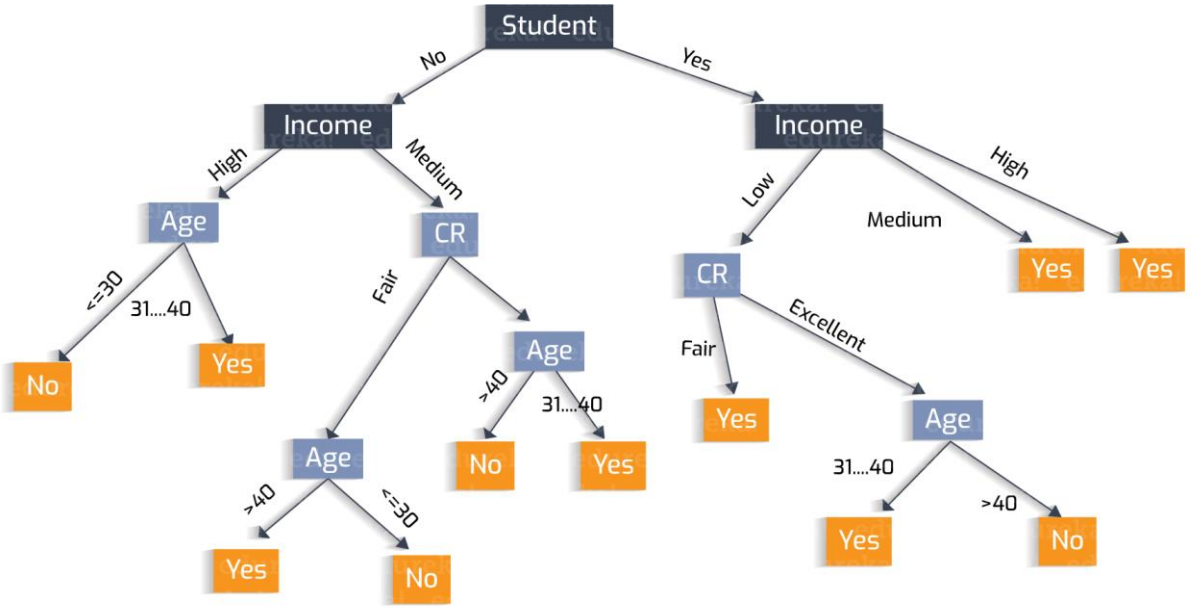
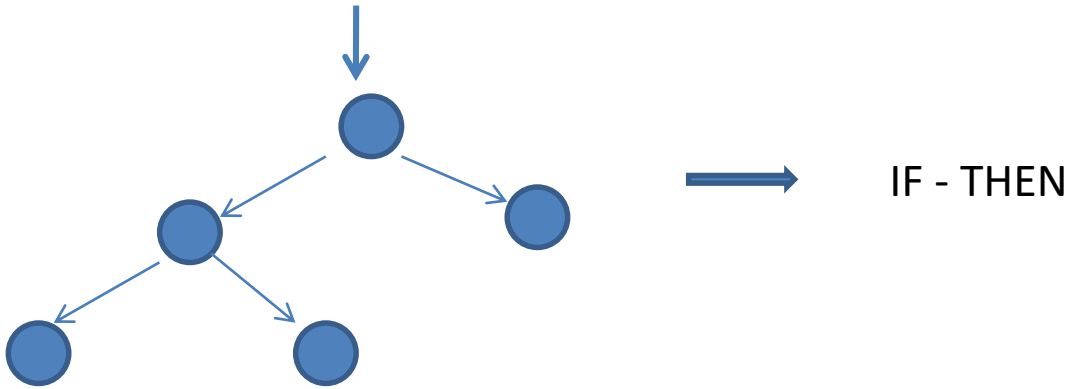
Decision Trees



Decision Trees

- Decision tree is a learning algorithm in which the source data are reduced to a tree that correctly classify the examples given

