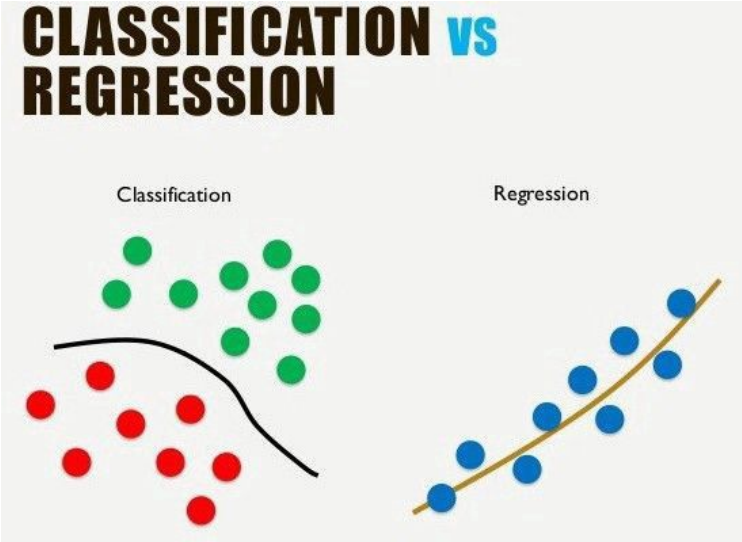
Decision Trees

TJ Machine Learning Club

Classification vs. Regression

- Classification
  - Classifying photos of fruits
  - Determining whether tumor is benign or malignant
- Regression
  - Predicting COVID-19 cases given demographic data
  - Predicting house prices given house features



Source: https://medium.com/datasoc/whats-the-problem-1ff8b338094b

Features vs. Labels

Features (like x): Characteristics of the input

- In the picture, features are whether or not patient smokes (smoke), consumes alcohol (alco), and performs physical activity (active)

Label (like y): The prediction or classification of the input

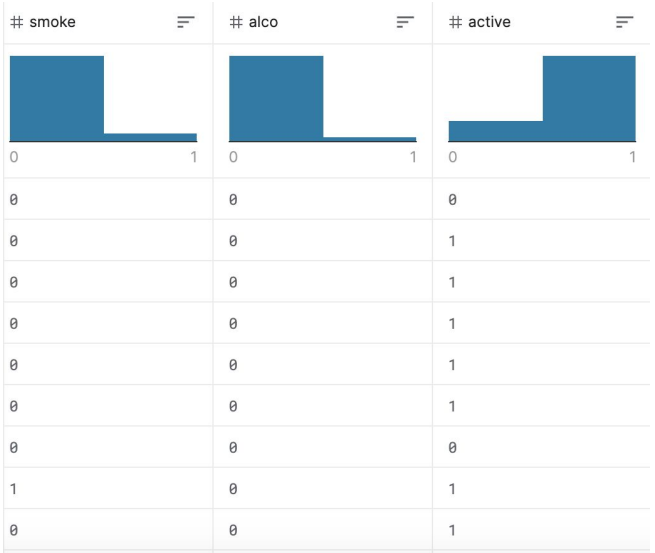
- Whether or not patient has cardiovascular disease (cardio)

Training and Testing Datasets

Training data has both features and labels

# smoke	# alco	# active	# cardio
0 1	0 1	0 1	0 1
0	1	0	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	1	1
0	0	1	1
0	0	1	0
0	0	1	0

Testing data only has the features



# smoke	# alco	# active
0 1	0 1	0 1
0	0	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
0	0	0
1	0	1
0	0	1

Need to predict cardio

What is a Decision Tree?

