CMPT 310 Assignment 4 Report

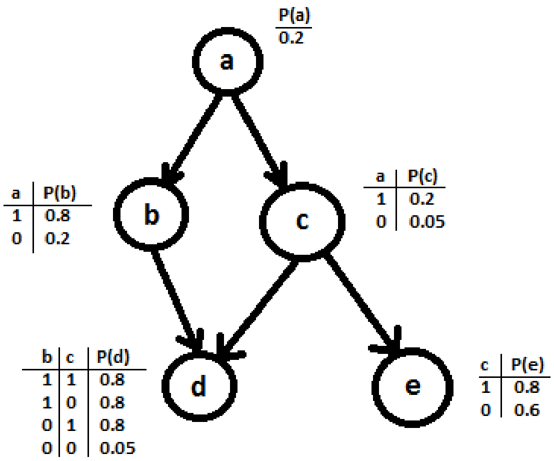
Joshua Campbell

301266191

Part 1:

a)

Belief Network:



P(a)
0.2

a	P(b)
1	0.8
0	0.2

a	P(c)
1	0.2
0	0.05

b	c	P(d)
1	1	0.8
1	0	0.8
0	1	0.8
0	0	0.05

c	P(e)
1	0.8
0	0.6

a = meta-static cancer
b = increased total serum calcium
c = brain tumor
d = occasional coma
e = severe headaches

b) Since each variable is conditionally independent of its non-descendants given the value of its parents, an example of an implicit independence assumption would be that node e is conditionally independent of d given the value of c. For example, P(e|c,d) = P(e|c).

c)

$$\boldsymbol{Formula\ for\ P(\neg B|A) = \frac{P(A \cap \neg B)}{P(A)} = \frac{P(A) - P(A \cap B)}{P(A)} = 1 - \frac{P(A \cap B)}{P(A)} = 1 - P(B|A)}$$

$$P(a, b, c, \neg d, e) = P(a)P(b|a)P(c|a)P(\neg d|b, c)P(e|c) = (0.2) * (0.8) * (0.2) * (1 - 0.8) * (0.8) = 0.00512$$

$$P(\neg a, b, c, \neg d, e) = P(\neg a)P(b|\neg a)P(c|\neg a)P(\neg d|b, c)P(e|c) = (1 - 0.2) * (0.2) * (0.05) * (1 - 0.8) * (0.8) = 0.00128$$

$$P(a, \neg b, c, \neg d, e) = P(a)P(\neg b|a)P(c|a)P(\neg d|\neg b, c)P(e|c) = (0.2) * (1 - 0.8) * (0.2) * (1 - 0.8) * (0.8) = 0.00128$$

$$P(a, b, \neg c, \neg d, e) = P(a)P(b|a)P(\neg c|a)P(\neg d|b, \neg c)P(e|\neg c) = (0.2) * (0.8) * (1 - 0.8) * (1 - 0.8) * (0.6) = 0.00384$$

$$\begin{aligned} P(\neg a, \neg b, c, \neg d, e) &= P(\neg a)P(\neg b|\neg a)P(c|\neg a)P(\neg d|\neg b, c)P(e|c) \\ &= (1-0.2)*(1-0.2)*(0.05)*(1-0.8)*(0.8) = 0.0000512 \end{aligned}$$

$$\begin{aligned} P(\neg a, b, \neg c, \neg d, e) &= P(\neg a)P(b|\neg a)P(\neg c|\neg a)P(\neg d|b, \neg c)P(e|\neg c) \\ &= (1-0.2)*(0.2)*(1-0.05)*(1-0.8)*(0.6) = 0.01824 \end{aligned}$$

$$\begin{aligned} P(a, \neg b, \neg c, \neg d, e) &= P(a)P(\neg b|a)P(\neg c|a)P(\neg d|\neg b, \neg c)P(e|\neg c) \\ &= (0.2)*(1-0.8)*(1-0.2)*(1-0.05)*(0.6) = 0.01824 \end{aligned}$$

$$\begin{aligned} P(a\neg, \neg b, \neg c, \neg d, e) &= P(\neg a)P(\neg b|\neg a)P(\neg c|\neg a)P(\neg d|\neg b, \neg c)P(e|\neg c) \\ &= (1-0.2)*(1-0.2)*(1-0.05)*(1-0.05)*(0.6) = 0.34656 \end{aligned}$$

d)

In order to determine if the patient is more likely to have meta-static cancer given that they are not experiencing occasional comas, but they are experiencing severe headaches, we can use Bayes' Rule and inference by enumeration to perform diagnostic analysis.

