1. For the following dataset:

ID	*Outl*	*Temp*	*Humi*	*Wind*	PLAY
TRAINING INSTANCES					
A	s	h	h	F	N
B	s	h	h	T	N
C	o	h	h	F	Y
D	r	m	h	F	Y
E	r	c	n	F	Y
F	r	c	n	T	N
TEST INSTANCES					
G	o	c	n	T	?
H	s	m	h	F	?

Classify the test instances using the **ID3 Decision Tree** method:

a) Using the **Information Gain** as a splitting criterion

b) Using the **Gain Ratio** as a splitting criterion

(a) IG: At each level of DT, choose attribute with largest IG

$$IG(A|R) = H(R) - \sum_{i \in A} P(A=i)\,H(A=i)$$

weight: fraction of instance at child node i

$H(R)$: entropy of parent node

$\sum_{i \in A} P(A=i)\,H(A=i)$: weighted average entropy across child nodes (Mean Information: MI)

H: entropy (impurity) of a node

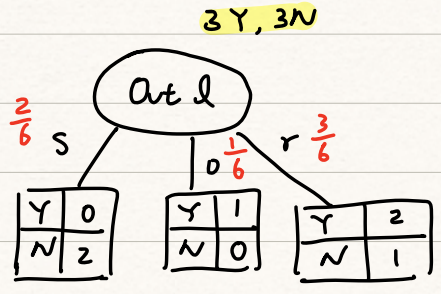
$$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$$

Root node: 3Y, 3N

$$H(R) = -\left\{ \tfrac{1}{2}\log_2 \tfrac{1}{2} + \tfrac{1}{2}\log_2 \tfrac{1}{2} \right\} = 1$$

($\frac{1}{2}\log_2\frac{1}{2}$: Y; $\frac{1}{2}\log_2\frac{1}{2}$: N)

Example: For Outl:



$$H(Outl=s) = -\{0\log_2 0 + 1\log_2 1\} = 0$$

$$H(Outl=o) = -\{1\log_2 1 + 0\log_2 0\} = 0$$

$$H(Outl=r) = -\left\{\tfrac{2}{3}\log_2\tfrac{2}{3} + \tfrac{1}{3}\log_2\tfrac{1}{3}\right\} = 0.9183$$

$$MI(Outl) = \tfrac{2}{6}\times 0 + \tfrac{1}{6}\times 0 + \tfrac{3}{6}\times 0.9183 = 0.4592$$

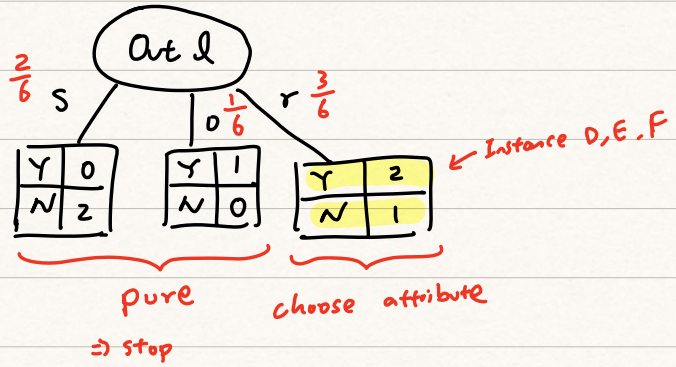
$$IG(Outl) = H(R) - MI(Outl) = 1 - 0.4592 = 0.5408$$

will get useless classifier

	R	Outl			Temp			H		Wind		ID					
		s	o	r	h	m	c	h	n	T	F	A	B	C	D	E	F
Y	3	0	1	2	1	1	1	2	1	0	3	0	0	1	1	1	0
N	3	2	0	1	2	0	1	2	1	2	1	1	1	0	0	0	1
Total	6	2	1	3	3	1	2	4	2	2	4	1	1	1	1	1	1
P(Y)	1/2	0	1	2/3	1/3	1	1/2	1/2	1/2	0	3/4	0	0	1	1	1	0
P(N)	1/2	1	0	1/3	2/3	0	1/2	1/2	1/2	1	1/4	1	1	0	0	0	1
H	1	0	0	0.9183	0.9183	0	1	1	1	0	0.8112	0	0	0	0	0	0
MI		0.4592			0.7924			1		0.5408		0					
IG		0.5408			0.2076			0		0.4592		1					
SI		1.459			1.459			0.9183		0.9183		2.585					
GR		0.3707			0.1423			0		0.5001		0.3868					

