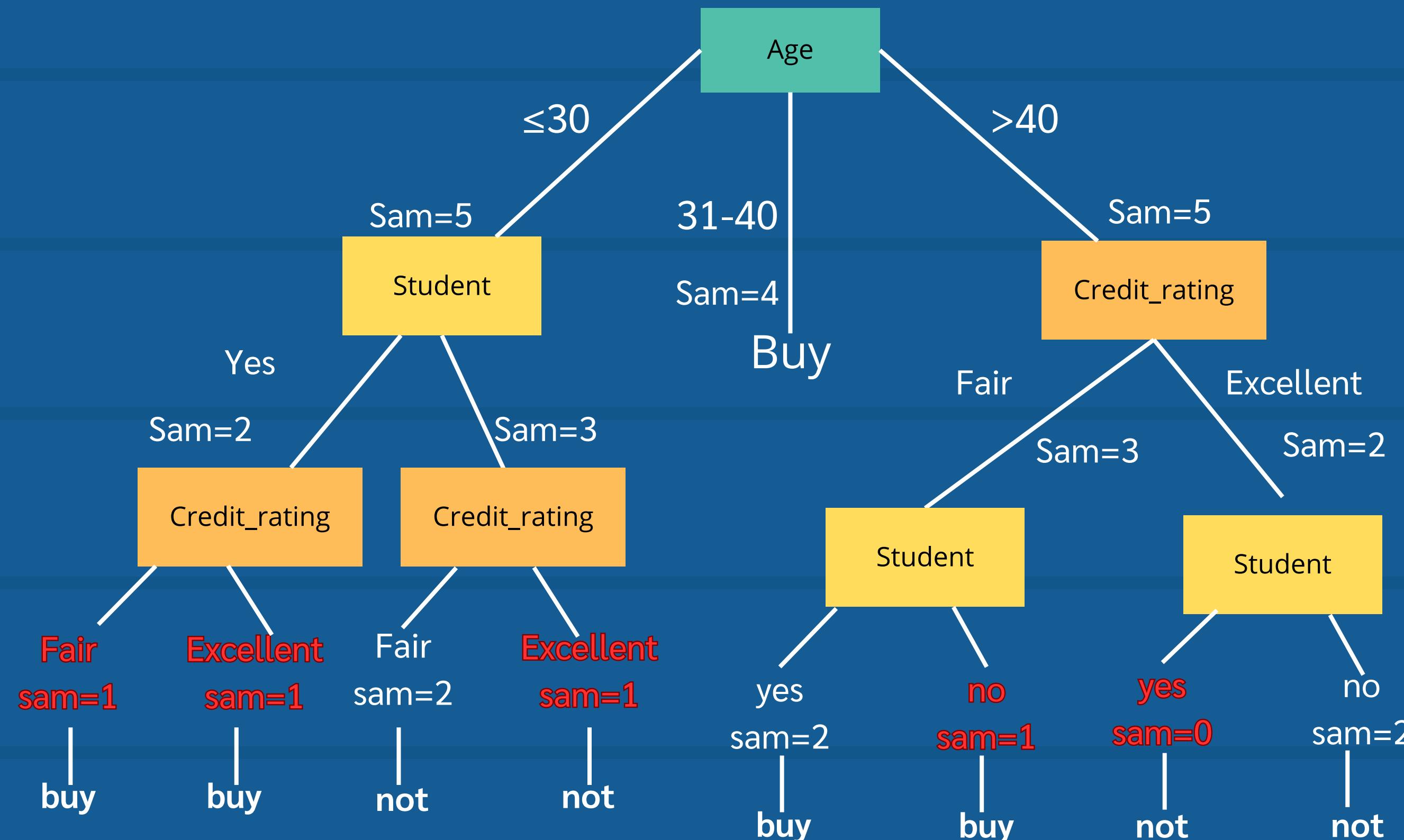
$$= 1 - 0.450$$

$$= 0.595$$



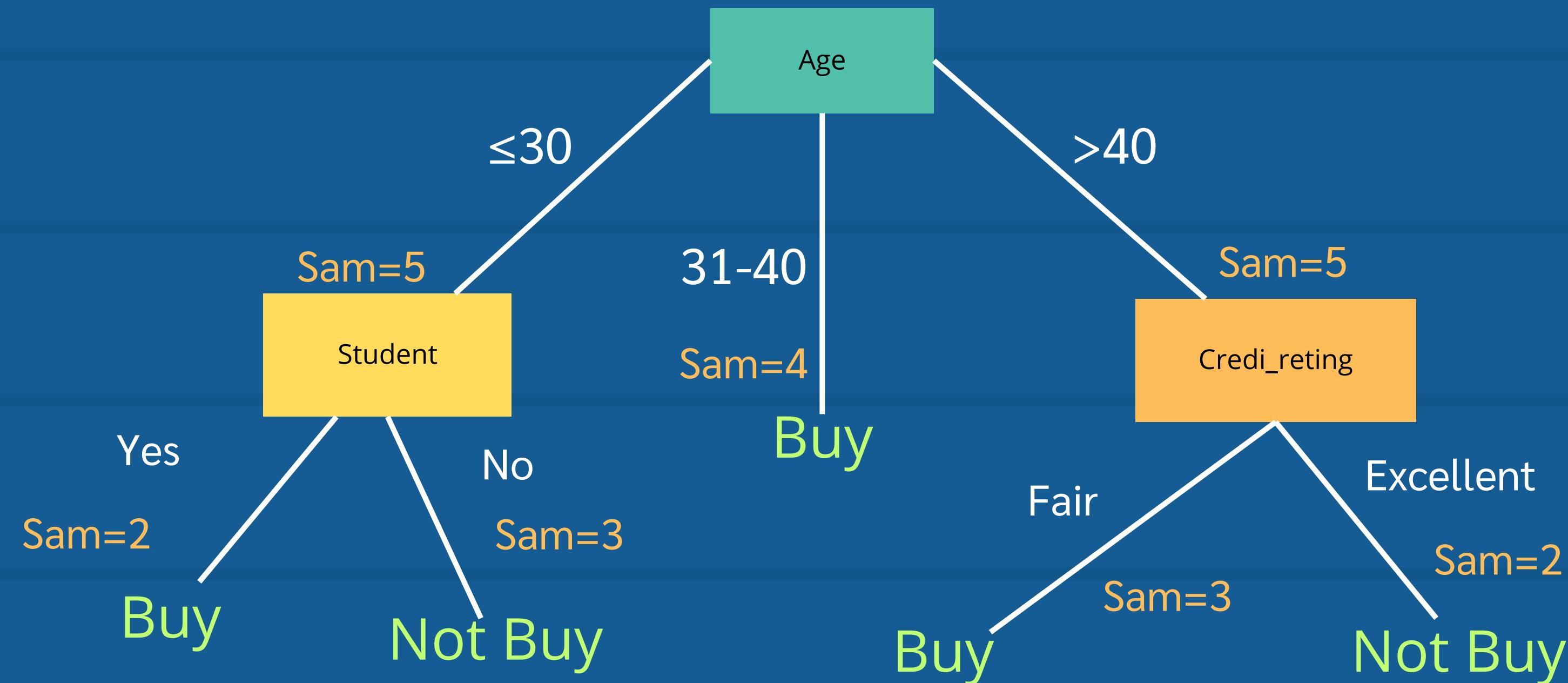
ดังนั้นจะได้ Decision Node ตัวที่ 7 คือ Student เพราะมีค่า Gain มากที่สุด

