Problem# 1(5 pts): Consider the following dataset, where we want to predict if a student will get an A in the course. Given the five attributes on the left, we want to predict if the student got an A in the course.

Early registration	Finished homework II	Senior	Likes Coffee	Liked The Last homework	A
1	1	0	0	1	1
1	1	1	0	1	1
0	0	1	0	0	0
0	1	1	0	1	0
0	1	1	0	0	1
0	0	1	1	1	1
1	0	0	0	1	0
0	1	0	1	1	1
0	0	1	0	1	1
1	0	0	0	0	0
1	1	1	0	0	1
0	1	1	1	1	0
0	0	0	0	1	0
1	0	0	1	0	1

[1] Create 2 decision trees for this dataset. For the first, only go to depth 1. For the second go to depth 2. For all trees, use the ID3 entropy algorithm from class. For each node of the tree, show the decision, the number of positive and negative examples and show the entropy at that node.

[2] Recommend another type of trees than ID3 that would build a less deep tree than ID3, Assume that you are building a complete ID3 tree. Justify your choice.

problem #1:

[1]

* $E(S)_{total} = -\frac{8}{14}\log_2\frac{8}{14} - \frac{6}{14}\log_2\frac{6}{14}$

$E(S) = 0.98$

* first attribute (Early registration

→ E(Early registration = 1) =

$-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.91$

→ E(Early registration = 0) =

$-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1$

→ Information Gain =

$0.98 - [\frac{6}{14} * 0.91 + \frac{8}{14} * 1]$

$= 0.0186$

Subject: ____ / / ____ موضوع الدرس:

* Second Attribute (finished homework

→ E (finished homework = 1) =

$$\frac{-5}{7} \log_2 \frac{5}{7} - \frac{2}{7} \log_2 \frac{2}{7} = 0,86$$

→ E (finished homework = 0) =

$$\frac{-3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0,98$$

→ Information Gain =

$$0,98 - \left(\frac{7}{14} * 0,86 + \frac{7}{14} * 0,98 \right)$$

$$= 0,06$$

Subject: _____ / / _____

* Third Attribute (Senior):

→ $E(\text{Senior}=1) =$

$$\frac{-5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} = 0.95$$

→ $E(\text{Senior}=0) =$

$$\frac{-3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

→ $\text{Gain} = 0.98 - \left[\frac{8}{14} * 0.95 + \frac{6}{14} * 1\right]$

$= 0.00875$

* Fourth attribute (likes coffee):

→ $E(\text{likes coffee}=1) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.81$

→ $E(\text{likes coffee}=0) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1$

→ $\text{Gain} = 0.98 - [\frac{4}{14} * 0.81 + \frac{10}{14} * 1] = 0.034$

---

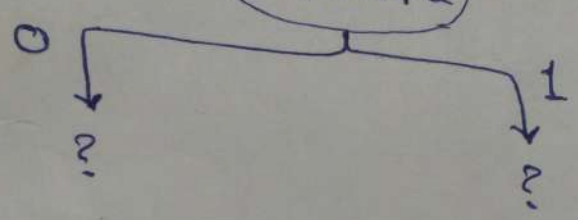
* 5th attribute (liked the last homework):

→ $E(\text{liked}=1) = -\frac{5}{9}\log_2\frac{5}{9} - \frac{4}{9}\log_2\frac{4}{9} = 0.99$

→ $E(\text{liked}=0) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

→ $\text{Gain} = 0.98 - [\frac{9}{14} * 0.99 + \frac{5}{14} * 0.97] = 0.01$

∴ root → (Finished homework)



* Total Entropy of (finished-homework = 1) = 0.86

* 1st attribute (Early registration) =▷

→ $E(\text{Early registration} = 1) = -\frac{3}{3}\log_2\frac{3}{3} - 0 = 0$

→ $E(\text{Early registration} = 0) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

→ $\text{Gain} = 0.86 - [\frac{3}{7} * 0 + \frac{4}{7} * 1] = 0.289$

Subject: / / موضوع الدرس:

* Second attribute (Senior):

