

Andrew Tran

CS1675

Assignment 7 Report

Due: 3/21/19

1a) I think restricted would be better because having too many nodes may lead to overfitting.

1b)

Max Num Splits	Min Parent Size	Min Leaf Size	Error
Unrestricted	Unrestricted	Unrestricted	0.2751
25	10	2	0.2576
2	10	2	0.2707
25	50	2	0.2314
25	10	10	0.2489
50	2	16	0.2140

2) Has Disease: D, test Positive: +

$$= P(D|+) = P(+|D) * P(D) / [P(+|D) * P(D) + P(+|D') * P(D')]$$

$$= 0.99 * 0.0001 / [(0.99 * 0.0001) + (0.01 * 0.9999)]$$

$$= \mathbf{0.98\% \Rightarrow \text{Do not recommend this test}}$$

$$3) \quad \mathbf{P(X, Y|Z) = P(X|Z)P(Y|Z)}$$

$$\mathbf{P(X|Y, Z) = P(X|Z)}$$

$$P(X|Y) = P(X, Y) / P(Y)$$

$$P(X|Y, Z) = \mathbf{P(X, Y|Z)} / P(Y|Z)$$

$$P(X|Y, Z) = \mathbf{P(X|Z)P(Y|Z)} / P(Y|Z)$$

$$\mathbf{P(X|Y, Z) = P(X|Z)}$$