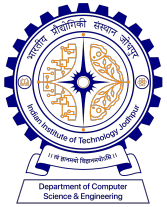
शिक्षा मंत्रालय
MINISTRY OF EDUCATION
सत्यमेव जयते

INDIAN INSTITUTE OF TECHNOLOGY
JODHPUR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NPTEL



PMRF
Prime Minister's Research Fellowship

Week 2 - Live Session

Data Mining

Swapnil S. Mane
mane.1@iitj.ac.in
PMRF Research Scholar

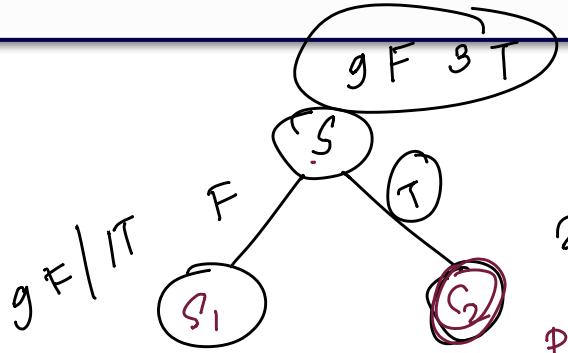
Week 2

Q1. A decision tree can be used to build models for:

a) Regression problems
b) Classification problems
c) Both of the above
d) None of the above

Q2. Entropy value of ____ represents that the data sample is pure or homogenous:

a) 1
b) 0
c) 0.5
d) None of the above.



pure Node

$$E(S_2) = -P_T \log_2 P_T - P_F \log_2 P_F$$

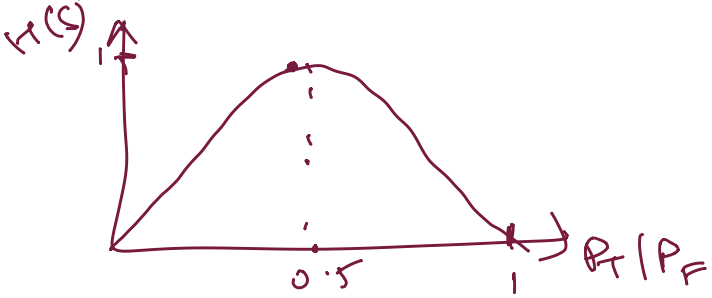
$$= -\frac{2}{2}\log_2 2/2 - \frac{0}{2}\cdots$$

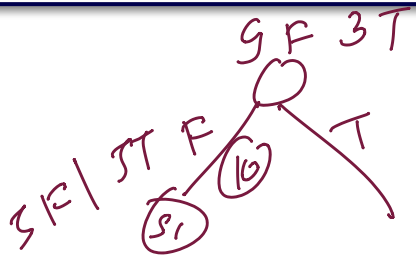
$$= 1 \quad 0 \quad - 0$$

$$= 0$$

Q3. Entropy value of _____ represents that the data sample has a 50-50 split belonging to two categories:

a) 1
b) 0
c) 0.5
d) None of the above

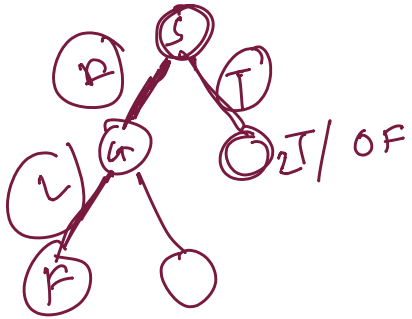


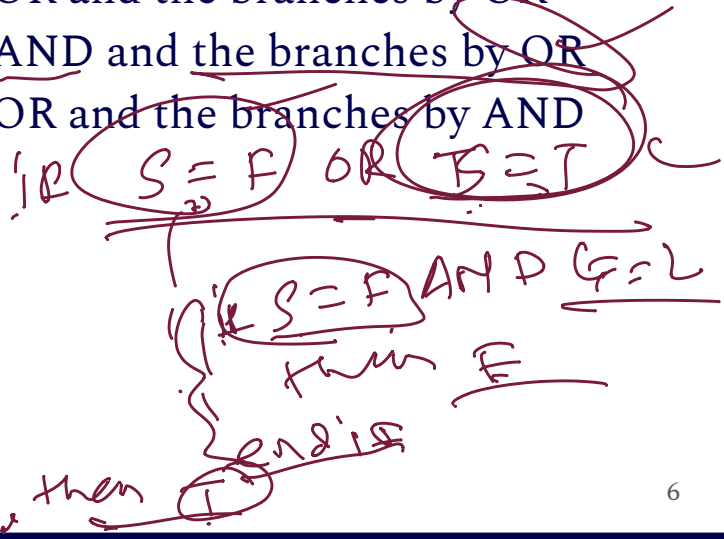


$$E(S_1) = -P_T \log_2 P_T - P_F \log_2 P_F$$
$$= -\frac{5}{10} \log \frac{5}{10} - \frac{5}{10} \log_2 \frac{5}{10}$$
$$= -0.5 * -1 - 0.5 * -1$$
$$= 1$$

Q4. If a decision tree is expressed as a set of logical rules, then:

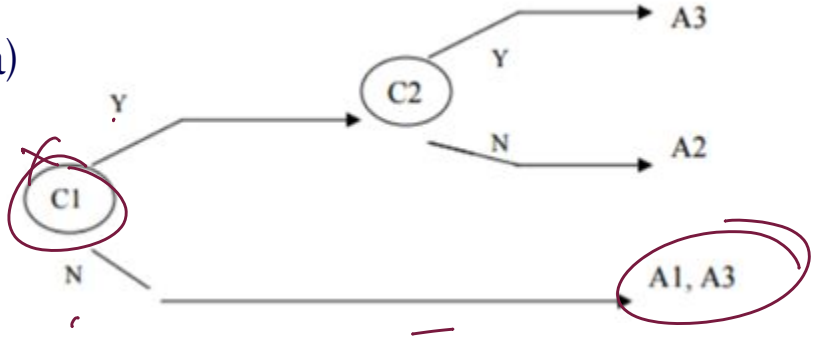
a) The internal nodes in a branch are connected by AND and the branches by AND
b) The internal nodes in a branch are connected by OR and the branches by OR
c) The internal nodes in a branch are connected by AND and the branches by OR
d) The internal nodes in a branch are connected by OR and the branches by AND



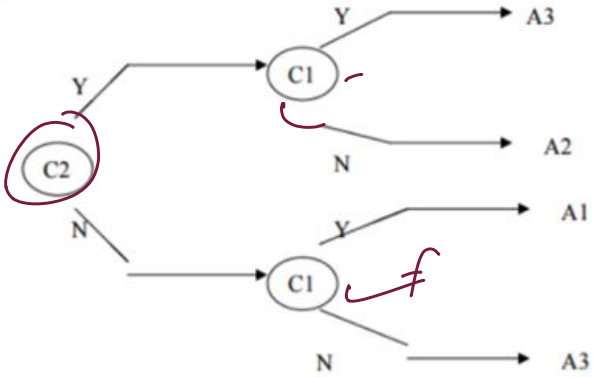


Q5. The Decision tree corresponding to the following is? (1 Mark) if C2 then if C1 then A3 else A2 endif else A1, A3 endif

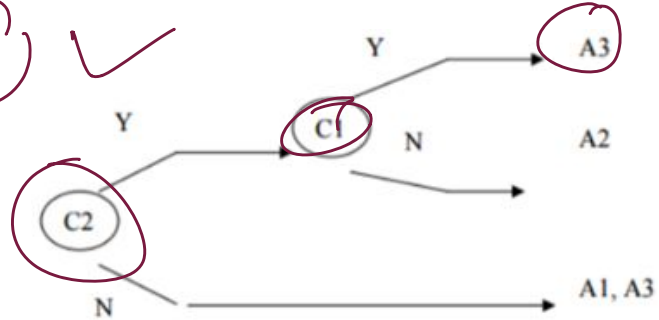
a)