Decision tree ที่ได้จากการคำนวณ

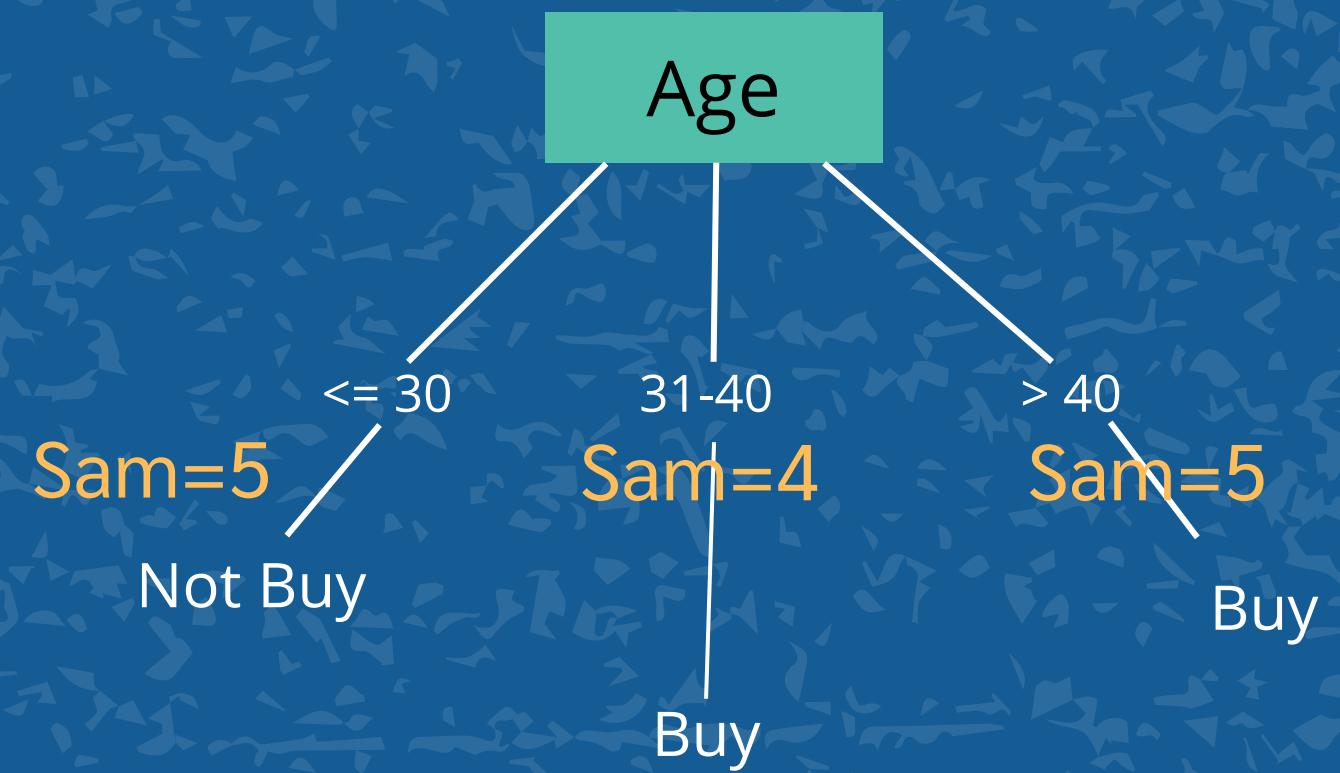


มี Node ที่มี Samples < 1.4 ดังนี้จึงไม่แตก Node นั้นต่อ
และค่าแต่ละโหนด มีความเป็นเอกลักษณ์แล้วจึงยุบคืน

