HỌC IT NHƯNG Đ BIT CODE

Presented by: Alexander Aronowitz

GINI INDEX

TÍNH TOÁN CHO THUỘC TÍNH CLASS: BUYS_COMPUTER

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

-Tổng số lượng dữ liệu: S = 14

-Số thuộc tính thuộc nhãn Class: buys_computer = yes: 9

-Số thuộc tính thuộc nhãn Class: buys_computer = no: 5

TÍNH TOÁN CHO THUỘC TÍNH AGE

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
youth	2	3	5
middle_aged	4	0	4
senior	3	2	5

- G(age=youth) = $1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$
- G(age=middle_aged) = $1 - (\frac{4}{4})^2 - (\frac{0}{4})^2 = 0$
- G(age=senior) = $1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$

→ L(age) = $\frac{5}{14} * 0.48 + \frac{4}{14} * 0 + \frac{5}{14} * 0.48 = 0.0294$

TÍNH TOÁN CHO THUỘC TÍNH INCOME

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
high	2	2	4
medium	4	2	6
low	3	1	4

- G(income=high) = $1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5$
- G(income=medium) = $1 - (\frac{4}{6})^2 - (\frac{2}{6})^2 = \frac{4}{9}$
- G(income=low) = $1 - (\frac{3}{4})^2 - (\frac{1}{4})^2 = 0.375$

➔ L(income) = $\frac{4}{14} * 0.5 + \frac{6}{14} * \frac{4}{9} + \frac{4}{14} * 0.375 = 0.4405$

TÍNH TOÁN CHO THUỘC TÍNH STUDENT

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
Yes	6	1	7
No	3	4	7

- G(student=no) = $1 - (\frac{6}{7})^2 - (\frac{1}{7})^2 = \frac{12}{49}$
- G(student=yes) = $1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = \frac{24}{49}$

➔ L(student) = $\frac{12}{49} * \frac{7}{14} + \frac{24}{49} * \frac{7}{14} = \frac{18}{49} = 0.3673$

TÍNH TOÁN CHO THUỘC TÍNH CREDIT_RATING

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
fair	6	2	8
excellent	3	3	6

- G(credit_rating=fair) = $1 - (\frac{6}{8})^2 - (\frac{2}{8})^2 = \frac{3}{8}$
- G(credit_rating=excellent) = $1 - (\frac{3}{6})^2 * 2 = 0.5$

➔ L(credit_rating) = $\frac{3}{8} * \frac{8}{14} + 0.5 * \frac{6}{14} = \frac{3}{7} = 0.4286$

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

-age: L(age) =0.0294

-income: L(income) =0.4405

-student: L(student) = 0.3673

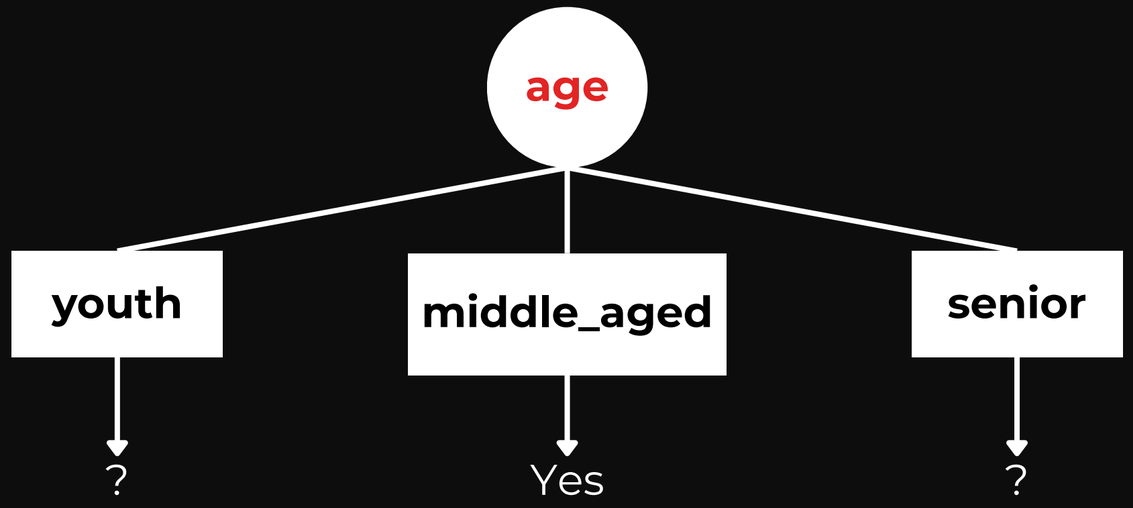
-credit_rating: L(credit_rating) 0.4286

➔Chọn age làm gốc (loss có giá trị nhỏ nhất)