$$P(a|\neg d, e) = \frac{P(a, \neg d, e)}{P(\neg d, e)} = \alpha P(a, \neg d, e), \alpha = \frac{1}{P(\neg d, e)}$$

$$\alpha P(a, \neg d, e) = \alpha \sum_b \sum_c P(a, b, c, \neg d, e) = \alpha \sum_b \sum_c P(a)P(b|a)P(c|a)P(\neg d|b, c)P(e|c)$$

$$\begin{aligned} &= \alpha\big(P(a)P(b|a)P(c|a)P(\neg d|b, c)P(e|c) + P(a)P(\neg b|a)P(c|a)P(\neg d|\neg b, c)P(e|c) \\ &\quad + P(a)P(b|a)P(\neg c|a)P(\neg d|b, \neg c)P(e|\neg c) + P(a)P(\neg b|a)P(\neg c|a)P(\neg d|\neg b, \neg c)P(e|\neg c)\big) \\ &= \alpha(0.00512 + 0.00128 + 0.00384 + 0.01824) = \alpha(0.02848) \end{aligned}$$

$$\alpha = \frac{1}{P(\neg d, e)};\ P(\neg d, e) = \sum_a \sum_b \sum_c P(a, b, c, \neg d, e)$$

$$\begin{aligned} &= P(a)P(b|a)P(c|a)P(\neg d|b, c)P(e|c) + P(a)P(\neg b|a)P(c|a)P(\neg d|\neg b, c)P(e|c) \\ &\quad + P(a)P(b|a)P(\neg c|a)P(\neg d|b, \neg c)P(e|\neg c) + P(a)P(\neg b|a)P(\neg c|a)P(\neg d|\neg b, \neg c)P(e|\neg c) \\ &\quad + P(\neg a)P(b|\neg a)P(c|\neg a)P(\neg d|b, c)P(e|c) + P(\neg a)P(\neg b|\neg a)P(c|\neg a)P(\neg d|\neg b, c)P(e|c) \\ &\quad + P(\neg a)P(b|\neg a)P(\neg c|\neg a)P(\neg d|b, \neg c)P(e|\neg c) \\ &\quad + P(\neg a)P(\neg b|\neg a)P(\neg c|\neg a)P(\neg d|\neg b, \neg c)P(e|\neg c) \\ &= 0.00512 + 0.00128 + 0.00384 + 0.01824 + 0.00128 + 0.0000512 + 0.01824 + 0.34656 \\ &= 0.3946112 \end{aligned}$$

$$\therefore \alpha = \frac{1}{0.3946112}$$

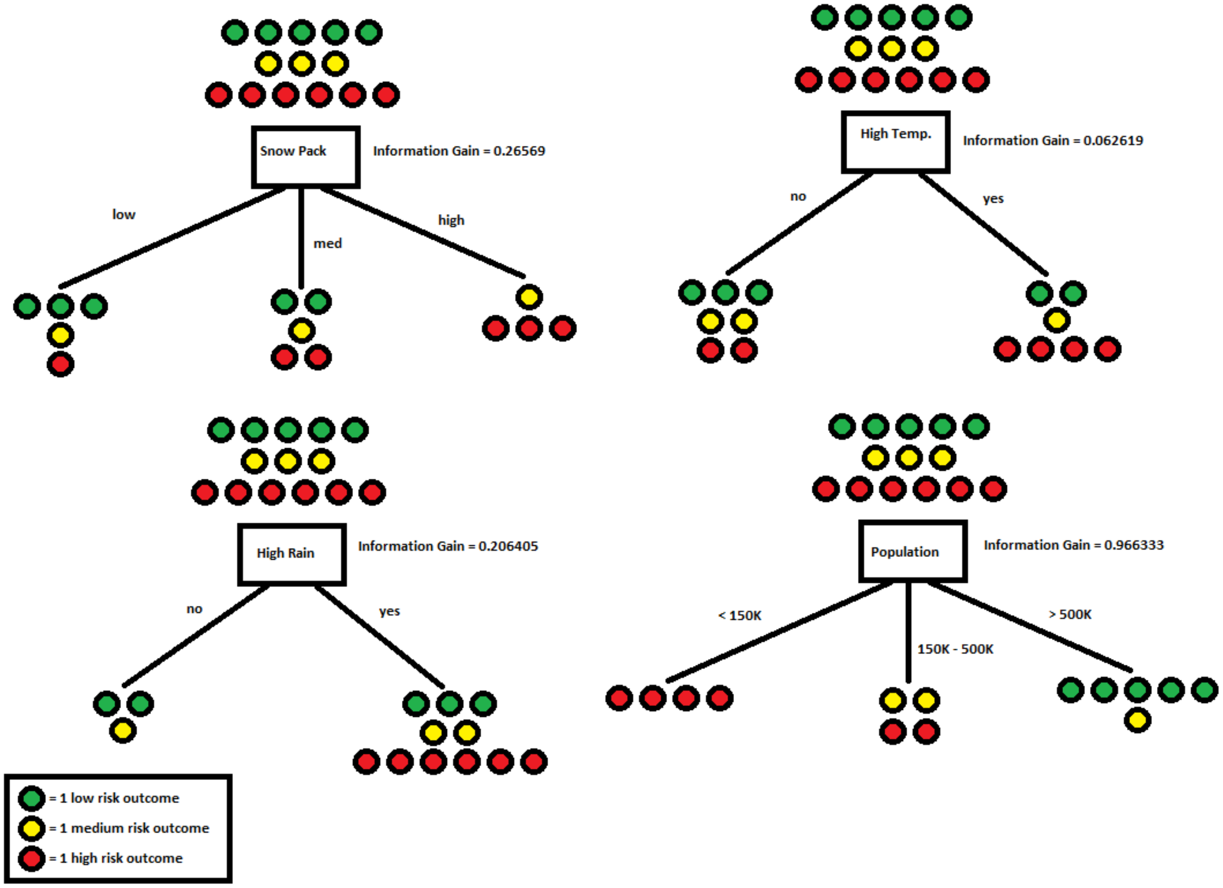
$$\therefore P(a|\neg d, e) = \frac{0.02848}{0.3946112} = 0.07217 < P(a) = 0.2$$

