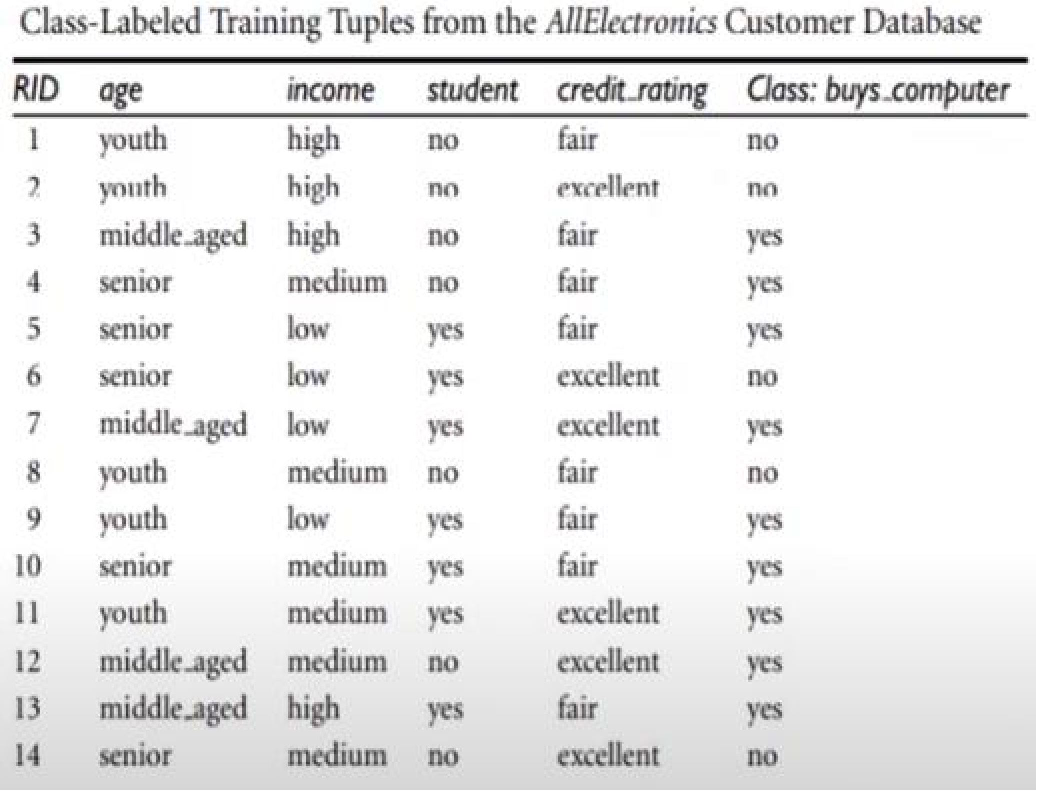
Decision Tree

Dataset:



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

Entropy of Dataset Class:

Entropy ([9+, 5-]) = $-(9/14*\log_2 9/14 + 5/14*\log_2 5/14) = 0.94$

Information Gain for (age):

S = [+9, -5]
S_{youth} = [+2, -3]
S_{middle_aged} = [+4, -0]
S_{senior} = [+3, -2]
Entropy = 0.94

$\text{Entropy}_{youth} = -(2/5*\log_2 2/5 + 3/5*\log_2 3/5) = 0.97$
$\text{Entropy}_{middle\_aged} = -(4/4*\log_2 4/4 + 0/4*\log_2 0/4) = 0$
$\text{Entropy}_{senior} = -(3/5*\log_2 3/5 + 2/5*\log_2 2/5) = 0.97$

Information Gain $_{\text{(age)}}$ = 0.94 - (5/14*0.97 + 4/14*0 + 5/14*0.97) = **0.24**

Information Gain for (income):

$S = [+9, -5]$
$S_{\text{high}} = [+2, -2]$
$S_{\text{medium}} = [+4, -2]$
$S_{\text{low}} = [+3, -1]$
$\text{Entropy} = 0.94$

$\text{Entropy}_{\text{high}} = -(2/4*\log_2(2/4) + 2/4*\log_2(2/4)) = 1$
$\text{Entropy}_{\text{medium}} = -(4/6*\log_2(4/6) + 2/6*\log_2(2/6)) = 0.92$
$\text{Entropy}_{\text{low}} = -(3/4*\log_2(3/4) + 1/4*\log_2(1/4)) = 0.81$

Information Gain $_{\textbf{(Income)}} = 0.94 - (4/14*1 + 6/14*0.92 + 4/14*0.81) = \mathbf{0.02}$

Information Gain for (student):

$S = [+9, -5]$
$S_{\text{no}} = [+3, -4]$
$S_{\text{yes}} = [+6, -1]$
$\text{Entropy} = 0.94$

$\text{Entropy}_{\text{no}} = -(3/7*\log_2(3/7) + 4/7*\log_2(4/7)) = 0.98$
$\text{Entropy}_{\text{yes}} = -(6/7*\log_2(6/7) + 1/7*\log_2(1/7)) = 0.59$

Information Gain $_{\textbf{(student)}} = 0.94 - (7/14*0.98 + 7/14*0.59) = \mathbf{0.15}$

Information Gain for (credit_rating):

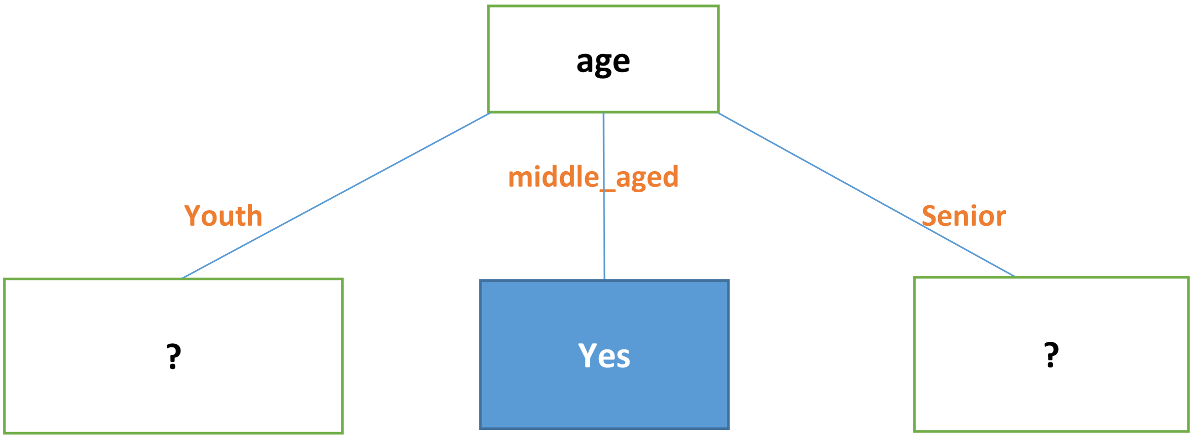
$S = [+9, -5]$
$S_{\text{fair}} = [+6, -2]$
$S_{\text{excellent}} = [+3, -3]$
$\text{Entropy} = 0.94$

$\text{Entropy}_{\text{fair}} = -(6/8*\log_2(6/8) + 2/8*\log_2(2/8)) = 0.81$
$\text{Entropy}_{\text{excellent}} = -(3/6*\log_2(3/6) + 3/6*\log_2(3/6)) = 1$

Information Gain $_{\textbf{(credit\_rating)}} = 0.94 - (8/14*0.81 + 6/14*1) = \mathbf{0.04}$

Information Gain for four attributes are as follow:

Gain (S, age) = 0.24
Gain (S, student) = 0.15
Gain (S, credit_rating) = 0.04
Gain (S, income) = 0.02



S_{youth} = [+2, -3]
Entropy_{youth} = $-(2/5 * \log_2 (2/5) + (3/5) * \log_2 (3/5)) = 0.97$

Information Gain (Youth) with respect to Income attribute:

Income = high, medium, low
Income_{high} = [+0, -2]
Income_{medium} = [+1, -1]
Income_{low} = [+1, -0]

Entropy Income_{high} = $-(0/2*\log_2 (0/2) + 2/2*\log_2 (2/2)) = 0$
Entropy Income_{medium} = $-(1/2*\log_2 (1/2) + 1/2*\log_2 (1/2)) = 1$
Entropy Income_{low} = $-(1/1*\log_2 (1/1) + 0/1*\log_2 (0/1)) = 0$

Information Gain (S_{youth}, Income) = 0.97 - (2/5*0 + 2/5*1 + 1/5*0) = **0.57**

Information Gain (Youth) with respect to Student attribute:

Student = yes, no

Student_{no} = [+0, -3]
Student_{yes} = [+2, -0]

Entropy student_{no} = $-(0/3*\log_2(0/3) + 3/3*\log_2(3/3)) = 0$
Entropy student_{yes} = $-(2/2*\log_2(2/2) + 0/2*\log_2(0/2)) = 0$

Information Gain (S_{youth}, student) = 0.97 - (3/5*0 + 2/5*0) = **0.97**

Information Gain (Youth) with respect to credit_rating attribute:

credit_rating = fair, excellent

$\text{credit\_rating}_{fair}$ = [+1, -2]
$\text{credit\_rating}_{excellent}$ = [+1, -1]

