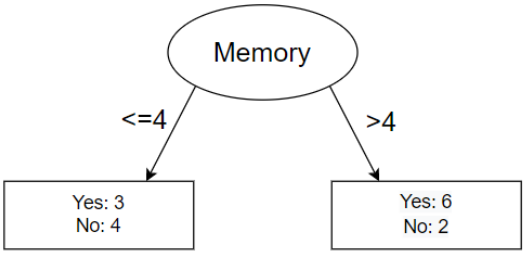
تمرین و پروژه شماره ۱ درس داده کاوی، پاسخ سؤال ۲

تکین جزایری، ۹۸۱۳۰۰۶ – امیرحسین رجبی، ۹۸۱۳۰۱۳

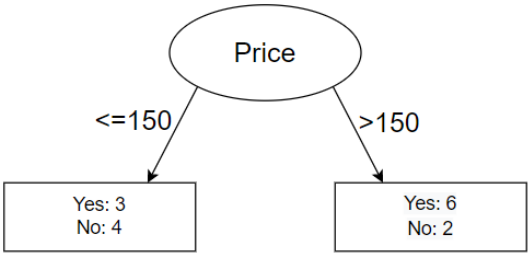
در حالت پایه، یعنی زمانی که تنها یک برگ داریم، میزان ناخالصی به صورت زیر است:

$$I(parent) = -\frac{9}{15}\log\frac{9}{15} - \frac{6}{15}\log\frac{6}{15} = 0.971$$

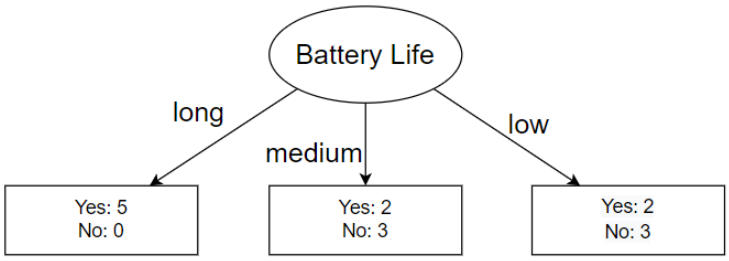
در ادامه به سه حالت می توانیم درخت را گسترش دهیم:



$$I(Memory \le 4) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} = 0.985$$
$$I(Memory > 4) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8} = 0.811$$
$$I(Memory) = (\frac{7}{15} \times 0.985) + (\frac{8}{15} \times 0.811) = 0.892$$
$$Gain = 0.971 - 0.892 = 0.079$$
$$\text{Split Info} = -\frac{7}{15}\log\frac{7}{15} - \frac{8}{15}\log\frac{8}{15} = 0.997$$
$$Gain\ ration = \frac{0.079}{0.997} = 0.079$$

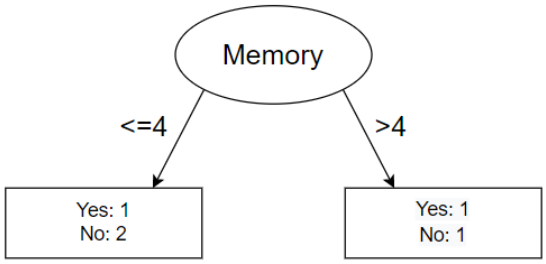


$$I(Price \le 150) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} = 0.985$$
$$I(Price > 150) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8} = 0.811$$
$$I(Price) = (\frac{7}{15} \times 0.985) + (\frac{8}{15} \times 0.811) = 0.892$$
$$Gain = 0.971 - 0.892 = 0.079$$
$$\text{Split Info} = -\frac{7}{15}\log\frac{7}{15} - \frac{8}{15}\log\frac{8}{15} = 0.997$$
$$Gain\ ration = \frac{0.079}{0.997} = 0.079$$

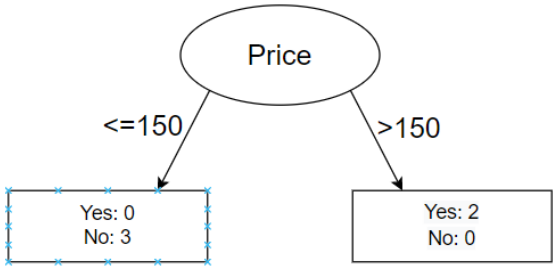


$$I(Battery\ life = long) = 0$$
$$I(Battery\ life = medium) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.971$$
$$I(Battery\ life = low) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.971$$
$$I(Battery\ life) = \frac{10}{15} \times 0.971 = 0.647$$
$$Gain = 0.971 - 0.647 = 0.324$$
$$\text{Split Info} = -\frac{5}{15}\log\frac{5}{15} - \frac{5}{15}\log\frac{5}{15} - \frac{5}{15}\log\frac{5}{15} = 0.477$$
$$Gain\ ration = \frac{0.324}{0.477} = 0.679$$

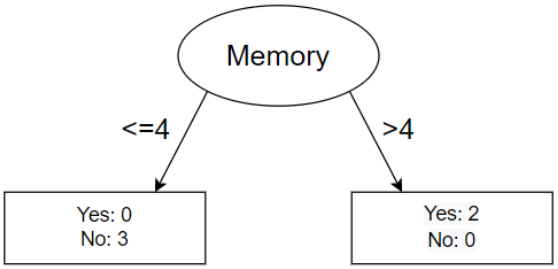
از آنجا که گسترش جدول بر اساس Battery life بیشترین Gain ration را حاصل می کند، این تقسیم بندی را به درخت تصمیم اضافه می کنیم. در ادامه باید تلاش کنیم تا دو برگ Battery life = medium و Battery life = low را گسترش دهیم.



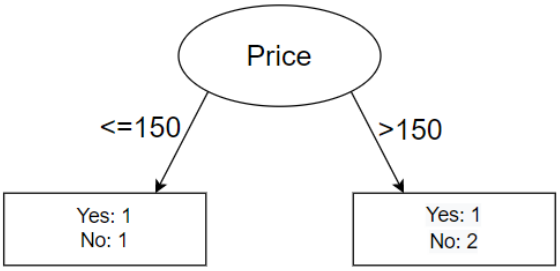
$$I(Memory \leq 4) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918$$
$$I(Memory > 4) = 1$$
$$I(Memory) = (\frac{3}{5} \times 0.918) + (\frac{2}{5} \times 1) = 0.951$$
$$Gain = 0.971 - 0.951 = 0.020$$
$$\text{Split Info} = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$
$$Gain\ ration = \frac{0.02}{0.971} = 0.021$$



$$I(Price) = 0$$
$$Gain = 0.971 - 0 = 0.971$$
$$\text{Split Info} = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$
$$Gain\ ration = \frac{0.971}{0.971} = 1$$

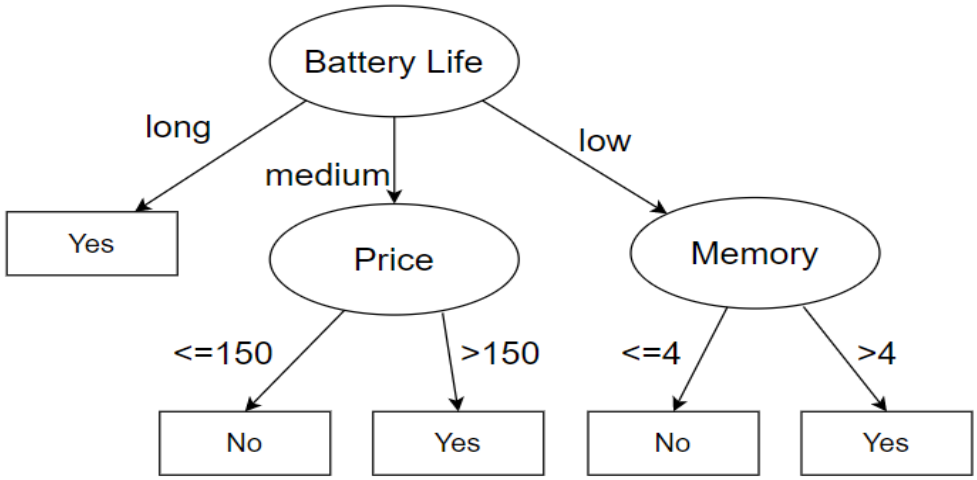


$$I(Memory) = 0$$
$$Gain = 0.971 - 0 = 0.971$$
$$\text{Split Info} = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$
$$Gain\ ration = \frac{0.971}{0.971} = 1$$



$$I(Memory \leq 4) = 1$$
$$I(Memory > 4) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918$$
$$I(Memory) = (\frac{2}{5} \times 1) + (\frac{3}{5} \times 0.918) = 0.951$$
$$Gain = 0.971 - 0.951 = 0.020$$
$$\text{Split Info} = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.971$$
$$Gain\ ration = \frac{0.02}{0.971} = 0.021$$

در نهایت درخت تصمیم به صورت زیر حاصل می شود:


Battery Life
long
medium
low
Yes
Price
Memory
<=150
>150
<=4
>4
No
Yes
No
Yes