Introduction to Artificial Intelligence

Quiz 7 – Naïve Bayes Classifier & ID3 tree

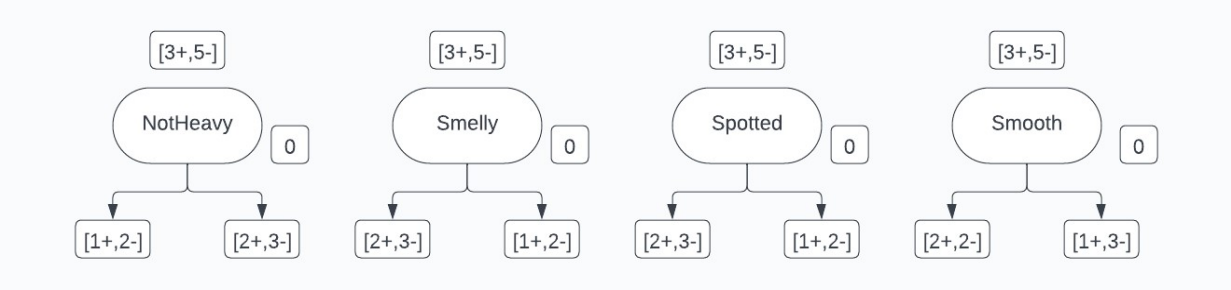
1. Problem 1:

You are stranded on a deserted island. Mushrooms of various types grow widely all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous and others as not (determined by your former companions' trial and error). You are the only one remaining on the island. You have the following data to consider:

Example	NotHeavy	Smelly	Spotted	Smooth	Edible
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1
D	0	0	0	1	0
E	1	1	1	0	0
F	1	0	1	1	0
G	1	0	0	1	0
H	0	1	0	0	0
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

You know whether or not mushrooms A through H are poisonous, but you do not know about U through W.

1. Build a ID3 decision tree to classify mushrooms as poisonous or not.



$H_{Edible} = H[3+, 5-] = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = \frac{3}{8}\log_2\frac{8}{3} + \frac{5}{8}\log_2\frac{8}{5} = \frac{3}{8} * 3 - \frac{3}{8}\log_2 3 + \frac{5}{8} * 3 + \frac{5}{8}\log_2 5 = 3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5 = 0.9544$

$H_{0/Smooth} = \frac{4}{8}H[2+, 2-] + \frac{4}{8}H[1+, 3-] = \frac{1}{2} + \frac{1}{2} * (\frac{1}{4}\log_2\frac{4}{1} + \frac{3}{4}\log_2\frac{4}{3}) = \frac{1}{2} + \frac{1}{2} * (\frac{1}{4} * 2 + \frac{3}{4} * 2 - \frac{3}{4}\log_2 3) = \frac{1}{2} + \frac{1}{2} * (2 - \frac{3}{4}\log_2 3) = \frac{1}{2} + 1 - \frac{3}{8} * \log_2 3 = \frac{3}{2} - \frac{3}{8}\log_2 3 = 0.9056$