- A decision tree is just a series of questions
- The key in creating a decision tree is asking the right questions

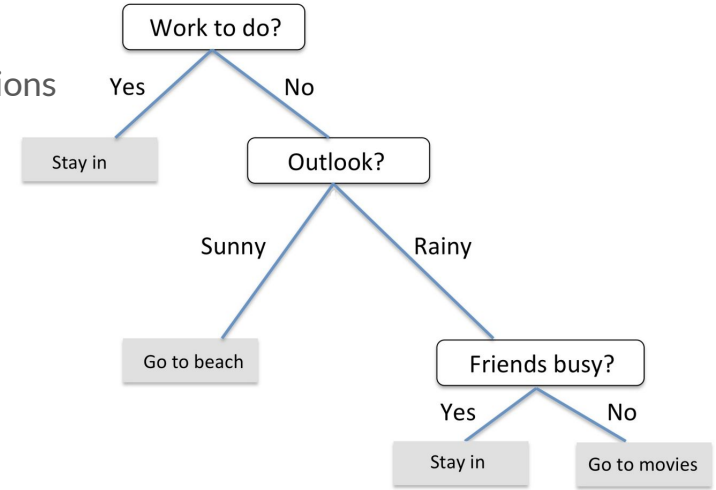


Figure 1: Real Life Decision Tree

Gini Impurity

- Measure of how “messy” some collection of data is

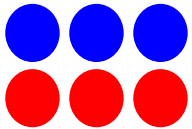
$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$

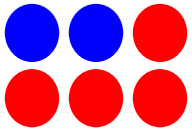
i = some data
k = class index
c = total number of classes
p(k|i) = probability of randomly selecting item of class k from data

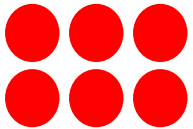
Ex. Gini Impurity

$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$

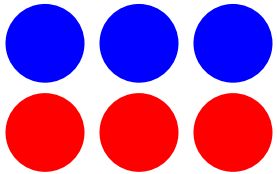
Let’s calculate the Gini Impurity for these groups of data, where the two possible classes are blue or red:



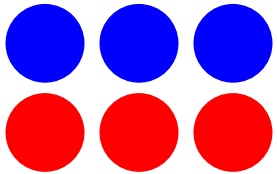




$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$

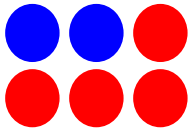


$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$

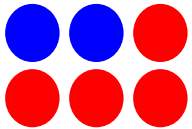


$$1 - (3/6)^2 - (3/6)^2 = \boxed{1/2}$$

$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$

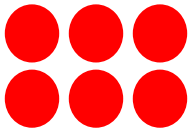


$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$

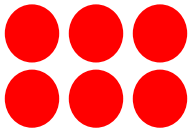


$$1 - (2/6)^2 - (4/6)^3 = 4/9 = \boxed{0.4444}$$

$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$



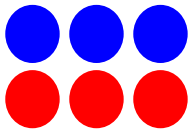
$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$



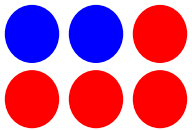
$$1 - (6/6)^2 = \boxed{0}$$

Ex. Gini Impurity

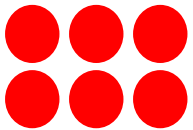
$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$



