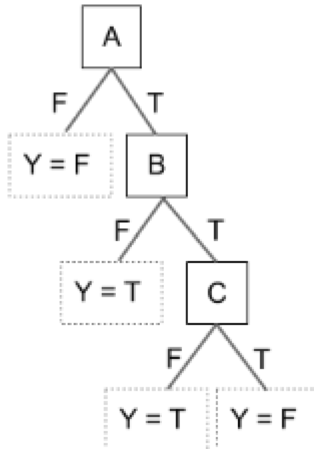
Question 5

a) Theoretical/Reasoning

b)



c)

Example	Color	Height	Width	Class
A	Red	Short	Thin	NO
B	Blue	Tall	Fat	YES
C	Green	Short	Fat	NO
D	Green	Tall	Thin	YES
E	Blue	Short	Thin	NO

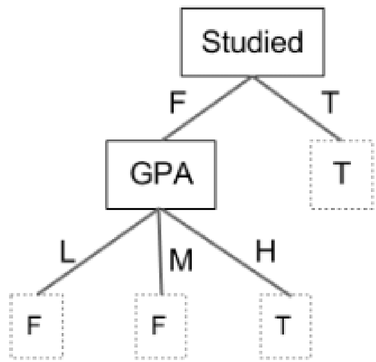
d)

$$H(Passed) = -(\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6})$$
$$H(Passed) = -(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3})$$
$$H(Passed) = \boxed{\log_2 3 - \frac{2}{3} \approx 0.92}$$

$$H(Passed|GPA) = -\frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(1\log_2 1)$$
$$H(Passed|GPA) = \frac{1}{3}(1) + \frac{1}{3}(1) + \frac{1}{3}(0)$$
$$H(Passed|GPA) = \boxed{\frac{2}{3} \approx 0.66}$$

$$H(Passed|Studied) = -\frac{1}{2}(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}) - \frac{1}{2}(1\log_2 1)$$
$$H(Passed|Studied) = \frac{1}{2}(\log_2 3 - \frac{2}{3}$$
$$H(Passed|Studied) = \boxed{\frac{1}{2}\log_2 3 - \frac{1}{3} \approx 0.46}$$

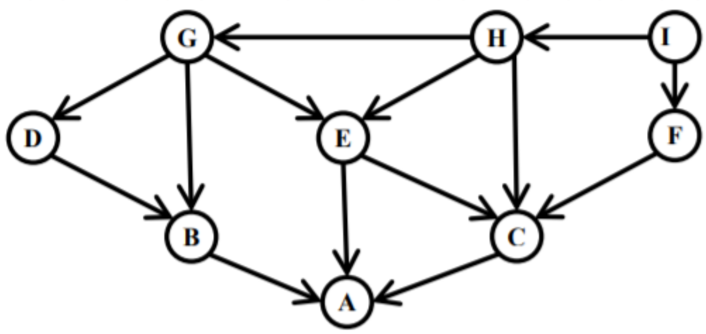
★ **ANSWER:** We want to split first on the variable which maximizes the information gain $H(Passed) - H(Passed|A)$. This is equivalent to minimizing $H(Passed|A)$, so we should split on "Studied?" first.



Question 6

a)

P(A | B,C,E) P(B | D,G) P(C | E,F,H) P(D | G) P(E| G,H) P(F | I) P(G | H) P(H | I) P(I)



b) Using Bayes' theorem, we know that:

P (Location|Observation) = P (Observation|Location)*P (Location) \ P(Observation).

c)

1. P (C)=
P (C|P,S)*P(P)*P(S)+P(C|~P,S)*P(~P)*P(S)+P(C|P,~S)*P(P)*P(~S)+P(C|~P,~S)*P(~P)*P(~S)
= 0.010
2. P(P|C) = P(C|P)*P(P)\P(C)

3. P(X|C) = P(X|C)*P(C)+P(X|~C)*P(~C) = 0.20

4. P(D|C) = P(D|C)*P(C)+P(D|~C)*P(~C) = 0.296