1.

Tìm thuộc tính tại node gốc:

Tính entropy của PATIENT ID:

Với mỗi PATIENT ID thì HEART ATTACK chỉ có thể là yes hoặc no nên ta có:

$$\begin{aligned} H(PATIENT\ ID = i) &= -p(\text{yes}) \log_2 p(\text{yes}) - p(\text{no}) \log_2 p(\text{no}) \\ &= -\frac{1}{1} \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1} \\ &= 0 \end{aligned}$$

Có 7 PATIENT ID. Vậy entropy của PATIENT ID là:

$$\begin{aligned} H(PATIENT\ ID) &= 7 \left(\frac{1}{7} * 0 \right) \\ &= 0 \end{aligned}$$

Tính entropy của CHEST PAIN:

$$\begin{aligned} H(CHEST\ PAIN = yes) &= -p(\text{yes}) \log_2 p(\text{yes}) - p(\text{no}) \log_2 p(\text{no}) \\ &= -\frac{3}{3} \log_2 \frac{3}{3} - \frac{0}{3} \log_2 \frac{0}{3} \\ &= 0 \\ H(CHEST\ PAIN = no) &= -p(\text{yes}) \log_2 p(\text{yes}) - p(\text{no}) \log_2 p(\text{no}) \\ &= -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} \\ &= 0.81 \\ H(CHEST\ PAIN) &= \frac{3}{7} * 0 + \frac{4}{7} * 0.81 \\ &= 0.46 \end{aligned}$$

Tính entropy của GENDER:

$$
\begin{aligned}
H(GENDER = male) &= -p(\text{yes})\log_2 p(\text{yes}) - p(\text{no})\log_2 p(\text{no}) \\
&= -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} \\
&= 1 \\
H(GENDER = female) &= -p(\text{yes})\log_2 p(\text{yes}) - p(\text{no})\log_2 p(\text{no}) \\
&= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \\
&= 0.92 \\
H(GENDER) &= \frac{4}{7} * 1 + \frac{3}{7} * 0.92 \\
&= 0.97
\end{aligned}
$$

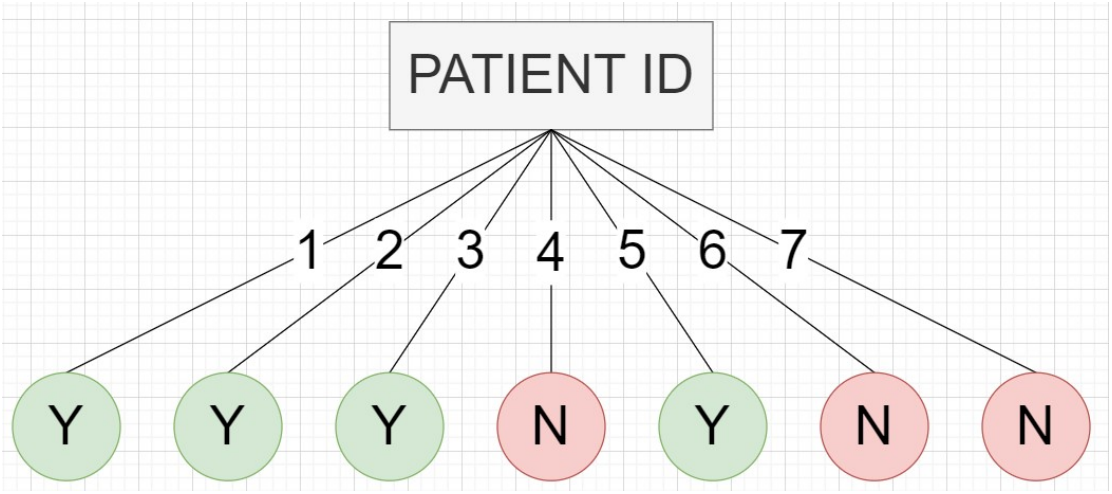
Tính entropy của SMOKES:

$$
\begin{aligned}
H(SMOKES = yes) &= -p(\text{yes})\log_2 p(\text{yes}) - p(\text{no})\log_2 p(\text{no}) \\
&= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} \\
&= 0.81 \\
H(SMOKES = no) &= -p(\text{yes})\log_2 p(\text{yes}) - p(\text{no})\log_2 p(\text{no}) \\
&= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} \\
&= 0.92 \\
H(SMOKES) &= \frac{4}{7} * 0.81 + \frac{3}{7} * 0.92 \\
&= 0.86
\end{aligned}
$$

Tính entropy của EXERCISES:

$$
\begin{aligned}
H(EXERCISES = yes) &= -p(\text{yes})\log_2 p(\text{yes}) - p(\text{no})\log_2 p(\text{no}) \\
&= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \\
&= 0.97 \\
H(EXERCISES = no) &= -p(\text{yes})\log_2 p(\text{yes}) - p(\text{no})\log_2 p(\text{no}) \\
&= -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} \\
&= 0 \\
H(EXERCISES) &= \frac{5}{7} * 0.97 + \frac{2}{7} * 0 \\
&= 0.69
\end{aligned}
$$

Entropy của PATIENT ID nhỏ nhất nên ta chọn PATIENT ID ở node gốc:



2.