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No



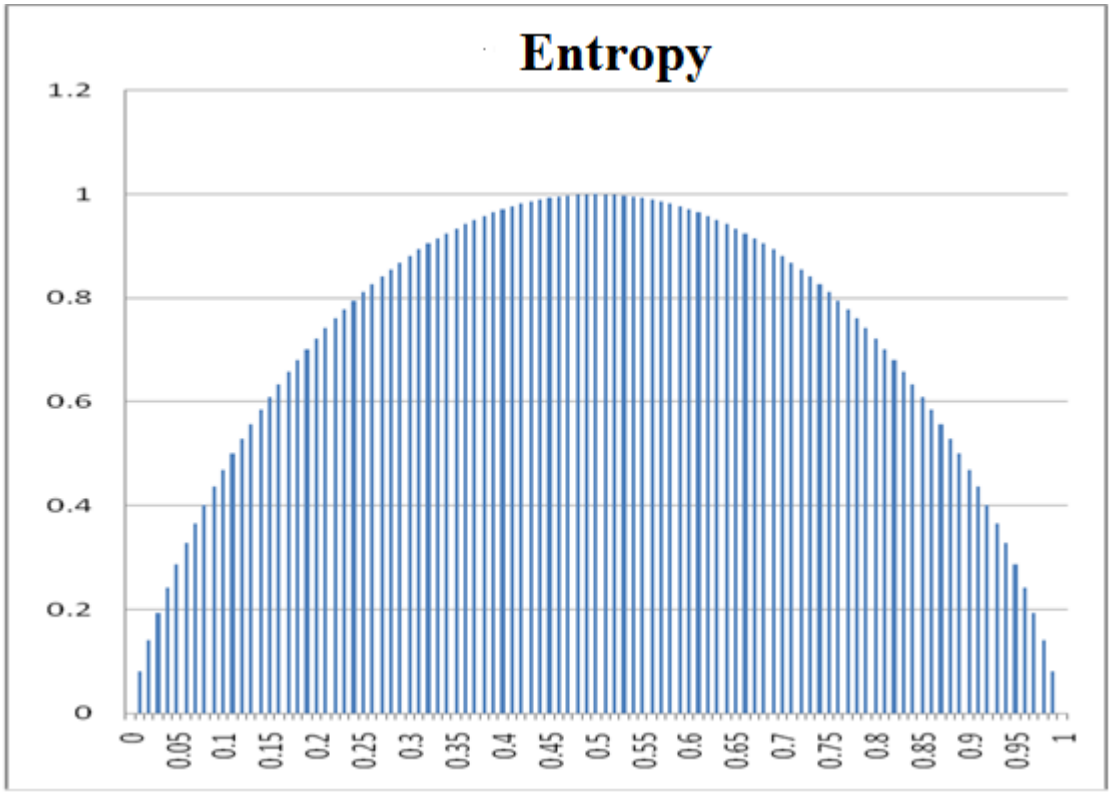
- In the construction of a decision tree the *statistical gain* and the *entropy* are used

- The *entropy* can be understand as the level of uncertainty in the data

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No

- Given a set of examples *S*, containing positive and negative examples of a target concept, the *entropy* of *S* relative to this Boolean classification is:

$$Entropy(S) = -(p_+ \log_2 p_+) - (p_- \log_2 p_-)$$



	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No

$$Entropy(S) = -(p_+ \log_2 p_+) - (p_- \log_2 p_-)$$

$$Entropy(S) = Entropy(2+,4-) = -(p_+ \log_2 p_+) - (p_- \log_2 p_-) =$$

$$= -(2/6)\log_2(2/6) - (4/6)\log_2(4/6) =$$

$$= -(0.33333)(-1.58496) - (0.66666)(-0.58496) =$$

$$= 0.52833 + 0.38997 = 0.9183$$

UPAEP

- The *information gain* is the measure of effectiveness of an attribute to classify the given examples

- The information gain of an attribute is simply the reduction of entropy caused by the division of the examples according to this attribute

- More precisely, the information gain ***Gain (S, A)*** of an attribute ***A***, relative to the collection of examples ***S***, is defined as:

$$Gain(S,A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_{A=v}|}{|S|} Entropy(S_{A=v})$$

$$Gain(S,F) = Entropy(S) - \sum_{v \in \{yes,no\}} \frac{|S_{F=v}|}{|S|} Entropy(S_{F=v})$$

