Team Members:

Shanmukha Datta Sai Sasank Tadepalli - NetID: na2500
Bhavana Kamasani - NetID: tl1593
Neeraj Kumar Maraka - NetID: ho4913

1. Continuing with our earlier fixation on our roommate's activities, construct a decision tree to predict her activities based upon the weather, how much money she currently has available, and whether her parents are visiting.

Weekend (Example)	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Sunny	Yes	Rich	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Answer:

1)

$\rightarrow$ $\text{Entropy}(S) = \sum -p_i \log p_i$

$$= -\frac{6}{10}\log\frac{6}{10} - \frac{2}{10}\log\frac{2}{10} - \frac{1}{10}\log\frac{1}{10} - \frac{1}{10}\log\frac{1}{10}$$

$$= 1.570$$

$\rightarrow$ $\text{Gain}(S, \text{Weather}) = \text{Entropy}(S) - \sum \frac{|S_v|}{|S|}\text{Entropy}(S_v)$

$$\text{Entropy}(S_{sunny}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$$

$$\text{Entropy}(S_{Windy}) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.811$$

$$\text{Entropy}(S_{Rainy}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

$$\text{Gain}(S, \text{Weather}) = 1.570 - \left[\frac{4}{10} + \frac{4}{10}\times 0.811 + \frac{2}{10}\right]$$

$$= 0.646$$

$\rightarrow$ $\text{Gain}(S, \text{Parents})$

$$\text{Entropy}(S_{Yes}) = -\frac{5}{5}\log\frac{5}{5} = 0$$

$$\text{Entropy}(S_{No}) = -\frac{2}{5}\log\frac{2}{5} - \frac{1}{5}\log\frac{1}{5} - \frac{1}{5}\log\frac{1}{5} - \frac{1}{5}\log\frac{1}{5}$$

$$= 1.921$$

$$\text{Gain}(S, \text{Parents}) = 1.57 - \left[\frac{5}{10}\times 0 + 1.921\times\frac{5}{10}\right] = 0.609$$

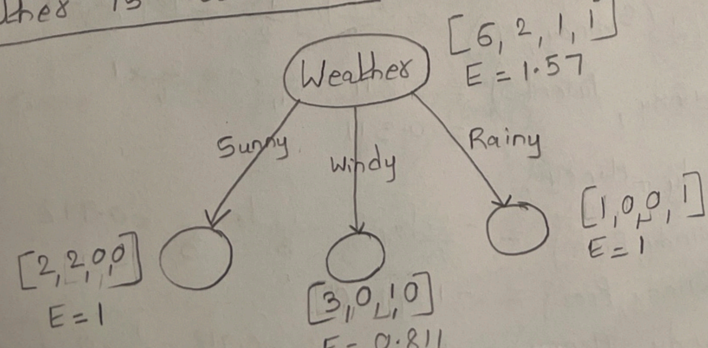
$\rightarrow$ $\text{Gain}(S, \text{Money})$

$$\text{Entropy}(S_{Rich}) = -\frac{4}{8}\log\frac{4}{8} - \frac{2}{8}\log\frac{2}{8} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = 1.75$$

$$\text{Entropy}(S_{poor}) = -\frac{2}{2}\log\frac{2}{2} = 0$$

$$\text{Gain}(S, \text{Money}) = 0.17$$

$\boxed{\text{Weather is root node}}$



$\rightarrow$ Entropy $(S_{sunny}) = 1$

$\rightarrow$ Gain$(S_{Sunny}, \text{Parents})$

$$\text{Entropy}(S_{Sunny, yes}) = -\frac{2}{2}\log\frac{2}{2} = 0$$

$$\text{Entropy}(S_{Sunny, No}) = -\frac{2}{2}\log\frac{2}{2} = 0$$

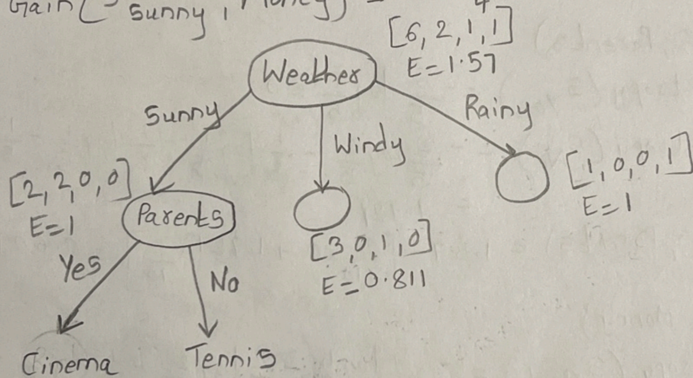
$$\text{Gain}(S_{Sunny}, \text{Parents}) = 1$$