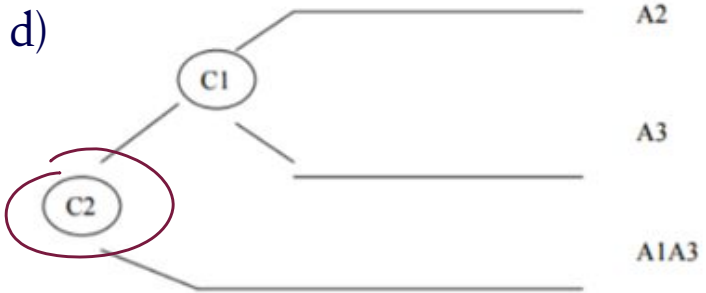
b)



c)

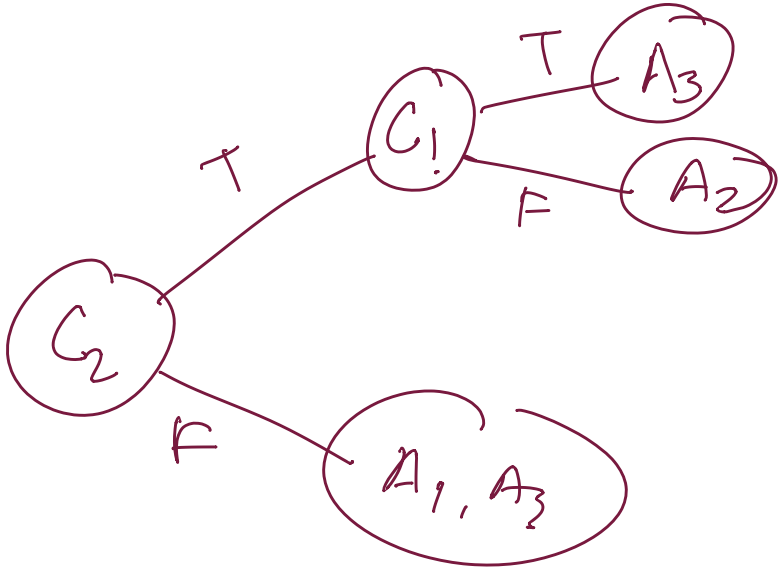


d)



Q5.

```
if C2 then
    if C1 then
        A3
    else A2
    endif
else A1, A3
endif
```



Q6. What is the entropy of the dataset?