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no



XÉT NHÁNH AGE = YOUTH

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

-Tổng số lượng dữ liệu: S = 5

-Số thuộc tính thuộc nhãn Class: buys_computer = yes: 2

-Số thuộc tính thuộc nhãn Class: buys_computer = no: 3

TÍNH TOÁN CHO THUỘC TÍNH INCOME

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
high	0	2	2
medium	1	1	2
low	1	0	1

- $G(income=high) = 1 - (\frac{0}{2})^2 - (\frac{2}{2})^2 = 0$
- $G(income=medium) = 1 - (\frac{1}{2})^2 * 2 = 0.5$
- $G(income=low) = 1 - (\frac{1}{1})^2 - (\frac{0}{1})^2 = 0.$

➔ $L(income) = \frac{2}{5} * 0 + \frac{2}{5} * 0.5 + \frac{1}{5} * 0 = 0.2$

TÍNH TOÁN CHO THUỘC TÍNH STUDENT

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
Yes	2	0	2
No	0	3	3

- $G(\text{student=no}) = 1 - (\frac{2}{2})^2 - (\frac{0}{2})^2 = 0$
- $G(\text{student=yes}) = 1 - (\frac{3}{3})^2 - (\frac{0}{3})^2 = 0$

→ $L(\text{student}) = = \frac{2}{5} * 0 + \frac{3}{5} * 0 = 0.$

TÍNH TOÁN CHO THUỘC TÍNH CREDIT_RATING

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
fair	1	2	3
excellent	1	1	2

- G(credit_rating=fair) = $1 - (\frac{1}{3})^2 - (\frac{2}{3})^2 = \frac{4}{9}$
- G(credit_rating=excellent) = $1 - (\frac{1}{2})^2 * 2 = 0.5$

➔ L(credit_rating) = $\frac{2}{5} * 0.5 + \frac{3}{5} * \frac{4}{9} = \frac{7}{15} = 0.4667$

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

-income: L(income) =0.2

-student: L(student) = 0

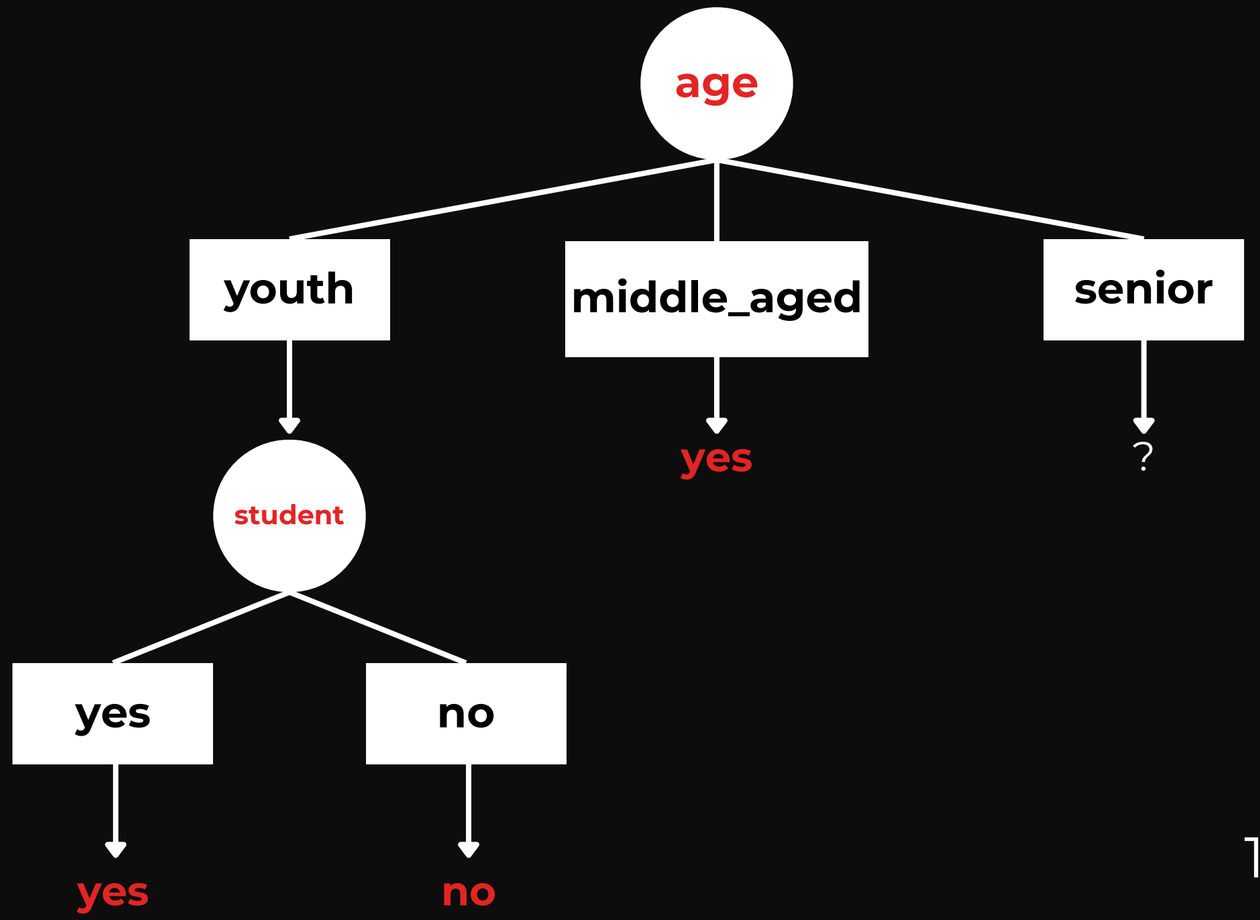
-credit_rating: L(credit_rating)= 0.4667

