ARAVIND
2019102014

Given the following dataset,

Review	Smell	Taste	Portion
Negative	Woody	Sweet	Small
Negative	Fruity	Salty	Large
Negative	Fruity	Salty	Large
Positive	Fruity	Sour	Small
Positive	Woody	Sour	Small
Negative	Woody	Sweet	Large
Positive	Woody	Sour	Large
Positive	Fruity	Salty	Small
Positive	Fruity	Salty	Small
Negative	Woody	Sweet	Large

Let Negative = neg
Positive = pos
Woody = W_{smell}
Fruity = F_{smell}
Sweet = Sw_{Taste}
Salty = Sa_{Taste}
Sour = So_{Taste}
Small = $S_{postion}$
Large = $L_{postion}$

(a)

(1) First stage:

There are 3 cases:

(i) Split by taste

(ii) Split by smell

(iii) Split by postion

Solving for (i),

a) For sweet taste, all 3 data is -ve review

$$E = -\sum p_i \log p_i = -1\log(1) = 0$$

b) For salty taste, 4 reviews [2 +ve, 2 -ve]

$$E = \left[\frac{1}{2}\log\left(\frac{1}{2}\right)\right] + \left[\frac{-1}{2}\log\left(\frac{1}{2}\right)\right]$$

$$E = \log 2 = 1$$

c) For sour taste, all 3 positive

$$E = 0$$

Sum of all 3 ⟵ $E_{taste} = \frac{4}{10} = 0.4$

Solving for (ii)

a) For woody smell _ (3 -ve + 2 +ve)

$$E = \frac{-3}{5}\log\left(\frac{3}{5}\right) + \left(\frac{-2}{5}\right)\log\left(\frac{2}{5}\right)$$

$$E = \frac{-3}{5}(-0.7369) + \left(\frac{-2}{5}\right)(-1.321)$$

$$E = 0.97054 \text{ [approx]}$$

b) For fruity smell, (2 -ve, 3 +ve)

$$E = \frac{-2}{5}\log\left(\frac{2}{5}\right) + \left(\frac{-3}{5}\right)\log\left(\frac{3}{5}\right)$$

$$E = 0.9705 \text{ (approx)}$$

$E_{smell} = 0.97054$

Solving for (iii)

a) For small position (1 -ve, 4 +ve)

$$E = \frac{-1}{5}\log\left(\frac{1}{5}\right) + \left(\frac{-4}{5}\right)\log\left(\frac{4}{5}\right)$$

$$E = 0.722 \text{ (approx)}$$

b) For large position (1 +ve, 4 -ve)

$$E = \left(\frac{-4}{5}\right)\log\left(\frac{4}{5}\right) + \frac{-1}{5}\log\left(\frac{1}{5}\right)$$

$$E = 0.722 \text{ (approx)}$$

$E_{position} = 0.722$

(b)

From above cases, we calculate the information gain,

(1) $I_{taste} = 1 - 0 - 1\left[\frac{4}{10}\right] - 0 = 0.6$

(2) $I_{smell} = 1 - \frac{1}{2}(0.97054) - \frac{1}{2}(0.97054) = 1 - 0.97054 = 0.02946$

(3) $I_{position} = 1 - \frac{1}{2}(0.722) - \frac{1}{2}(0.722) = 1 - 0.722 = 0.278$

→ Most information gain is from taste, so that would be first split

→ This already sorts the sugary and sour taste series, leaving us with only salty taste

→ When split based on portion, salty taste gets sorted.

∴ Decision Tree is as followings: