IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Measures of Attribute Selection

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

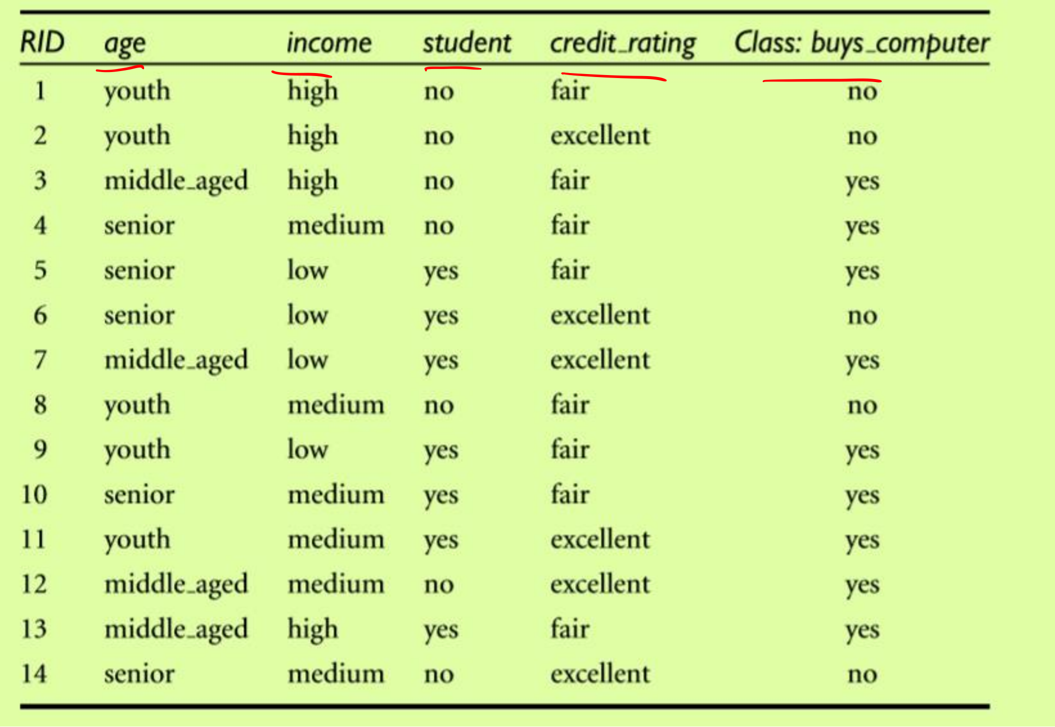
Agenda

- Measures of attribute selection using
  - Information Gain
  - Gain ratio
  - Gini Index

Example

- The following Table presents a training set, D, of class-labeled tuples randomly selected from the AllElectronics customer database

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.



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

Example

- In this example, each attribute is discrete-valued
- Continuous-valued attributes have been generalized
- The class label attribute, buys computer, has two distinct values (namely, {yes, no}); therefore, there are two distinct classes (that is, $m = 2$)
- Let class C_1 correspond to ‘yes’ and class C_2 correspond to ‘no’.
- There are nine tuples of class ‘yes’ and five tuples of class ‘no’.
- A (root) node N is created for the tuples in D

Expected information needed to classify a tuple in *D*

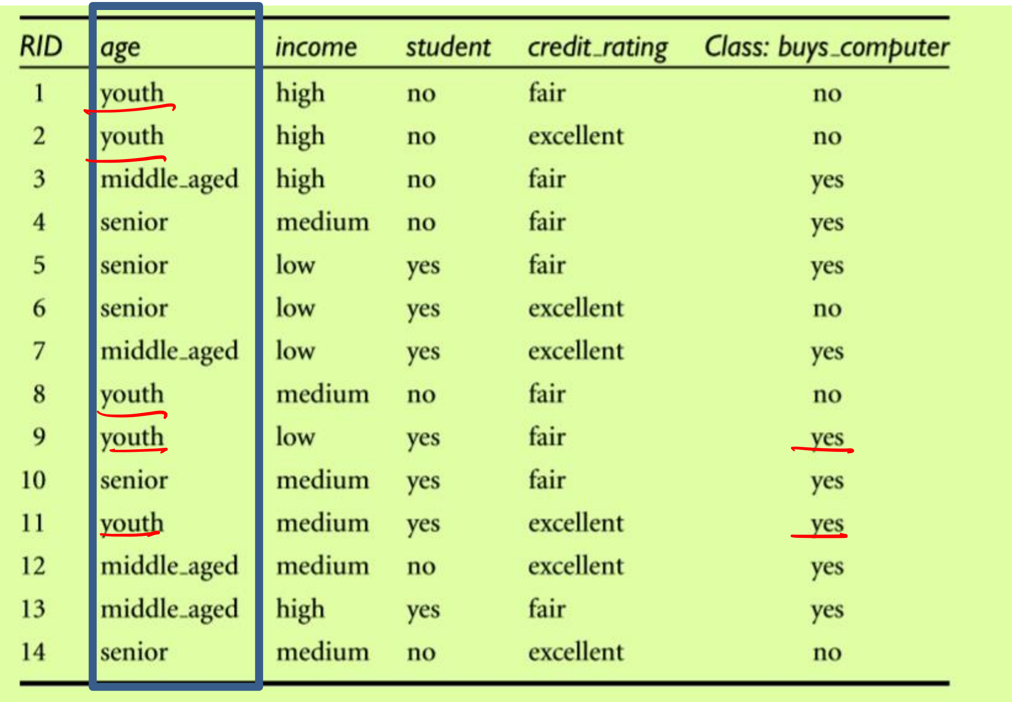
- To find the splitting criterion for these tuples, we must compute the information gain of each attribute
- Let us consider Class: buys computer as decision criteria D
- Calculate information:
- = $-p_y \log_2 (p_y) - p_n \log_2 (p_n)$
- Where p_y is probability of ‘yes’ and p_n is probability of ‘no’

