Assignment: Decision Trees & Random Forest

ID: 2021701010

Question 5:

	Review	Smell	Taste	Portion
1	Negative	Woody	Sweet	Small
2	Negative	Fruity	Salty	Large
3	Negative	Fruity	Salty	Large
4	Positive	Fruity	Sour	Small
5	Positive	Woody	Sour	Small
6	Negative	Woody	Sweet	Large
7	Positive	Woody	Sour	Large
8	Positive	Fruity	Salty	Small
9	Positive	Fruity	Salty	Small
10	Negative	Woody	Sweet	Large

↑ To predict

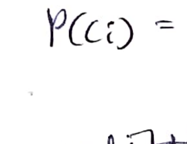
a

↳ Compute a 'decision- Tree' with the goal to predict the food review based on its smell, taste & portion size.

a) Compute the entropy of each rule in the first stage.

Ans

Root node



The "entropy" at root node $= -\sum_{i=1}^{k} P(C_i)\log_2[P(C_i)] = H_{root}$

→ where C_i = ith class

$P(C_i) = \dfrac{\#\text{ of samples in Class 'i'}}{\#\text{ total samples}}$

And the entropy after splitting (based on some feature & a particular question about that feature)

$= \sum_{j=1}^{no\_split} P(D_j)\,H(D_j)$, where $P(D_j) = \dfrac{|D_j|}{|D|} = \dfrac{\#\text{ of samples in } D_j}{\text{total samples in } D}$

& $H(D_j)$ = Entropy of the node containing D_j.
= Entropy of each child nodes

~~& Information gation :~~

& no_split = total # splits (= total child nodes formed).

- Information gain : = Δi

Δi = (Entropy before splitting) − (Entropy after splitting)

= (Entropy of parent node) − (weighted combinations of Entropy of child nodes).

×

Calculation time :

- Entropy at root node (Entropy before splitting)

$$= -P(C_1)\log_2[P(C_1)] + -P(C_2)\log_2[P(C_2)]$$

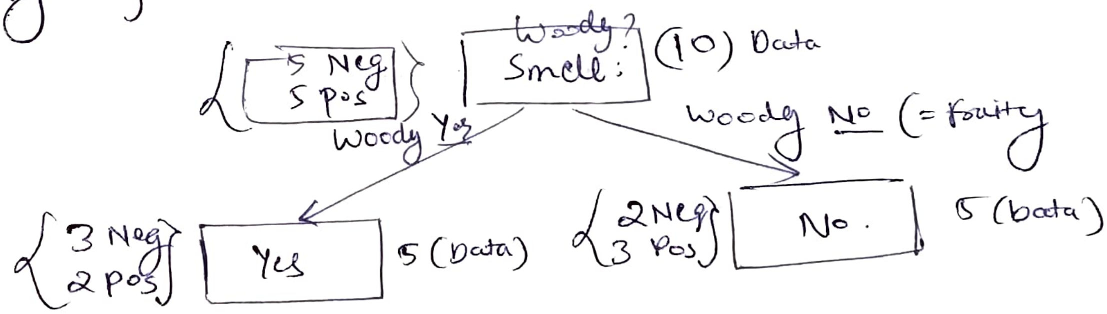
where C_1 = +ve class, C_2 = −ve class → for Review

$$= -p_1\log_2 p_1 - p_2\log_2 p_2$$

$$= -\left(\frac{5}{10}\right)\log_2\left(5/10\right) - \left(5/10\right)\log_2\left(5/10\right)$$

$$= 1$$

- Entropy "if data are splitted on the basis of Smell" :



$$\text{Total\_H} = \underbrace{\frac{5}{10}}_{P(D_1)}\underbrace{\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]}_{H(D_1)} + \underbrace{\frac{5}{10}}_{P(D_2)}\underbrace{\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]}_{H(D_2)}$$