→ $E(\text{Senior} = 1) = \frac{-3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

// $= 0.97$

→ $E(\text{Senior} = 0) = \frac{-2}{2}\log_2\frac{2}{2} - 0 = 0$

→ $\text{Gain} = 0.86 - [\frac{5}{7} \times 0.97 + 0]$

$= 0.167$

---

* 3rd attribute (likes coffee):

→ $E(\text{likes coffee} = 1) =$

$-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

→ $E(\text{likes coffee} = 0) =$

$-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.72$

→ $\text{Gain} = 0.86 - [\frac{2}{7} \times 1 + \frac{5}{7} \times 0.72]$

$= 0.06$

Subject: / /

موضوع الدرس :

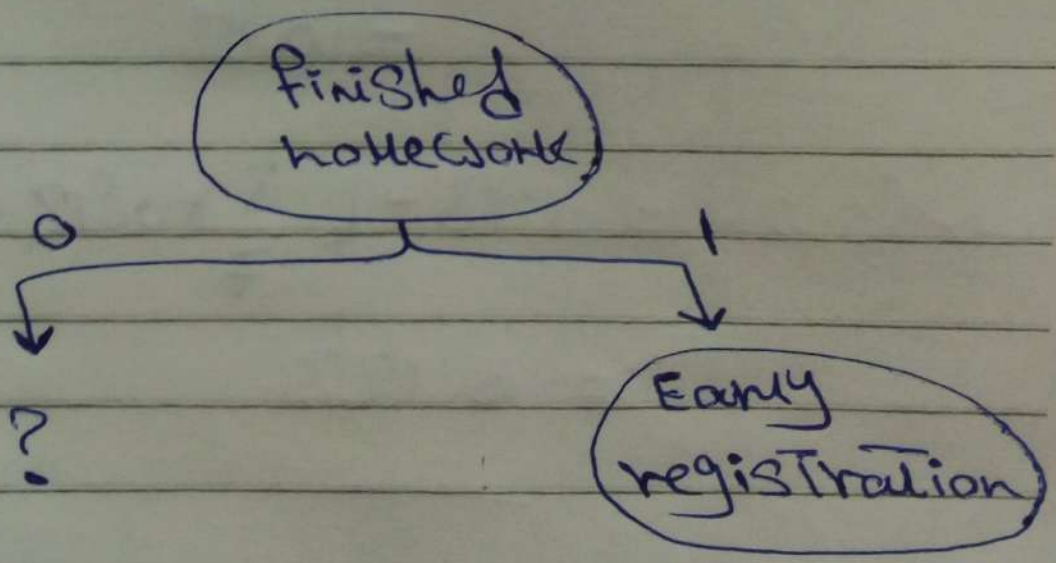
*4th attribute (liked last homework):

→ $E(\text{liked} = 1) = \frac{-3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.97$

→ $E(4^{th}\ \text{attribute} = 0) = \frac{-2}{2}\log_2\frac{2}{2} - 0 = 0$

→ $\text{Gain} = 0.86 - \left(\frac{5}{7} \times 0.97 + 0\right) = 0.167$



Subject: _____ / / _____ موضوع الدرس:

* Total Entropy of (Finished Homeworks) equal $\frac{9}{14}$ = 0.98

* 1st attribute (Senior):

→ $E(\text{Senior} = 1) = \frac{-3}{3}\log_2\frac{3}{3} - \frac{1}{3}\log_2\frac{1}{3}$

" = 0.92

→ $E(\text{Senior} = 0) = \frac{-3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$

" = 0.81

→ $\text{Gain} = 0.98 - \left[\frac{3}{7} * 0.92 + \frac{4}{7} * 0.81\right]$

= 0.123

Subject:

* 2nd attribute (likes coffee):

→ E(likes coffee = 0) = $-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}$

= 0.72

→ E(likes coffee = 1) = $-\frac{2}{2}\log_2\frac{2}{2} - 0 = 0$

→ Gain = $0.98 - \left(\frac{5}{7} \times 0.72 + 0\right) = 0.47$

* 3rd (liked last homework):

