Assignment - 1 (ML)

Name- Sourab Nad
Roll- 191333
CS-65

Decision tree using entropy and information gain

→ Entropy (S) of entire dataset

$$S = [9^+, 5^-] = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.94$$

→ Find entropy of individual attribute (age, income, student, credit_rating)

i) age → youth, middle-aged, senior

youth $= [2^+, 3^-]$ entropy $= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.971$

middle-aged $= [4^+, 0^-]$ entropy $= -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log$

$= 0$

senior $= [3^+, 2^-]$ entropy $= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{3}{5}$

$= 0.971$

Gain (S, Age) = entropy (S) - $\sum \frac{|S_v|}{|S|}$ entropy (S_v)

$= 0.94 - \frac{5}{14} * 0.971 - \frac{4}{14} * 0 - \frac{5}{14} * 0.971 =$ **0.2464**

ii) Income → low, medium, high

low = [3+, 1-] $e = \frac{-3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$

medium = [4+, 2-] $e = \frac{-4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.9183$

high = [2+, 2-] $e = \frac{-2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

$Gain(S, income) = 0.94 - \frac{4}{14} * 0.8113 - \frac{6}{14} * 0.9183 - \frac{4}{14} * 1$

Gain(income) = 0.0289

iii) Student → NO, yes

no = [3+, 4-] $e = \frac{-3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.9852$

yes = [6+, 1-] $e = \frac{-6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.5916$

$Gain(S, student) = entropy(S) - \frac{7}{14} * 0.9852 - \frac{7}{14} * 0.5916$

$= 0.01516$

iv) Credit-rating → fair, excellent

fair = $[6^+, 2^-]$ $\quad e = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.8113$

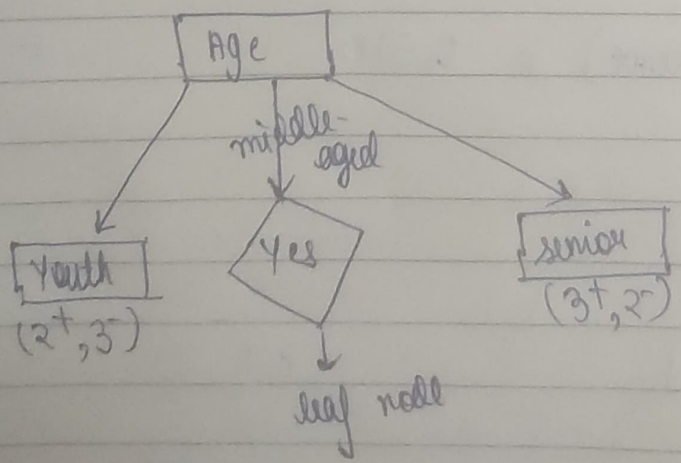
excellent = $[3^+, 3^-]$ $\quad e = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$

$Gain(S, \text{credit\_rating}) = 0.94 - \frac{6}{14} * 1 - \frac{8}{14} * 0.8113 = 0.0478$

Gain (S, Age) = 0.2464
Gain (S, income) = 0.0289
Gain (S, student) = 0.01516
Gain (S, credit_rating) = 0.0478

Since, gain of Age is maximum.
∴ Age is our root node.

Age → youth $(D_1, D_2, D_8, D_9, D_{11})$ → 2^+, 3^-
Age → middle-aged $(D_3, D_7, D_{12}, D_{13})$ → 4^+, 0^-
Age → senior $(D_4, D_5, D_6, D_{10}, D_{14})$ → 3^+, 2^-



Now, consider only 'youth' and 'senior'.

For youth

Data point	Income	student	credit rating	Buys
D_1	high	no	fair	no
D_2	high	no	excellent	no
D_8	medium	no	fair	no
D_9	low	yes	fair	yes
D_{11}	medium	yes	excellent	yes

$S_{youth} = [2^+, 3^-]$ $\quad e = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.97$

i) income

high = $[0^+, 2^-]$ $\quad e = -\frac{2}{2}\log_1\frac{2}{2} = 0$

medium = $[1^+, 1^-]$ $\quad e = 1$

low = $[1^+, 0^-]$ $\quad e = 0$

Gain (S_{youth}, income) = 0.570

ii) student

no = $[0^+, 3^-]$ $\quad e = 0$

yes = $[2^+, 0^-]$ $\quad e = 0$

Gain (S_{youth}, student) = 0.97

iii) Credit_rating

fair = $[1^+, 2^-]$ $\quad e = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.9183$