$IG_{0/Smooth} = H_{Edible} - H_{0/Smooth} = 0.9544 - 0.9056 = 0.048$

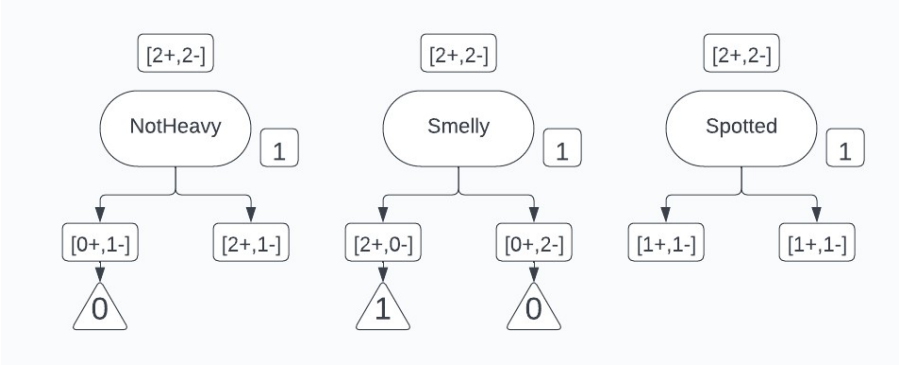
$H_{0/NotHeavy} = \frac{3}{8}H[1+,2-] + \frac{5}{8}H[2+,3-] = \frac{3}{8}(\frac{1}{3}\log_2\frac{3}{1} + \frac{2}{3}\log_2\frac{3}{2}) + \frac{5}{8}\left(\frac{2}{5}\log_2\frac{5}{2} + \frac{3}{5}\log_2\frac{5}{3}\right)$
$= \frac{3}{8}(\frac{1}{3}\log_2 3 + \frac{2}{3}\log_2 3 - \frac{2}{3}) + \frac{5}{8}\left(\frac{2}{5}\log_2 5 - \frac{2}{5} + \frac{3}{5}\log_2 5 - \frac{3}{5}\log_2 3\right) = \frac{3}{8}(\log_2 3 - \frac{2}{3}) +$
$\frac{5}{8}\left(\log_2 5 - \frac{2}{5} - \frac{3}{5}\log_2 3 + \frac{2}{5}\right) = \frac{3}{8}\log_2 3 - \frac{2}{8} + \frac{5}{8}\log_2 5 - \frac{3}{8}\log_2 3 - \frac{2}{8} = 0.9512$

=> $IG_{0/NotHeavy} = H_{Edible} - H_{0/NotHeavy} = 0.9544 - 0.9512 = 0.032 = 0.048$

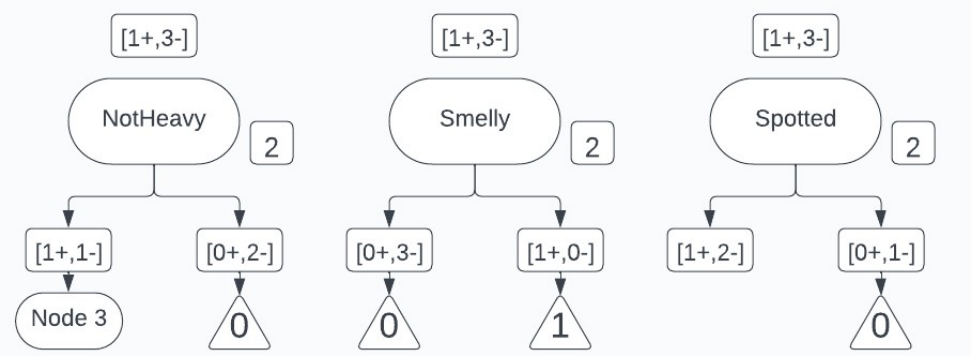
$IG_{0/NotHeavy} = IG_{0/Smelly} = IG_{0/Spotted} = 0.032 < IG_{0/Smooth} = 0.048$