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No

$$Gain(S,F) = Entropy(S) - \left(\frac{3}{6}\right) Entropy(S_{F=yes}) - \left(\frac{3}{6}\right) Entropy(S_{F=no})$$

$$Gain(S,M) = Entropy(S) - \left(\frac{4}{6}\right) Entropy(S_{M=yes}) - \left(\frac{2}{6}\right) Entropy(S_{M=no})$$

$$Gain(S,W) = Entropy(S) - \left(\frac{4}{6}\right) Entropy(S_{W=yes}) - \left(\frac{2}{6}\right) Entropy(S_{W=no})$$

$$Gain(S,D) = Entropy(S) - \left(\frac{4}{6}\right) Entropy(S_{D=yes}) - \left(\frac{2}{6}\right) Entropy(S_{D=no})$$

$$Gain(S,F) = Entropy(S) - \left(\frac{3}{6}\right)Entropy(S_{F=yes}) - \left(\frac{3}{6}\right)Entropy(S_{F=no})$$

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No

$$Entropy(S_{F=yes}) = Entropy(2+,1-) = -\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) =$$

$$= -(0.66666)(-0.58496) - (0.33333)(-1.58496) = 0.38997 + 0.52833 = 0.9183$$

$$Entropy(S_{F=no}) = Entropy(0+,3-) = -\left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right) - \left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right) = 0 + 0 = 0$$

So That

$$Gain(S,F) = 0.9183 - \left(\frac{3}{6}\right)(0.9183) - \left(\frac{3}{6}\right)(0) = 0.45915$$

$$Gain(S,M) = Entropy(S) - \left(\frac{4}{6}\right) Entropy(S_{M=yes}) - \left(\frac{2}{6}\right) Entropy(S_{M=no})$$

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No

$$Entropy(S_{M=yes}) = Entropy(1+,3-) = -\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) =$$

$$= -(0.25)(-2) - (0.75)(-0.41504) = 0.5 + 0.31128 = 0.81128$$

$$Entropy(S_{M=no}) = Entropy(1+,1-) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

$$= -(0.5)(-1) - (0.5)(-1) = 1$$

So That

$$Gain(S,M) = 0.9183 - \left(\frac{4}{6}\right)(0.81128) - \left(\frac{2}{6}\right)(1) = 0.044115$$

$$Gain(S,W) = Entropy(S) - \left(\frac{4}{6}\right) Entropy(S_{W=yes}) - \left(\frac{2}{6}\right) Entropy(S_{W=no})$$

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No

$$Entropy(S_{W=yes}) = Entropy(2+,2-) = -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) =$$

$$= -(0.5)(-1) - (0.5)(-0.1) = 1$$

$$Entropy(S_{W=no}) = Entropy(0+,2-) = -\left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) - \left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) = 0$$

So That

$$Gain(S,W) = 0.9183 - \left(\frac{4}{6}\right)(1) - \left(\frac{2}{6}\right)(0) = 0.25163$$

$$Gain(S,D) = Entropy(S) - \left(\frac{4}{6}\right) Entropy(S_{D=yes}) - \left(\frac{2}{6}\right) Entropy(S_{D=no})$$

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No

$$Entropy(S_{D=yes}) = Entropy(1+,3-) = -\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) =$$

$$= -(0.25)(-2) - (0.75)(-0.41504) = 0.5 + 0.31128 = 0.81128$$

$$Entropy(S_{D=no}) = Entropy(1+,1-) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

So That

$$Gain(S,D) = 0.9183 - \left(\frac{4}{6}\right)(0.81128) - \left(\frac{2}{6}\right)(1) = 0.044115$$