Tìm thuộc tính tại node gốc:

Tính gini của Type of restaurant:

$$
\begin{aligned}
gini(Fast\ food) &= 1 - p(OK)^2 - p(Not\ OK)^2 \\
&= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \\
&= 0.44 \\
gini(Casual\ dining) &= 1 - p(OK)^2 - p(Not\ OK)^2 \\
&= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 \\
&= 0.48 \\
gini(Ethnic) &= 1 - p(OK)^2 - p(Not\ OK)^2 \\
&= 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \\
&= 0.44 \\
gini(Type\ of\ restaurant) &= \frac{3}{11} * 0.44 + \frac{5}{11} * 0.48 + \frac{3}{11} * 0.44 \\
&= 0.46
\end{aligned}
$$

Tính gini của Neighborhood:

$$
\begin{aligned}
gini(Oakland) &= 1 - p(OK)^2 - p(Not\ OK)^2 \\
&= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \\
&= 0.44 \\
gini(Squirrel\ Hill) &= 1 - p(OK)^2 - p(Not\ OK)^2 \\
&= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 \\
&= 0.5 \\
gini(Shadyside) &= 1 - p(OK)^2 - p(Not\ OK)^2 \\
&= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 \\
&= 0.5
\end{aligned}
$$

$$gini(Neighborhood) = \frac{3}{11} * 0.44 + \frac{4}{11} * 0.5 + \frac{4}{11} * 0.5$$
$$= 0.48$$

Tính gini của Restriction:

$$gini(Vegetarian) = 1 - p(OK)^2 - p(Not\ OK)^2$$
$$= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$$
$$= 0.5$$
$$gini(Gluten\ Free) = 1 - p(OK)^2 - p(Not\ OK)^2$$
$$= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2$$
$$= 0.44$$
$$gini(None) = 1 - p(OK)^2 - p(Not\ OK)^2$$
$$= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$$
$$= 0.5$$
$$gini(Restriction) = \frac{4}{11} * 0.5 + \frac{3}{11} * 0.44 + \frac{4}{11} * 0.5$$
$$= 0.48$$

Tính gini của Price:

Price	OK?
5	Not OK
40	Not OK
35	Not OK
80	Not OK
11	OK
31	OK
10	OK
22	Not OK
140	Not OK
68	OK
57	OK

chuyển thành

Price	OK?
5	Not OK
10	OK
11	OK
22	Not OK
31	OK
35	Not OK
40	Not OK
57	OK
68	OK
80	Not OK
140	Not OK

Ta tính gini cho từng mức Price nằm giữa 2 giá trị liên tiếp:

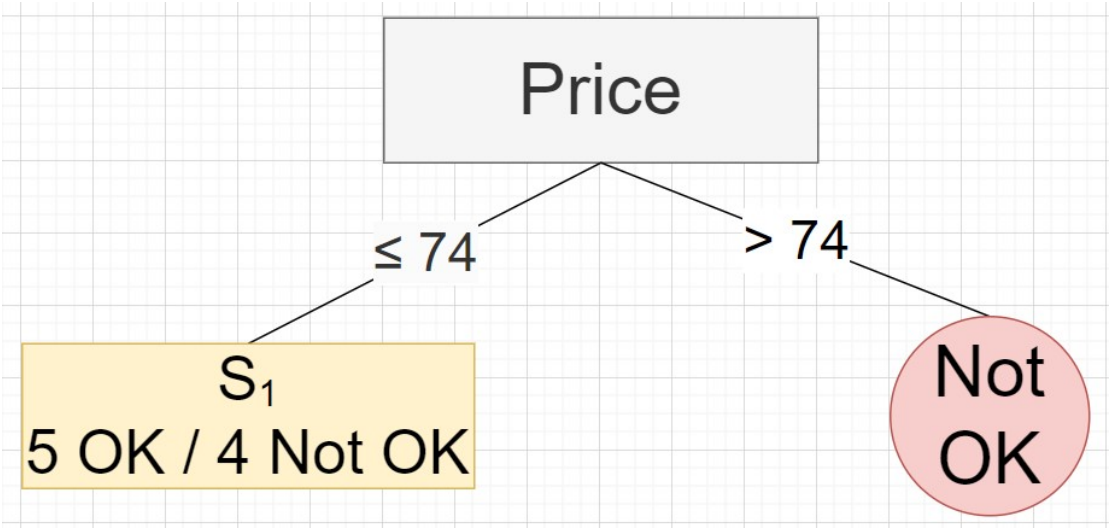
$$
\begin{aligned}
gini(Price \leq 7.5) &= 1 - p(OK)^2 - p(Not\ OK)^2 \\
&= 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2 \\
&= 0 \\
gini(Price > 7.5) &= 1 - p(OK)^2 - p(Not\ OK)^2 \\
&= 1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2 \\
&= 0.5 \\
gini(Price <> 7.5) &= \frac{1}{11} * 0 + \frac{10}{11} * 0.5 \\
&= 0.45
\end{aligned}
$$

Tính tương tự cho 10.5, 16.5, 26.5, 33, 37.5, 48.5, 62.5, 74, 110 ta có:

$$
\begin{aligned}
gini(Price <> 10.5) &= 0.49 \\
gini(Price <> 16.5) &= 0.46 \\
gini(Price <> 26.5) &= 0.49 \\
gini(Price <> 33) &= 0.46 \\
gini(Price <> 37.5) &= 0.49 \\
gini(Price <> 48.5) &= 0.49 \\
gini(Price <> 62.5) &= 0.48 \\
gini(Price <> 74) &= 0.4 \\
gini(Price <> 110) &= 0.45
\end{aligned}
$$

Ta thấy $gini(Price <> 74)$ nhỏ nhất nên thuộc tính Price có gini là 0.4

gini của Price bé nhất nên ta chọn Price ở node gốc:



Tính tương tự ta có cây: