DATE. NO.

< Decision Tree 과제 1 >

① C4.5 알고리즘.

처음 entropy $= -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right) = 0.940286$

1) outlook 로 나눌 때

sunny $-(Y_2\ N_3) = -\frac{5}{14}\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.3467681$

rain $= (Y_3\ N_2) = -\frac{5}{14}\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.3467681$

overcast $= (Y_4\ N_0) = 0$

$\therefore E(S, \text{outlook}) > 0.693536$ ↗ $IGR = \frac{IG}{IV} =$ **0.156428**

$IG = 0.940286 - 0.693536$, $IV = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{5}{14}\log_2\frac{5}{14} + \frac{4}{14}\log_2\frac{4}{14}\right) = 1.577406$,

2) temp 로 나눌 때

high $= (Y_2 N_2) = -\frac{4}{14}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.285714$

mild $= (Y_4 N_2) = -\frac{6}{14}\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.342643$

cool $= (Y_3 N_1) = -\frac{4}{14}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = 0.231784$ $\therefore E(S, \text{temp}) = 0.851123$

$IG = 0.940286 - 0.851123$, $IV = -\left(\frac{4}{14}\log_2\frac{4}{14} + \frac{6}{14}\log_2\frac{6}{14} + \frac{4}{14}\log_2\frac{4}{14}\right)$ $\therefore IGR =$ **0.0568614**

3) humidity 로 나눌 때

$E(S, \text{humidity}) = -\frac{7}{14}\left(\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right) - \frac{7}{14}\left(\frac{6}{7}\log_2\frac{6}{7} + \frac{1}{7}\log_2\frac{1}{7}\right) = 0.788505$

$IV = -\left(\frac{7}{14}\log_2\frac{7}{14} + \frac{7}{14}\log_2\frac{7}{14}\right) = 1$. $\therefore IGR =$ **0.151836**

4) Windy 로 나눌 때

$E(S, \text{windy}) = -\frac{6}{14}\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) - \frac{8}{14}\left(\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right) = 0.892158$

$IV = -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = 0.985228$. $\therefore IGR =$ **0.048484**

∴ 처음에 기준은 **outlook** (IGR maximize)

두번째 기준 찾기. — $Y_2\ N_3$.

① outlook - sunny 인 때 entropy $= -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.970956$

1) temp 로 나눌 때

high $= (Y_0\ N_2) =$ 0.

mild $= (Y_1\ N_1) = -\frac{2}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.4$ $E(S_1, \text{temp}) = 0.4$

cool $= (Y_1\ N_0) = 0$.

$IV = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 1.521928$, $\therefore IGR =$ **0.375146**

2) humidity 로 나눌 때

high $= (Y_0\ N_3)$, normal $= (Y_2\ N_0) \Rightarrow E(S_1, \text{humidity}) =$ **0** : best!

$IG = 0.970956$, $IV = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.970956$ $IGR =$ **1** : best

3) Windy 로 나눌 때

True $= (Y_1\ N_1)$, False $= (Y_1\ N_2)$.

$E(S_1, \text{windy}) = -\frac{2}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{3}{5}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.950995$

$IG = 0.019961$, $IV = 0.970956$, $IGR =$ **0.020100**

∴ outlook - sunny : humidity 로 분리.

② outlook - rain 인 경우. entropy $= 0.970956$.

1) temp 로

high $= \times$, mild $= (Y_2\ N_1)$, cool $= (Y_1\ N_1)$

$E(S_2, \text{temp}) = -\frac{3}{5}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.950995$

$IG = 0.019961$, $IV = 0.970956$, $IGR =$ **0.020029**

2) humidity

high $= (Y_1\ N_1)$, normal $= (Y_2\ N_1)$. $\Rightarrow IGR = 0.020029$

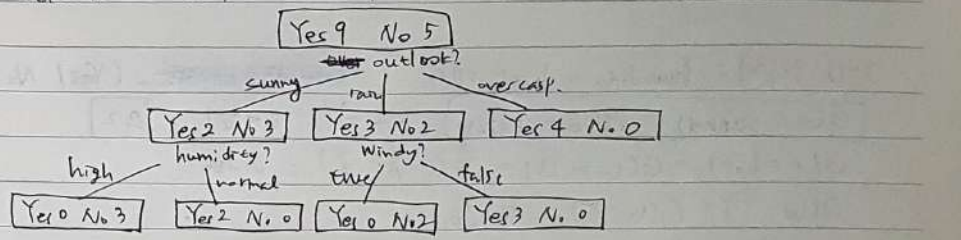
3) windy

true $= (Y_0\ N_2)$, false $= (Y_3\ N_0)$ $E(S_2, \text{windy}) = 0$.

$IV = 0.970956$, $IGR =$ **1**

∴ outlook - rain 은 windy 로 분리

③ outlook - overcast 는 끝.



② CART.

1) 1단계

$G(outlook = sunny) = \frac{5}{14}(1-\frac{4}{25}-\frac{9}{25}) + \frac{9}{14}(1-\frac{4}{81}-\frac{49}{81}) = 0.39365$

$G(\ \text{"} = rain) = 0.39365$

$\boxed{G(\ \text{"} = overcast)} = \frac{4}{14}\times 0 + \frac{10}{14}(1-\frac{9}{100}-\frac{49}{100}) = \boxed{0.357}$ → minimum.

$G(temp = hot) = \frac{4}{14}(1-\frac{1}{4}-\frac{1}{4}) + \frac{10}{14}(1-\frac{9}{100}-\frac{49}{100}) = 0.4428\cdots$

$G(\ \text{"} = mild) = \frac{6}{14}(1-\frac{1}{9}-\frac{4}{9}) + \frac{8}{14}(1-\frac{9}{64}-\frac{25}{64}) = 0.4581\cdots$

$G(\ \text{"} = cool) = \frac{4}{14}(1-\frac{1}{16}-\frac{9}{16}) + \frac{10}{14}(1-\frac{36}{100}-\frac{16}{100}) = 0.45$

$G(windy = T) = \frac{6}{14}(1-\frac{1}{4}-\frac{1}{4}) + \frac{8}{14}(1-\frac{4}{64}-\frac{36}{64}) = 0.4285\cdots$

$G(\ \text{"} = F) = 0.4285\cdots$

$G(humidity = high) = G(\ \text{"} = normal) = \frac{1}{2}(1-\frac{9}{49}-\frac{16}{49}) + \frac{1}{2}(1-\frac{1}{49}-\frac{36}{49}) = 0.3673\cdots$

⇒ outlook이 overcast인지 아니지로 첫 분기

2) 2단계 (outlook이 sunny·rain인 가지만 진행)

$G(o = sunny) = G(o = rain) = \frac{1}{2}(1-\frac{4}{25}-\frac{9}{25}) + \frac{1}{2}(1-\frac{4}{25}-\frac{9}{25}) = 0.48$

$G(t = hot) = 0 + \frac{8}{10}(1-\frac{9}{64}-\frac{25}{64}) = 0.375$

$G(t = mild) = 0.48, \quad G(t = cool) = \frac{3}{10}(1-\frac{1}{9}-\frac{4}{9}) + \frac{7}{10}(1-\frac{9}{49}-\frac{16}{49}) = 0.4761\cdots$

$G(w = T) = G(w = F) = \frac{4}{10}(1-\frac{1}{16}-\frac{9}{16}) + \frac{6}{10}(1-\frac{4}{36}-\frac{16}{36}) = 0.4166\cdots$

$\boxed{G(h = high) = G(h = normal)} = \frac{1}{2}(1-\frac{1}{25}-\frac{16}{25}) + \frac{1}{2}(1-\frac{1}{25}-\frac{16}{25}) = \boxed{0.32}$

⇒ humidity가 high인지 normal인지로 두번째 분기.

3-1) 3단계 - humidity = high에서. ~~(Yes 5 No 5)~~ (Yes 1 No 4)

$\boxed{G(o = sunny) = G(o = rain)} = 0 + \frac{2}{5}(1-\frac{1}{4}-\frac{1}{4}) = \boxed{0.2}$

$G(t = hot) = G(t = mild) = \frac{3}{5}(1-\frac{1}{9}-\frac{4}{9}) = 0.266\cdots$

$G(w = T) = G(w = F) = 0.266\cdots$

⇒ outlook이 sunny인지 rain인지로 분기.

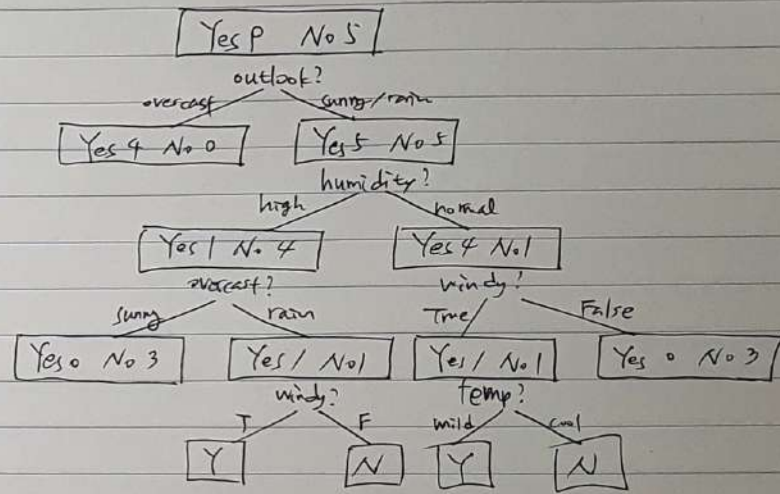
3-2) 3단계 - humidity = normal에서.

$G(o = sunny) = G(o = rain) = 0.266\cdots$

$G(t = mild) = G(t = high) = 0.266\cdots$

$\boxed{G(w = T) = G(w = F)} = \frac{2}{5}\times\frac{1}{2} + \frac{3}{5}\times 0 = \frac{1}{5} = \boxed{0.2}$

⇒ Windy가 True인지 False인지로 분기.



* 4단계는 구분시킬 수 있는 기준을 전체 표를 보고 직접 찾음
(단, 가지치기할 수도.)