=> Choose Smooth as root node

If node 1: Smooth = 0

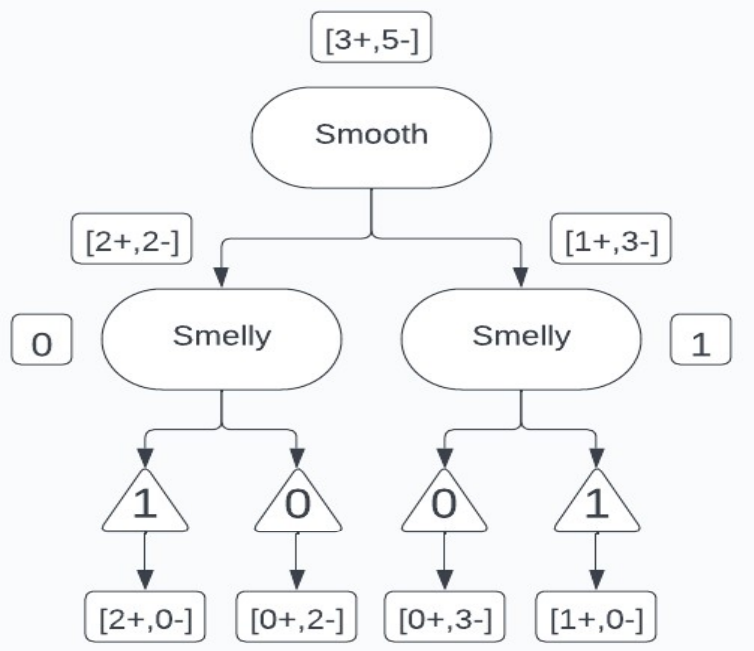


If node 2: Smooth = 1



=> Choose Smelly as second node

The resulting ID3 tree:



2. Classify mushrooms U, V and W using the decision tree as poisonous or not poisonous.
   U: Smooth = 0, Smelly =1 => Edible = 1
   V: Smooth = 1, Smelly =1 => Edible = 1
   W: Smooth = 0, Smelly =1 => Edible = 1

   U and V both classify as not poisonous
   W classifies as poisonous
3. If the mushrooms A through H that you know are not poisonous suddenly became scarce, should you consider trying U, V and W? Which one(s) and why? Or if none of them, then why not?

If the mushrooms A through H that are not poisonous suddenly became scarce, I would consider trying W based on the decision tree.

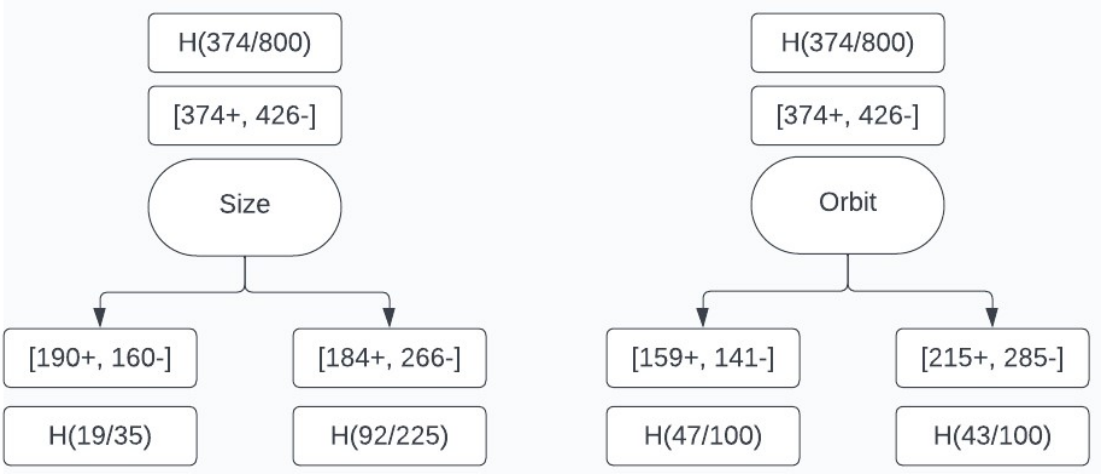
2. Problem 2:

As of September 2012, 800 extrasolar planets have been identified in our galaxy. Supersecret surveying spaceships sent to all these planets have established whether they are habitable for humans or not, but sending a spaceship to each planet is expensive. In this problem, you will come up with decision trees to predict if a planet is habitable based only on features observable using telescopes.