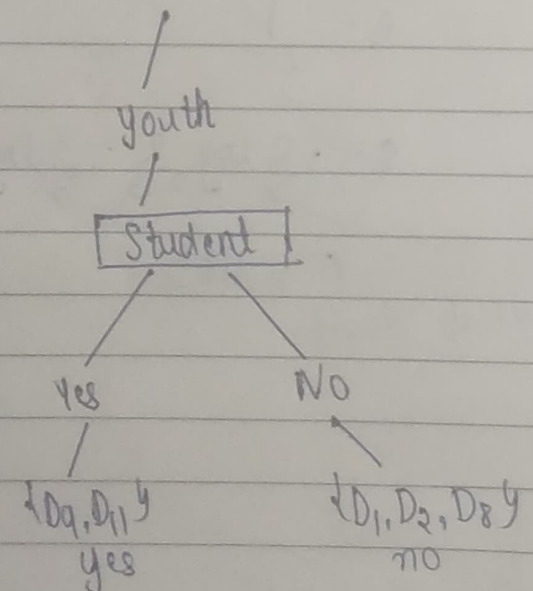
excellent = $[1^+, 1^-]$ $\quad e = 1$

Gain (S_{youth}, cred.) = 0.0192

Gain (S_{youth}, income) = 0.570

Gain (S_{youth}, student) = 0.97

∴ Gain of student is maximum
∴ student will be considered as node.



For senior

Data point	income	student	credit_rating	buys
D_4	medium	no	fair	yes
D_5	low	yes	fair	yes
D_6	low	yes	excellent	no
D_{10}	medium	yes	fair	yes
D_{14}	medium	no	excellent	no

$S_{senior} = [3^+, 2^-]$ $\quad e = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.97$

i) income

low $= [1^+, 1^-]$ $\quad e = 1$

medium $= [2^+, 1^-]$ $\quad e = -\frac{2}{3}\log\frac{2}{2\,3} - \frac{1}{3}\log\frac{1}{2\,3}$

$= 0.9183$

Gain $(S_{senior}, income) = 0.0192$

ii) student

no $= [1^+, 1^-]$ $\quad e = 1$

yes $= [2^+, 1^-]$ $\quad e = 0.9183$

Gain $(S_{senior}, student) = 0.0192$

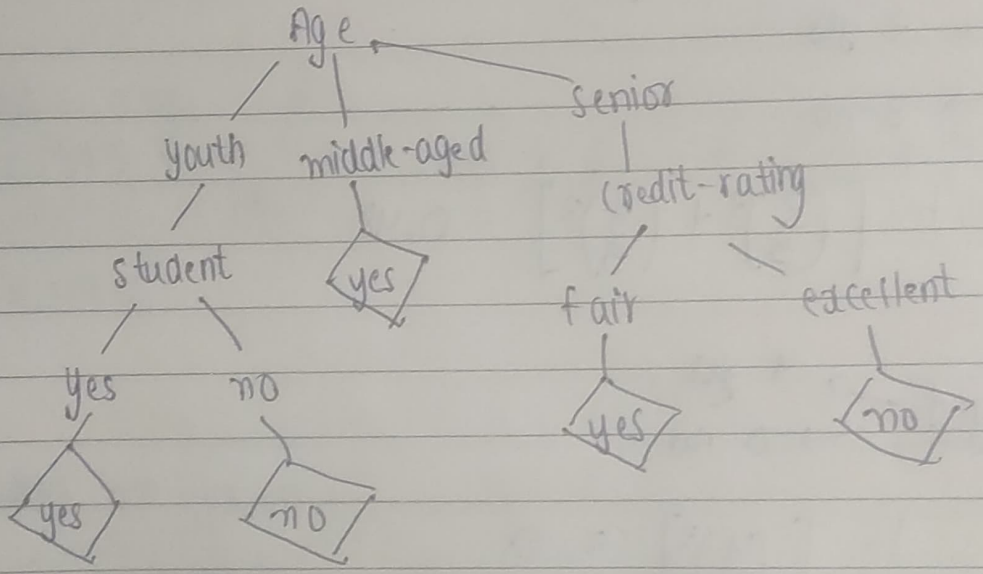
iii) credit_rating

fair = [3+, 0-] $e = 0$

excellent = [0+, 2-] $e = 0$

Gain(S_{senior}, credit_rating) = 0.97

or, next node is credit_rating



↑ final decision tree

Decision tree using 'gini index'

→ Gini index for overall dataset

Buys computer

$$Gini(S) = 1 - \left[\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right] = 0.459$$

