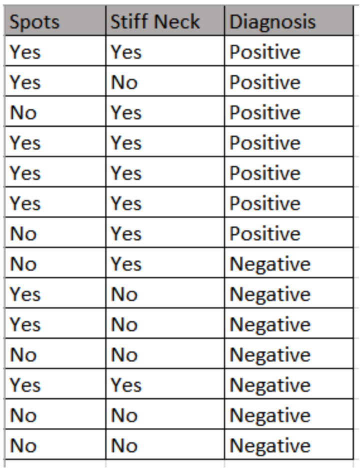
Gini Index Decision Tree

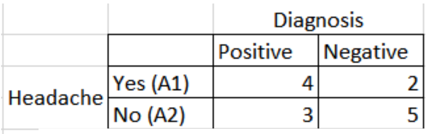


Spots	Stiff Neck	Diagnosis
Yes	Yes	Positive
Yes	No	Positive
No	Yes	Positive
Yes	Yes	Positive
Yes	Yes	Positive
Yes	Yes	Positive
No	Yes	Positive
No	Yes	Negative
Yes	No	Negative
Yes	No	Negative
No	No	Negative
Yes	Yes	Negative
No	No	Negative
No	No	Negative

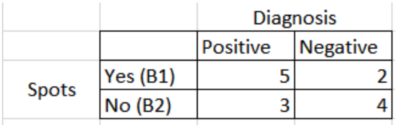
$$\text{Impurity} = 1 - \sum_{i=1}^{c} P(x=i)^2$$

All calculations to 2dp

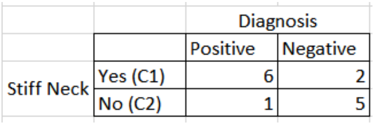
$$\text{Parent Impurity} = 1 - \left(\left(\frac{7}{14}\right)^2 + \left(\frac{7}{14}\right)^2\right) = 0.5$$



		Diagnosis	
		Positive	Negative
Headache	Yes (A1)	4	2
	No (A2)	3	5



		Diagnosis	
		Positive	Negative
Spots	Yes (B1)	5	2
	No (B2)	3	4



		Diagnosis	
		Positive	Negative
Stiff Neck	Yes (C1)	6	2
	No (C2)	1	5

$$A1 = 1 - \left(\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right) = 0.44$$

$$A2 = 1 - \left(\left(\frac{3}{8}\right)^2 + \left(\frac{5}{8}\right)^2\right) = 0.47$$

$$\text{Gini}(X, \text{Headache}) = 0.5 - \frac{6}{14}(0.44) - \frac{8}{14}(0.47) = \underline{\underline{0.04}}$$

$$B1 = 1 - \left(\left(\frac{5}{7}\right)^2 + \left(\frac{2}{7}\right)^2\right) = 0.41$$

$$B2 = 1 - \left(\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right) = 0.49$$

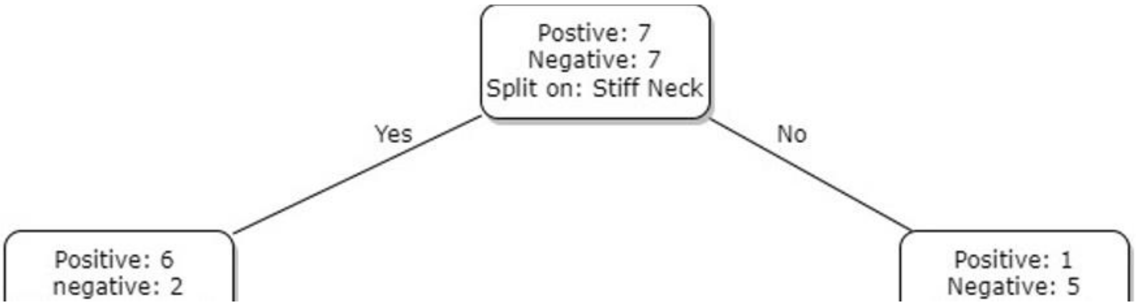
$$\text{Gini}(X, \text{Spots}) = 0.5 - \frac{7}{14}(0.41) - \frac{7}{14}(0.49)\ \underline{\underline{0.05}}$$

$$C1 = 1 - \left(\left(\frac{6}{8}\right)^2 + \left(\frac{2}{8}\right)^2\right) = 0.38$$

$$C2 = 1\ \left(\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right) = 0.27$$

$$\text{Gini}(X, \text{StiffNeck}) = 0.5 - \frac{8}{14}(0.38) - \frac{6}{14}(0.27) = \underline{\underline{0.17}}$$

Stiff Neck has highest value so split on that.