1. In below table you are given the data from all 800 planets surveyed so far. The features observed by telescope are Size ("Big" or "Small"), and Orbit ("Near" or "Far"). Each row indicates the values of the features and habitability, and how many times that set of values was observed. So, for example, there were 20 "Big" planets "Near" their star that were habitable

Size	Orbit	Habitable	Count
Big	Near	Yes	20
Big	Far	Yes	170
Small	Near	Yes	139
Small	Far	Yes	45
Big	Near	No	130
Big	Far	No	30
Small	Near	No	11
Small	Far	No	255

Derive and draw the decision tree learned by ID3 on this data. Make sure to clearly mark at each node what attribute you are splitting on, and which value corresponds to which branch. By each leaf node of the tree, write in the number of habitable and inhabitable planets in the training data that belong to that node



$$H_{\text{Habitable}} = H[374+, 426-] = -\frac{374}{800}\log_2\frac{374}{800} - \frac{426}{800}\log_2\frac{426}{800} = 0.9969$$

If the mushrooms A through H that are not poisonous suddenly became scarce, I would consider trying W based on the decision tree.

$H_{(Habitual/Size)} = \frac{5}{8} * H\left[\frac{19}{35}\right] + \frac{45}{80} * H\left[\frac{92}{225}\right] = \frac{35}{80} * 0.9946 + \frac{45}{80} * 0.9759 = 0.9841$

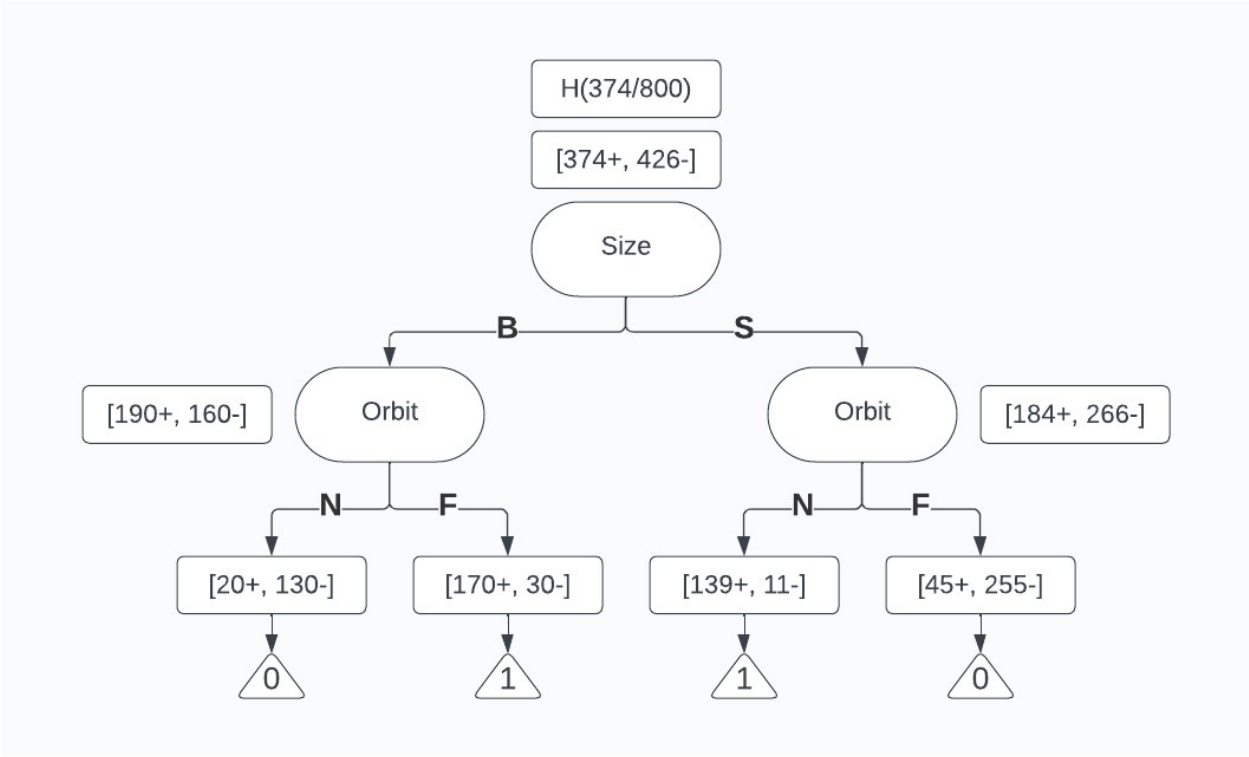
$H_{(Habitual/Orbit)} = \frac{3}{8} * H\left[\frac{47}{100}\right] + \frac{5}{8} * H\left[\frac{43}{100}\right] = \frac{3}{8} * 0.9974 + \frac{5}{8} * 0.9858 = 0.9901$