➔Chọn **student** làm gốc (loss có giá trị nhỏ nhất)

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no



XÉT NHÁNH AGE = SENIOR

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

-Tổng số lượng dữ liệu: S = 5

-Số thuộc tính thuộc nhãn Class: buys_computer = yes: 3

-Số thuộc tính thuộc nhãn Class: buys_computer = no: 2

TÍNH TOÁN CHO THUỘC TÍNH INCOME

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
high	0	0	0
medium	2	1	3
low	1	1	2

- G(income=high) = $1 - (\frac{0}{0})^2 - (\frac{0}{0})^2 = 1$
- G(income=medium) = $1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 = \frac{4}{9}$
- G(income=low) = $1 - (\frac{1}{1})^2 * 2 = 0.5$

➔ L(income)= $1*0 + \frac{3}{5} * \frac{4}{9} + \frac{2}{5} * 0.5 = \frac{7}{15} = 0.4667$

TÍNH TOÁN CHO THUỘC TÍNH CREDIT_RATING

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
fair	3	0	3
excellent	0	2	2

- G(credit_rating=fair) = $1 - (\frac{0}{3})^2 - (\frac{3}{3})^2 = 0$
- G(credit_rating=excellent) = $1 - (\frac{0}{2})^2 - (\frac{2}{2})^2 = 0$