Left Subtree

Headache	Spots	Diagnosis
Yes	Yes	Positive
Yes	No	Positive
No	Yes	Positive
Yes	Yes	Positive
No	Yes	Positive
Yes	No	Positive
No	No	Negative
No	Yes	Negative

$$\text{Impurity} = 1-\left(\left(\tfrac{6}{8}\right)^2 + \left(\tfrac{2}{8}\right)^2 = 0.38$$

		Diagnosis	
		Positive	Negative
Headache	Yes (D1)	4	0
	No (D2)	2	2

		Diagnosis	
		Positive	Negative
Spots	Yes (E1)	4	1
	No (E2)	2	1

$$D1 = 1 - \left(\left(\tfrac{4}{4}\right)^2 + \left(\tfrac{0}{4}\right)^2\right) = 0$$

$$D2 = 1 - \left(\left(\tfrac{2}{4}\right)^2 + \left(\tfrac{2}{4}\right)^2\right) = 0.5$$

$$\text{Gini}\left(X, \text{Headache}\right) = 0.38 - \tfrac{4}{8}(0) - \tfrac{4}{8}(0.5) = \underline{\underline{0.13}}$$

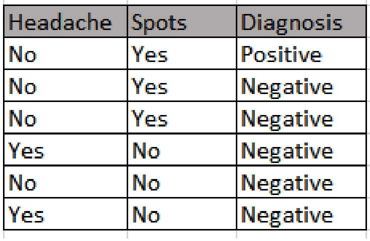
$$E1 = 1 - \left(\left(\tfrac{4}{5}\right)^2 + \left(\tfrac{1}{5}\right)^2\right) = 0.32$$

$$E2 = 1 - \left(\left(\tfrac{2}{3}\right)^2 + \left(\tfrac{1}{3}\right)^2\right) = 0.44$$

$$\text{Gini}\left(X, \text{Spots}\right) = 0.38 - \tfrac{5}{8}(0.32) - \tfrac{3}{8}(0.44) = \underline{\underline{0.015}}$$

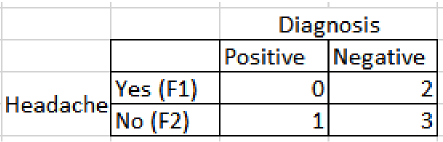
Headache is a higher value so split on that.

Right Subtree

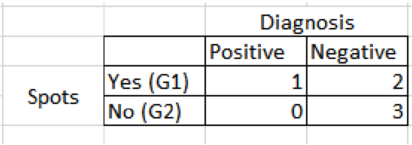


Headache	Spots	Diagnosis
No	Yes	Positive
No	Yes	Negative
No	Yes	Negative
Yes	No	Negative
No	No	Negative
Yes	No	Negative

Impurity $= 1-\left(\left(\frac{1}{6}\right)^2+\left(\frac{5}{6}\right)^2\right) = 0.28$



		Diagnosis	
		Positive	Negative
Headache	Yes (F1)	0	2
	No (F2)	1	3



		Diagnosis	
		Positive	Negative
Spots	Yes (G1)	1	2
	No (G2)	0	3

$$F1 = 1-\left(\left(\frac{0}{2}\right)^2+\left(\frac{2}{2}\right)^2\right)=0$$

$$f2 = 1-\left(\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right)= 0.38$$

