Q.

| Deadline | is there a Party ? | Lazy | Activity |
|---|---|---|---|
| Urgent | yes | yes | party |
| Urgent | no | yes | study |
| Near | yes | yes | party |
| None | yes | no | party |
| None | no | yes | pub |
| None | yes | no | party |
| Near | no | no | study |
| Near | no | yes | TV |
| Near | yes | yes | party |
| Urgent | no | no | study |

A:

→ $\text{Entropy}(s)$ or $E(s) = -P_p \log_2 P_p - P_s \log_2 P_s - P_{pub} \log_2 P_{pub} - P_{TV} \log_2 P_{TV}$

$$E(s) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{1}{10}\log_2\frac{1}{10} - \frac{1}{10}\log_2\frac{1}{10}$$

$$= 0.5 + 0.5211 + 0.3322 + 0.3322$$

$$= 1.6855$$

→ To find feature with max. info. gain:

$$\text{Gain}(S, \text{deadline}) \text{ or } G(S, \text{deadline}) = E(s) - \frac{|S_{urg}|}{|S|}E(S_{urg}) - \frac{|S_{near}|}{|S|}E(S_{near}) - \frac{|S_{none}|}{|S|}E(S_{none})$$

$$= 1.6855 - \frac{3}{10}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{4}{10}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4}\right] - \frac{3}{10}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)$$

$$\therefore G(S, \text{deadline}) = 1.6855 - 0.2755 - 0.6 - 0.2755$$

$$= 0.5345$$

$$G(S, \text{party}) = 1.6855 - \frac{5}{10}\left(-\frac{5}{5}\log_2\frac{5}{5}\right) - \frac{5}{10}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{1}{5}\log_2\frac{1}{5} - \frac{1}{5}\log_2\frac{1}{5}\right)$$

(yes (party); party (activity); no; study; pub; TV)

$$= 1.6855 - 0 - 0.6855$$

$$\therefore G(S, \text{party}) = 1.0$$

$$G(S, \text{lazy}) = 1.6855 - \frac{6}{10}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{1}{6}\log_2\frac{1}{6} - \frac{1}{6}\log_2\frac{1}{6} - \frac{1}{6}\log_2\frac{1}{6}\right) - \frac{4}{10}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right)$$

(yes (lazy); no (lazy); party; study (activity))

$$= 1.6855 - 1.0755 - 0.4$$

$$\therefore G(S, \text{lazy}) = 0.21$$

→ Since $G(S, \text{party})$ is max, party feature will be the root node with possible values "yes" & "no". Here, for all 5 "yes" in the data, the activity is "party". For the "no"s we need further features to decide

→ After removing the "yes" for Party rows from the data :

| Deadline | is there a Party ? | Lazy | Activity |
|---|---|---|---|
| Urgent | No | yes | study |
| None | No | yes | Pub |
| Near | No | no | study |
| Near | No | yes | TV |
| Urgent | No | ~~yes~~ no | study |

→ Calculating entropy again :

$$E(S) = -P_S \log_2 P_S - P_{pub} \log_2 P_{pub} - P_{TV} \log_2 P_{TV}$$

$$= -\frac{3}{5}\log_2\frac{3}{5} - \frac{1}{5}\log_2\frac{1}{5} - \frac{1}{5}\log_2\frac{1}{5}$$

$$= 0.4421 + 0.4643 + 0.4643$$

$$\therefore E(S) = 1.371$$

→ Calculating info. gain :

$$G(S, \text{deadline}) = 1.371 - \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2}\right) \quad \text{(urgent)}$$

$$- \frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) \quad \text{(near)}$$

$$- \frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) \quad \text{(none)}$$

$$= 1.371 - 0 - 0.4 - 0$$

$$\therefore G(S, \text{deadline}) = 0.971$$

$$G(S, \text{lazy}) = 1.371 - \frac{3}{5}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) \quad \text{(yes)}$$

$$- \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2}\right) \quad \text{(no)}$$

$$= 1.371 - 0.9509 - 0$$

$$\therefore G(S, \text{lazy}) = 0.42$$

→ Since $G(S, \text{deadline})$ is max, it is chosen as the next node. Wherever deadline is "urgent", activity is "study" so that will lead to a leaf node. For ~~none~~ ~~and~~ "near" we need to calculate further. For "none" it leads to "pub" as a leaf node.

| Deadline | Party? | Lazy | Activity |
|---|---|---|---|
| near | no | no | study |
| near | no | yes | TV |

→ From the simplified table, it is obvious that when lazy is "yes" activity is TV & lazy is "no", activity is study

→ Complete decision tree :