$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

Calculation of entropy for ‘ Youth’

- Age can be:
  - youth
  - Middle_aged
  - Senior
- Youth

Youth	Class: buys computer
Yes	2
No	3



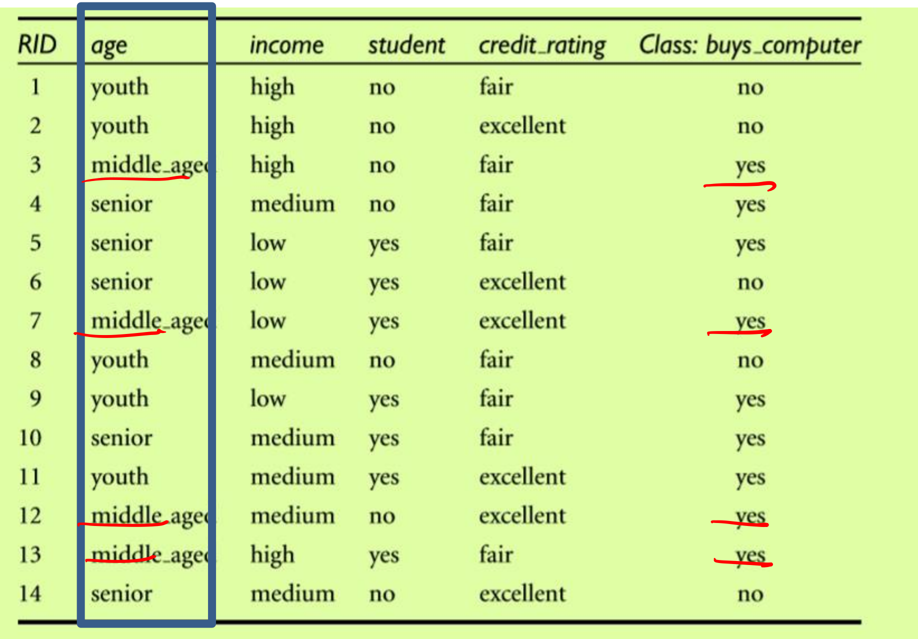
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes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

Calculation of entropy for ‘ Youth’

- Calculate Entropy for youth:
- Entropy for youth = $-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}$

- Middle_aged

middle	Class: buys computer
Yes	4
No	0



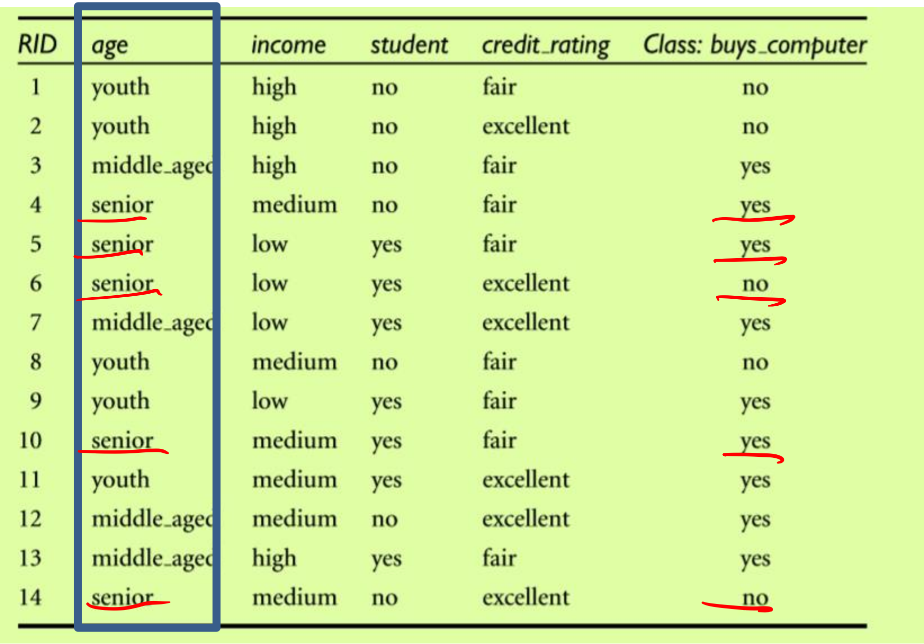
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes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