$\rightarrow$ Gain$(S_{Sunny}, \text{Money})$

$$\text{Entropy}(S_{Sunny, Rich}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$$

$$\text{Entropy}(S_{Sunny, Poor}) = 0$$

$$\text{Gain}(S_{sunny}, \text{Money}) = 1 - \frac{4}{4}\times 1 - 0 = 0$$



$\rightarrow$ Entropy $(S_{Windy}) = 0.811$

$\rightarrow$ Gain$(S_{Windy}, \text{Parents})$

$$\text{Entropy}(S_{Windy}, yes) = -\frac{2}{2}\log\frac{2}{2} = 0$$

$$\text{Entropy}(S_{Windy}, No) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

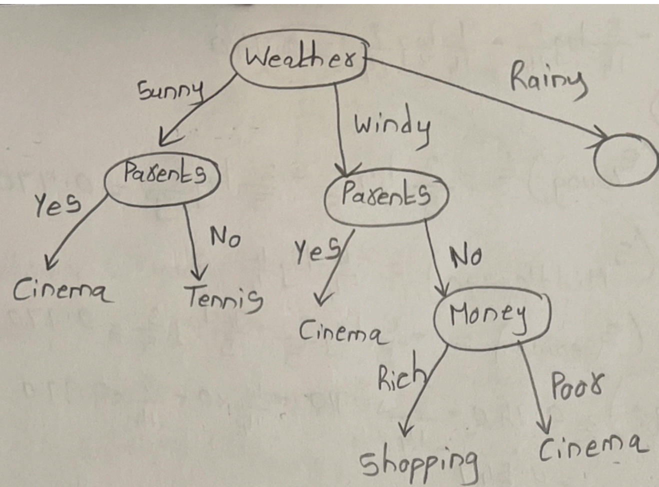
$$\text{Gain}(S_{Windy}, \text{Parents}) = 0.811 - \frac{2}{4}\times 0 - \frac{2}{4}\times 1$$
$$= 0.311$$

$\rightarrow$ Gain$(S_{Windy}, \text{Money})$

$$E(S_{Windy}, Rich) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$$

$$E(S_{Windy}, Poor) = -1\log 1 = 0$$

$$\text{Gain}(S_{Windy}, \text{Money}) = 0.811 - \frac{3}{4}\times 0.918 - \frac{1}{4}\times 0 = 0.122$$