➔ L(credit_rating) = $\frac{3}{5} * 0 + \frac{2}{5} * 0 = 0$

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

-income: L(income) = 0.4667

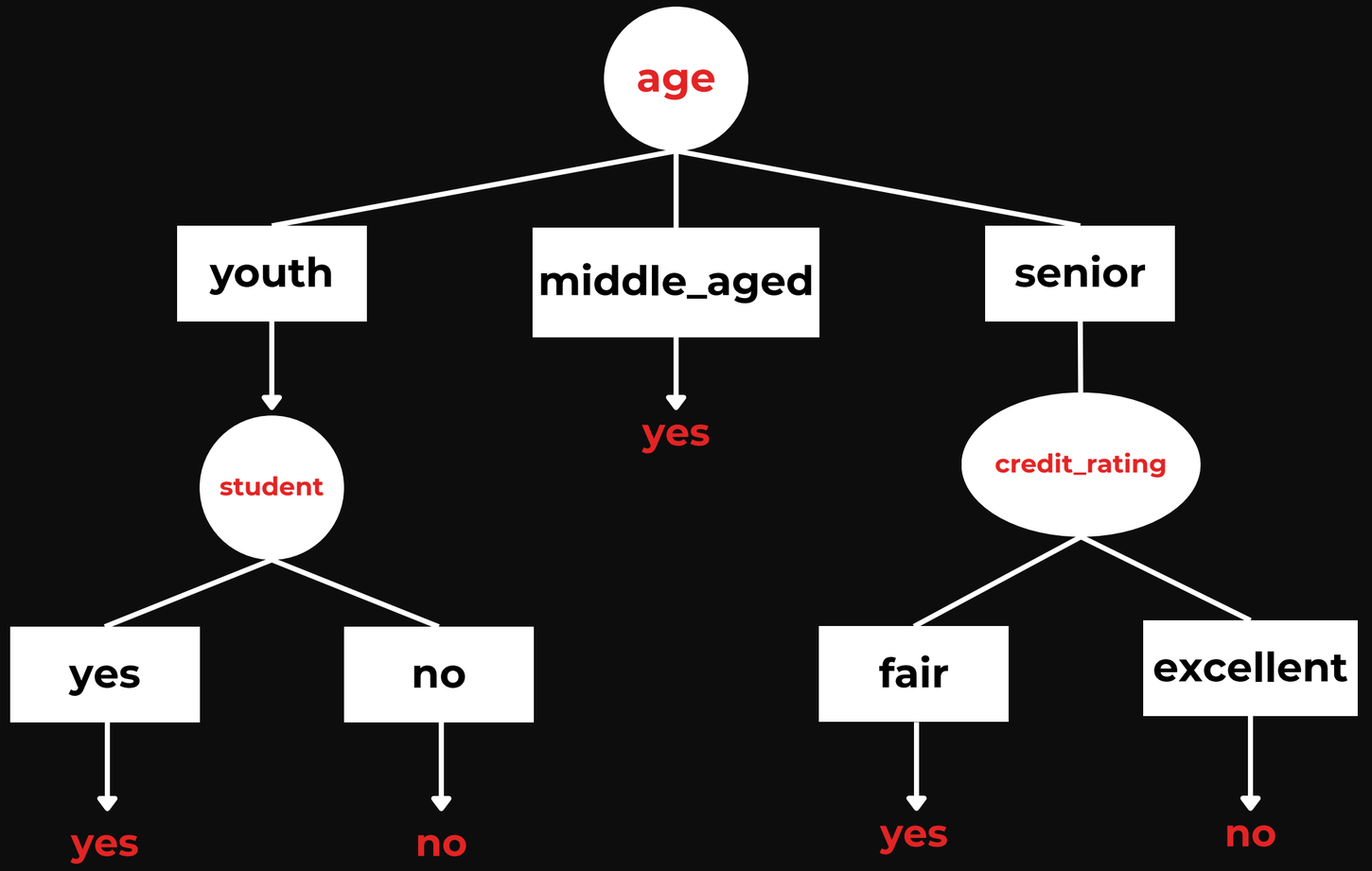
-credit_rating: L(credit_rating)= 0

➔ Chọn credit_rating làm gốc (loss có giá trị nhỏ nhất)

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no



INFORMATION GAIN

TÍNH TOÁN CHO THUỘC TÍNH CLASS: BUYS_COMPUTER

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

- Tổng số lượng dữ liệu: S=14
- Số thuộc tính thuộc nhãn Class: buys_computer = yes: 9
- Số thuộc tính thuộc nhãn Class: buys_computer = no: 5

➔ $H(S) = -(\frac{9}{14}\log_2(\frac{9}{14}) + \frac{5}{14}\log_2(\frac{5}{14})) = 0.94$

TÍNH TOÁN CHO THUỘC TÍNH AGE

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
youth	2	3	5
middle_aged	4	0	4
senior	3	2	5

$$H(S, age) = -\frac{5}{14}\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) - \frac{4}{14}\left(\frac{4}{4}\log_2\left(\frac{4}{4}\right)\right) - \frac{5}{14}\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) = 0.69$$

➔ $G(S,age) = 0.94 - 0.69 = 0.25$

TÍNH TOÁN CHO THUỘC TÍNH INCOME

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
high	2	2	4
medium	4	2	6
low	3	1	4

$H(S, income) = -\frac{4}{14}\left(\frac{2}{4}\log_2\left(\frac{2}{4}\right) * 2\right) - \frac{6}{14}\left(\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) - \frac{4}{14}\left(\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 0.91$

➔ $G(S, income) = 0.94 - 0.91 = 0.03$

TÍNH TOÁN CHO THUỘC TÍNH STUDENT

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
Yes	6	1	7
No	3	4	7

$H(S, student) = -\frac{7}{14}\left(\frac{6}{7}\log_2\left(\frac{6}{7}\right) + \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right) - \frac{7}{14}\left(\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right) = 0.79$