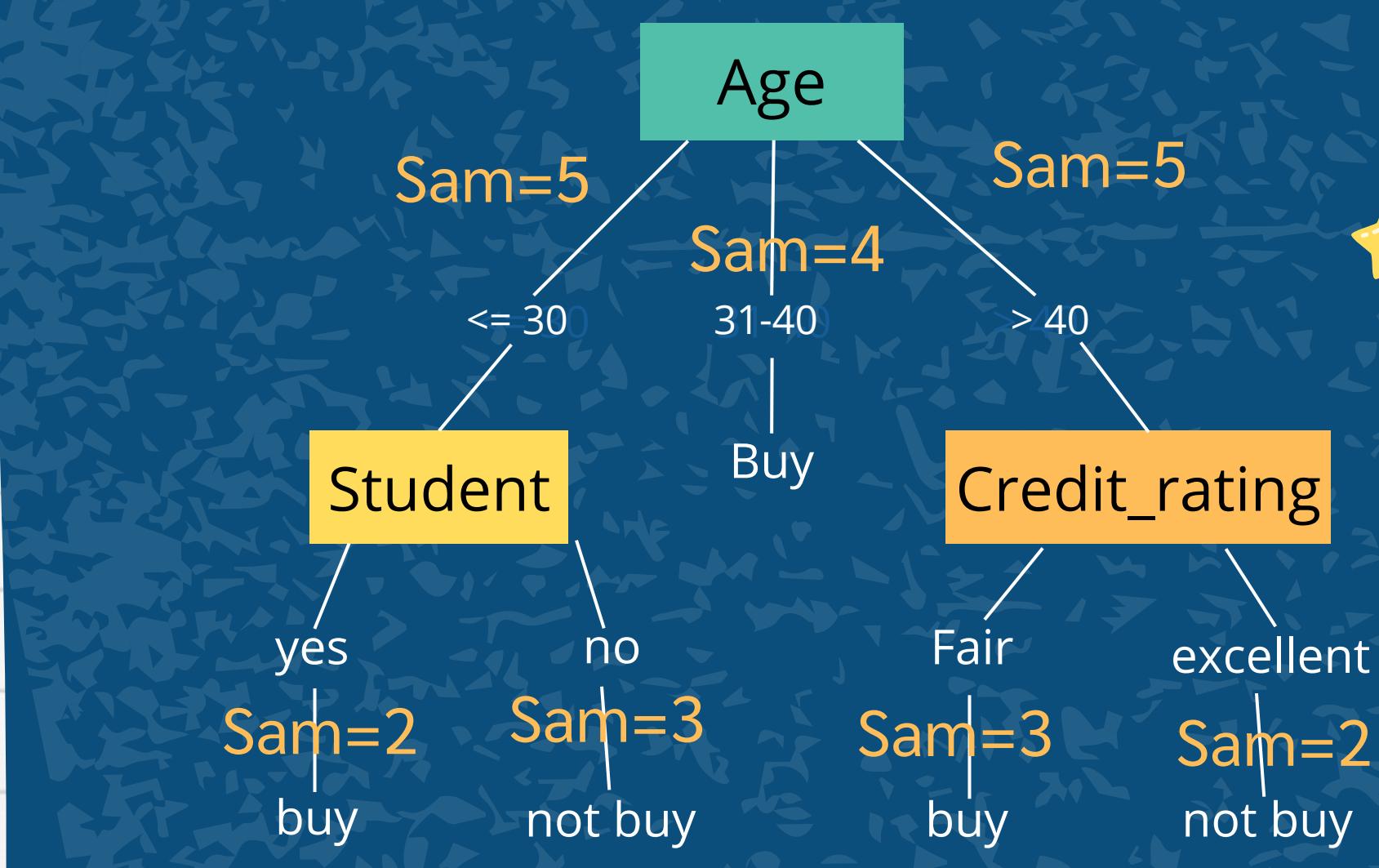
Decision tree ที่ได้จาก กำหนดพารามิเตอร์ min_weight_fraction_leaf = 0.1



min_weight_fraction_leaf = 0.28



min_weight_fraction_leaf = 0.14



หมายความว่า node sample สุดท้ายต้องมากกว่าหรือเท่ากับ 3.92 หรือ 4 samples

หมายความว่า node sample สุดท้ายต้องมากกว่าหรือเท่ากับ 1.96 หรือ 2 samples

**THANK
YOU VERY
MUCH!**