$\rightarrow$ Entropy $(S_{Rainy}) = 1$

$\rightarrow$ Gain $(S_{Rainy}, \text{Parents})$

Entropy $(S_{Rainy, Yes}) = 0$

Entropy $(S_{Rainy, No}) = 0$

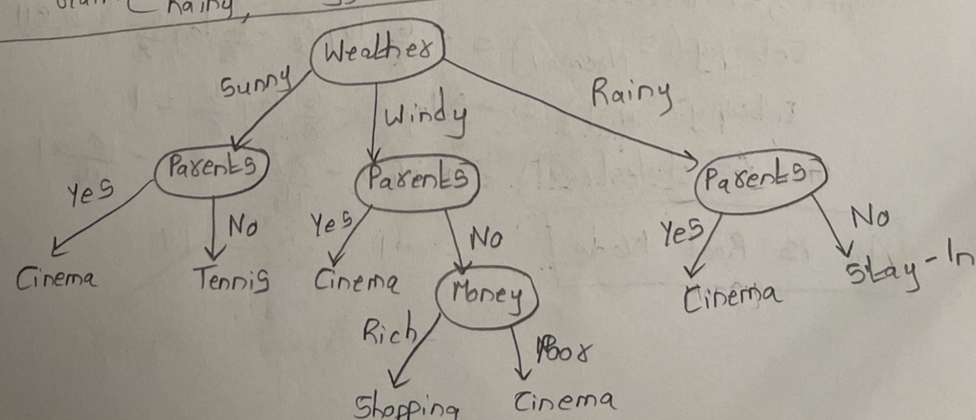
Gain $(S_{Rainy}, \text{Parents}) = 1$

$\rightarrow$ Gain $(S_{Rainy}, \text{Money})$

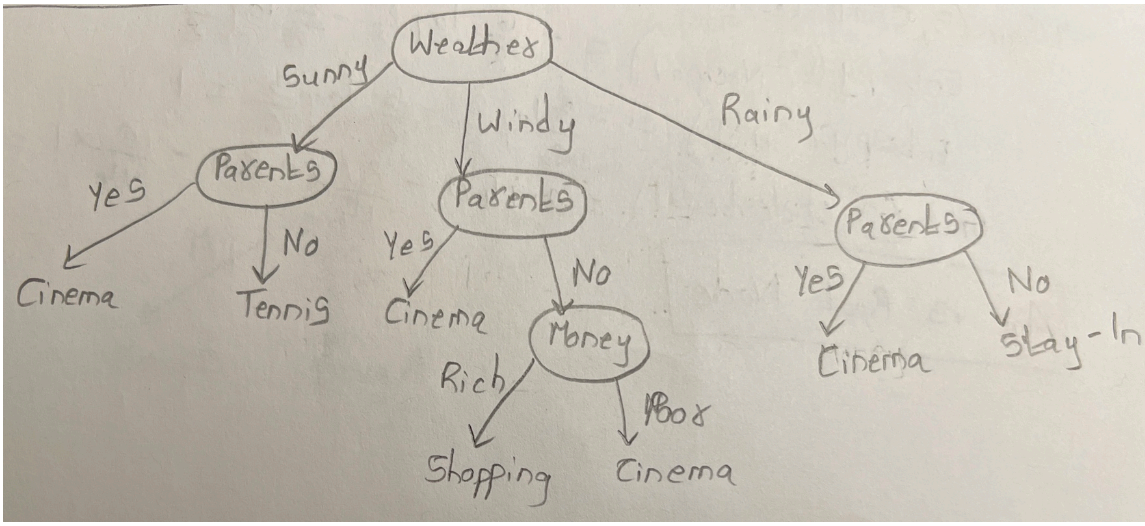
Entropy $(S_{Rainy, Rich}) = 0$

Entropy $(S_{Rainy, Poor}) = 0$

Gain $(S_{Rainy}, \text{Money}) = 1$



The decision tree for the given data is



2. Given the 14 patient examples given below, construct a ID3 decision tree to determine whether patient 15 should be given either drug A or drug B.

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

Answer:

2) $\rightarrow$ Entropy (S) $= -\frac{5}{14}\log\frac{5}{14} - \frac{9}{14}\log\frac{9}{14} = 0.940$

$\rightarrow$ Gain (S, Age)

Entropy $(S_{Young}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.970$

Entropy $(S_{Middle\text{-}age}) = -\frac{4}{4}\log\frac{4}{4} = 0$

Entropy $(S_{Senior}) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.970$

Gain $(S, Age) = 0.940 - \frac{5}{14}\times 0.970 - \frac{4}{14}\times 0 - \frac{5}{14}\times 0.970$

$= 0.247$

$\rightarrow$ Gain (S, Sex)

Entropy $(S_M) = -\frac{1}{7}\log\frac{1}{7} - \frac{6}{7}\log\frac{6}{7} = 0.591$

Entropy $(S_F) = -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0.985$

Gain $(S, Sex) = 0.940 - \frac{7}{14}\times 0.591 - \frac{7}{14}\times 0.985 = 0.152$

$\rightarrow$ Gain (S, BP)

Entropy $(S_{High}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$

Entropy $(S_{Normal}) = -\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6} = 0.918$

Entropy $(S_{Low}) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.811$

Gain $(S, BP) = 0.940 - \frac{4}{14}\times 1 - \frac{6}{14}\times 0.918 - \frac{4}{14}\times 0.811 = 0.029$

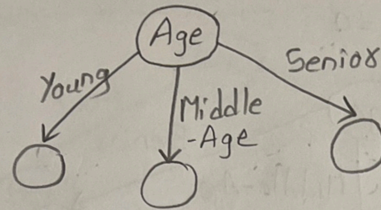
$\rightarrow$ Gain (S, Cholesterol)