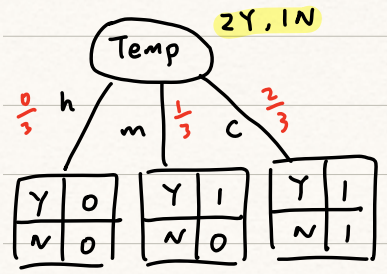
largest

H : 3Y, 3N



parent : H(Outl = r) = 0.9183

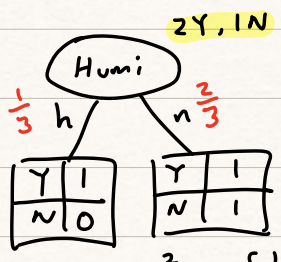
=> Find IG for Temp, Humi, Wind



$MI(T) = \frac{2}{3} \times -\{\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\}$

$= \frac{2}{3}$

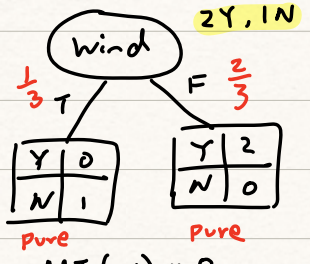
$IG(T) = 0.9183 - \frac{2}{3} = 0.2516$



$MI(w) = \frac{2}{3} \times -\{\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\}$

$= \frac{2}{3}$

$IG(w) = 0.2516$

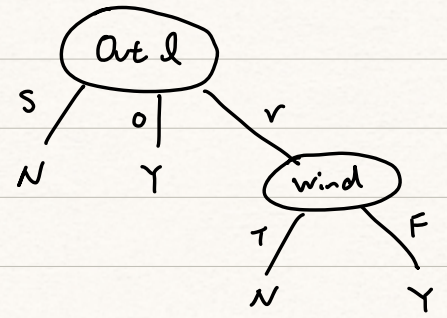


$MI(w) = 0$

(largest)

$IG(w) = 0.9183 - 0 = 0.9183$

DT :



Classify G: Outl = O => Y

Classify H: Outl = S => N

1. For the following dataset:

ID	*Outl*	*Temp*	*Humi*	*Wind*	PLAY
TRAINING INSTANCES					
A	s	h	h	F	N
B	s	h	h	T	N
C	o	h	h	F	Y
D	r	m	h	F	Y
E	r	c	n	F	Y
F	r	c	n	T	N
TEST INSTANCES					
G	o	c	n	T	?
H	s	m	h	F	?

Classify the test instances using the **ID3 Decision Tree** method:

a) Using the **Information Gain** as a splitting criterion

b) Using the **Gain Ratio** as a splitting criterion

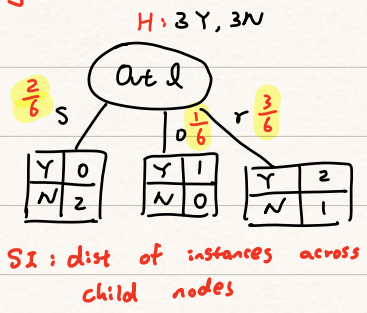
(b) Choose attribute with largest GR.

$$GR(A) = \frac{IG(A)}{SI(A)}$$ ← split info (entropy)

$$SI(Outl) = -\left\{\frac{2}{6}\log_2\frac{2}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{3}{6}\log_2\frac{3}{6}\right\}$$

$$= 1.459$$

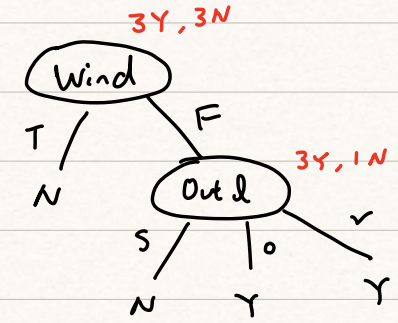
$$\Rightarrow GR(Outl) = \frac{0.5408}{1.459} = 0.3707$$



SI : dist of instances across child nodes

	R	Outl			Temp			H		Wind		ID					
		s	o	r	h	m	c	h	n	T	F	A	B	C	D	E	F
Y	3	0	1	2	1	1	1	2	1	0	3	0	0	1	1	1	0
N	3	2	0	1	2	0	1	2	1	2	1	1	1	0	0	0	1
Total	6	2	1	3	3	1	2	4	2	2	4	1	1	1	1	1	1
P(Y)	1/2	0	1	2/3	1/3	1	1/2	1/2	1/2	0	3/4	0	0	1	1	1	0
P(N)	1/2	1	0	1/3	2/3	0	1/2	1/2	1/2	1	1/4	1	1	0	0	0	1
H	1	0	0	0.9183	0.9183	0	1	1	1	0	0.8112	0	0	0	0	0	0
MI		0.4592			0.7924			1		0.5408		0					
IG		0.5408			0.2076			0		0.4592		1					
SI		1.459			1.459			0.9183		0.9183		2.585					
GR		0.3707			0.1423			0		0.5001		0.3868					