0.5



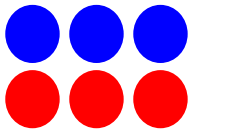
0.444



0

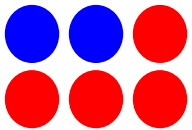
Ex. Gini Impurity

$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$

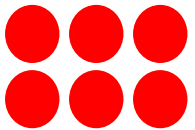


0.5

Maximum possible impurity



0.444



0

Minimum possible impurity

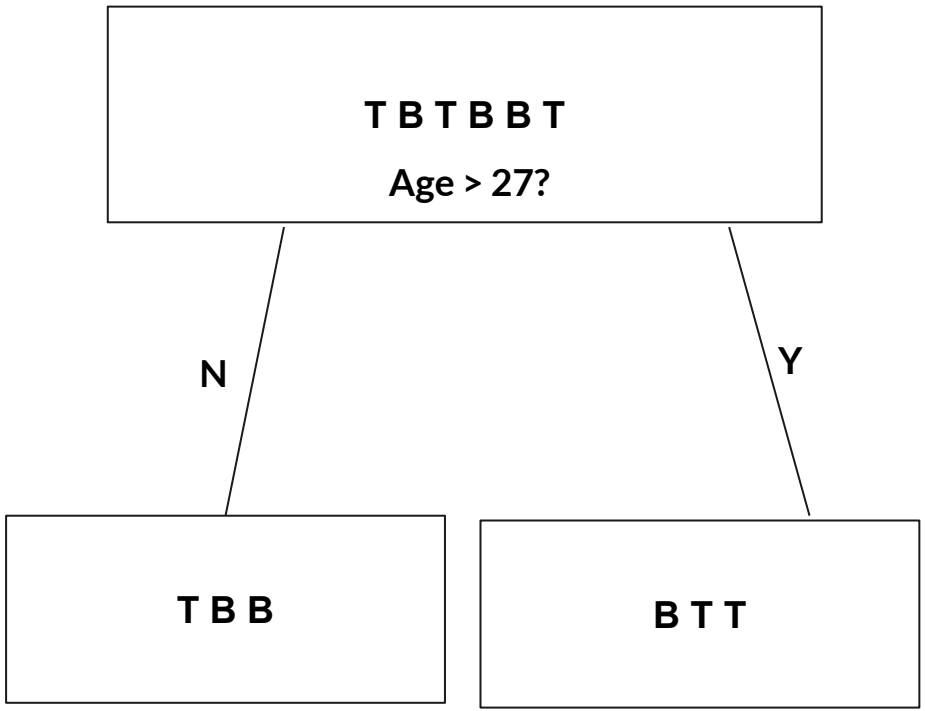
Information Gain

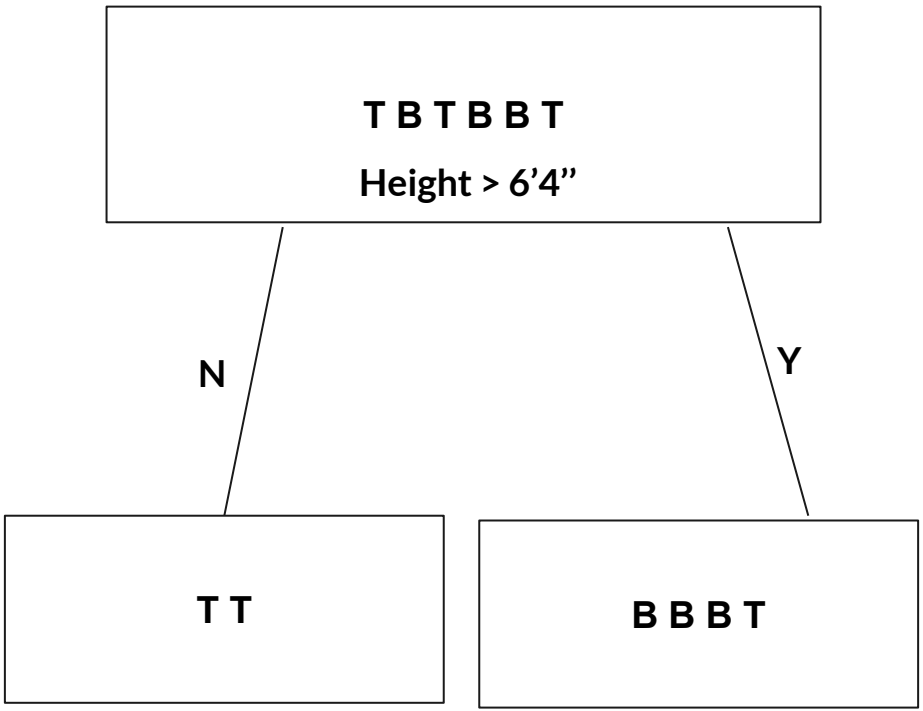
$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

- D_p , D_{left}, D_{right} are the parent node, left node dataset, and right node dataset respectively
- I is a measure of impurity (like Gini Impurity)
- N_p, N_{left}, and N_{right} are the number of items in the parent, left, and right nodes respectively
- f is the question you are asking to create the split

Let's figure out which question is a better question to ask to split the athletes according to sport