Entropy $(S_{Normal}) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8} = 0.811$

Entropy $(S_{High}) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$

Gain $(S, \text{cholesterol}) = 0.940 - \frac{8}{14}\times 0.811 - \frac{6}{14}\times 1 = 0.048$

$\boxed{\text{Age is Root Node}}$



→ Entropy (S_{young}) = 0.970

→ Gain(S_{Young}, Sex)

$$\text{Entropy}(S_{Young}, M) = -\frac{2}{2}\log\frac{2}{2} = 0$$

$$\text{Entropy}(S_{Young}, F) = -\frac{3}{3}\log\frac{3}{3} = 0$$

$$\text{Gain}(S_{Young}, \text{Sex}) = 0.970$$

→ Gain(S_{Young}, BP)

$$\text{Entropy}(S_{Young}, \text{High}) = -\frac{2}{2}\log\frac{2}{2} = 0$$

$$\text{Entropy}(S_{Young}, \text{Normal}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

$$\text{Entropy}(S_{Young}, \text{Low}) = -\frac{1}{1}\log\frac{1}{1} = 0$$

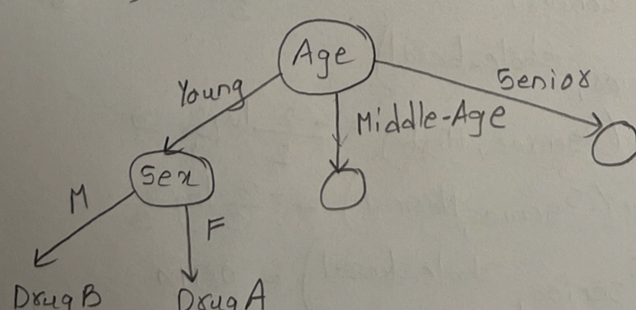
$$\text{Gain}(S_{Young}, BP) = 0.970 - \frac{2}{5}\times 0 - \frac{2}{5}\times 1 - \frac{1}{5}\times 0 = 0.57$$

→ Gain(S_{Young}, cholesterol)

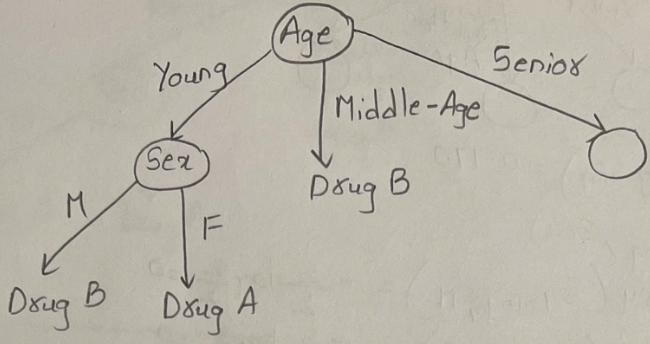
$$\text{Entropy}(S_{Young}, \text{Normal}) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$$

$$\text{Entropy}(S_{Young}, \text{High}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{3} = 1$$

$$\text{Gain}(S_{Young}, \text{cholesterol}) = 0.970 - \frac{3}{5}\times 0.918 - \frac{2}{5}\times 1$$
$$= 0.0192$$



→ $\text{Entropy}(S_{\text{Middle-Age}}) = 0$



→ $\text{Entropy}(S_{\text{senior}}) = 0.970$

→ $\text{Gain}(S_{\text{senior}}, \text{sex})$

$$\text{Entropy}(S_{\text{senior}, M}) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$$

$$\text{Entropy}(S_{\text{senior}, F}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

$$\text{Gain}(S_{\text{senior}}, \text{sex}) = 0.970 - \frac{3}{5}\times 0.918 - \frac{2}{5}\times 1 = 0.0192$$

→ $\text{Gain}(S_{\text{senior}}, BP)$

$$\text{Entropy}(S_{\text{senior}, \text{High}}) = 0$$

$$\text{Entropy}(S_{\text{senior}, \text{Normal}}) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$$

$$\text{Entropy}(S_{\text{senior}, \text{Low}}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