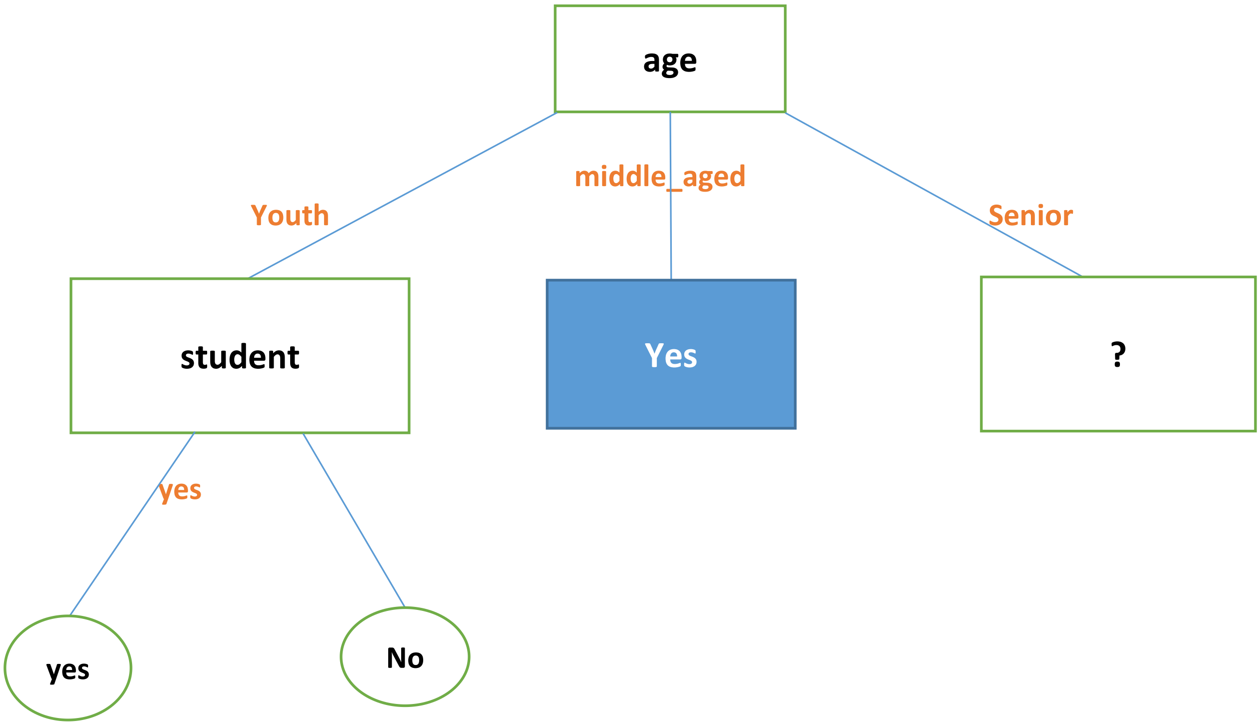
Entropy $\text{credit\_rating}_{fair}$ = $-(1/3*\log_2(1/3) + (2/3*\log_2(2/3)) = 0.91$
Entropy $\text{credit\_rating}_{excellent}$ = $-(1/2*\log_2(1/2) + (1/2*\log_2(1/2)) = 1$

Information Gain (S_{youth}, credit_rating) = 0.97 - (3/5*0.91 + 2/5*1) = **0.02**

Finally we get:

Gain (S_{youth}, student) = 0.97
Gain (S_{youth}, Income) = 0.57
Gain (S_{youth}, credit_rating) = 0.02

So highest information is gain by student attribute relative to age (youth), so left node will be of student attribute.



S_{senior} = [+3, -2]

$Entropy_{senior}$ = $-(3/5*\log_2(3/5) + 2/5*\log_2(2/5)) = 0.97$

Information Gain (Senior) with respect to Income attribute:

Income = high, medium, low

$Income_{high}$ = [+0, -0]

$Income_{medium}$ = [+2, -1]

$Income_{low}$ = [+1, -1]

Entropy $Income_{high}$ = $-(0/0*\log_2(0/0) + 0/0*\log_2(0/0)) = 0$

Entropy $Income_{medium}$ = $-(2/3*\log_2(2/3) + 1/3*\log_2(1/3)) = 0.91$

Entropy $Income_{low}$ = $-(1/2*\log_2(1/2) + 1/2*\log_2(1/2)) = 1$

Information Gain (S_{senior}, income) = 0.97 - (0/5*0 + 3/5*0.91 + 2/5*1) = **0.02**

Information Gain (Senior) with respect to student attribute:

student = no, yes

$student_{no}$ = [+1, -1]

$student_{yes}$ = [+2, -1]