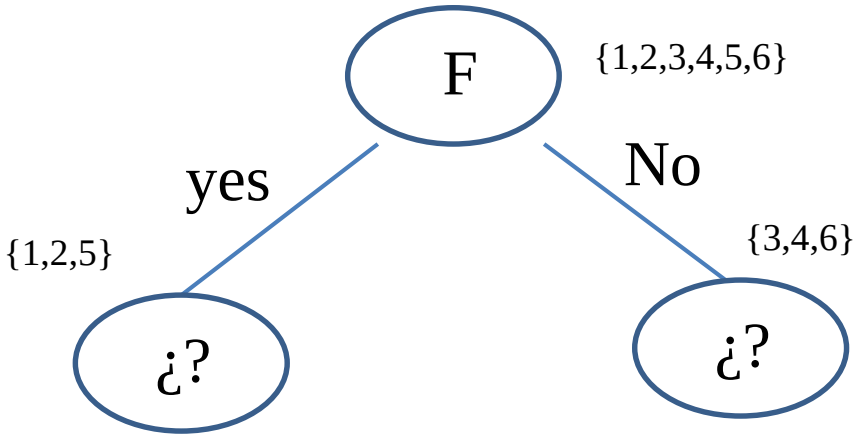
$$Gain(S,F) = 0.45915$$

$$Gain(S,M) = 0.044115$$

$$Gain(S,W) = 0.25163$$

$$Gain(S,D) = 0.044115$$

The attribute that has the highest gain is F ,therefore, the root of the tree is F (First last year?)



	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No

Subset of examples where F = yes {1, 2, 5}

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Gail	Yes	No	Yes	Yes	Yes

$$Entropy(S_{F=yes}) = Entropy(2+,1-) = -\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) =$$

$$= -(0.66666)(-0.58496) - (0.33333)(-1.58496) = 0.38997 + 0.52833 = 0.9183$$

$$Gain(S,M) = Entropy(S_{F=yes}) - \left(\frac{2}{3}\right) Entropy(S_{M=yes}) - \left(\frac{1}{3}\right) Entropy(S_{M=no})$$

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Gail	Yes	No	Yes	Yes	Yes

$$Entropy(S_{M=yes}) = Entropy(1+,1-) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

$$Entropy(S_{M=no}) = Entropy(1+,0-) = -\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) - \left(\frac{0}{1}\right)\log_2\left(\frac{0}{1}\right) = 0$$

So That

$$Gain(S_{F=yes}, M) = 0.9183 - \left(\frac{2}{3}\right)(1) - \left(\frac{1}{3}\right)(0) = .25163$$

$$Gain(S,W) = Entropy(S_{F=yes}) - \left(\frac{2}{3}\right) Entropy(S_{W=yes}) - \left(\frac{1}{3}\right) Entropy(S_{W=no})$$

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Gail	Yes	No	Yes	Yes	Yes

$$Entropy(S_{W=yes}) = Entropy(2+,0-) = -\left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) - \left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) = 0$$

$$Entropy(S_{W=no}) = Entropy(0+,1-) = -\left(\frac{0}{1}\right)\log_2\left(\frac{0}{1}\right) - \left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) = 0$$

So That

$$Gain(S_{F=yes},W) = 0.9183 - \left(\frac{2}{3}\right)(0) - \left(\frac{1}{3}\right)(0) = .9183$$

$$Gain(S,D) = Entropy(S_{F=yes}) - \left(\frac{2}{3}\right) Entropy(S_{D=yes}) - \left(\frac{1}{3}\right) Entropy(S_{D=no})$$

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Gail	Yes	No	Yes	Yes	Yes

$$Entropy(S_{D=yes}) = Entropy(1+,1-) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