$IG_{(Habitual/Size)} = 0.9969 - 0.9841 = 0.0128$

$IG_{(Habitual/Orbit)} = 0.9969 - 0.9901 = 0.0068$

$IG_{(Habitual/\ Orbit)} = 0.0068 < IG_{(Habitual/Size)} = 0.\ 0128$

Therefore, we choose size as root node, the ID3 decision tree will become:



2. For just 9 of the planets, a third feature, Temperature (in Kelvin degrees), has been measured, as shown in the nearby table. Redo all the steps from part 1 on this data using all three features. For the Temperature feature, in each iteration you must maximize over all possible binary thresholding splits (such as T ≤ 250 vs. T > 250, for example).

Size	Orbit	Temperature	Habitable
Big	Far	205	No
Big	Near	205	No
Big	Near	260	Yes
Big	Near	380	Yes
Small	Far	205	No
Small	Far	260	Yes
Small	Near	260	Yes
Small	Near	380	No

The 9th row is small, near, 380 and no

Binary threshold splits for the continuous attribute temperature

205°C: [0+, 3-] < 232.5°C < 260°C: [3+, 0-] < 320°C <380°C: [1+, 2-]

$H_{Habitable} = H\left[\frac{4}{9}\right] = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.9912$

$H_{(Habitable/Size)} = \frac{4}{9} + \frac{5}{9} * H\left[\frac{2}{5}\right] = \frac{4}{9} + \frac{5}{9} * 0.9709 = 0.9838$

$H_{(Habitable/Temp<=232.5)} = \frac{2}{3} * H\left[\frac{1}{3}\right] = \frac{2}{3} * 0.9182 = 0.6121$

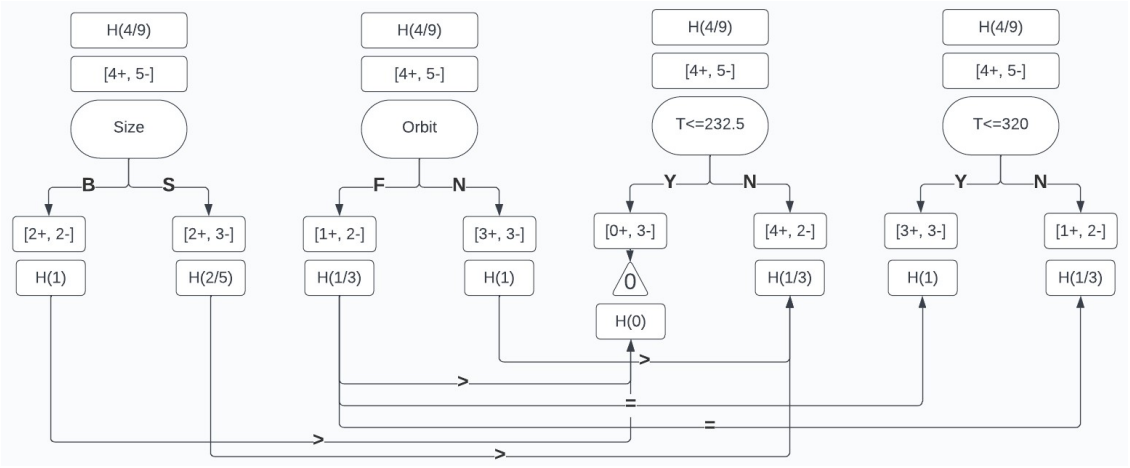
$H_{(Habitable/Orbit)} = \frac{1}{3} * H\left[\frac{1}{3}\right] + \frac{2}{3} = \frac{1}{3} * 0.9182 + \frac{2}{3} = 0.9272$