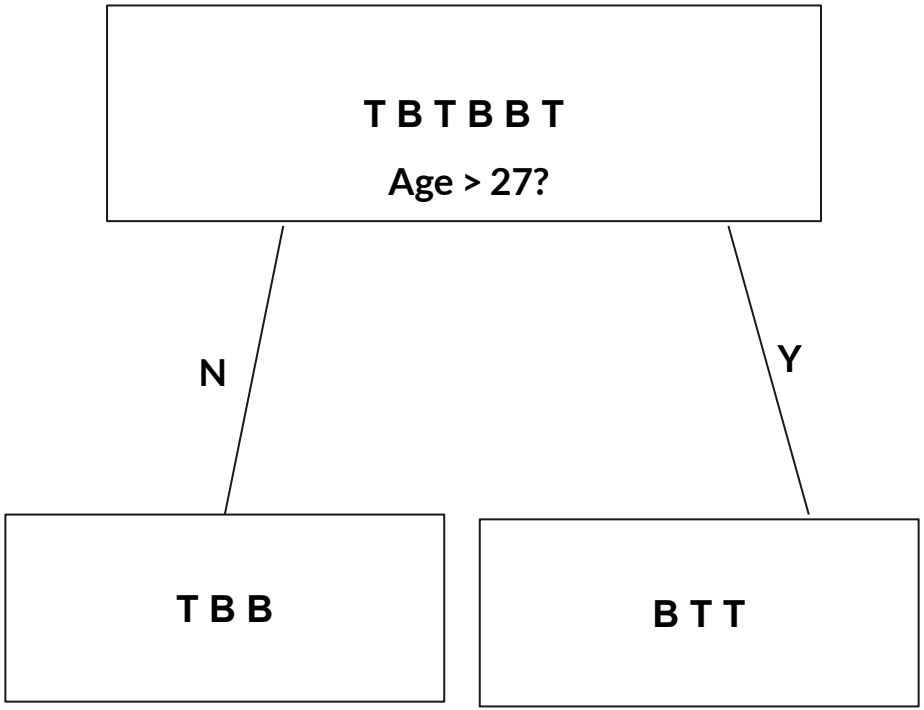
T = Tennis Player
B = Basketball Player



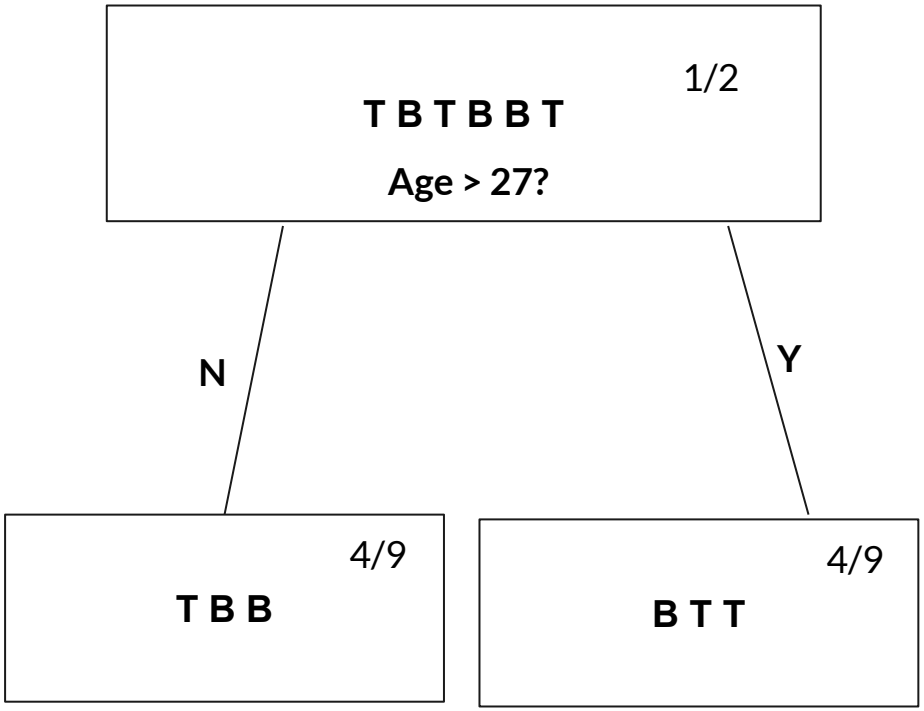


$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

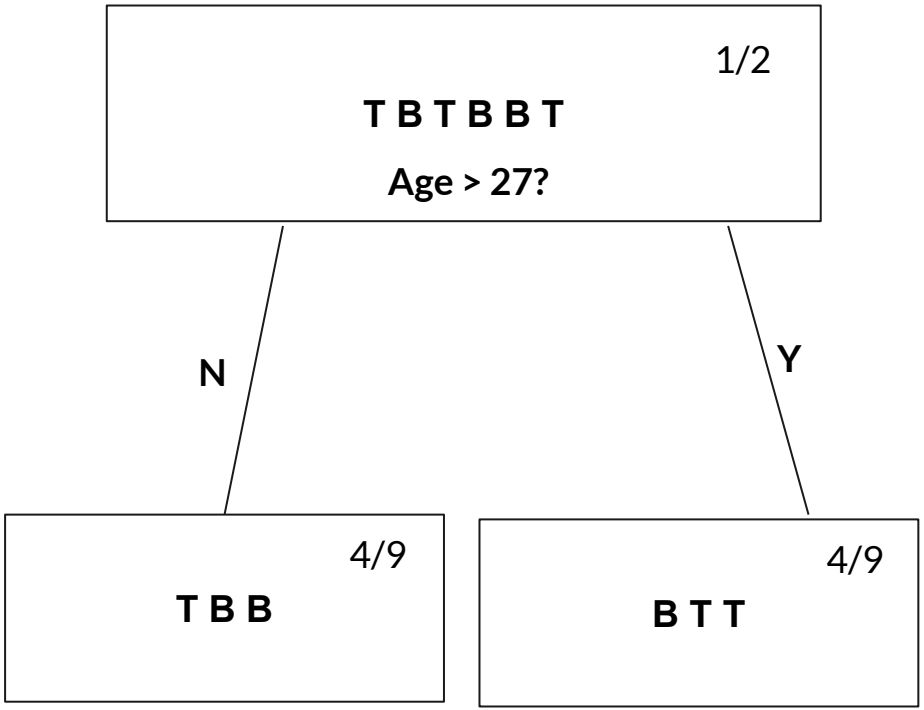