Since $P(a|\neg d, e) = 0.07217 < 0.2 = P(a)$ we are less inclined to believe that the patient has cancer given the available information.

Part 2:

Initial Calculations (L = number of Low Risk of flooding outcomes, M = number of Medium Risk of flooding outcomes, H = number of High Risk of flooding outcomes):

Total Outcomes: 5 Low risks of flooding, 3 Medium risks of flooding, 6 High risks of flooding



Functions Used:

$$Entropy = I\left(\frac{L}{L+M+H}, \frac{M}{L+M+H}, \frac{H}{L+M+H}\right) = -\frac{L}{L+M+H}\log_2\left(\frac{L}{L+M+H}\right) - \frac{M}{L+M+H}\log_2\left(\frac{M}{L+M+H}\right) - \frac{H}{L+M+H}\log_2\left(\frac{H}{L+M+H}\right)$$

$$Remainder(attribute) = \frac{\sum_{i=1}^{n\,(branches\ of\ attribute)}(L_i + M_i + H_i) * I(\frac{L_i}{L_i + M_i + H_i}, \frac{M_i}{L_i + M_i + H_i}, \frac{H_i}{L_i + M_i + H_i})}{L + M + H}$$

$$Information\ Gain = IG(attribute) = I\left(\frac{L}{L + M + H}, \frac{M}{L + M + H}, \frac{H}{L + M + H}\right) - Remainder(attribute)$$

Node 1:

Population:

$$entropy < 150K\ branch\colon\ -\frac{4}{4}\log_2\left(\frac{4}{4}\right)-\frac{0}{4}-\frac{0}{4}=0$$

$$entropy\ 150K-500K\ branch\colon\ -\frac{0}{4}-\frac{2}{4}\log_2\left(\frac{2}{4}\right)-\frac{2}{4}\log_2\left(\frac{2}{4}\right)=1$$

$$entropy\ >500K\ branch\colon\ -\frac{5}{6}\log_2\left(\frac{5}{6}\right)-\frac{1}{6}\log_2\left(\frac{1}{6}\right)-\frac{0}{6}=0.65$$

The above branch entropies are used to calculate remainder:

$Remainder(Pop)=\frac{4}{14}\left(-\frac{4}{4}\log_2\left(\frac{4}{4}\right)\right)+\frac{4}{14}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right)-\frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)+\frac{6}{14}\left(-\frac{5}{6}\log_2\left(\frac{5}{6}\right)-\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right)=\frac{4}{14}(0)+\frac{4}{14}(1)+\frac{6}{14}(0.65)=\ 0.564286$

$$IG(Population)=-\frac{5}{14}\log_2(\frac{5}{14})-\frac{3}{14}\log_2(\frac{3}{14})-\frac{6}{14}\log_2(\frac{6}{14})-Remainder(Population)=\ 0.966333$$

Snow Pack:

$Remainder(Snow)=\frac{5}{14}\left(-\frac{3}{5}\log_2\left(\frac{3}{5}\right)-\frac{1}{5}\log_2\left(\frac{1}{5}\right)-\frac{1}{5}\log_2\left(\frac{1}{5}\right)\right)+\frac{5}{14}\left(-\frac{2}{5}\log_2\left(\frac{2}{5}\right)-\frac{1}{5}\log_2\left(\frac{1}{5}\right)-\frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)+\frac{4}{14}\left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right)-\frac{3}{4}\log_2\left(\frac{3}{4}\right)\right)=\ 1.264929$

$$IG(Snow)=-\frac{5}{14}\log_2(\frac{5}{14})-\frac{3}{14}\log_2(\frac{3}{14})-\frac{6}{14}\log_2(\frac{6}{14})-Remainder(Snow)=\ 0.26569$$

High Temp:

$Remainder(\mathrm{Temp})=\frac{7}{14}\left(-\frac{3}{7}\log_2\left(\frac{3}{7}\right)-\frac{2}{7}\log_2\left(\frac{2}{7}\right)-\frac{2}{7}\log_2\left(\frac{2}{7}\right)\right)+\frac{7}{14}\left(-\frac{2}{7}\log_2\left(\frac{2}{7}\right)-\frac{1}{7}\log_2\left(\frac{1}{7}\right)-\frac{4}{7}\log_2\left(\frac{4}{7}\right)\right)=\ 1.468$

$$IG(\mathrm{Temp})=-\frac{5}{14}\log_2(\frac{5}{14})-\frac{3}{14}\log_2(\frac{3}{14})-\frac{6}{14}\log_2(\frac{6}{14})-Remainder(\mathrm{Temp})=\ 0.062619$$

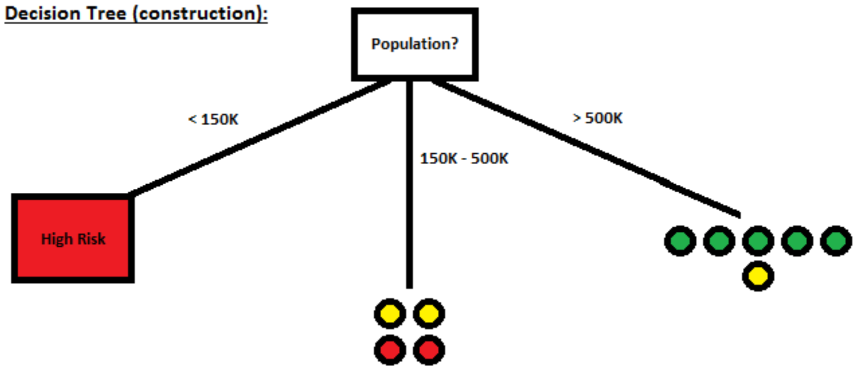
High Rain:

$Remainder(Rain)=\frac{3}{14}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right)-\frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)+\frac{11}{14}\left(-\frac{3}{11}\log_2\left(\frac{3}{11}\right)-\frac{2}{11}\log_2\left(\frac{2}{11}\right)-\frac{6}{11}\log_2\left(\frac{6}{11}\right)\right)=\ 0.564286$

$$IG(Rain)=-\frac{5}{14}\log_2(\frac{5}{14})-\frac{3}{14}\log_2(\frac{3}{14})-\frac{6}{14}\log_2(\frac{6}{14})-Remainder(Rain)=\ 0.206405$$

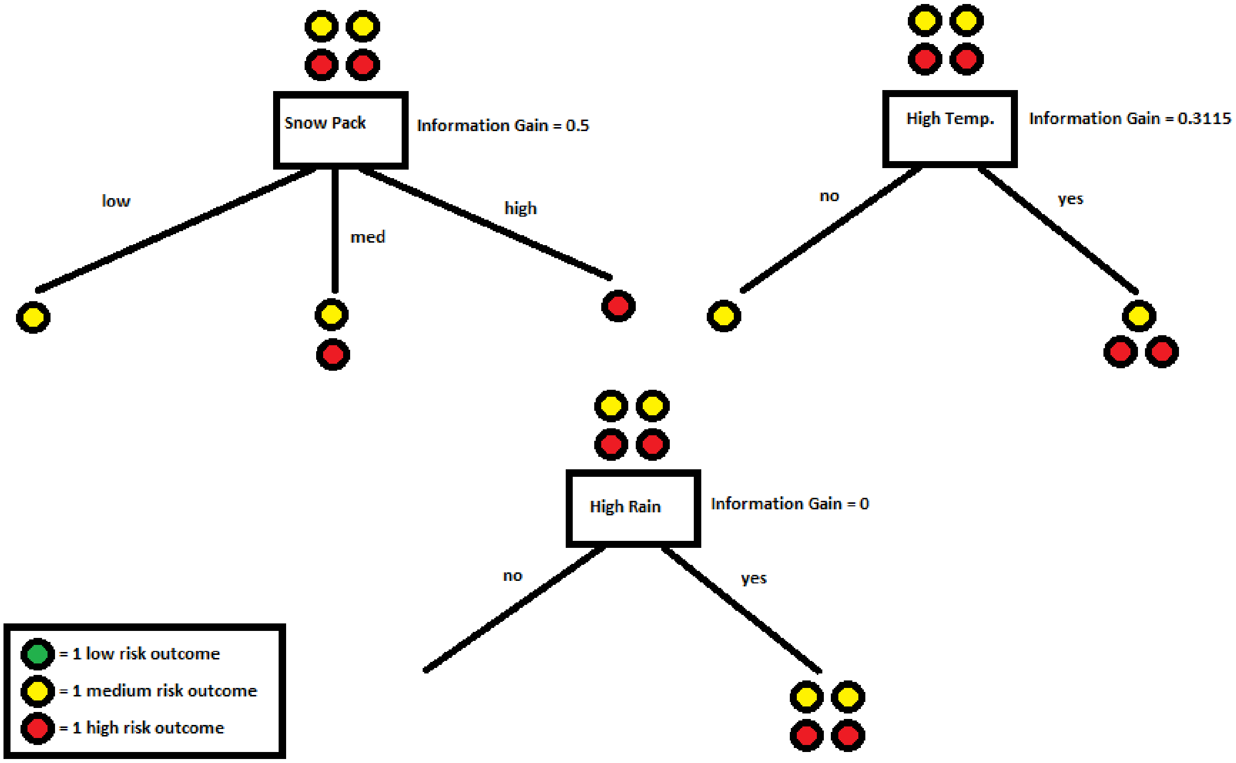
Therefore the first node of the Decision Tree is Population since it has the most Information Gain.