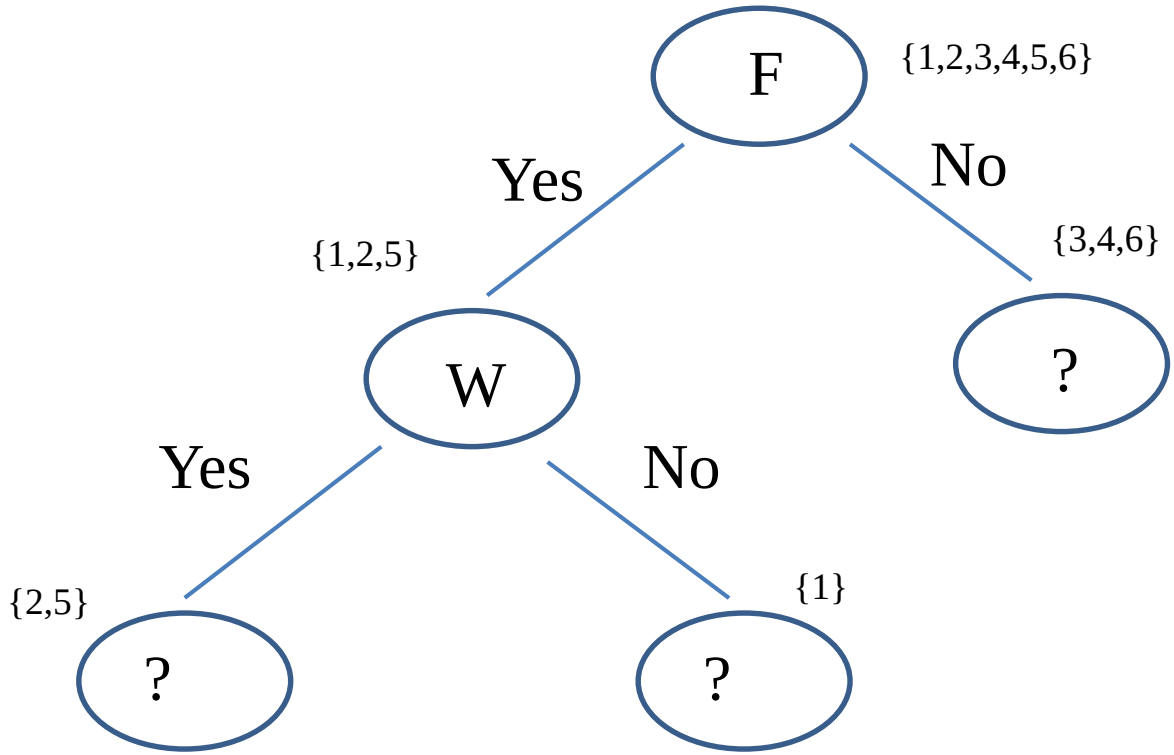
$$Entropy(S_{D=no}) = Entropy(1+,0-) = -\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) - \left(\frac{0}{1}\right)\log_2\left(\frac{0}{1}\right) = 0$$

So That

$$Gain(S_{F=yes}, D) = 0.9183 - \left(\frac{2}{3}\right)(1) - \left(\frac{1}{3}\right)(0) = .25163$$

$$Gain(S_{F=yes}, M) = 0.25163$$
$$Gain(S_{F=yes}, W) = 0.9183$$
$$Gain(S_{F=yes}, D) = 0.25163$$

The attribute that has the highest gain is W ,therefore, the root of the Sub tree is W (Works hard?)