$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$



$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$



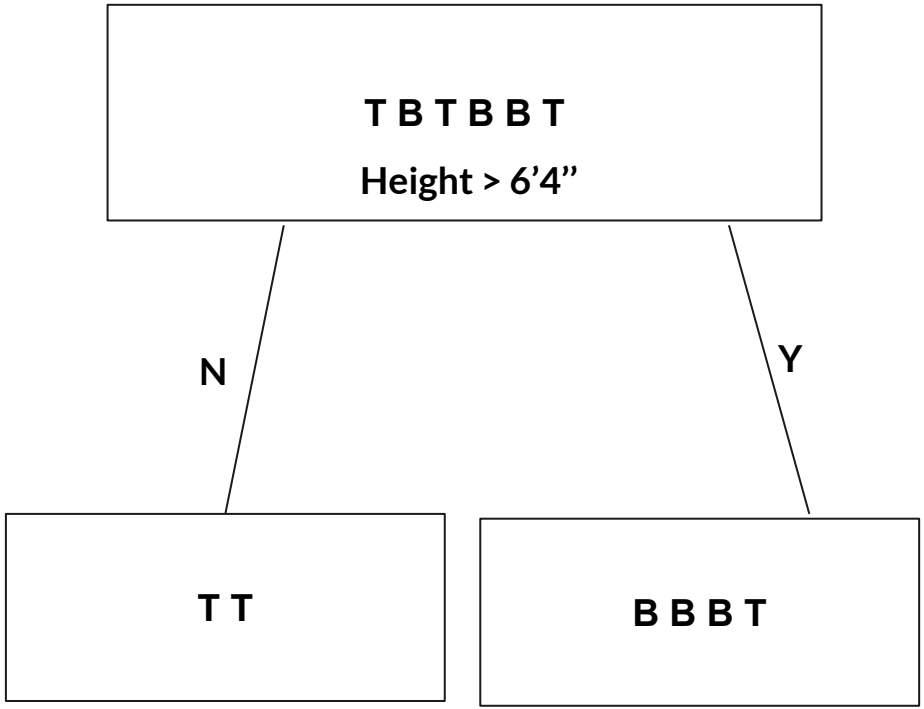
$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$