Calculation of entropy for ‘ Middle Age’

- Calculate Entropy for middle_aged
- $= -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}$
- $= \underline{0}$

- For Senior

Senior	Class: buys computer
Yes	3
No	2



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

Calculate Entropy for senior

Calculate Entropy for senior

$$= \quad -\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}$$

The expected information needed to classify a tuple in D according to age

The expected information needed to classify a tuple in D if the tuples are partitioned according to age is

$$
\begin{aligned}
Info_{age}(D) &= \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \\
&\quad + \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) \\
&\quad + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) \\
&= 0.694 \text{ bits.}
\end{aligned}
$$

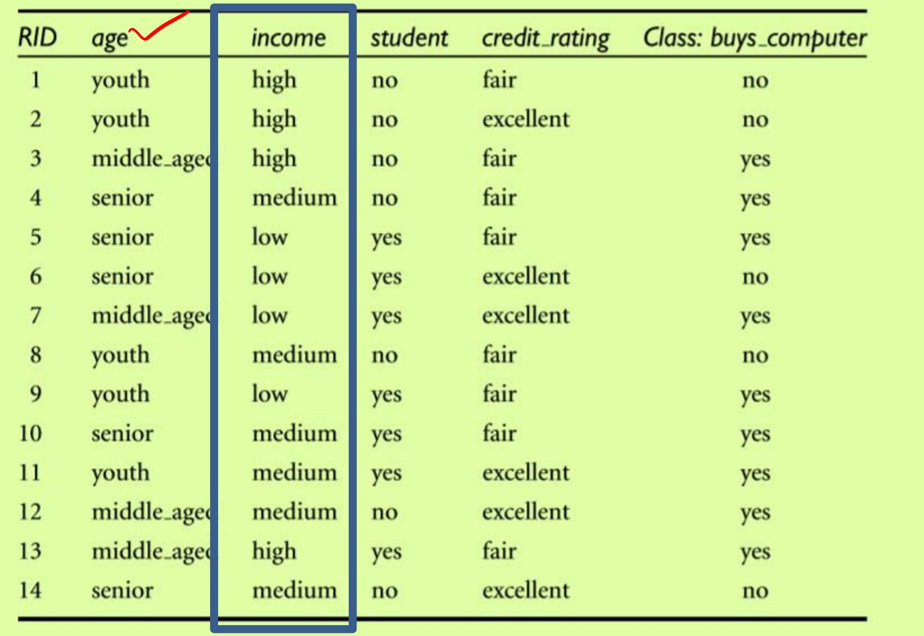
Calculation information Gain of Age

- Gain of Age:

$$Gain(age) = \underline{Info(D)} - Info_{age}(D) = \underline{0.940} - \underline{0.694} = 0.246 \text{ bits.}$$

Calculation information Gain of Income

- Calculation of gain for income:
- Income cane be:
  - High
  - Medium
  - Low



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

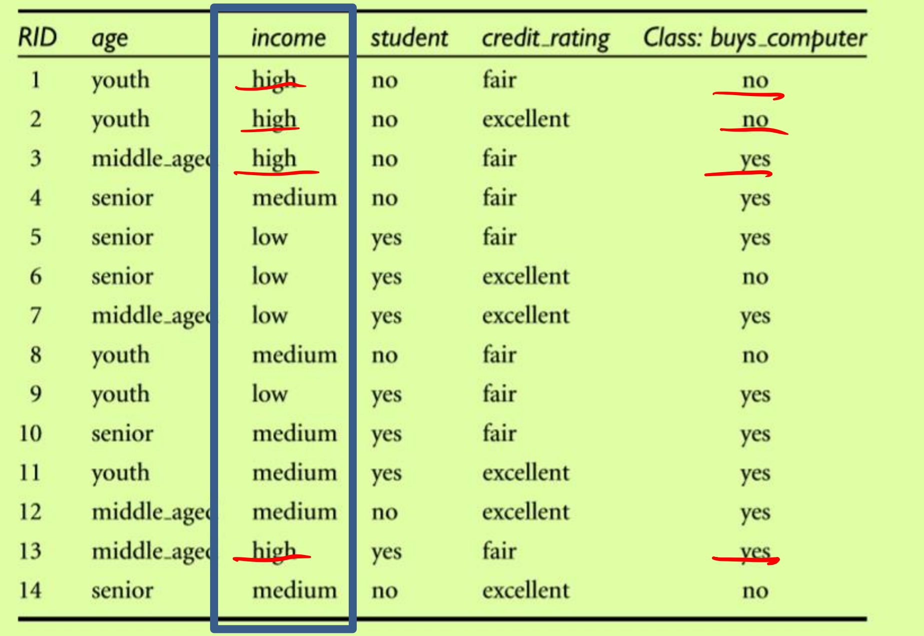
Calculate Entropy for high

- High :

High	Class: buys computer
Yes	2
No	2

- Calculate Entropy for high:

$$= -(2/4)\log_2(2/4) - (2/4)\log_2(2/4)$$



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

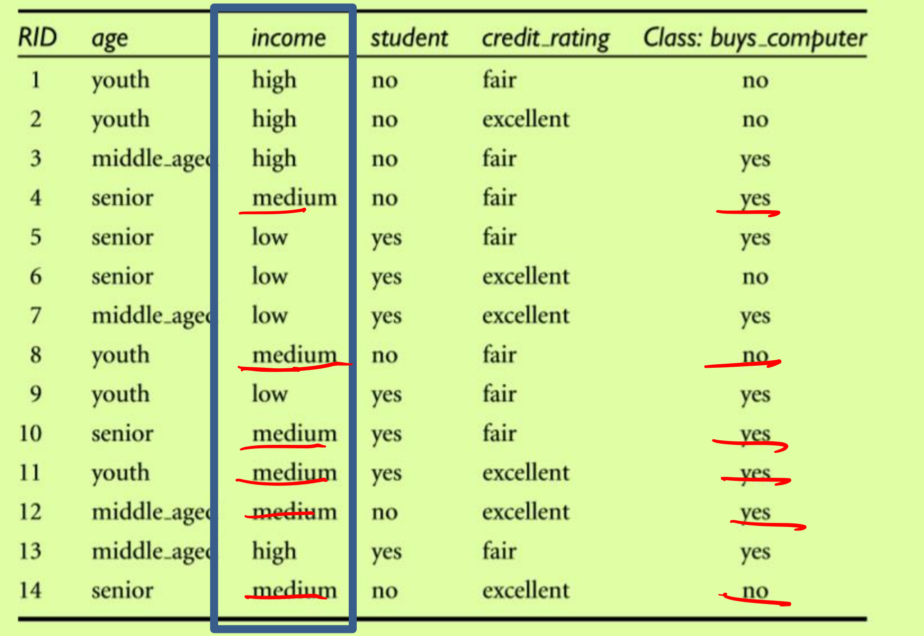
Calculate Entropy for ‘medium’

- Medium:

Medium	Class: buys computer
Yes	4
No	2

- Calculate Entropy for Medium:

$$= -(4/6)\log_2(4/6) - (2/6)\log_2(2/6)$$



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

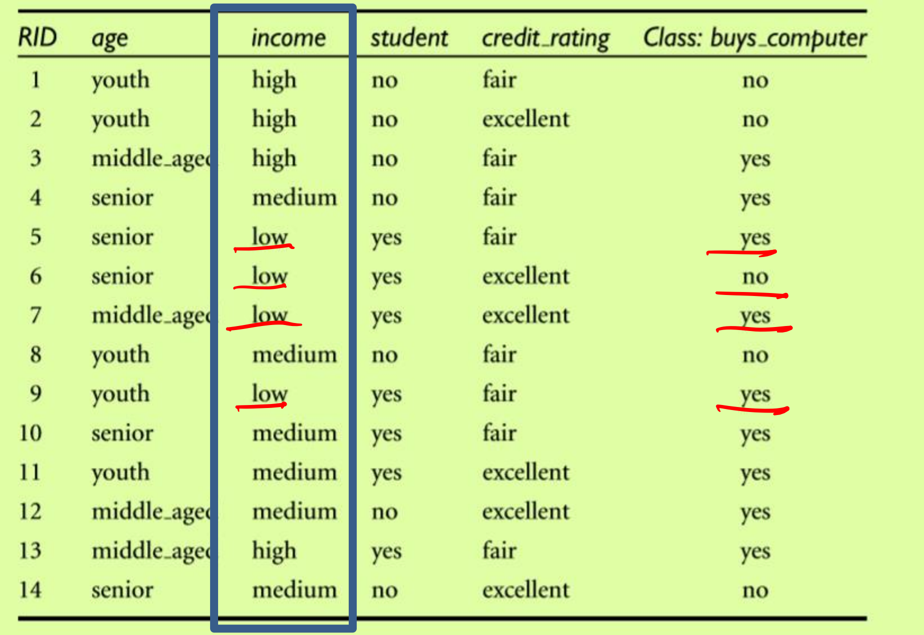
Calculate Entropy for ‘low’

- Low :

Low	Class: buys computer
No	1
Yes	3

- Calculate Entropy for Low:

$$= -(1/4)\log_2(1/4) - (3/4)\log_2(3/4)$$



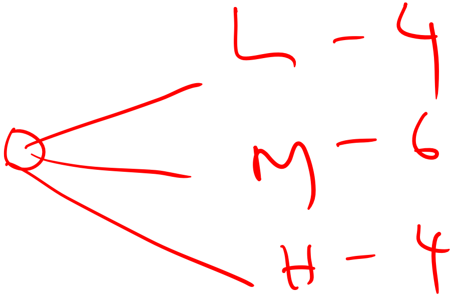
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes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

Gain of income

- The expected information needed to classify a tuple in D if the tuples are partitioned according to income is:
- $\text{Info}_{\text{income}}(D) = (4/14)(\ -(2/4)\log_2(2/4) - (2/4)\log_2(2/4)) +$
  $(6/14)(\ -(4/6)\log_2(4/6) - (2/6)\log_2(2/6)) +$
  $(4/14) -(1/4)\log_2(1/4) - (3/4)\log_2(3/4)$
  $= 0.911$

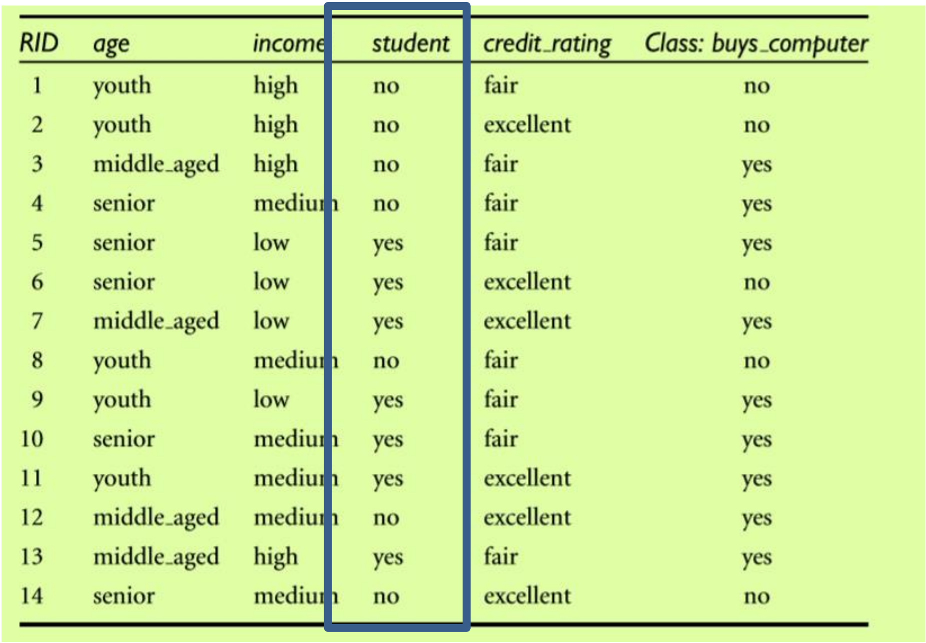
Gain of income : $\text{Info}(D) - \text{Info}_{\text{income}}(D)$

$= 0.94 – 0.911 = 0.029$



Calculation of gain for student

- Calculation of gain for student
- Student can be:
  - Yes (7)
  - No (7)



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

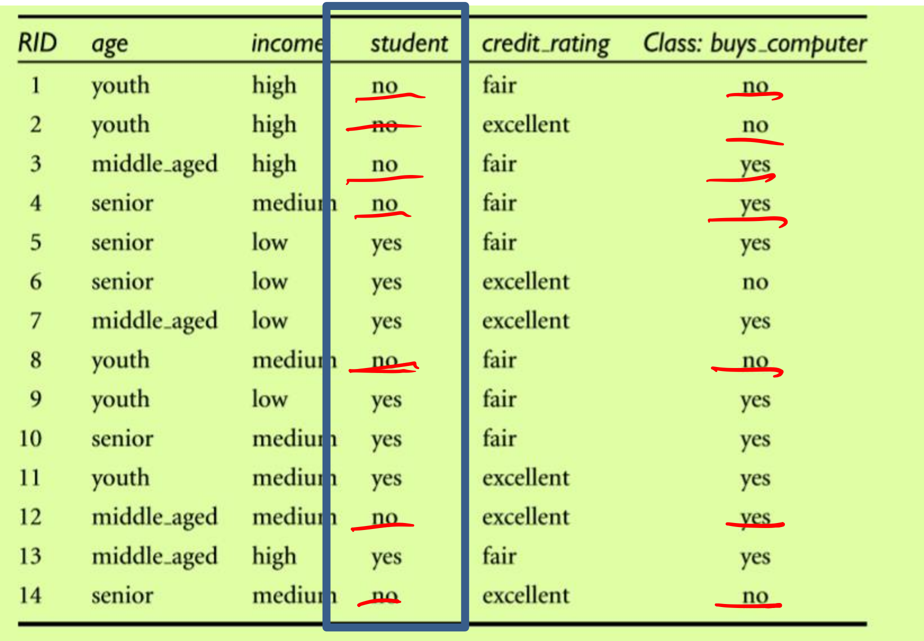
Calculate Entropy for No

- No :

No	Class: buys computer
Yes	3
No	4

- Calculate Entropy for No:

$$= -(3/7)\log_2(3/7) - (4/7)\log_2(4/7)$$



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

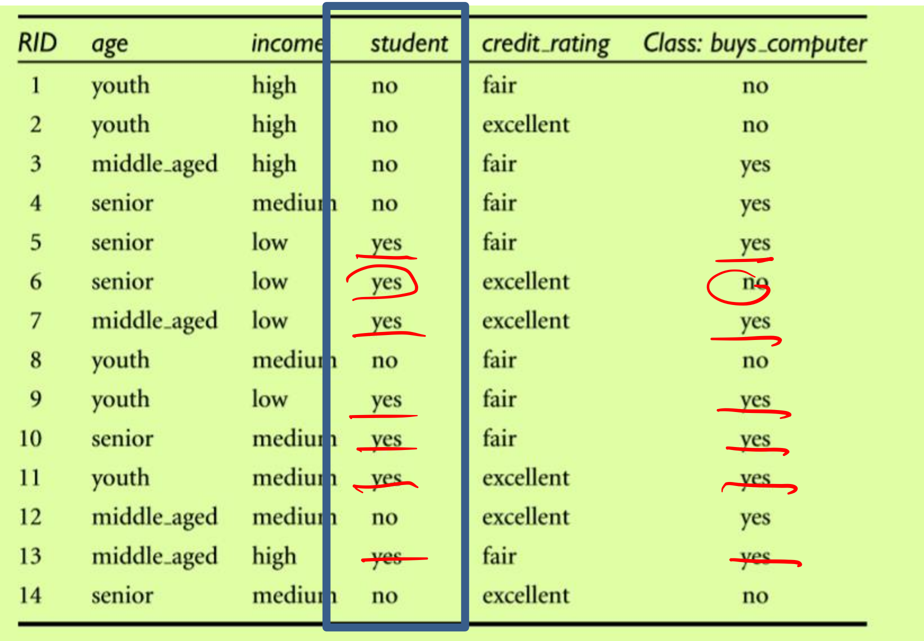
Calculate Entropy for ‘Yes’

- Yes :

Yes	Class: buys computer
Yes	6
No	1

- Calculate Entropy for Yes:

$$= -(6/7)\log_2(6/7) - (1/7)\log_2(1/7)$$



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

Gain of student

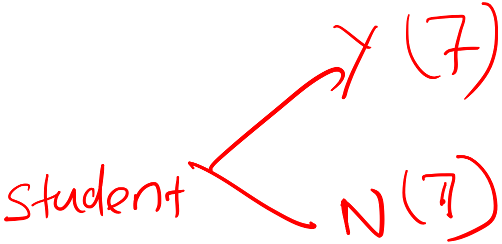
- The expected information needed to classify a tuple in D if the tuples are partitioned according to student is:
- $\text{Info}_{\text{Student}}(D) = \underline{(7/14)}\,(-(3/7)\log_2(3/7) - (4/7)\log_2(4/7)) +$

  $(7/14)\,(-(6/7)\log_2(6/7) - (1/7)\log_2(1/7))$

  $\underline{=0.789}$
- Gain(student) :

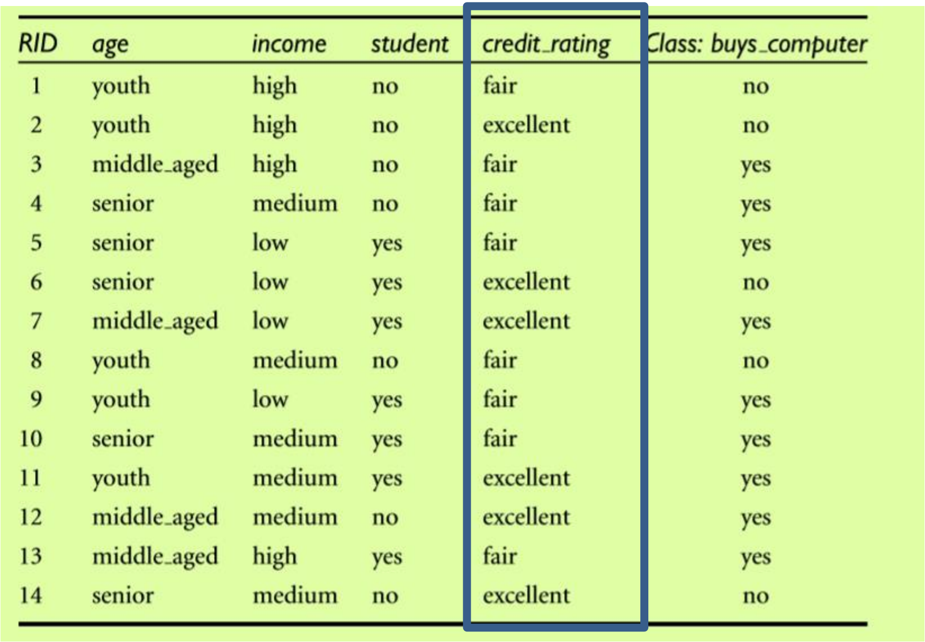
  $\text{Info}(D) - \text{Info}_{\text{student}}(D)$

  $= \underline{0.94} - \underline{0.789} = \underline{0.151}$



Calculation of gain for credit rating

- Calculation of gain for credit rating
- Credit rating can be:
  - Fair — 8
  - Excellent - 6



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

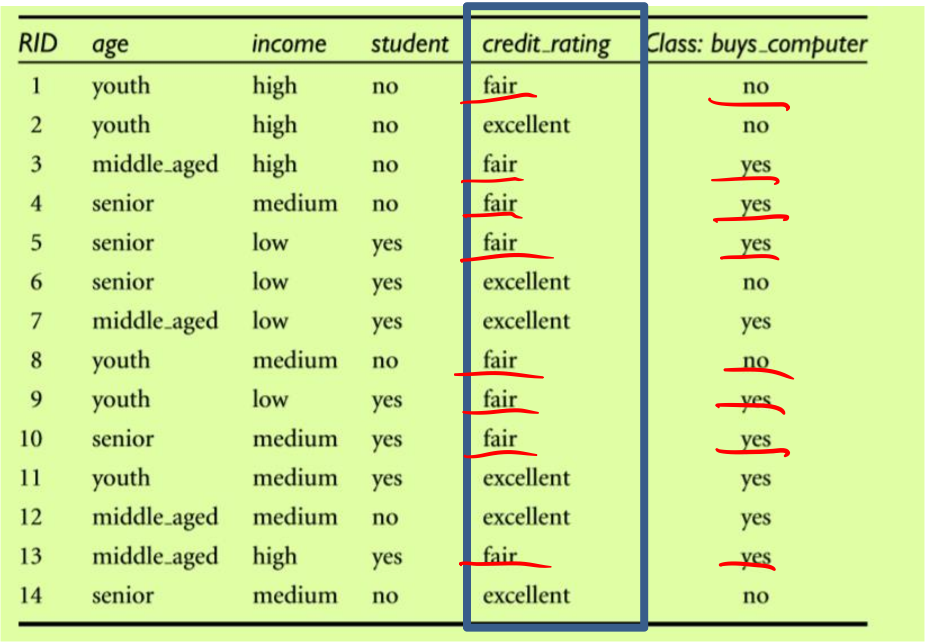
Calculate Entropy for Fair

- Fair :

Fair	Class: buys computer
Yes	6
No	2

- Calculate Entropy for Fair:

$$= -(6/8)\log_2(6/8) - (2/8)\log_2(2/8)$$



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

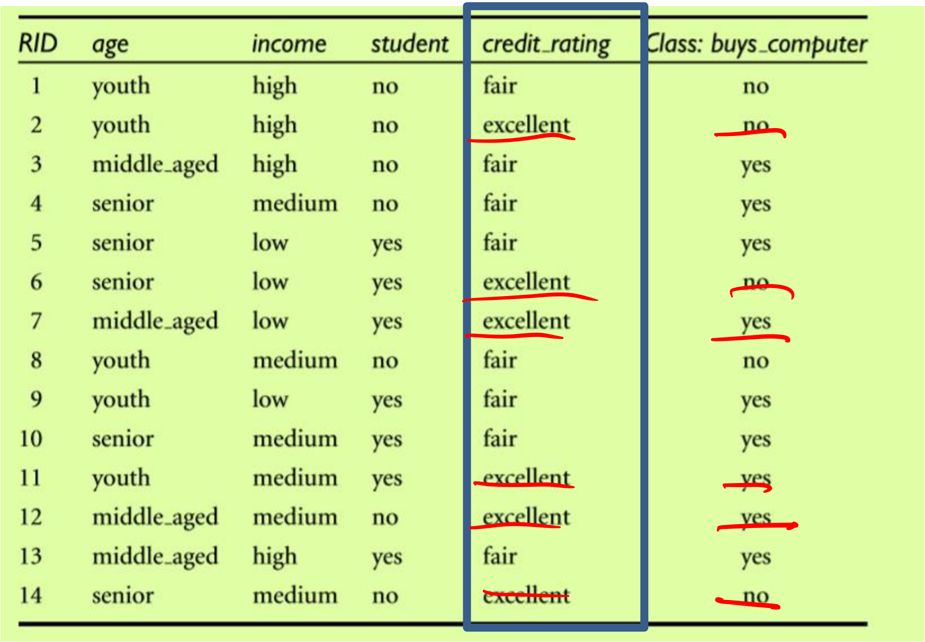
Calculate Entropy for Excellent

- Excellent :

Yes	Class: buys computer
Yes	3
No	3

- Calculate Entropy for Excellent:

$$= -(3/6)\log_2(3/6) - (3/6)\log_2(3/6)$$



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

Gain for credit rating

- The expected information needed to classify a tuple in D if the tuples are partitioned according to Credit rating is:
- $\text{Info}_{\text{Credit rating}}(D) = (8/14)\,(-(6/8)\log_2(6/8) - (2/8)\log_2(2/8)) +$