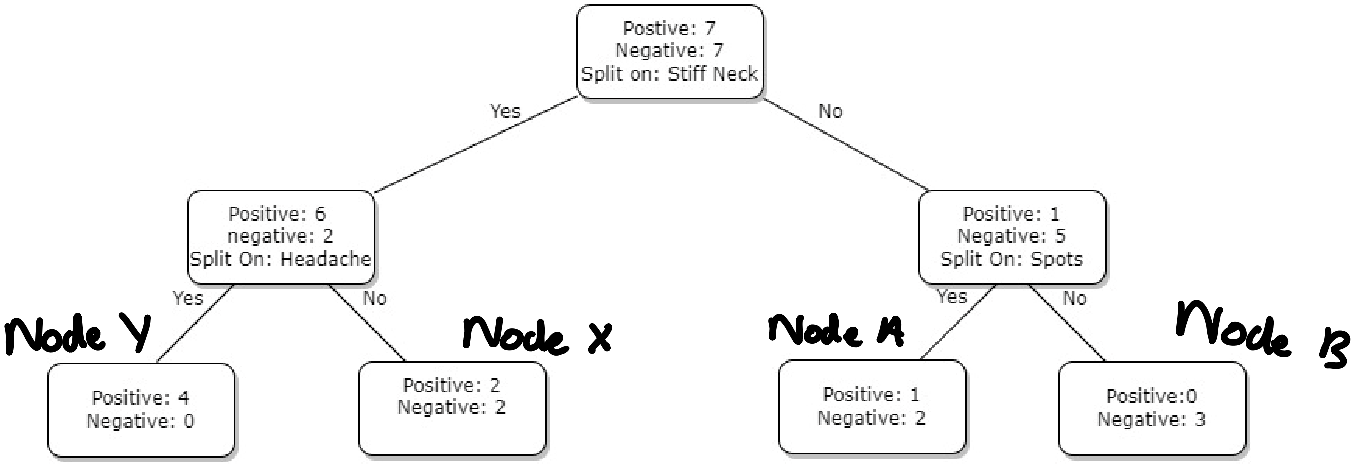
$$\text{Gini}\left(X, \text{Headache}\right) = 0.28 - \frac{2}{6}(0) - \frac{4}{6}(0.38) = \underline{0.03}$$

$$G1 = 1-\left(\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2 = 0.44\right.$$

$$G2 = 1-\left(\left(\frac{0}{3}\right)^2+\left(\frac{3}{3}\right)^2\right)= 0$$

$$\text{Gini}\left(X, \text{Spots}\right) = 0.28 - \frac{3}{6}(0.44) - \frac{3}{6}(0) = \underline{0.06}$$

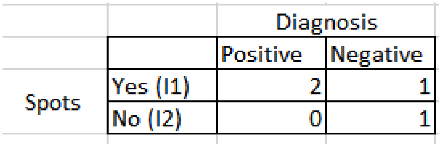
Spots is higher, so split on spots



Left Left Subtree (Node Y)

- Node is pure, so it no longer needs to be split

Left Right Subtree (Node X)



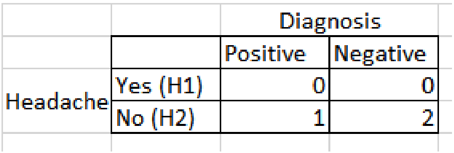
		Diagnosis	
		Positive	Negative
Spots	Yes (I1)	2	1
	No (I2)	0	1

Only spots left to split on, so split the node on spots

Right Right subtree (Node B)