→ Age → youth, middle-aged, senior

youth → 2 yes
→ 3 no

$$Gini = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right] = 0.48$$

middle-aged → 4 yes
→ 0 no

$$Gini = 1 - \left[\left(\frac{4}{4}\right)^2\right] = 0$$

senior → 3 yes
→ 2 no

$$Gini = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0.48$$

$$\text{Weighted avg. } g = 0.48 * \left(\frac{5}{14}\right) + 0 + 0.48 * \left(\frac{5}{14}\right)$$

$$= \frac{4.8}{14} = 0.343$$

ii) Income → low, medium, high

low → 3 yes, 1 no Gini(g) = $1-\left[\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right] = 0.375$

medium → 4 yes, 2 no $g = 1-\left[\left(\frac{4}{6}\right)^2+\left(\frac{2}{6}\right)^2\right] = 0.444$

high → 2 yes, 2 no $g = 1-\left[2*\left(\frac{2}{4}\right)^2\right] = 0.5$

weighted avg. g

$= 0.375 * \frac{4}{14} + 0.444 * \frac{6}{14} + \frac{0.5 * 4}{14}$

$= 0.440$

iii) student → yes, no

yes → 6 yes, 1 no $g = 1-\left[\left(\frac{6}{7}\right)^2+\left(\frac{1}{7}\right)^2\right] = 0.245$

no → 3 yes, 4 no $g = 1-\left[\left(\frac{3}{7}\right)^2+\left(\frac{4}{7}\right)^2\right] = 0.490$

weighted avg. g

$= 0.245 * \frac{7}{14} + 0.49 * \frac{7}{14}$ 0.735

$= 0.3675$

iv) credit_rating → fair, excellent

fair → 6 yes, 2 no

$g = 1 - \left[\left(\frac{6}{8}\right)^2 + \left(\frac{2}{8}\right)^2\right] = 0.375$

excellent → 3 yes, 3 no

$g = 0.5$

Weighted avg. g

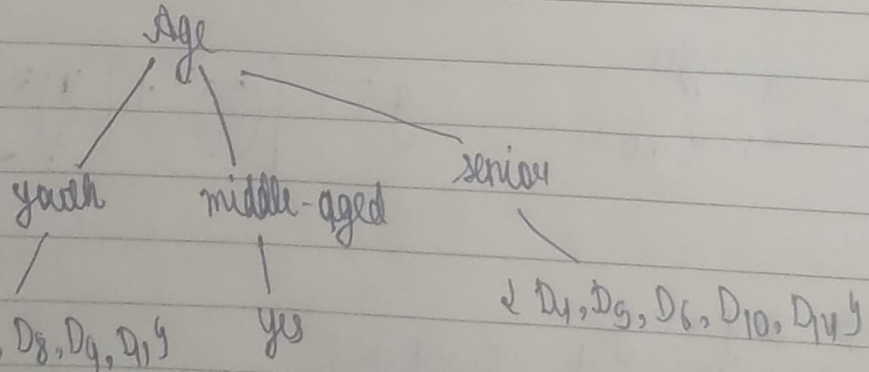
$= 0.375 * \frac{8}{14} + 0.5 * \frac{6}{14}$