Entropy $student_{no}$ = $-(1/2*\log_2(1/2) + (1/2*\log_2(1/2)) = 1$

Entropy $student_{yes}$ = $-(2/3*\log_2(2/3) + (1/3*\log_2(1/3)) = 0.91$

Information Gain (S_{senior}, student) = 0.97 - (2/5*1 + 3/5*0.91) = **0.02**

Information Gain (Senior) with respect to credit_rating attribute:

credit_rating = fair, excellent
$\text{credit\_rating}_{\text{fair}}$ = [+3, -0]
$\text{credit\_rating}_{\text{excellent}}$ = [+0, -2]

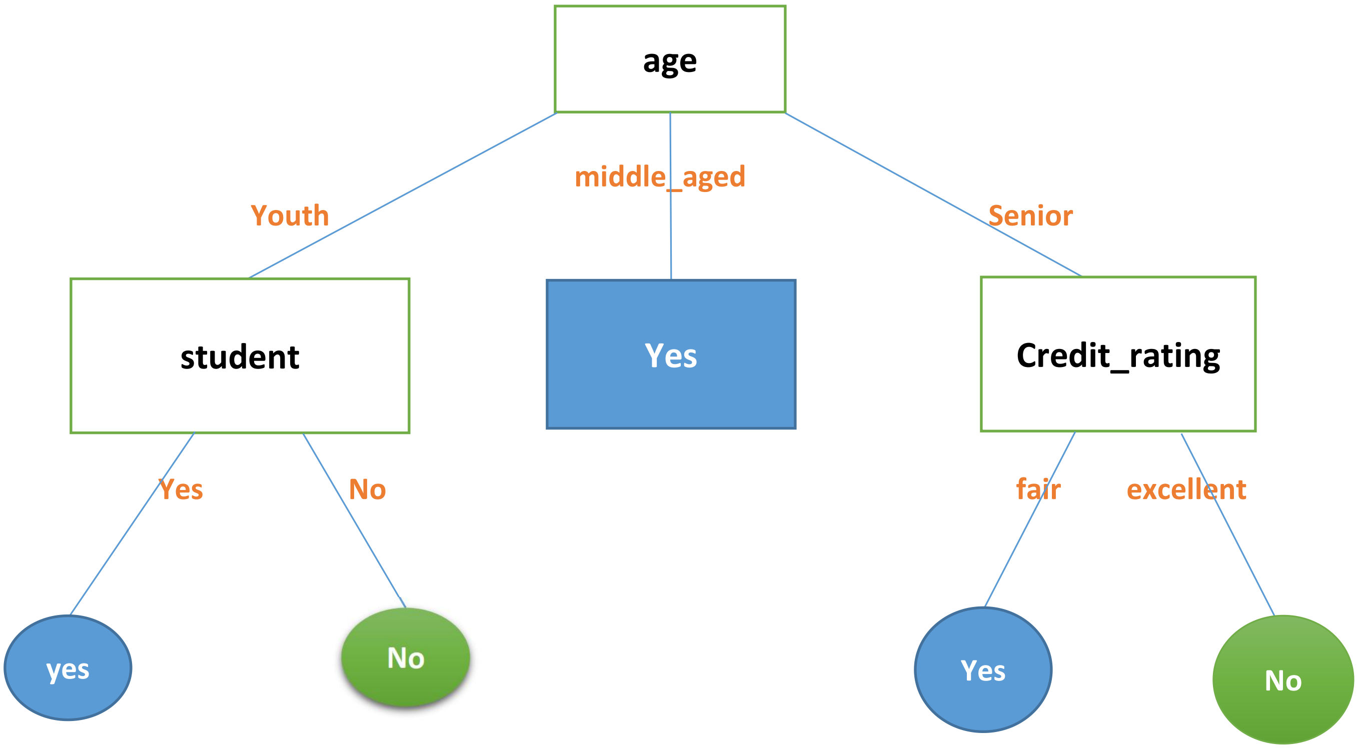
Entropy $\text{credit\_rating}_{\text{fair}}$ $= -(3/3*\log_2 (3/3) + (0/3*\log_2 (0/3)) = 0$
Entropy $\text{credit\_rating}_{\text{excellent}}$ $= -(0/2*\log_2 (0/2) + (2/2*\log_2 (2/2)) = 0$

Information Gain (S_{senior}, credit_rating) = 0.97 - (3/5*0 + 2/5*0) = **0.97**

Finally we get:

Gain (S_{senior}, student) = 0.02
Gain (S_{senior}, Income) = 0.02
Gain (S_{senior}, credit_rating) = 0.97

So highest information is gain by credit_rating attribute relative to age (senior), so right node will be of credit_rating attribute.



So Decision Tree is constructed. When age = middle_edge, customer buy computer, when age = youth and student = yes then customer buy computer. When age = senior and credit_rating = fair then customer buy computer