- Node is pure, so it no longer needs to be split

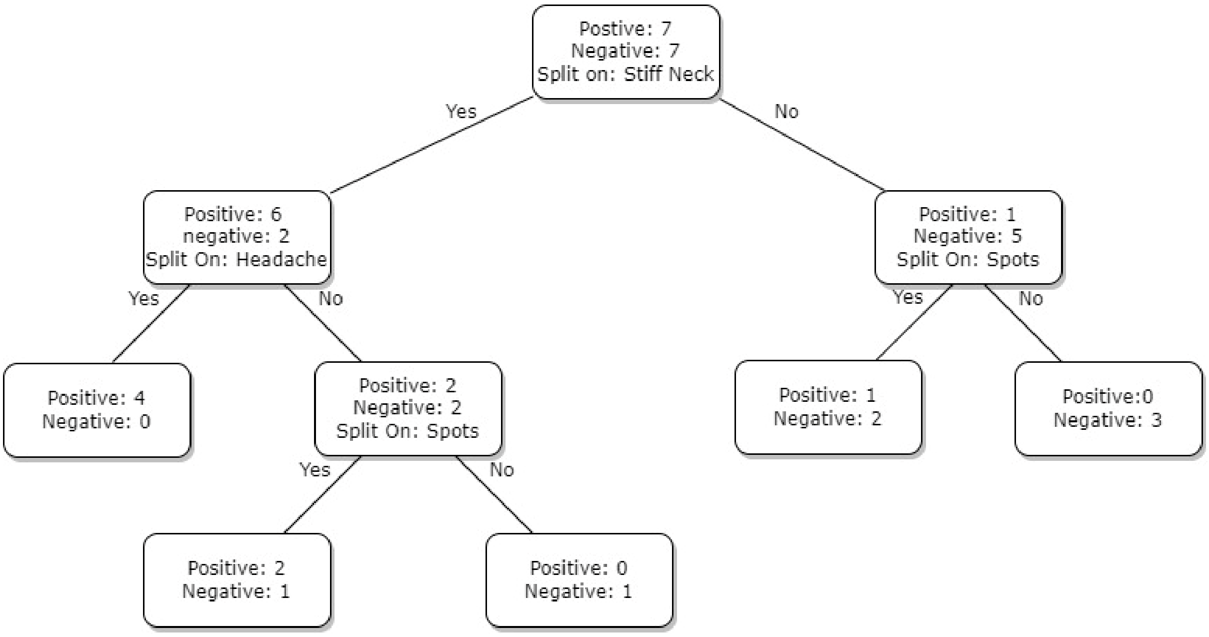
Right Left Subtree (Node A)



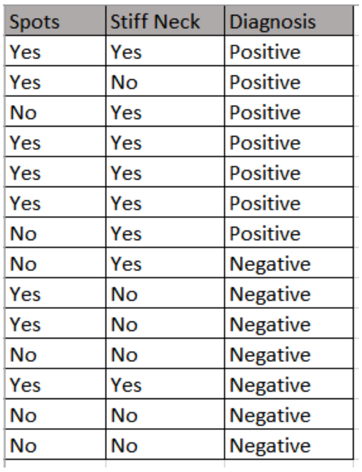
		Diagnosis	
		Positive	Negative
Headache	Yes (H1)	0	0
	No (H2)	1	2

Only headache left to split on, however splitting on this attribute makes no difference to data since all three cases of data will be on the no side, so this attribute does not need be branched on.

Final Gini Decision Tree



Chi Squared Decision Tree



Spots	Stiff Neck	Diagnosis
Yes	Yes	Positive
Yes	No	Positive
No	Yes	Positive
Yes	Yes	Positive
Yes	Yes	Positive
Yes	Yes	Positive
No	Yes	Positive
No	Yes	Negative
Yes	No	Negative
Yes	No	Negative
No	No	Negative
Yes	Yes	Negative
No	No	Negative
No	No	Negative

formula to get Chi Squared result:

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

values inside brackets mean (expected values)

Headache

	Positive	Negative	Total
Yes	$4(6 * (\frac{7}{14}) = 3)$	$2(6 * \left(\frac{7}{14}\right) = 3)$	6
No	$3(8 * \left(\frac{7}{14}\right) = 4)$	$5(8 * \left(\frac{7}{14}\right) = 4)$	8
Total	7	7	14
Probability	7/14	7/14	

$$\text{Headache } x^2 = \frac{(4-3)^2}{3} + \frac{(2-3)^2}{3} + \frac{(3-4)^2}{4} + \frac{(5-4)^2}{4}$$

$$= \underline{\underline{1.17}}$$

Spots

	Positive	Negative	Total
Yes	$5(7 * (\frac{8}{14}) =4)$	$2(7 * \left(\frac{6}{14}\right) = 3)$	7
No	$3(7 * \left(\frac{8}{14}\right) =4)$	$4(7 * \left(\frac{6}{14}\right) = 3)$	7
Total	8	6	14
Probability	8/14	6/14	