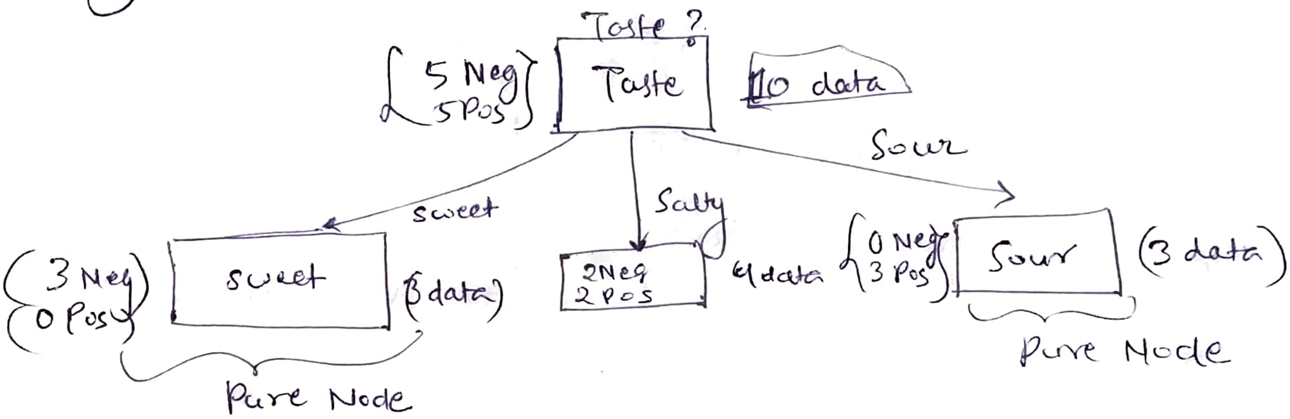
$$= -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$

$$= -\frac{3}{5}(-0.737) - \frac{2}{5}(-1.322)$$

$$= 0.971$$

$$\Delta i = 1 - 0.971 = \boxed{0.029} \quad \text{——} \; (1)$$

- Entropy "if data are splitted on the basis of Taste".



$$\text{Total } H = \frac{3}{10}\left[-\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right] +$$

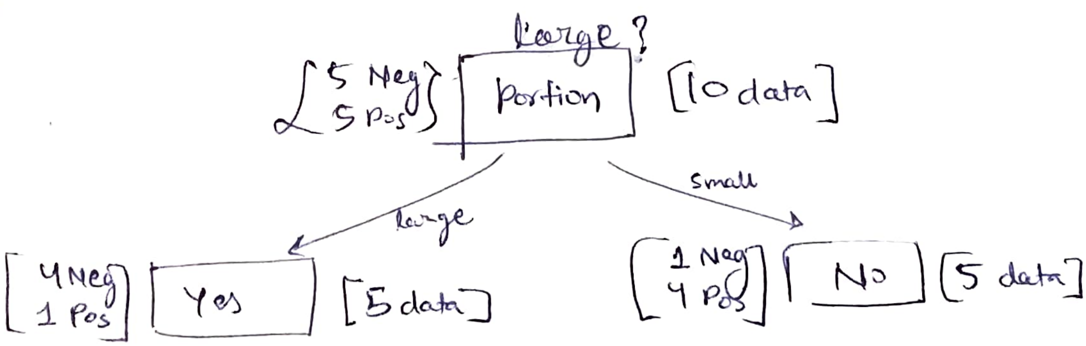
$$\frac{4}{10}\left[-\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] +$$

$$\frac{3}{10}\left[-\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2(3/3)\right]$$

$$= 0 + 0.4 + 0. = 0.4$$

$$\text{Then, } \Delta i = 1 - 0.4 = \boxed{0.6} \quad \text{——} \; (2)$$

- Entropy "if data are splitted on the basis of "portion".)