Since Populations with <150K residents are always at high risk of flooding, this will be the first leaf on the tree (High Risk).



Node 2:

For populations in the range of 150K-500K, the outcomes are 2 Medium risks of flooding and 2 High risks of flooding.



The Information Gain for attributes based on the 150K-500K branch are:

Snow Pack:

$$Remainder(Snow) = \frac{1}{4}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) + \frac{2}{4}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \frac{1}{4}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) = 0.5$$

$$IG(Snow) = -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4}) - Remainder(Snow) = 0.5$$

High Temp:

$$Remainder(\text{Temp}) = \frac{1}{4}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) + \frac{3}{4}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) = 0.6885$$

$$IG(\text{Temp}) = -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4}) - Remainder(\text{Temp}) = 0.3115$$

High Rain:

$$Remainder(Rain) = \frac{0}{4}(0) + \frac{4}{4}\left(-\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4})\right) = 1$$

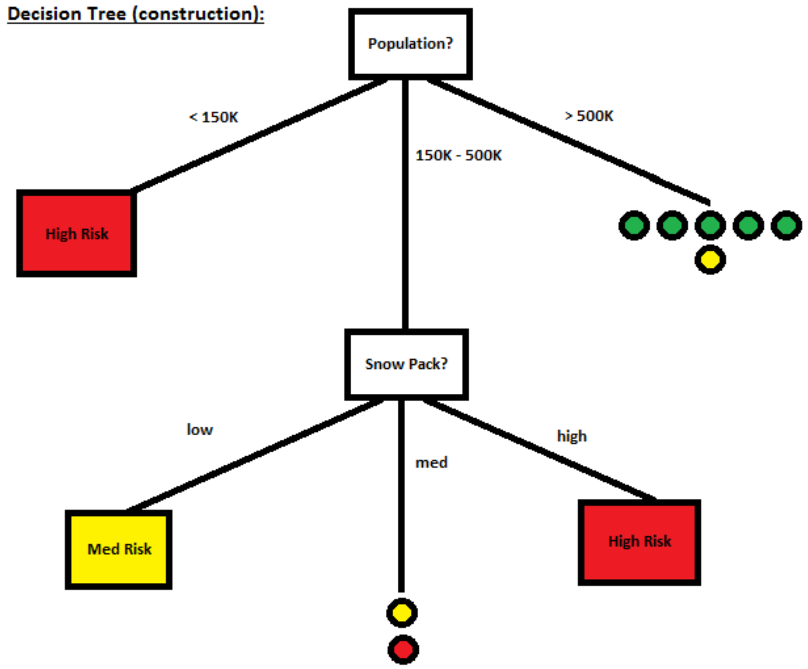
$$IG(Rain) = -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4}) - Remainder(Rain) = 0$$

Note: The following branches will use the same calculation (an attribute (child) node's number of input outcomes (low risk, medium risk, high risk) is equivalent to its parent branch's outcome outputs (low risk, medium risk, high risk)); in the interest of space, the summaries of each attribute's Information Gain will be included.