Skip to DT:



Classify G: Wind = T ⇒ N

Classify H: Wind = F → Outl = S ⇒ N

2. What is the difference between "model bias" and "model variance"?
   i. Why is a high bias, low variance classifier undesirable?
   ii. Why is a low bias, high variance classifier (usually) undesirable?

Bias: propensity of a classifier to systematically produce the same errors. $E[g(x) - f(x)]$ (average model approx error over all possible training sets)
If it doesn't produce error / produces different kinds of errors ⇒ unbiased (e.g. predict too many instances as the majority class)

Variance: propensity of a classifier to produce different classifications using different training set. (randomly sampled from same population)
Measure of the inconsistency of the classifier, from training set to training set.

$$E[\{f(x) - E(f(x))\}^2]$$

(i) High bias & low variance
⇒ Consistently wrong.

(ii) Low bias & high variance
low bias → may be correct predictions.
high variance → difficult to be certain about the performance of the classifier
If high variance, ER may be low on one set of data, and high on another set (not generalised)

3. Describe how validation set, and cross-validation can help reduce overfitting?

models usually have hyperparameter(s) ⟶ control model complexity

find best values ⟶ to achieve best predictive performance on new data.

(may also consider a range of different types of models ⇒ find best one)

performance on training data : not a good indicator on unseen data.
(may be overfitting)

2 ways :

① Validation set: we train models on training set, compare them on independent data (val set) ⇒ select best one.
evaluate the final model with test set.

② CV: If data is limited & want good models
⇒ use as much of the available data as possible for training.
⇒ small validation set ⇒ Use CV

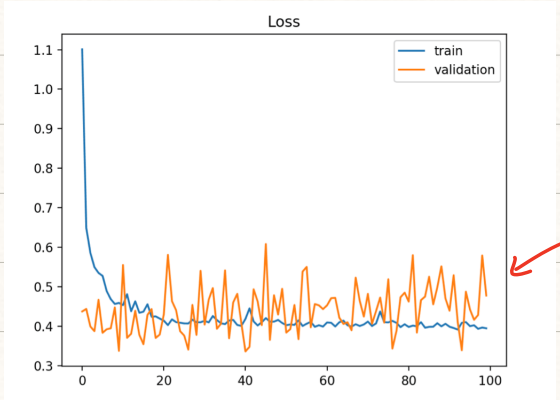
4. [OPTIONAL] Given a dataset with N instances, what is the expected impact on "model variance", if you increase k in k-fold cross validation? what is the expected impact on "evaluation variance"?

K-fold CV:

K ↑ ⇒ more training data (K-1 folds) & less test data (1 fold) in each iteration

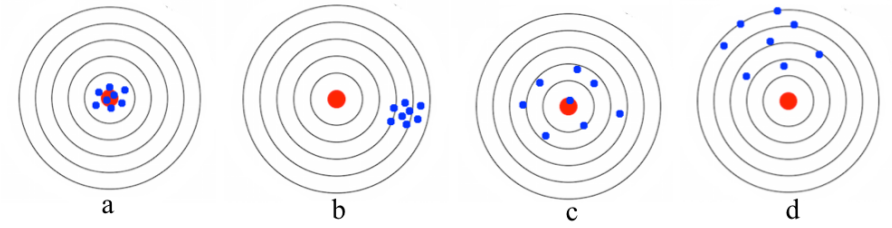
⇒ model variance ↓ as train data ↑

⇒ evaluation variance ↑ as test data ↓



validation set too small ⇒ high evaluation variance

5. [OPTIONAL] Considering the following model results where the blue dots represent the distribution of model predictions and the red dot is the actual (correct) answer, explain each model using the model bias and variance? Explain which of these models is valid and which one is reliable.



(a) Low bias & low variance

⇒ valid & reliable

(b) High bias & low variance (underfitting)

⇒ reliable (can predict its behaviour) but not valid

(c) Low bias & high variance (overfitting)

⇒ valid (can develop correct predictions) but not reliable.

(d) High bias & high variance

⇒ not reliable nor valid