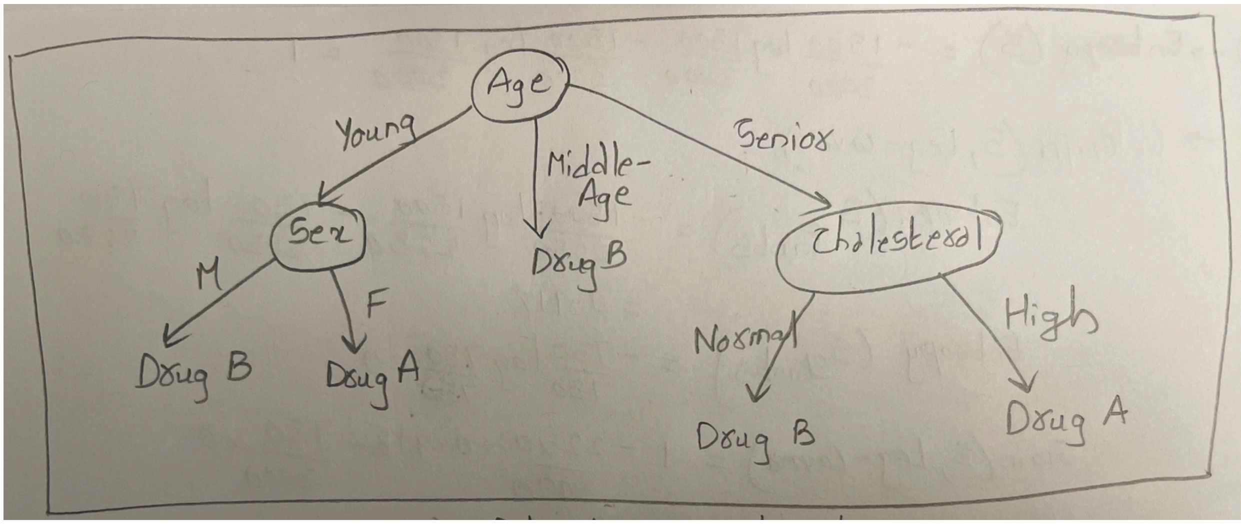
$$\text{Gain}(S_{\text{senior}}, BP) = 0.970 - \frac{3}{5}\times 0.918 - \frac{2}{5}\times 1 = 0.0192$$

→ $\text{Gain}(S_{\text{senior}}, \text{cholesterol})$

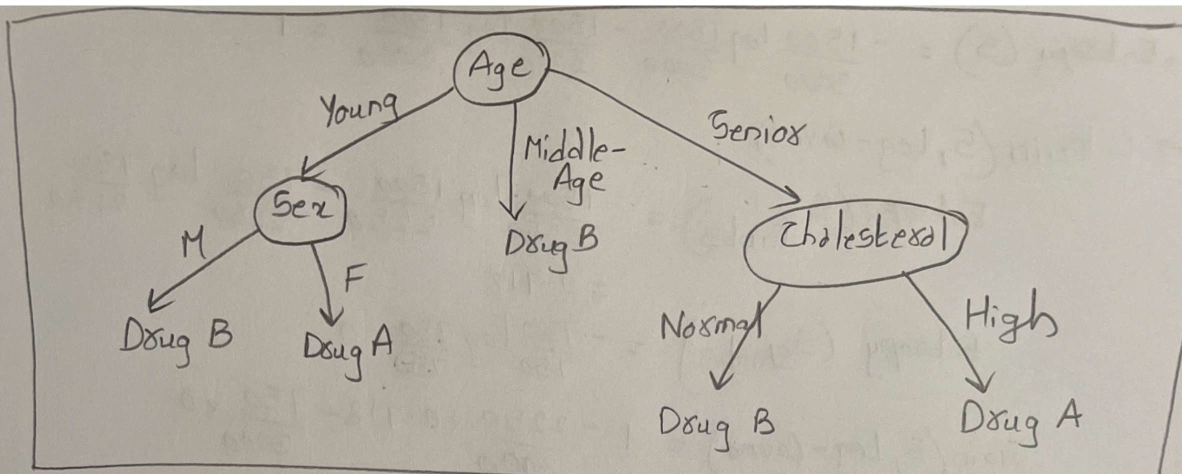
$$\text{Entropy}(S_{\text{senior}, \text{High}}) = -\frac{2}{2}\log\frac{2}{2} = 0$$