  $(6/14)\,(-(3/6)\log_2(3/6) - (3/6)\log_2(3/6))$

  $=0.892$

- Gain for credit rating :

  $\text{Info}(D) - \text{Info}_{\text{Credit rating}}(D)$

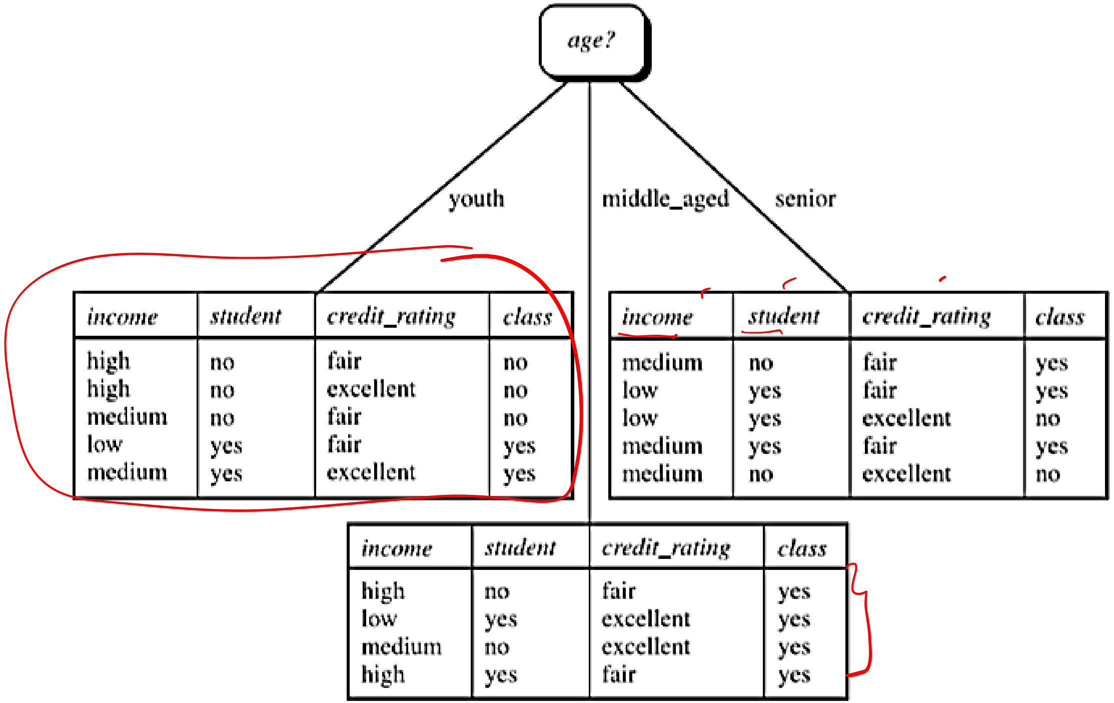
  $= \underline{0.94} – \underline{0.892} = \underline{0.048}$

Independent variable	Information gain
Age	0.246
Income	0.029
Student	0.151
Credit_rating	0.048

Selection of root classifier

- Because age has the highest information gain among the attributes, it is selected as the splitting attribute
- Node N is labelled with age, and branches are grown for each of the attribute’s values
- The tuples are then partitioned accordingly
- Notice that the tuples falling into the partition for age = middle aged all belong to the same class
- Because they all belong to class “yes,” a leaf should therefore be created at the end of this branch and labelled with “yes.”

Decision tree



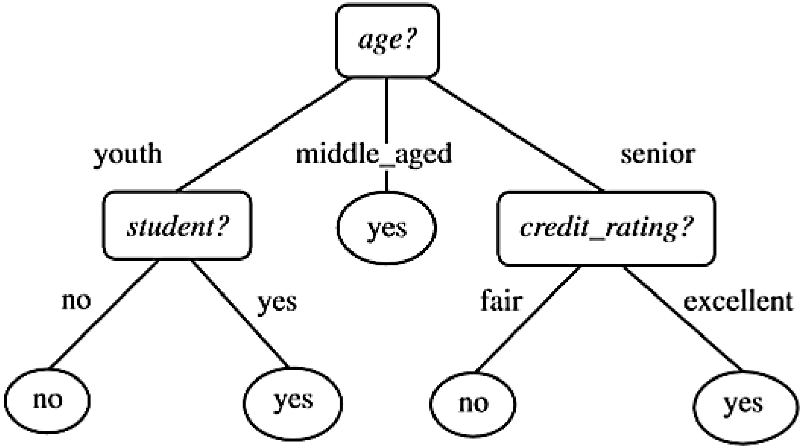
income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Decision tree

- The final decision tree returned by the algorithm is shown in Figure



Thank You