$H_{(Habitable/Temp<=320)} = \frac{2}{3} + \frac{1}{3} * H\left[\frac{1}{3}\right] = \frac{2}{3} + \frac{1}{3} * 0.9182 = 0.9272$

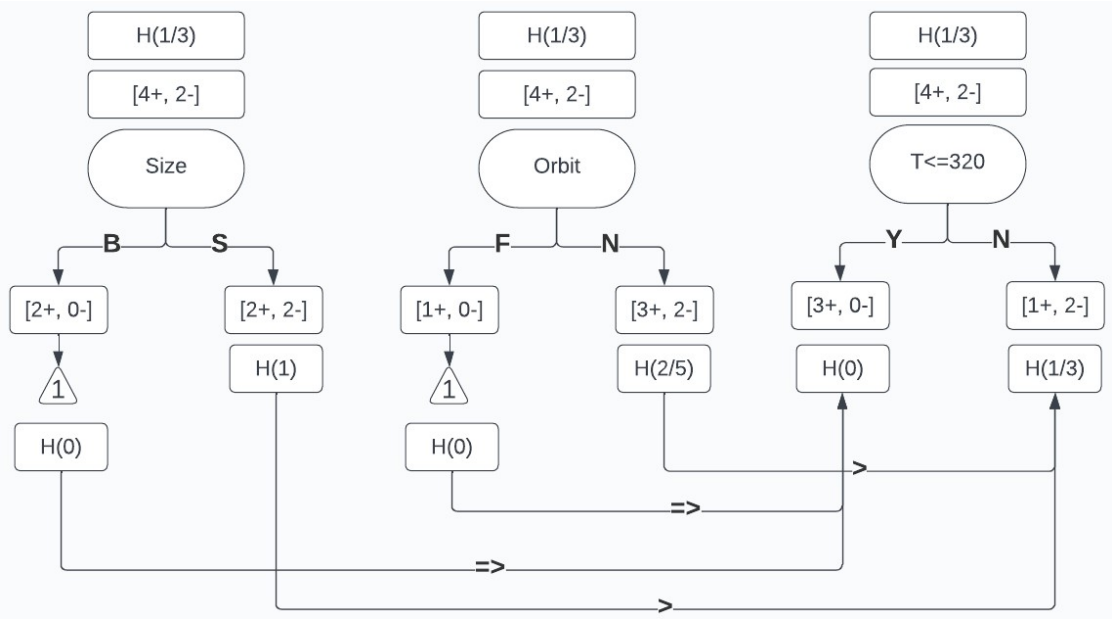
$IG_{(Habitable/Size)} = H\left[\frac{4}{9}\right] - 0.938 = 0.9912 - 0.9838 = 0.0074$

$IG_{(Habitable/Orbit)} = IG_{(Habitable/\ Temp<=320)} = H\left[\frac{4}{9}\right] - 0.9272 = 0.9912 - 0.9272 = 0.064$

$IG_{(Habitable/Temp<=232.5)} = 0.3791 > IG_{(Habitable/Size)} > IG_{(Habitable/\ Orbit)} = IG_{(Habitable/\ Temp<=320)}$



Thus, we choose (T <= 232.5) as root node



$H_{Temp<=232.5} = H\left[\frac{1}{3}\right] = 0.9182$

$H_{(Temp<=232.5\ /Size)} = \frac{2}{3} * H[1] = \frac{2}{3}$

$H_{(Temp<=232.5\ /Temp<=320)} = \frac{1}{2} * H\left[\frac{1}{3}\right] = \frac{1}{2} * 0.9182 = 0.4591$

$H_{(Temp<=232.5\ /Orbit)} = \frac{5}{6} * H\left[\frac{2}{5}\right] = \frac{5}{6} * 0.9709 = 0.8091$