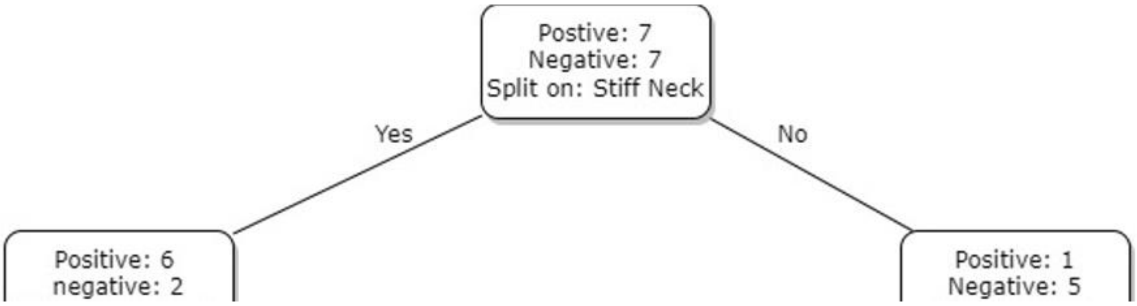
$$\text{Spots } x^2 = \frac{(5-4)^2}{4} + \frac{(2-3)^2}{3} + \frac{(3-4)^2}{4} + \frac{(4-3)^2}{3} = \underline{\underline{1.17}}$$

Stiff Neck

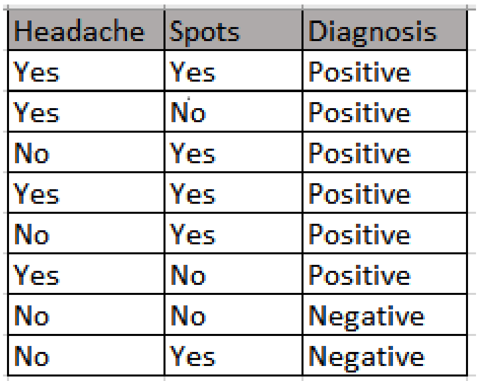
	Positive	Negative	Total
Yes	$6(8*(\frac{7}{14})=4)$	$2(8*\left(\frac{7}{14}\right)=4)$	8
No	$1(6*\left(\frac{7}{14}\right)=3)$	$5(6*\left(\frac{7}{14}\right)=3)$	6
Total	7	7	14
Probability	7/14	7/14	

Stiff neck $x^2 = \frac{(6-4)^2}{4} + \frac{(2-4)^2}{4} + \frac{(1-3)^2}{3} + \frac{(5-3)^2}{3} = \underline{\underline{4.67}}$

Split of Stiff neck sinces it's the highest.



Left Subtree



Headache	Spots	Diagnosis
Yes	Yes	Positive
Yes	No	Positive
No	Yes	Positive
Yes	Yes	Positive
No	Yes	Positive
Yes	No	Positive
No	No	Negative
No	Yes	Negative

Headache

	Positive	Negative	Total
Yes	$4(4 * (\frac{6}{8}) = 3)$	$0(4 * \left(\frac{2}{8}\right) = 1)$	4
No	$2(4 * \left(\frac{6}{8}\right) = 3)$	$2(4 * \left(\frac{2}{8}\right) = 1)$	4
Total	6	2	8
Probability	6/8	2/8	

$$\text{Headache } x^2 = \frac{(4-3)^2}{3} + \frac{(0-1)^2}{1} + \frac{(2-3)^2}{3} + \frac{(2-1)^2}{1} = \underline{\underline{2.67}}$$