a) 0.50
b) 0.92
c) 1
d) 0

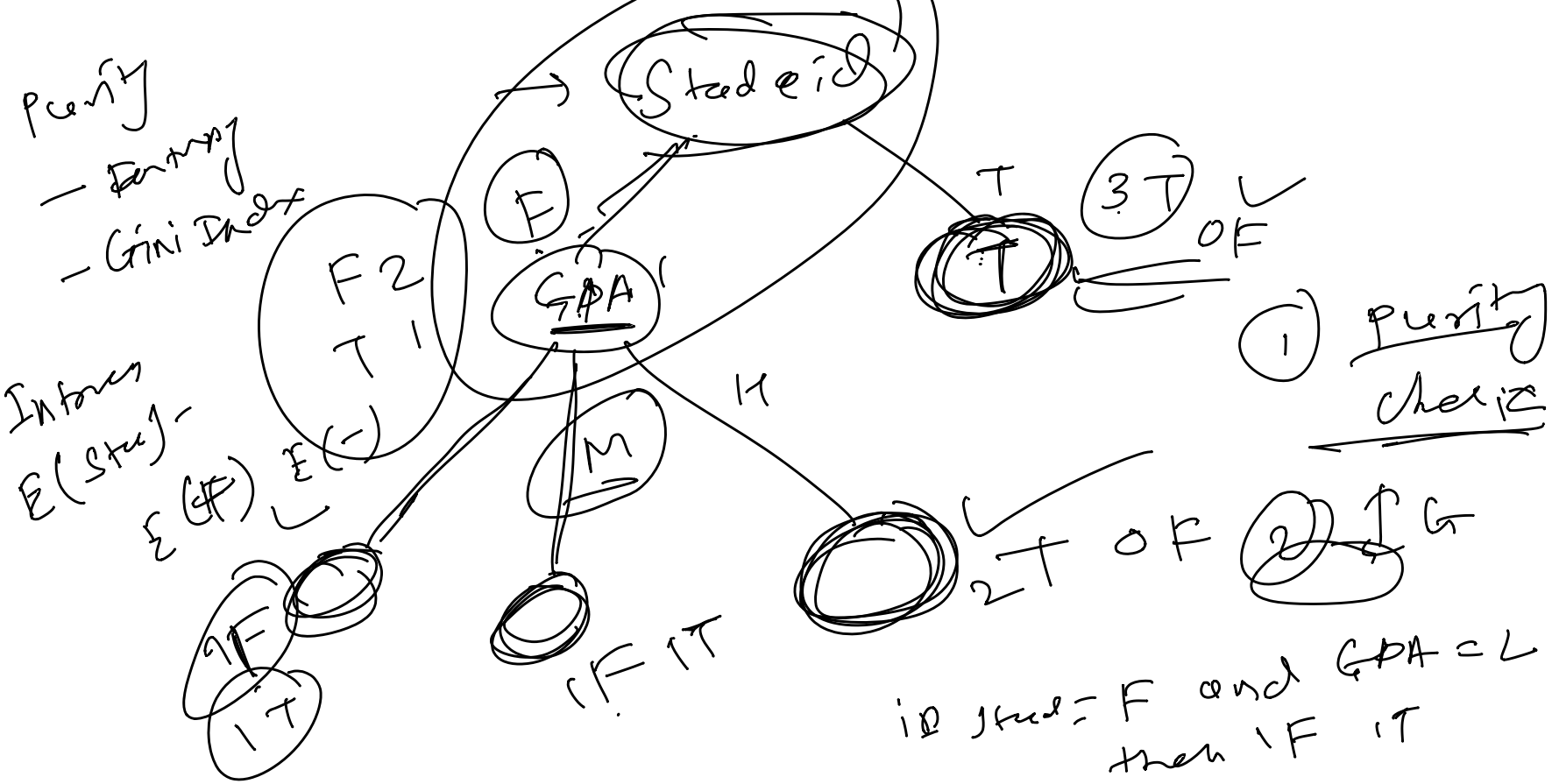
$T = 4$
$F = 2$

$E(S) = -P_T \log_2 P_T - P_F \log_2 P_F$

$E(\text{Passed}) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}$

$= 0.9182 \approx 0.92$

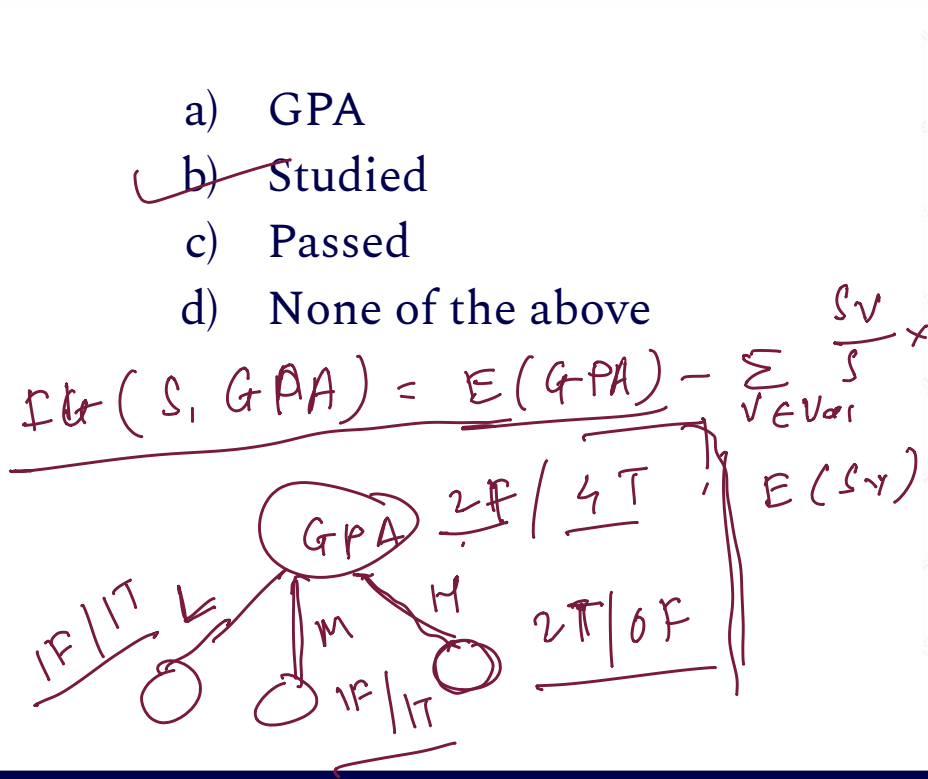
GPA	Studied	Passed
Low	F	F
Low	T	T
Medium	F	F
Medium	T	T
High	F	T
High	T	T