$= 0.429$

3	6 4 3,000

g (age) = 0.343
g (income) = 0.440
g (student) = 0.368
g (credit_rating) = 0.429

∴ g is smallest for 'age' attribute
∴ it will be root node



youth → 2 yes
→ 3 no

$g = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right]$

ii) income → low, medium, high

low → 1 yes
→ 0 no

$g = 0$

medium → 1 yes
→ 1 no

$g = 0.5$

high → 0 yes
→ 2 no

$g = 0$

weighted avg. $g = 0 + 0.5 * \frac{2}{5} + 0$

$= 0.2$

ii) student → yes, no

yes → 2 yes
→ 0 no

$g = 0$

no → 0 yes
→ 3 no

$g = 0$

Weighted avg. $g = 0$

iii) credit_rating → fair, excellent

fair → 1 yes
→ 2 no

$g = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.444$

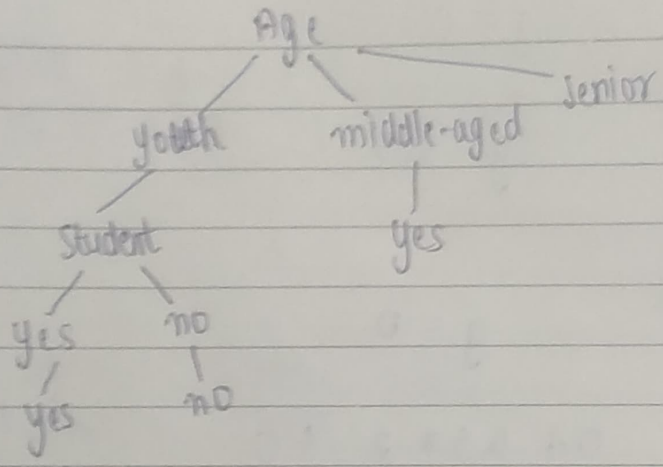
excellent → 1 yes
→ 1 no

$g = 0.5$

Weighted avg. $g = 0.444 * \frac{3}{5} + 0.5 * \frac{2}{5} = 0.466$

g(income) = 0.2
g(student) = 0
g(...) = 0.466

∴ g of student is smallest
∴ next node is student node



For senior

i) income

low → 1 yes, 1 no
medium → 2 yes, 1 no

$g = 0.5$

$g = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.444$

Weighted avg. g = 0.466

ii) student

yes → 2 yes, 1 no
no → 1 yes, 1 no

$g = 0.444$

$g = 0$

weighted avg. g = 0.3664

iii) credit_rating

fair → 3 yes
→ 0 no

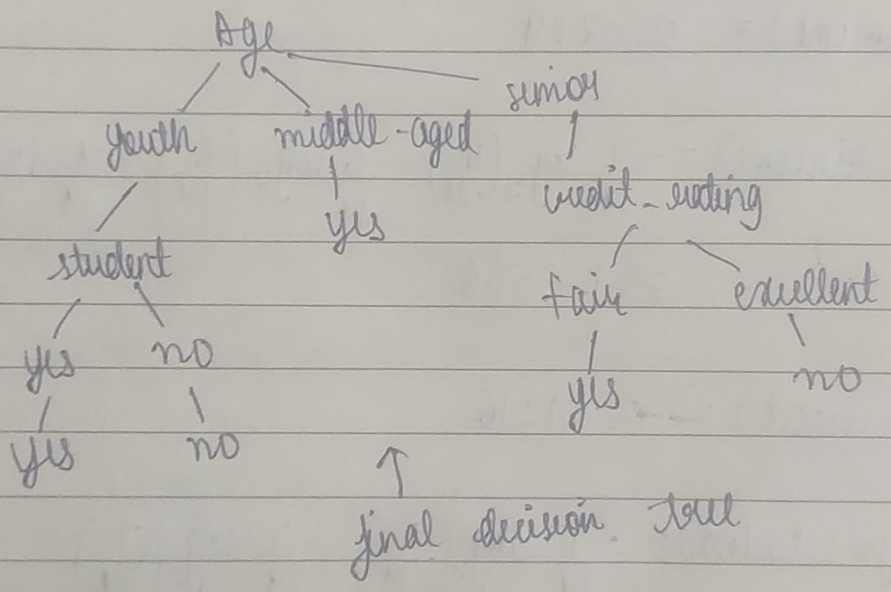
excellent → 0 yes
→ 2 no

$g = 0$

$g = 0$

Weighted avg. $g = 0$

or, next node is credit_rating



↑ final decision tree

Decision tree using gain ratio (information gain, split info)

$$GR \text{ (gain ratio)} = \frac{Gain(A)}{Split\ info\ (A)}$$

$$Split\ info\ (A) = -\sum_{j=1}^{y} \frac{|D_j|}{|D|} * \log\left(\frac{|D_j|}{|D|}\right)$$

↳ total no. of data points

i) Age

Gain(S, Age) = 0.2464 (calculated before)

Split info (Age) = $-\frac{5}{14}\log_2\left(\frac{5}{14}\right) - \frac{4}{14}\log_2\left(\frac{4}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$

= 1.5774

ii) Income

Gain(S, income) = 0.0289

Split info (income) = $-\frac{4}{14}\log_2\left(\frac{4}{14}\right) - \frac{6}{14}\log_2\left(\frac{6}{14}\right) - \frac{4}{14}\log_2\left(\frac{4}{14}\right)$

= 1.5566

iii) student

Gain (S, student) = 0.01516

Split info (student) = $-\frac{7}{14}\log_2\left(\frac{7}{14}\right) - \frac{7}{14}\log_2\left(\frac{7}{14}\right) = -\log_2\left(\frac{1}{2}\right)$

= 1

iv) credit_rating

Gain (S, credit_rating) = 0.0478

Split info (credit_rating) = $-\frac{8}{14}\log_2\left(\frac{8}{14}\right) - \frac{6}{14}\log_2\left(\frac{6}{14}\right)$

= 0.9851

Gain ratio

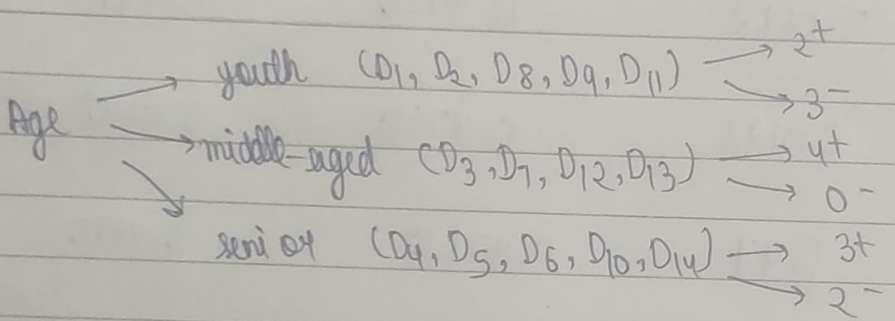
i) age = $\frac{\text{Gain}}{\text{split info}} = \frac{0.2464}{1.5774} = 0.1562$

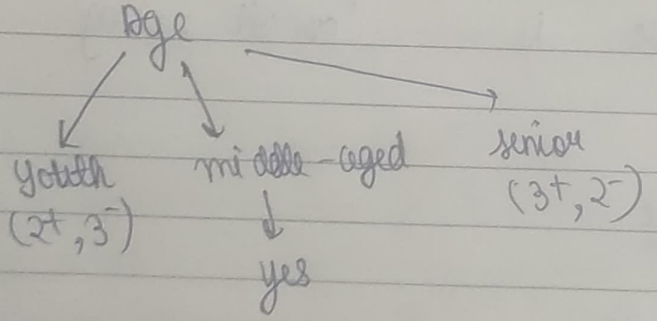
ii) income = 0.0186

iii) student = 0.01516

iv) credit. = 0.0485

∴ gain ratio of age is maximum.
∴ root node is age





Now, consider only youth and senior

For youth

i) income: Gain = 0.570 Split info = 1.5218
GR = 0.3746

ii) student: Gain = 0.97 Split info = 0.9710
GR = 0.9990

~~0.9710 * 3~~
5

iii) Credit_rating

Gain = 0.0192 Split info = 0.9710

GR = 0.0198

∴ GR of student is maximum

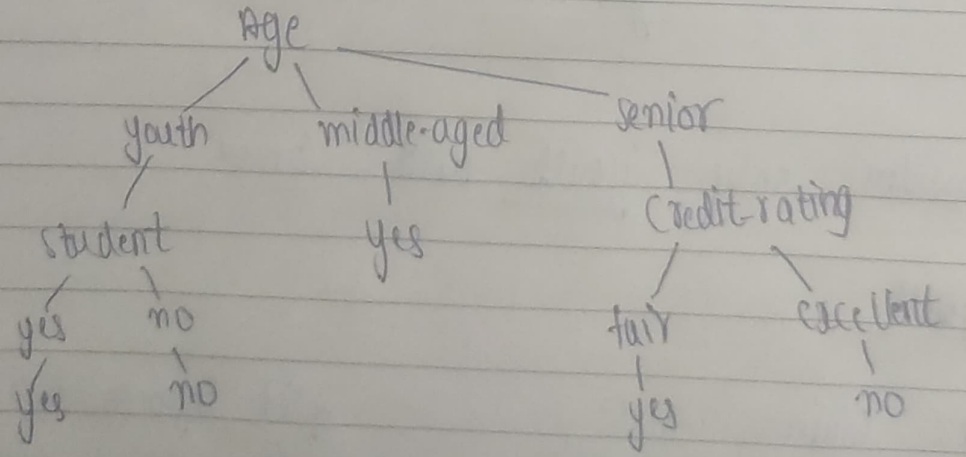
∴ Student node will be considered.

For senior

	Gain	Split info	GR
income	0.0192	0.9710	0.0198
student	0.0192	0.9710	0.0198
credit_rating	0.97	0.9710	0.9990

∴ next node is 'credit_rating'

Final decision tree



Shortage of ID3 (entropy + information gain)

→ Attributes must be nominal values

→ Dataset must not include missing data.

→ Uses a greedy approach that why it does not guarantee an optimal soln; it can get stuck in local optimums.

C4.5 algo (gain, split info, gain ratio)

→ C4.5 algorithm acts similar to ID3 but improves a few of ID3 behaviors:
- A possibility to use continuous data
- Using unknown (missing data) values

Gini index

- Has difficulties when the no. of classes is large.
- Favours tests that result in equal-sized partitions with purity.