$$\frac{1}{2} - \left(\frac{3}{6}\right)\left(\frac{4}{9}\right) - \left(\frac{3}{6}\right)\left(\frac{4}{9}\right) = \frac{1}{18} =$$

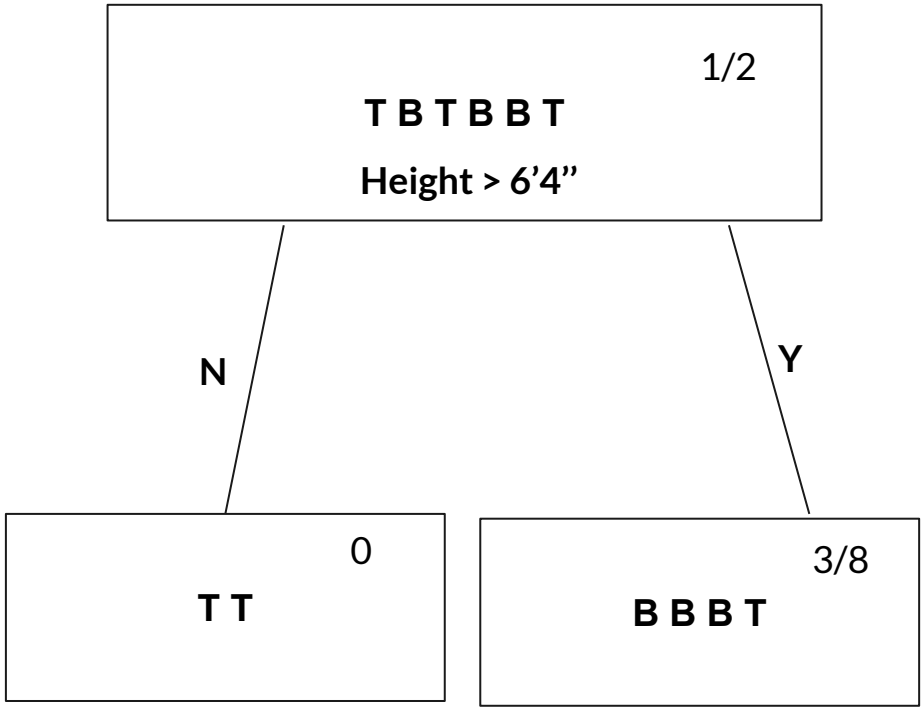
$0.05556\overline{6}$

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

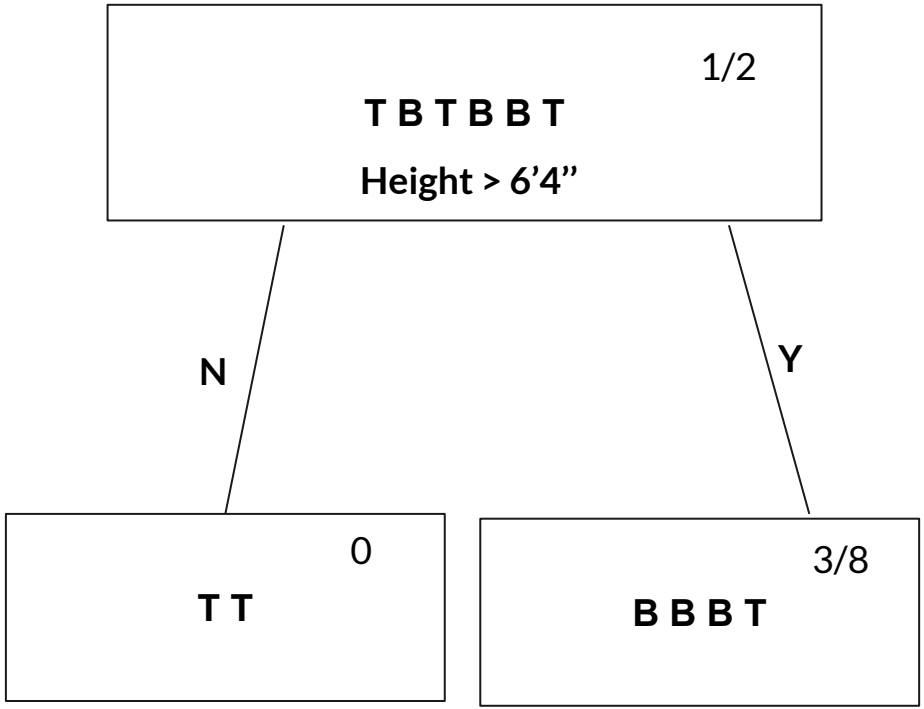


$$I_G(i) = 1 - \sum_{k=1}^{c} p(k|i)^2$$

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

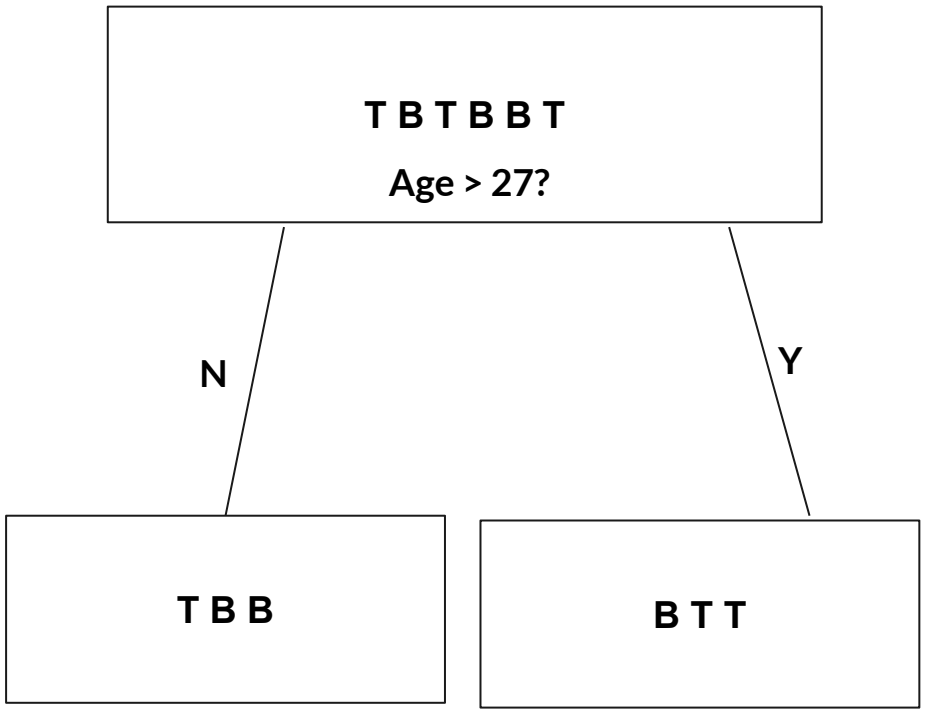