$IG_{(Temp<=232.5\ //Temp<=320)} = H\left[\frac{4}{9}\right] - 0.4591 = 0.9182 - 0.4591 = 0.4591$

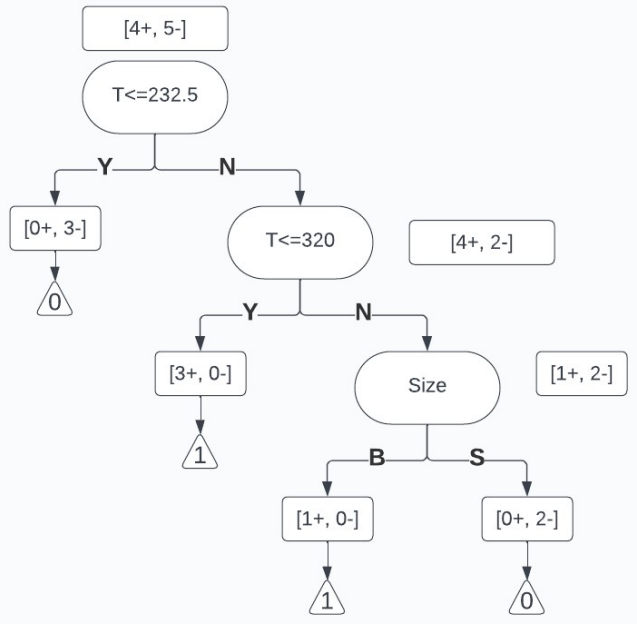
$IG_{(Temp<=232.5\ /Orbit)} = H\left[\frac{4}{9}\right] - 0.8091 = 0.9182 - 0.9272 = 0.1091$

$IG_{(Temp<=232.5\ /\ Size)} = H\left[\frac{4}{9}\right] - \frac{2}{3} = 0.2515$

$IG_{(Temp<=232.5\ /Temp<=320)} = 0.4591 > IG_{(Temp<=232.5\ /Size)} > IG_{(Temp<=232.5\ /Orbit)}$

Therefore the level 2 will become T <=320°C

So the final decision tree is:



According to my decision tree, the planet with the features (Big, Near, 280) is habitable.

3. Problem 3:

Given dataset about animal

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
Leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
Gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Using Naïve Bayes Classifier to find whether this animal is mammal or not (Using Laplacian smoothing if need)

	Give Birth	Can Fly	Live in Water	Have Legs	Class
1	Yes	No	Yes	No	?

$P_{mammal} = \frac{7}{20}$; $P_{non\text{-}mammal} = \frac{13}{20}$

	Give birth	Can fly	Live in water	Have legs	Total
	Yes	No	Yes	No	
Mammal	6	6	2	2	7
Non-mammal	1	10	3	4	13

$P_{(Mammal|X)} = P_{(X|Mammal)} \times P_{(Mammal)} = P_{(GiveBirth|Mammal)} \times P_{(CanFly|Mammal)} \times P_{(LiveInWater|Mammal)} \times P_{(HaveLegs|Mammal)} \times P_{(Mammal)} = \frac{6}{7} * \frac{6}{7} * \frac{2}{7} * \frac{2}{7} * \frac{7}{20} = 0.020991$

$P_{(Non\text{-}mammal|X)} = P_{(X|Non\text{-}mammal)} \times P_{(Non\text{-}mammal)} = P_{(GiveBirth|Non\text{-}mammal)} \times P_{(CanFly|Non\text{-}mammal)} \times P_{(LiveInWater|Non\text{-}mammal)} \times P_{(HaveLegs|Non\text{-}mammal)} \times P_{(Non\text{-}mammal)} = \frac{1}{13} * \frac{10}{13} * \frac{3}{13} * \frac{4}{13} * \frac{13}{20} = 0.002731$

Because $P_{(Mammal|X)} > P_{(Non\text{-}mammal|X)}$, we conclude that the animal is mammal