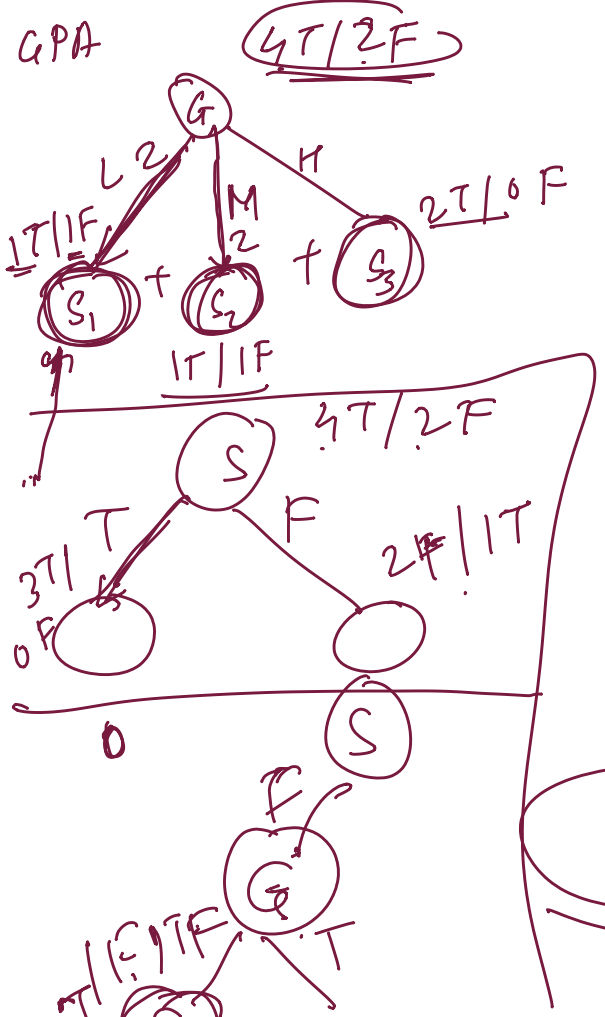

Purity
- Entropy
- Gini Index
Stud id
F
F2
T 1
GPA
M
T
3T
OF
Interview
E(Stu) -
E(GPA)
E(-)
1F
1T
1F 1T
1T
2T OF
2
G
1) Purity check
if stud = F and GPA = L
then 1F 1T
if s1 = F & G = TM
then 1F 1T

Q7. Which attribute would information gain choose as the root of the tree?

a) GPA
b) Studied
c) Passed
d) None of the above

GPA	Studied	Passed
Low	F	F
Low	T	T
Medium	F	F
Medium	T	T
High	F	T
High	T	T





$$IG(S_i, GPA) = E(GPA) - \sum_{v \in Val(S)} \frac{|S_{val}|}{|S|} * E(S_{val})$$

$$=$$

① $E(GPA) = -P_T \log_2 P_T - P_F \log_2 P_F$

$$= -\frac{4}{6}\log_2 - \frac{2}{6}\log_2 \frac{2}{6} = 0.92$$

② $E(S_1) = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log \frac{1}{2} = 1$

③ $E(S_2) = 1$

④ $E(S_3) = 0$

$$IG(GPA) = 0.92 - \left[\frac{2}{6} * 1 + \frac{2}{6} * 1 + 0\right]$$

$$= 0.92 - \left[\frac{1}{3} + \frac{1}{3}\right]$$

$$= -\frac{1}{3}\log_2 \frac{1}{3} - 2\left(\frac{1}{3}\log \frac{2}{3}\right)$$

$$= 0.91$$

$$IG(GPA) = 0.253$$

$$IG(S) = 0.92 - [(3/6 * 0) + 3/6 * 0.811)]$$

$$= 0.5145$$

$$IG(S, GA) = 0.2583$$

$$IG(S, S) = 0.545$$