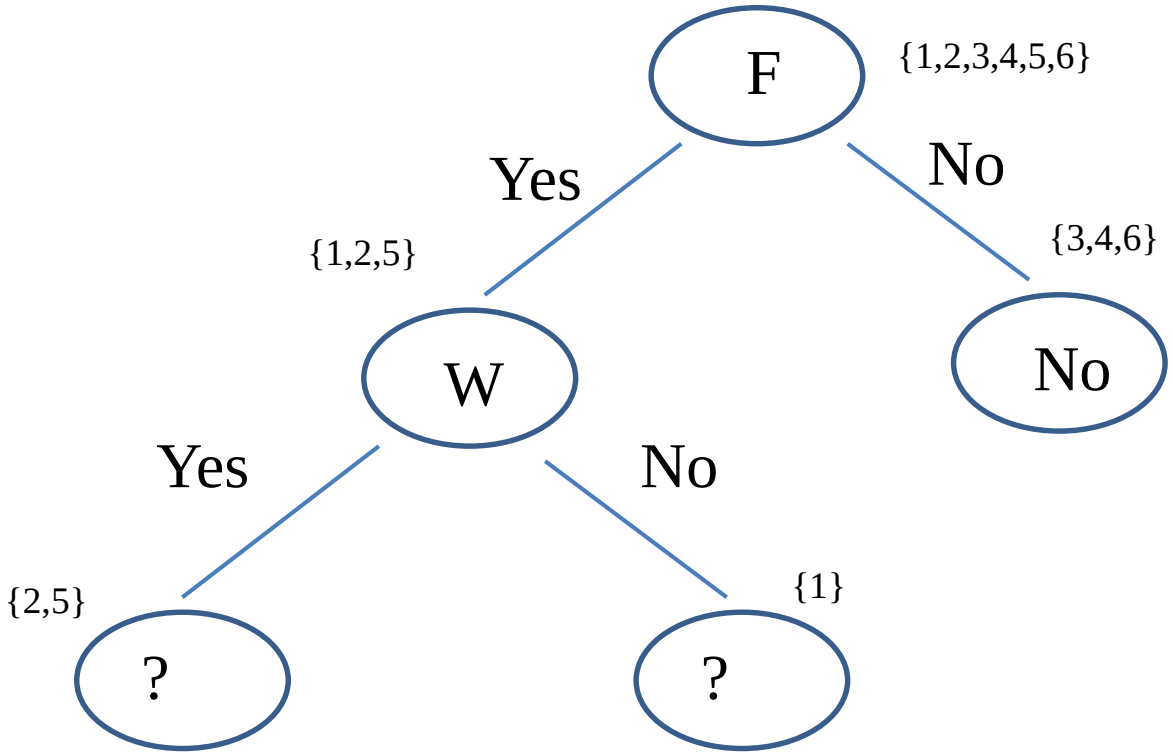


	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No

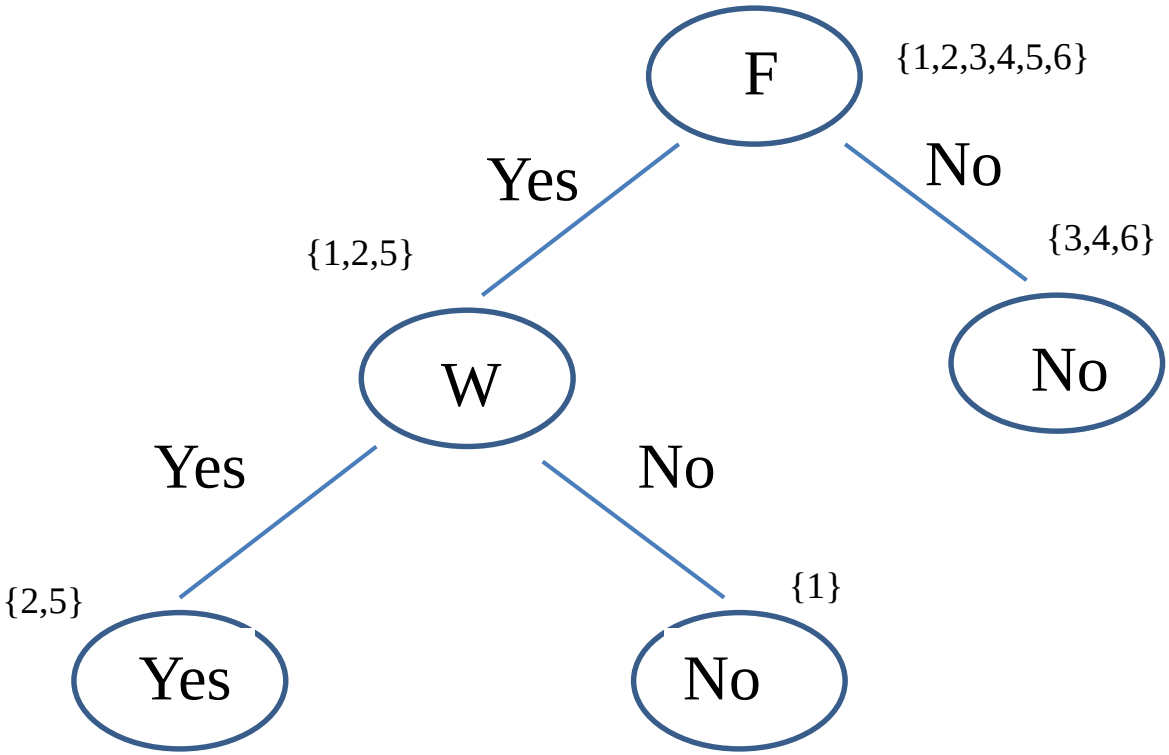
Subset of examples where F = No {3, 4, 6}