Total H = Total entropy

$$= \frac{5}{10}\underbrace{\left[-\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right]}_{\text{for (Yes) = Large}} + \frac{5}{10}\underbrace{\left[-\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right]}_{\text{for (No) = Small}}$$

$$= 0.72$$

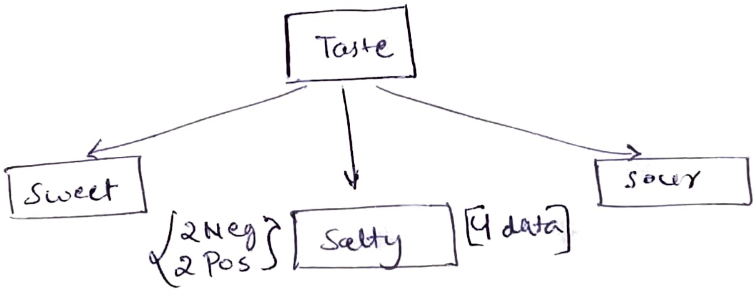
$$\Delta i = 1 - 0.72 = \boxed{0.28} \text{ —— (3)}$$

Comparing (1), (2) & (3); we have (2) > (3) > (1).

⇒ 'Taste' should be at the root node.

→ Now; once "taste" has took the root node, it's left & right node (i.e., sweet & sour are "pure" nodes & hence the branch won't be extended from them & they will be regarded as leaf nodes).

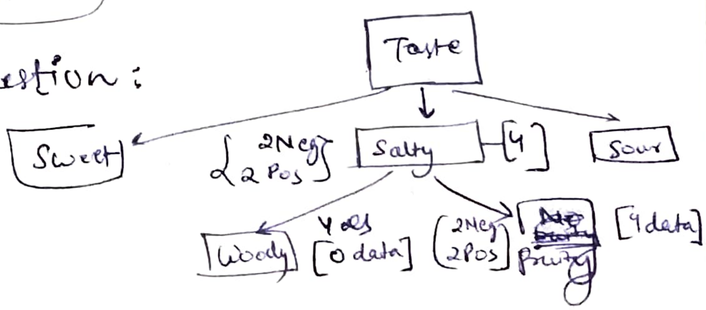
→ from "salty" again splitting would be done (with the same process & steps as done for root node) but with other features.



$$\text{Entropy[salty-node]} = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= 1.0$$

- Taking smell = woody ? question:

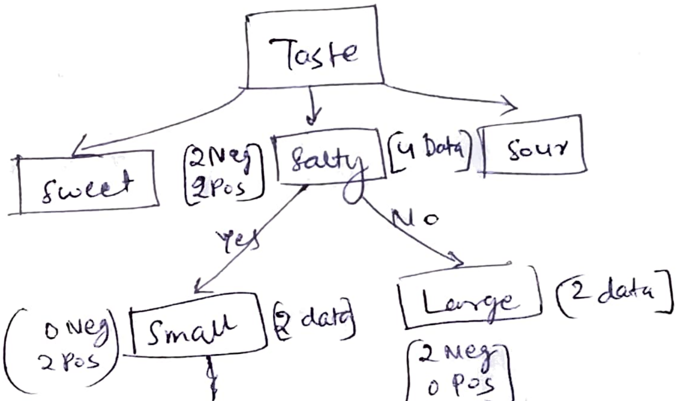


$$\text{Total-H} = \underbrace{\frac{0}{4}\left[\quad\right]}_{\text{for Woody}} + \underbrace{\frac{4}{4}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]}_{\text{(for Fruity)}}$$

$$= 1$$

$$\Rightarrow \Delta i = 1 - 1 = 0 \quad \text{(Not preferable)} \longleftarrow (4)$$

- Taking Portion = Small ?



Both are "pure" (hence "leaf") nodes → indicates termination of tree.