→ $G(S,student) = 0.94 - 0.79 = 0.15$

TÍNH TOÁN CHO THUỘC TÍNH CREDIT_RATING

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
fair	6	2	8
excellent	3	3	6

$$H(S, credit\_rating) = -\frac{8}{14}\left(\frac{6}{8}\log_2\left(\frac{6}{8}\right) + \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right) - \frac{6}{14}\left(\frac{3}{6}\log_2\left(\frac{3}{6}\right) * 2\right) = 0.89$$

➔ $G(S, credit\_rating) = 0.94 - 0.89 = 0.05$

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

-age: G(S,age) = 0.25

-income: G(S, income) = 0.03

-student: G(S,student) = 0.15

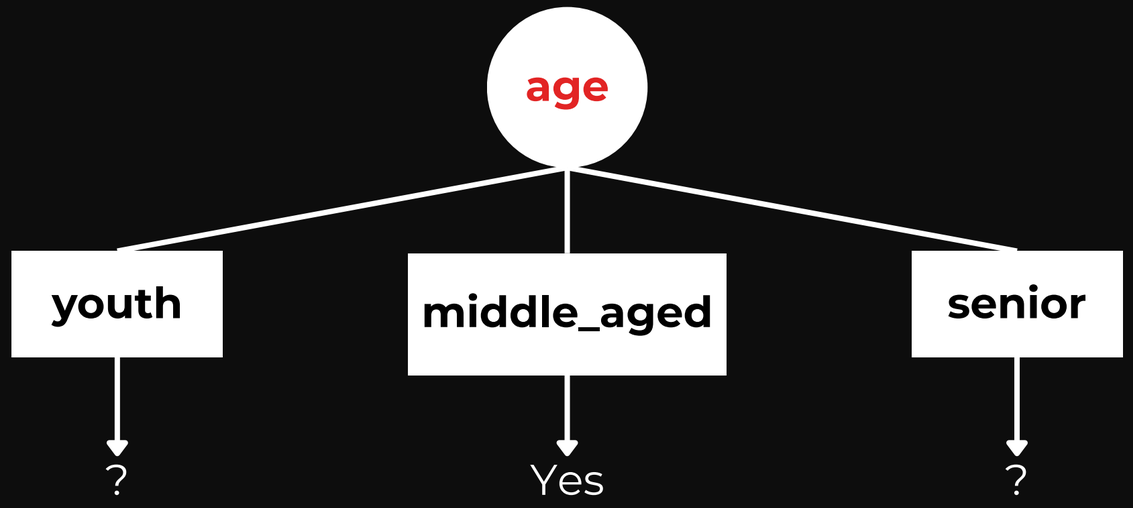
-credit_rating: G(S, credit_rating) = 0.05

➔Chọn **age** làm gốc.

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no



XÉT NHÁNH AGE = YOUTH

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

- Tổng số lượng dữ liệu: S=5
- Số thuộc tính thuộc nhãn Class: buys_computer = yes: 2
- Số thuộc tính thuộc nhãn Class: buys_computer = no: 3

→ $H(S) = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) = 0.97$

TÍNH TOÁN CHO THUỘC TÍNH INCOME

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
high	0	2	2
medium	1	1	2
low	1	0	1

$$H(S, income) = -\frac{2}{5}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) * 2\right) - \frac{1}{5}\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) = 0.4$$

→G(S, income) = 0.97 – 0.4 = 0.37

TÍNH TOÁN CHO THUỘC TÍNH STUDENT

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
Yes	2	0	2
No	0	3	3

$$H(S, student) = -\frac{2}{5}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) - \frac{3}{5}\left(\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right) = 0$$

→G(S, student) = 0.97 – 0 = 0.97

TÍNH TOÁN CHO THUỘC TÍNH CREDIT_RATING

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
fair	1	2	3
excellent	1	1	2

$$H(S, credit\_rating) = -\frac{3}{5}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) * 2\right) = 0.95$$