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Simon	No	Yes	Yes	Yes	No

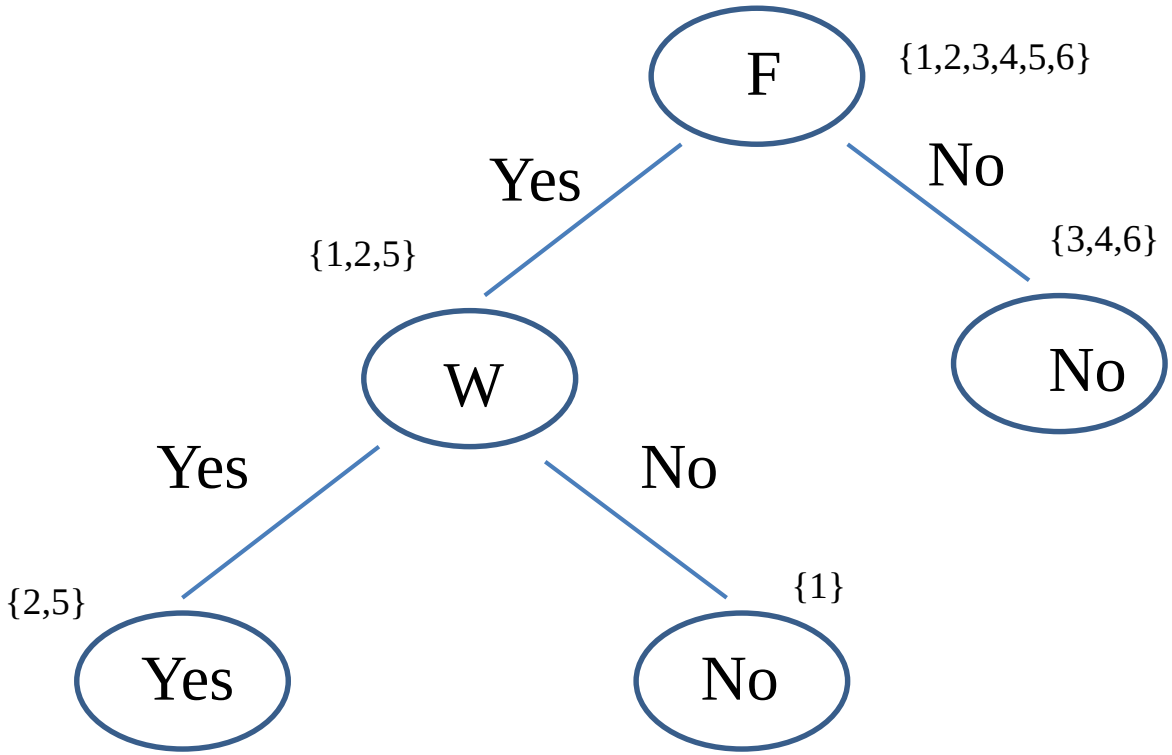
$$Entropy(S_{F=no}) = Entropy(0+,3-) = -\left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right) - \left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right) = 0$$


F
{1,2,3,4,5,6}
Yes
No
{1,2,5}
{3,4,6}
W
No
Yes
No
{2,5}
{1}
?
?

	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Gail	Yes	No	Yes	Yes	Yes



	F	M	W	D	
Student	**First last year?**	**Male?**	**Works hard?**	**Drinks?**	**First this year?**
Richard	Yes	Yes	No	Yes	No
Alan	Yes	Yes	Yes	No	Yes
Alison	No	No	Yes	No	No
Jeff	No	Yes	No	Yes	No
Gail	Yes	No	Yes	Yes	Yes
Simon	No	Yes	Yes	Yes	No



Confusion matrix

In the field of machine learning and specifically the problem of statistical classification, a **confusion matrix**, also known as an error matrix

Is a specific table layout that allows visualization of the performance of an algorithm

		Actual Class	
		Cat	Dog
Predicted Class	Cat	5	2
	Dog	3	4

	P	N
P	**TP**	FP
N	FN	**TN**

where:

P = Positive; N = Negative;

TP = True Positive; FP = False Positive;

TN = True Negative; FN = False Negative

	C1	C2	C3	C4	C5	C6
C1						
C2						
C3						
C4						
C5						
C6						

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

weather.nominal.arff