Total-H

$$= \frac{2}{4}\left[-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right]$$

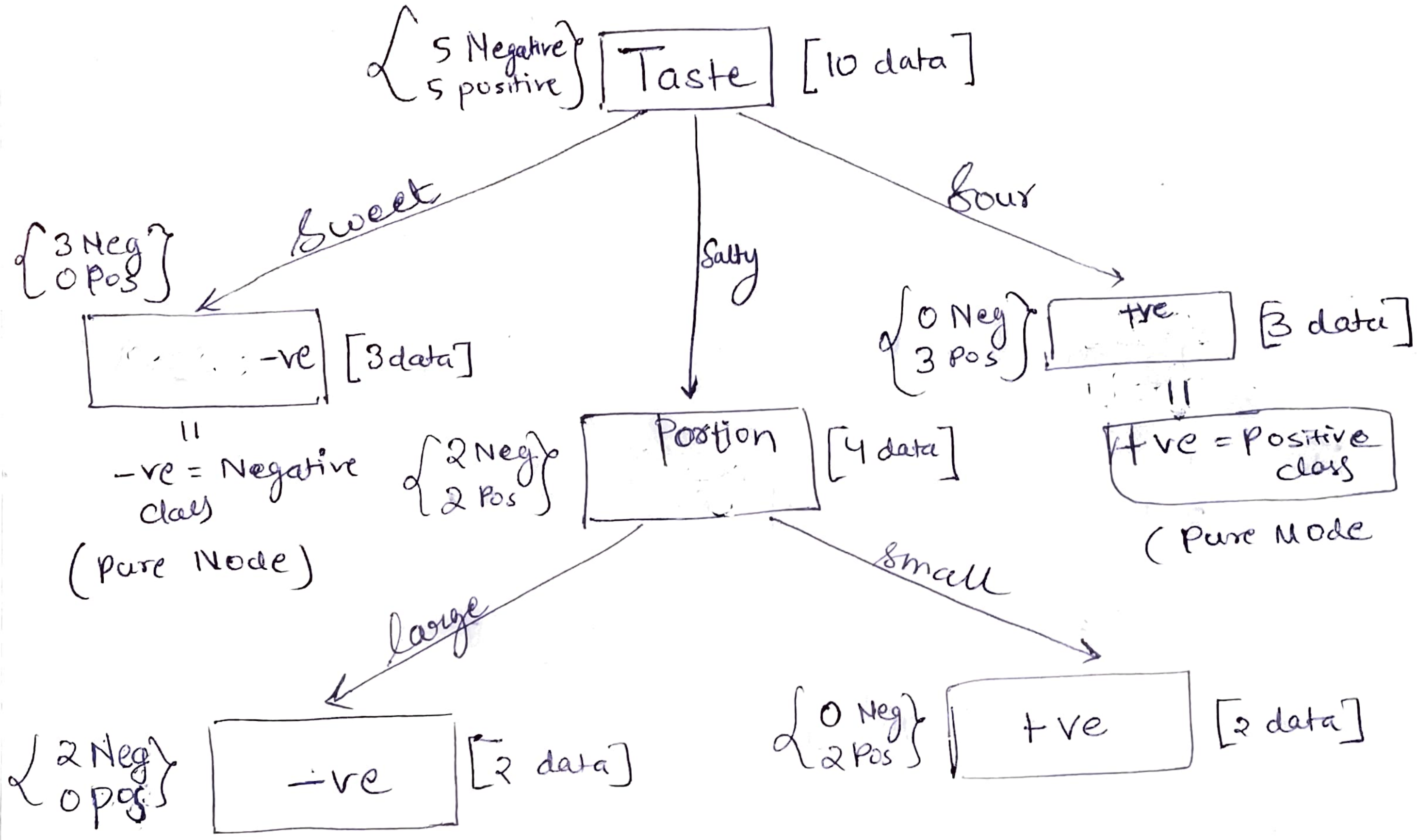
$$+ \frac{2}{4}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right]$$

$$= 0$$

$$\Rightarrow \Delta i = \underset{\text{(at Salty)}}{1} - 0 = 1 \longrightarrow (5)$$

☆ Since (5) > (4) ⇒ At "salty" node, the question which should be asked = (portion = small)? (i.e. splitting would be done based on feature 'portion')

→ Now, as we have got 'leaf' nodes (i.e pure nodes) our tree has fully grown. So final tree would look like something ~~at back~~ in next page.



Note: 'Small' feature remain useless (or doesn't have any impact on final ~~decision~~ label i.e.. Negative or positive).