Spots

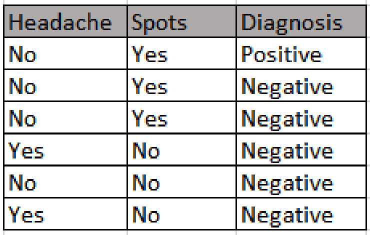
	Positive	Negative	Total
Yes	$4(5 * (\frac{6}{8}) = 3.75)$	$1(5 * \left(\frac{2}{8}\right) = 1.25)$	5
No	$2(3 * \left(\frac{6}{8}\right) = 2.25)$	$1(3 * \left(\frac{2}{8}\right) = 0.75)$	3
Total	6	2	8
Probability	6/8	2/8	

$$\text{Spots } x^2 = \frac{(4-3.75)^2}{3.75} + \frac{(1-1.25)^2}{1.25} + \frac{(2-2.25)^2}{2.25} + \frac{(1-0.75)^2}{0.75}$$

$$= \underline{\underline{0.17}}$$

headache is higher, split left subtree on it.

Right Subtree



Headache	Spots	Diagnosis
No	Yes	Positive
No	Yes	Negative
No	Yes	Negative
Yes	No	Negative
No	No	Negative
Yes	No	Negative

Headache

	Positive	Negative	Total
Yes	$0(2*(\frac{1}{6}) = 1/3)$	$2(2*\left(\frac{5}{6}\right) = 5/3)$	2
No	$1(4*\left(\frac{1}{6}\right) = 2/3)$	$3(4*\left(\frac{5}{6}\right) = 10/3)$	4
Total	1	5	6
Probability	1/6	5/6	

Headache $x^2 = \frac{(0-\frac{1}{3})^2}{\frac{1}{3}} + \frac{(2-\frac{5}{3})^2}{\frac{5}{3}} + \frac{(1-\frac{2}{3})^2}{\frac{2}{3}} + \frac{(3-\frac{10}{3})^2}{\frac{10}{3}}$

$= \underline{\underline{0.6}}$

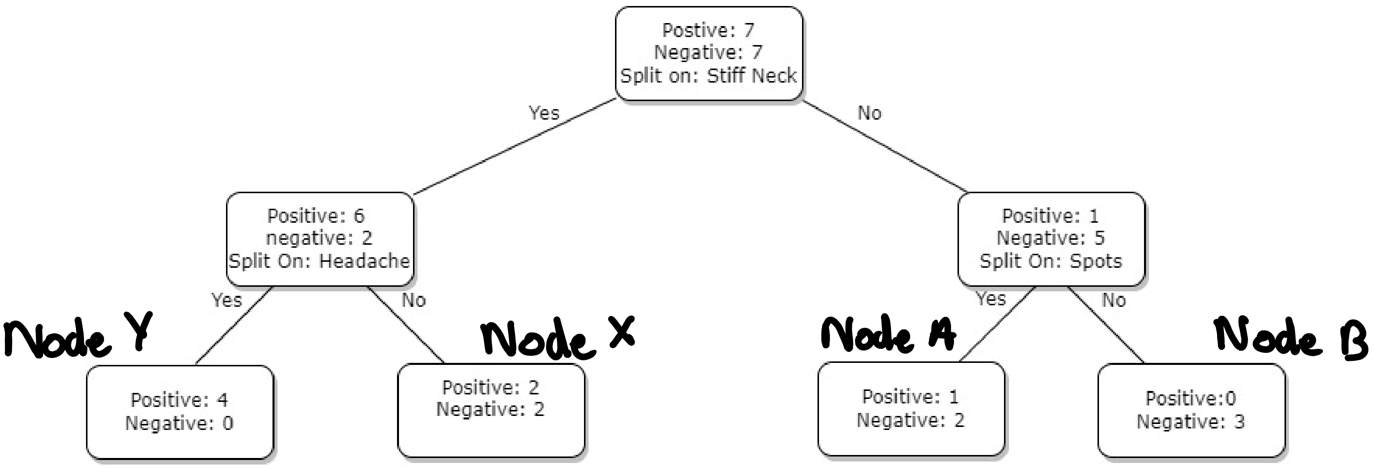
Spots

	Positive	Negative	Total
Yes	$1(3*(\frac{1}{6}) = 0.5)$	$2(3*\left(\frac{5}{6}\right) = 2.5)$	3
No	$0(3*\left(\frac{1}{6}\right) = 0.5)$	$3(3*\left(\frac{5}{6}\right) = 2.5)$	3
Total	1	5	6
Probability	1/6	5/6	

Spots $x^2 = \frac{(1-0.5)^2}{0.5} + \frac{(2-2.5)^2}{2.5} + \frac{(0-0.5)^2}{0.5} + \frac{(3-2.5)^2}{2.5}$

$= \underline{\underline{1.2}}$

Spots is greater, split right subtree on it.