- Snow Pack: IG(Snow) = I(Snow) – Remainder(Snow) = 1 – 0.5 = 0.5
- High Temp: IG(Temp) = I(Temp) – Remainder(Temp) = 1 – 0.6885 = 0.3115
- High Rain: IG(Rain) = I(Rain) – Remainder(Rain) = 1 – 1 = 0

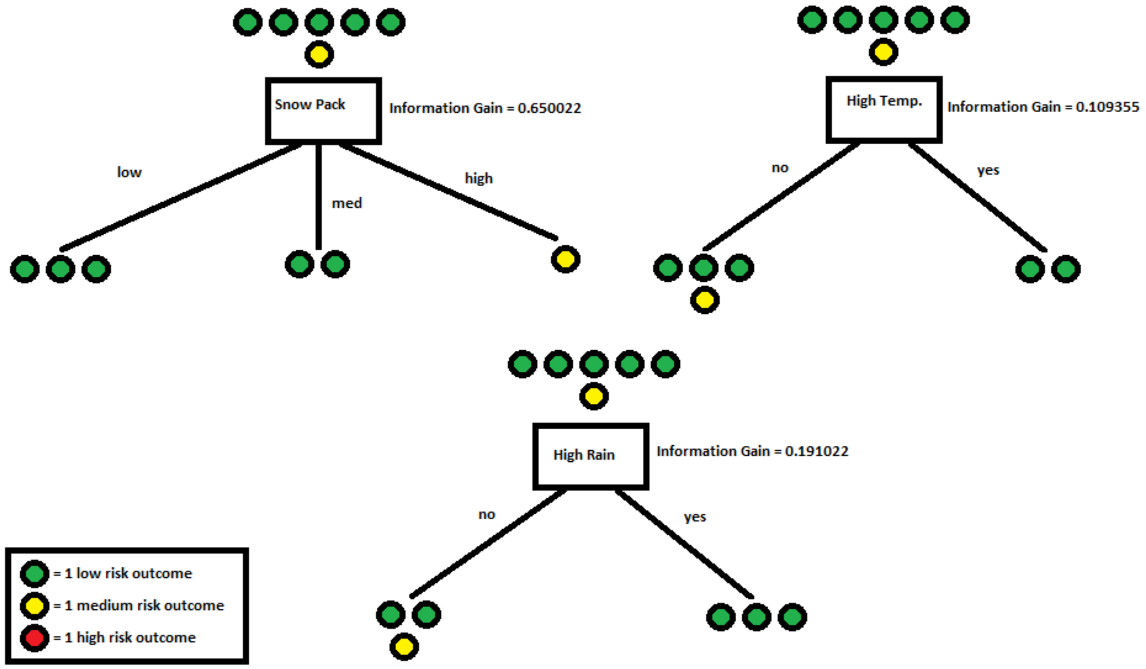
Therefore, the child node of the 150K-500K branch is Snow Pack since it has the most information gain.

Since the Low branch of Snow Pack only has one outcome, it will become a leaf node (Medium Risk). Since the High branch of Snow Pack only has one outcome, it will become a leaf node (High Risk).



Node 3:

For populations > 500K, the outcomes are 5 Low risks of flooding and 1 Medium risk of flooding.



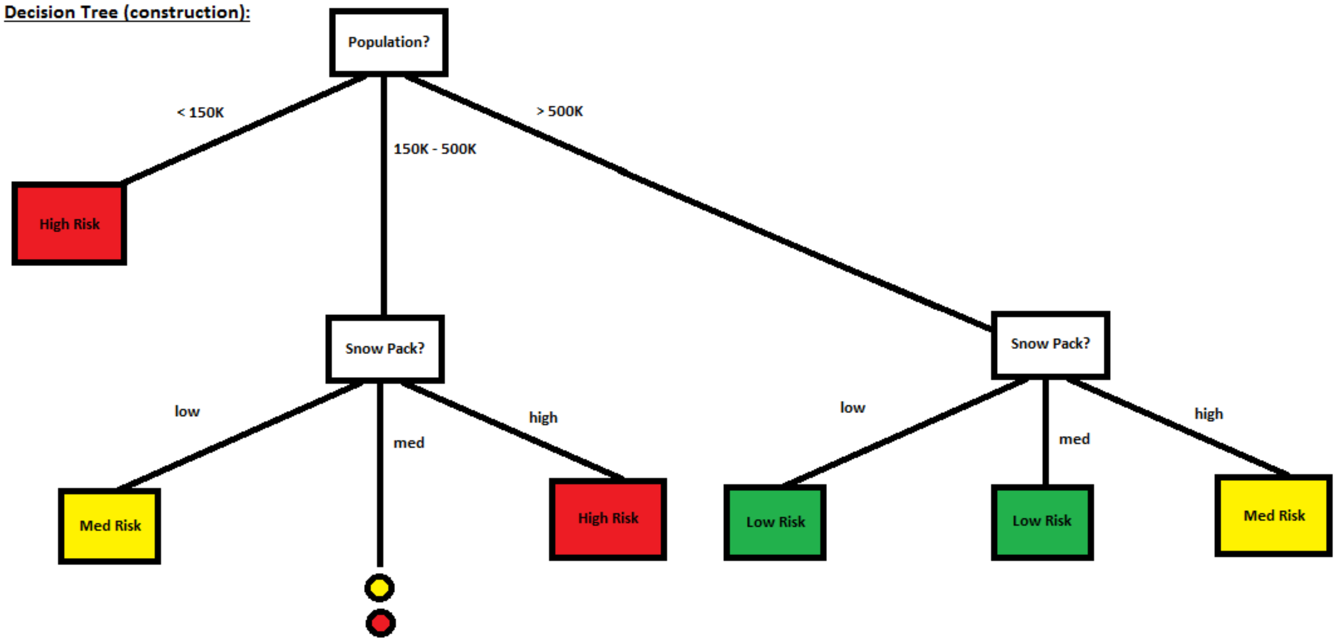
The Information Gain for attributes based on the > 500K branch are:

- Snow Pack: IG(Snow) = I(Snow) – Remainder(Snow) = 0.650022 – 0 = 0.650022
- High Temp: IG(Temp) = I(Temp) – Remainder(Temp) = 0.650022 – 0.540667 = 0.109355
- High Rain: IG(Rain) = I(Rain) – Remainder(Rain) = 0.650022 – 0.459 = 0.191022

Therefore, the child node of the > 500K branch is Snow Pack since it has the most information gain.

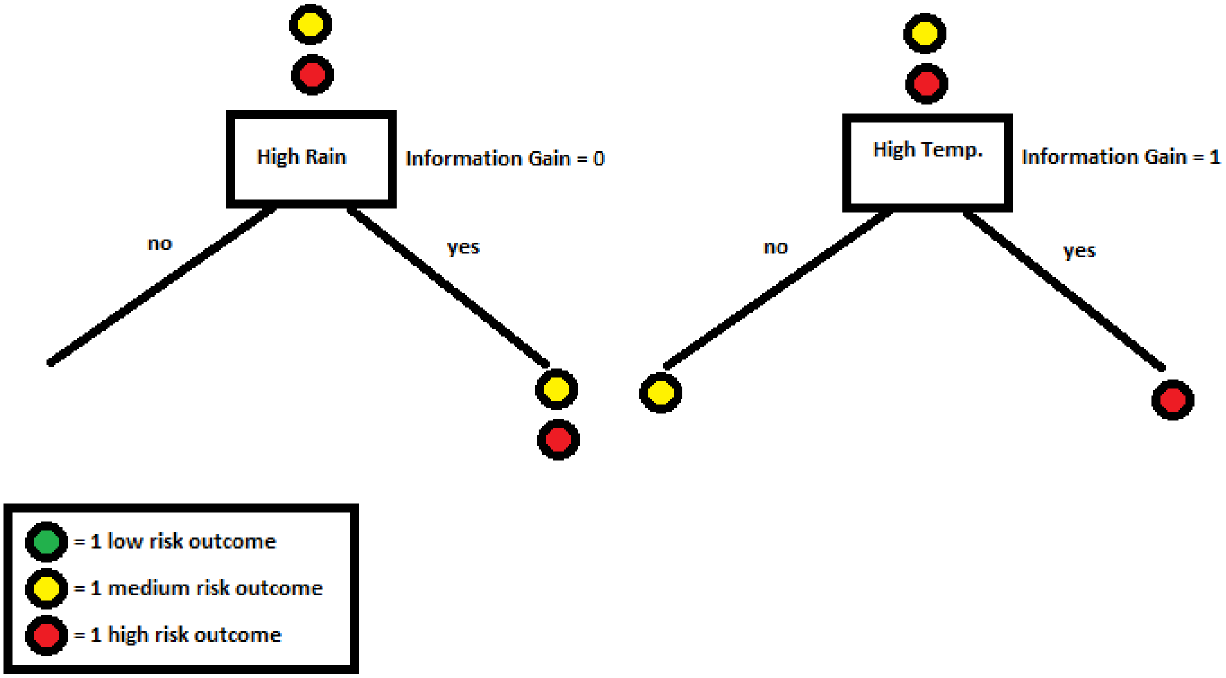
Since the Low branch of Snow Pack only has one outcome, it will become a leaf node (Low Risk). Since the Medium branch of Snow Pack only has one outcome, it will become a leaf node (Low Risk). Since the High branch of Snow Pack only has one outcome, it will become a leaf node (Medium Risk).