→ E(liked — = 0) = $-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

= 0.92

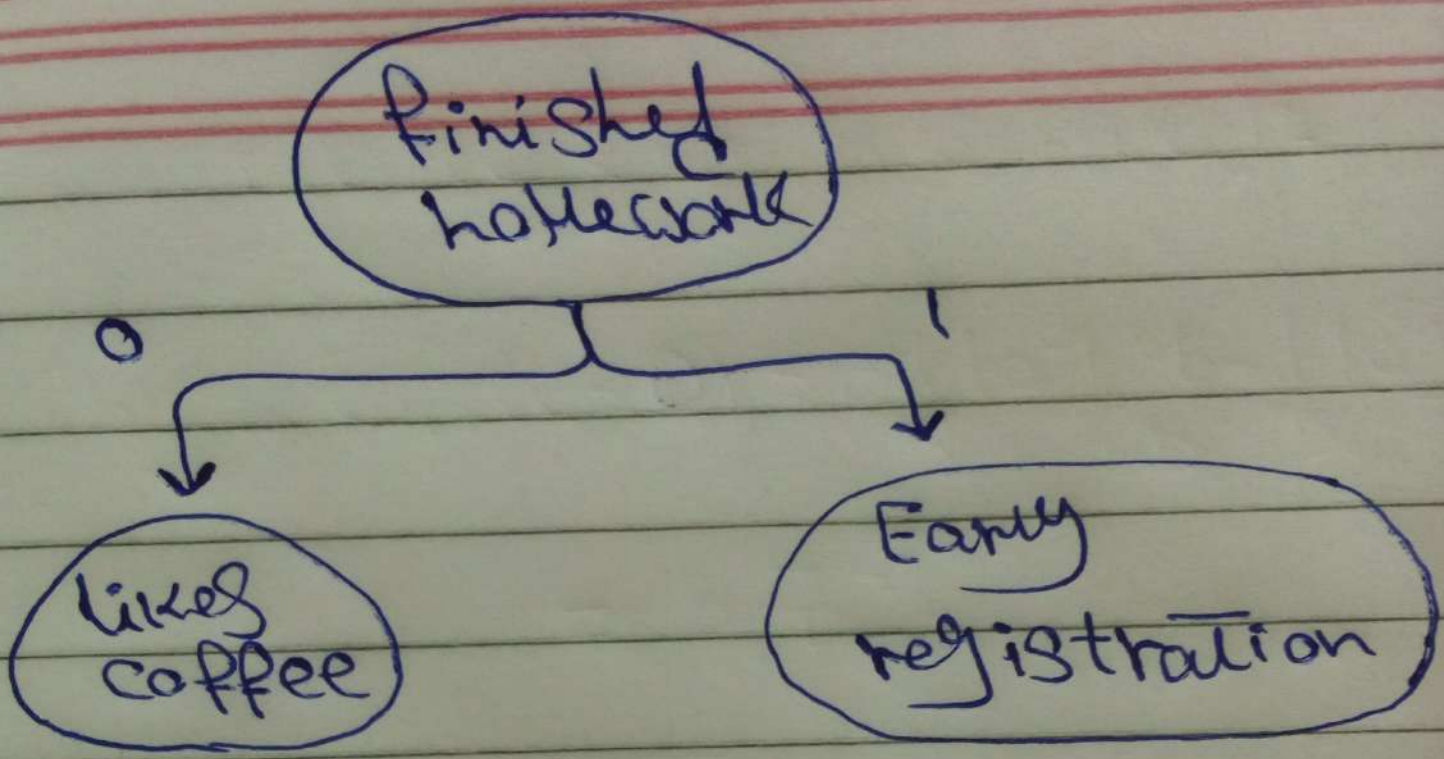
→ E(liked — = 1) = $-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$

= 1

→ Gain = $0.98 - \left[\frac{3}{7} \times 0.92 + \frac{4}{7} \times 1\right]$

= 0.0143

Subject: ____/__/____ موضوع الدرس:



Problem #1:

[2] Random forests: because they consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees and produce less depth and this prevents the probability of overfitting problem also.