Left Left Subtree (Node Y)

- Node is pure, so it no longer needs to be split

Left Right Subtree (Node X)

- Only spots left to split on, so split the node on spots

Spots

	Positive	Negative	Total
Yes	2	1	3
No	0	1	1
Total	2	2	4
Probability	2/4	2/4	

Right Right subtree (Node B)

- Node is pure, so it no longer needs to be split

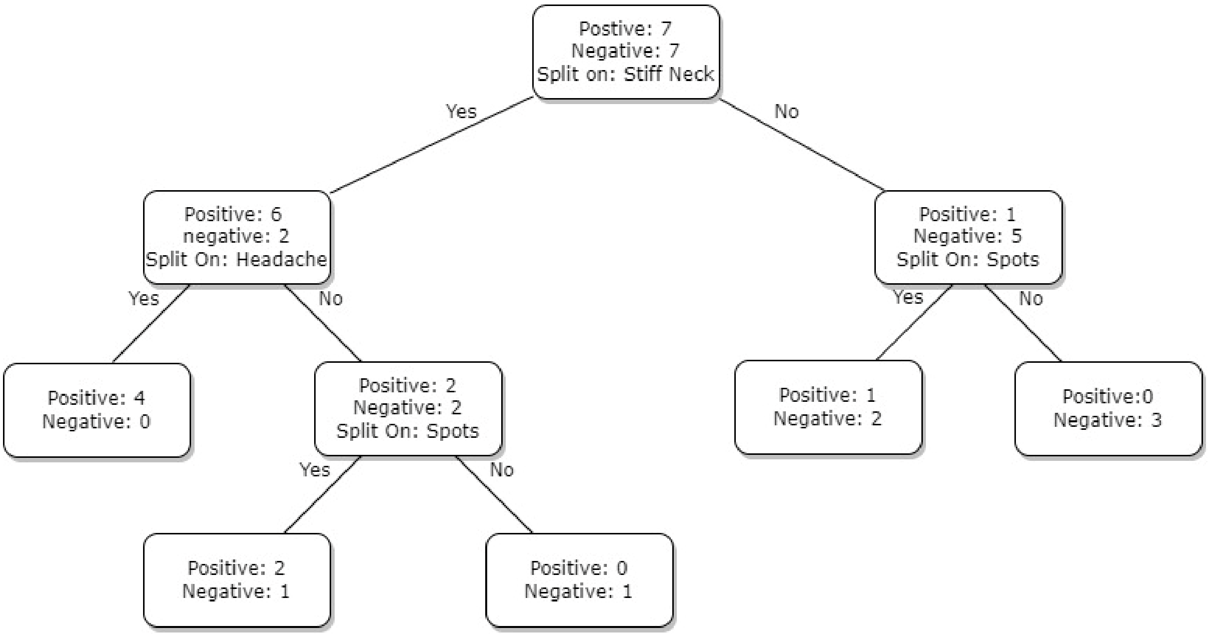
Right Left Subtree (Node A)

- Only left to split on headache, however splitting on this attribute makes no difference to data since all three cases of data will be on the no side, so this attribute does not need be branched on.

Headache

	Positive	Negative	Total
Yes	0	0	0
No	1	2	3
Total	1	2	3
Probability	1/3	2/3	

Final Chi squared decision tree


Postive: 7
Negative: 7
Split on: Stiff Neck
Yes
No
Positive: 6
negative: 2
Split On: Headache
Positive: 1
Negative: 5
Split On: Spots
Yes
No
Yes
No
Positive: 4
Negative: 0
Positive: 2
Negative: 2
Split On: Spots
Positive: 1
Negative: 2
Positive:0
Negative: 3
Yes
No
Positive: 2
Negative: 1
Positive: 0
Negative: 1