➔ **G(S, credit_rating) = 0.97 – 0.95 = 0.02**

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

-income: G(S, income) = 0.37

-student: G(S,student) = 0.97

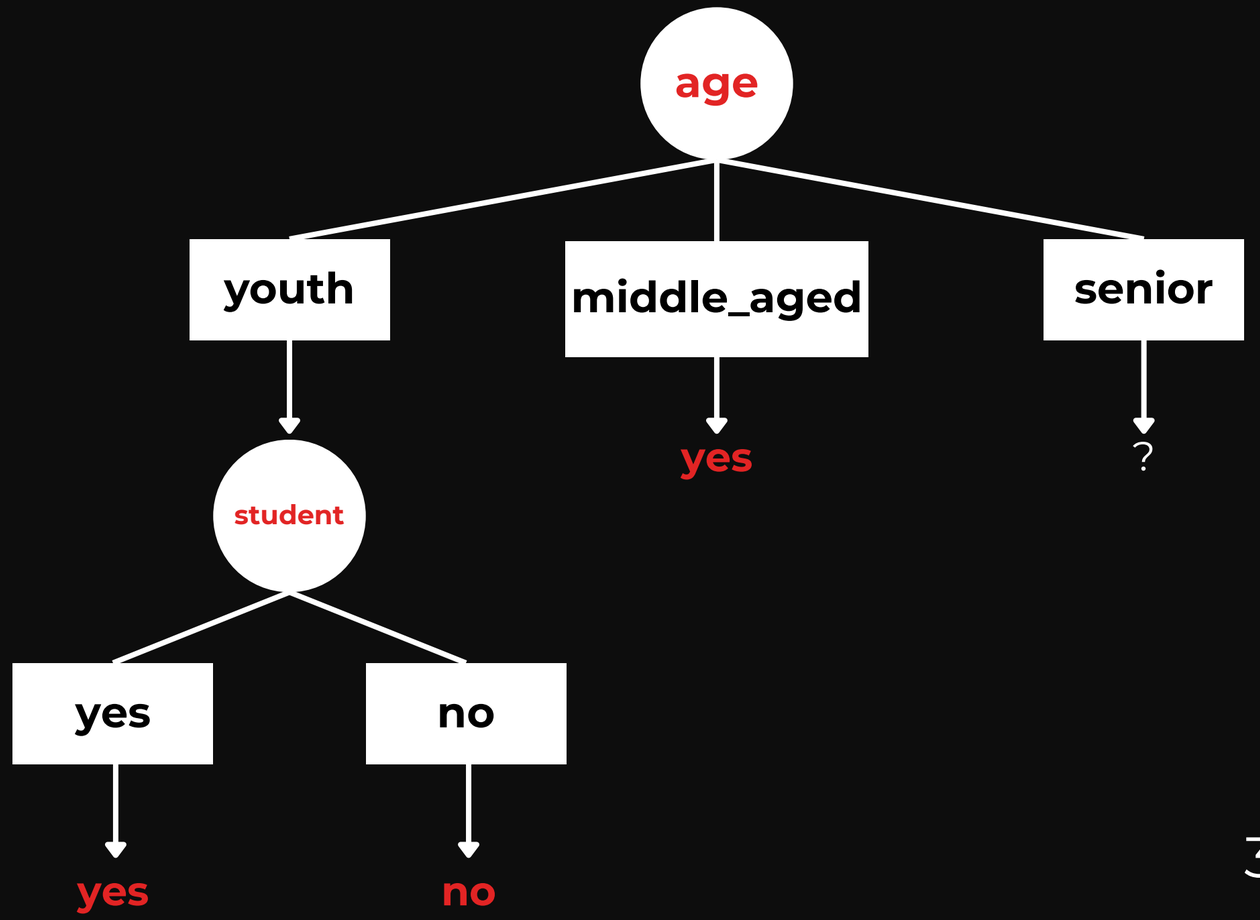
-credit_rating: G(S, credit_rating) = 0.02

➔Chọn **student** làm gốc.

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no



XÉT NHÁNH AGE = SENIOR

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

- Tổng số lượng dữ liệu: S=5
- Số thuộc tính thuộc nhãn Class: buys_computer = yes: 3
- Số thuộc tính thuộc nhãn Class: buys_computer = no: 2

➔ $H(S) = -(\frac{2}{5}\log_2(\frac{2}{5}) + \frac{3}{5}\log_2(\frac{3}{5})) = 0.97$

TÍNH TOÁN CHO THUỘC TÍNH INCOME

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
high	0	0	0
medium	2	1	3
low	1	1	2

$$H(S, income) = -\frac{3}{5}\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) * 2\right) = 0.95$$

➔**G(S, income) = 0.97 – 0.95 = 0.02**

TÍNH TOÁN CHO THUỘC TÍNH CREDIT_RATING

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

	Yes	No	Total
fair	3	0	3
excellent	0	2	2

$$H(S, \text{credit\_rating}) = -\frac{3}{5}\left(\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right) - \frac{2}{5}\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) = 0$$

→G(S, credit_rating) = 0.97 – 0 = 0.97

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

-income: G(S, income) = 0.02

-credit_rating: G(S, credit_rating) = 0.97

➔Chọn **credit_rating** làm gốc.

TỔNG HỢP

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no