$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$



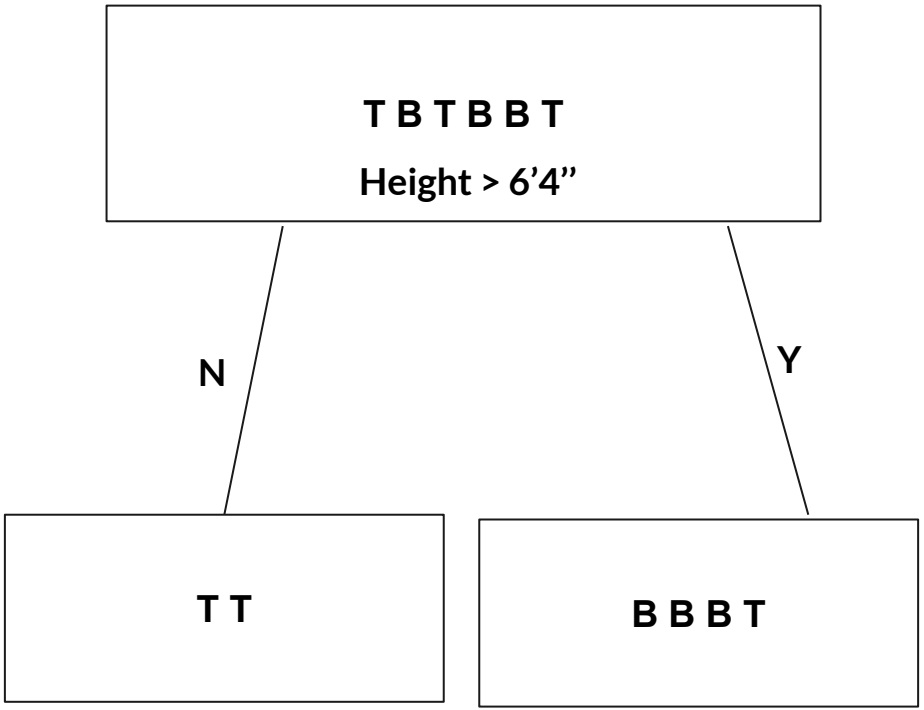
$$\frac{1}{2} - \left(\frac{2}{6}\right)(0) - \left(\frac{4}{6}\right)\left(\frac{3}{8}\right) = \frac{1}{4} = 0.25$$

Information Gain: 0.055556



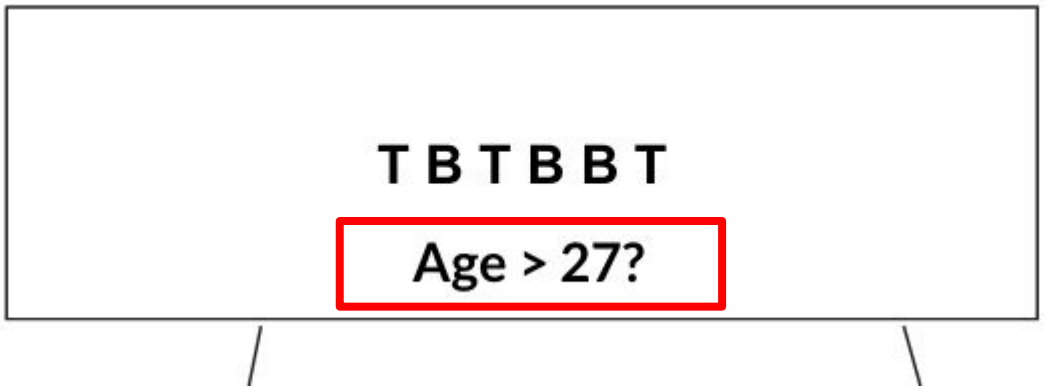
Information Gain: 0.25

Since Information Gain is higher, this the better question to ask to classify our athletes

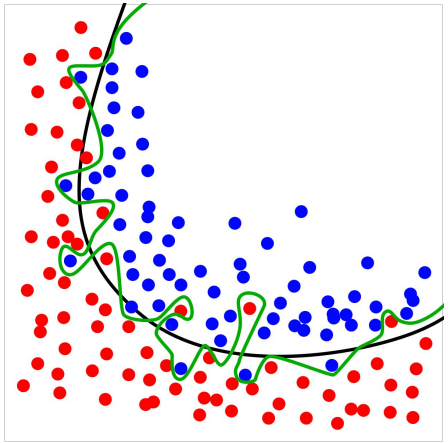


How to Come Up with Values for the Questions?

- The most straightforward way: Try out different values from the items in your training dataset



Overfitting



- **Techniques to prevent overfitting in decision trees:**
- Continue recursively generating nodes only if information gain is larger than some threshold (e.g. 0.1)
- After creating the tree, prune all nodes that are at a depth greater than some threshold