$$\text{Entropy}(S_{\text{senior}, \text{Normal}}) = -\frac{3}{3}\log\frac{3}{3} = 0$$

$$\text{Gain}(S_{\text{senior}}, \text{cholesterol}) = 0.970$$



The decision tree for the given data is



We need to classify Patient 15 who has

Age → Middle-Age
Sex → F
BP → Low
Cholesterol → Normal

From the Decision Tree above we classify Patient F under Drug-B

Therefore, we give Drug-B to Patient 15

We need to classify Patient 15 who has

Age == Middle-Age

Sex == F

BP == Low

Cholesterol == Normal

From the decision tree above since age is middle-age we go to that subtree and we reach leaf node with Drug B as the target classification.
Therefore, we give Drug-B to Patient 15.

3. We would like to predict the sex of a person based on two binary attributes: leg- cover (pants or skirts) and facial-hair (some or none). We have a data set of 3,000 individuals, half male and half female. 70% of the males have no facial hair. Skirts are worn by 50% of the females. All females are barefaced and no male wears a skirt. Using ID3, construct a decision tree for this classification problem.

Answer:

From the given data,

Totat Population = 3000

Number of Male = 1500

Number of Female = 1500

Number of male with no facial hair = 70 % of 1500
= (70 / 100) x 1500
= 1050

Number of female who wear skirts = 50 % of 1500
= (50 / 100) x 1500
= 750

Number of female who have facial hair = 0

Number of male who wear skirt = 0

So,

Number of male with facial hair = 1500 - 1050
= 450

Number of female who wear pants = 1500 - 750
= 750

3) $\rightarrow$ $\text{Entropy}(S) = -\frac{1500}{3000}\log\frac{1500}{3000} - \frac{1500}{3000}\log\frac{1500}{3000} = 1$

$\rightarrow$ Gain (S, Leg-Cover)

$$\text{Entropy}(S_{\text{Pants}}) = -\frac{1500}{2250}\log\frac{1500}{2250} - \frac{750}{2250}\log\frac{750}{2250}$$
$$= 0.918$$

$$\text{Entropy}(S_{\text{Skirts}}) = -\frac{750}{750}\log\frac{750}{750} = 0$$

$$\text{Gain}(S, \text{Leg-Cover}) = 1 - \frac{2250}{3000}\times 0.918 - \frac{750}{3000}\times 0$$
$$= 0.311$$

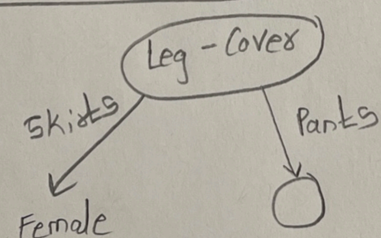
$\rightarrow$ Gain (S, Facial-Hair)

$$\text{Entropy}(S_{\text{Yes}}) = -\frac{450}{450}\log\frac{450}{450} = 0$$

$$\text{Entropy}(S_{\text{No}}) = -\frac{1500}{2550}\log\frac{1500}{2550} - \frac{1050}{2550}\log\frac{1050}{2550} = 0.977$$

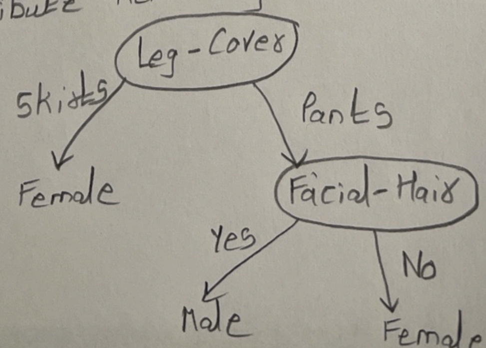
$$\text{Gain}(S, \text{Facial-Hair}) = 1 - \frac{450}{3000}\times 0 - \frac{2550}{3000}\times 0.977$$
$$= 0.169$$

$\boxed{\text{Leg-Cover is root Node}}$



$\rightarrow$ $\text{Entropy}(S_{\text{Pants}}) = 0.918$

Only One Attribute Remaining - "Facial Hair"



The decision tree for the given data is