Decision Tree (construction):



Node 4:

For populations in the range of 150K-500K with a medium snow pack, the outcomes are 1 Medium risk of flooding and 1 High risk of flooding.



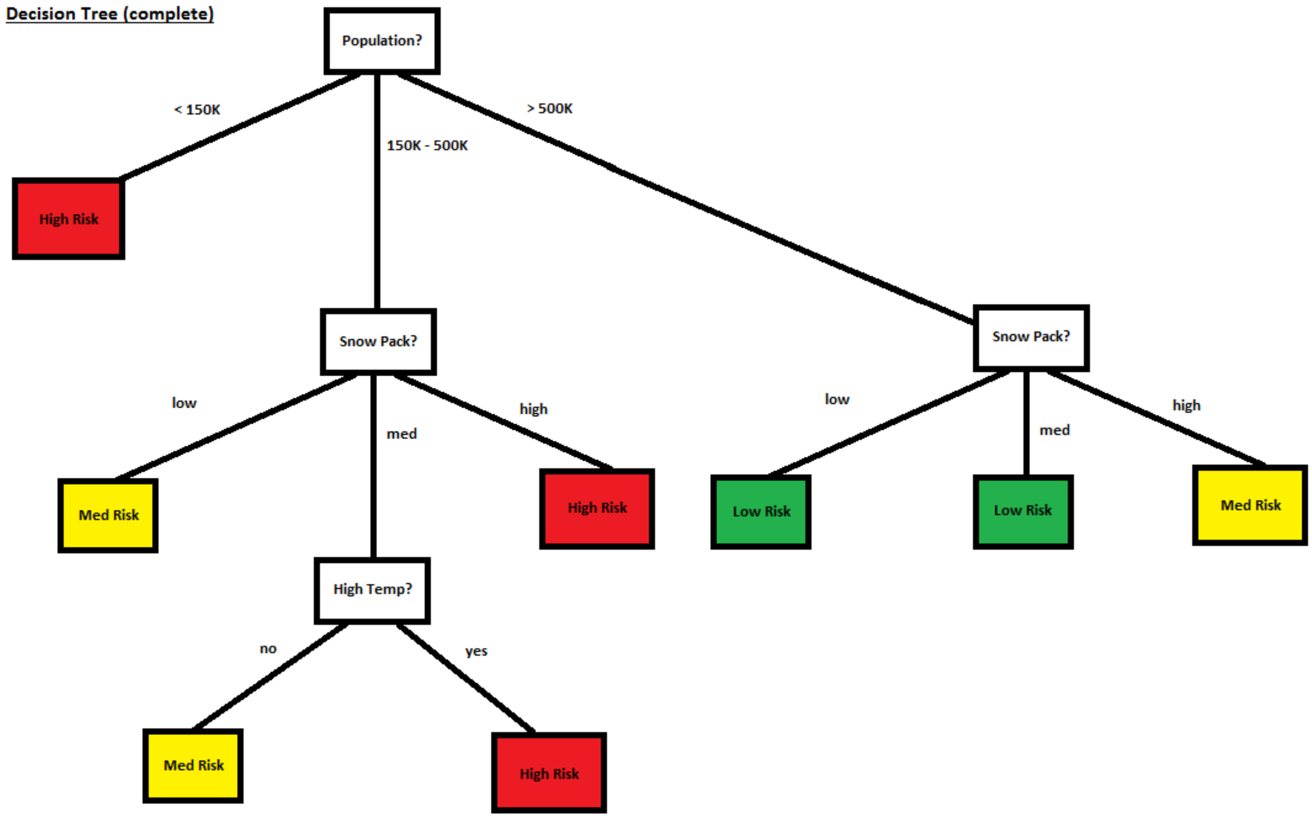
The Information Gain for attributes based on this branch path are:

- High Temp: IG(Temp) = I(Temp) – Remainder(Temp) = 1 – 0 = 1
- High Rain: IG(Rain) = I(Rain) – Remainder(Rain) = 1 – 1 = 0

Therefore, the child node for this branch path is High Temp since it has the most information gain.

Since the No branch of High Temp only has one outcome, it will become a leaf node (Medium Risk). Since the High branch of High Temp only has one outcome, it will become a leaf node (High Risk).

Since all branches now lead to leaves, the complete decision tree is:


Decision Tree (complete)
Population?
< 150K
150K - 500K
> 500K
High Risk
Snow Pack?
low
med
high
Med Risk
High Risk
High Temp?
no
yes
Med Risk
High Risk
Snow Pack?
low
med
high
Low Risk
Low Risk
Med Risk