

# Decision Trees

## Classification & Regression

**Dr. Saed Sayad**

University of Toronto

2010

saed.sayad@utoronto.ca


# Decision Tree

A set of training examples is broken down into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. At the end of the learning process, a decision tree covering the training set is returned.

Mitchell, 1997

# Decision Tree - Classification

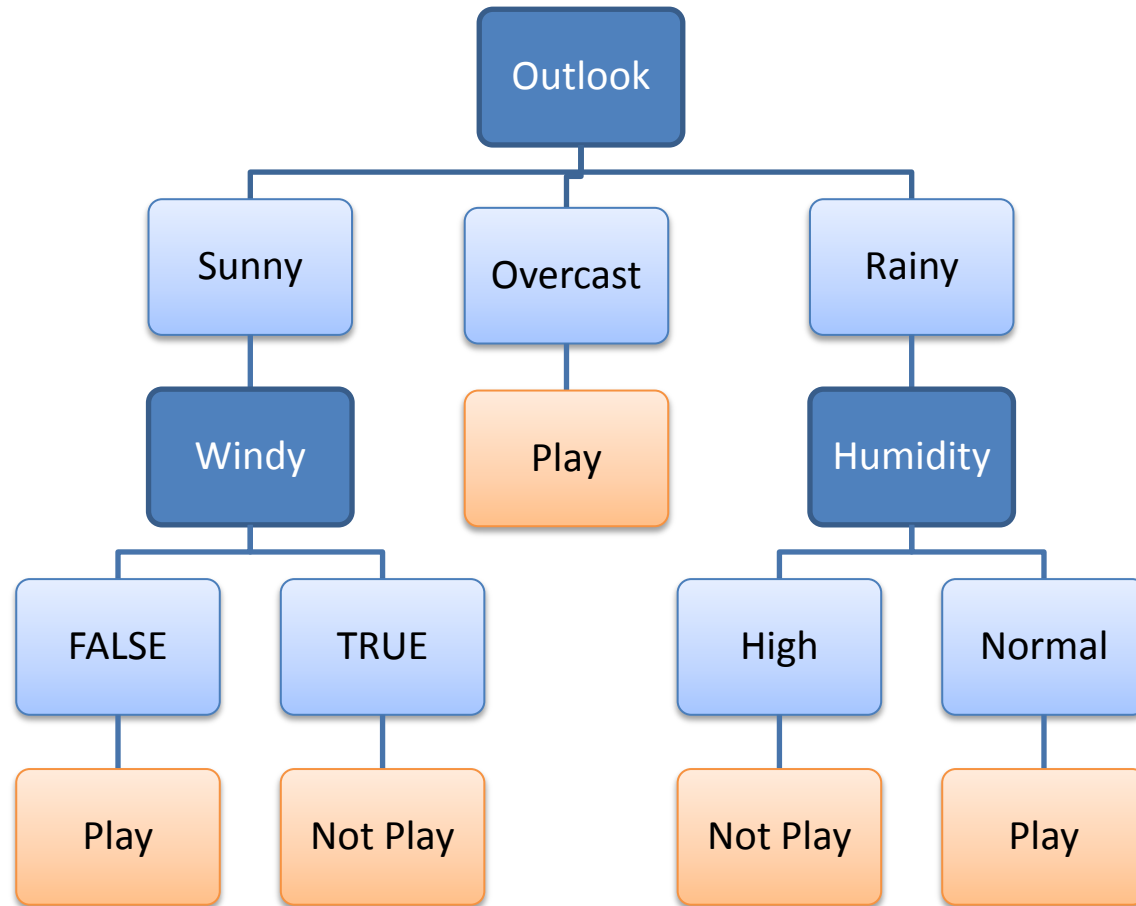
# Dataset



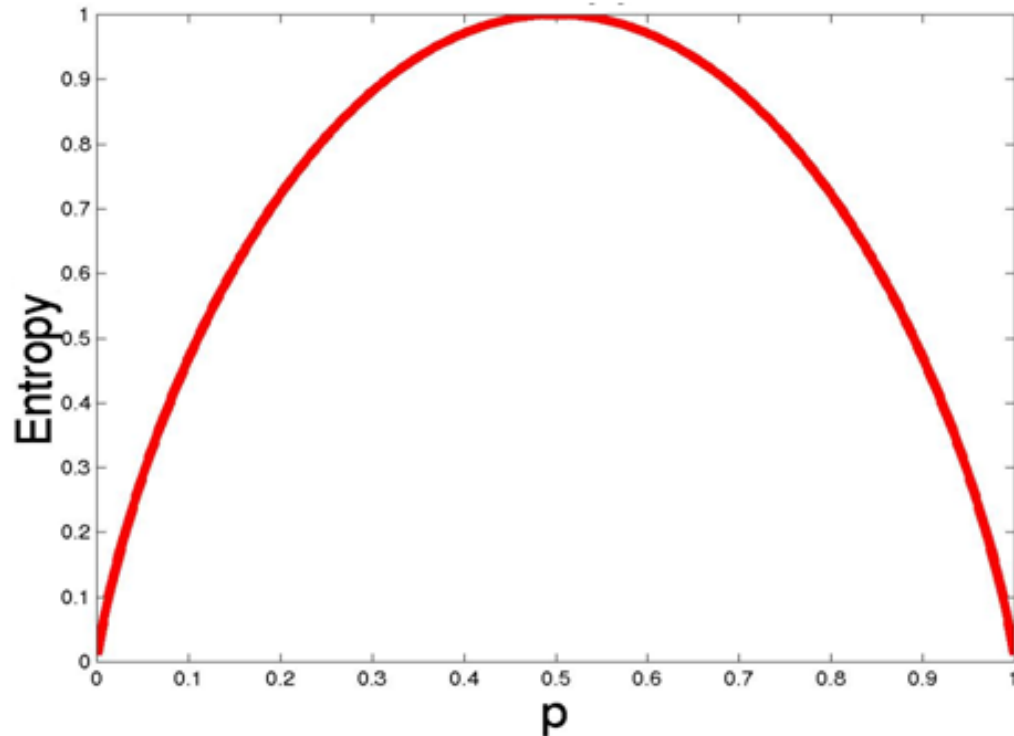
The diagram shows a table with five columns. The first four columns are grouped under a green bracket labeled 'Predictors', and the fifth column is grouped under an orange bracket labeled 'Target'.

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

# Decision Tree



# Entropy



$$\text{Entropy} = -p \log_2 p - q \log_2 q$$

$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

# Entropy – Frequency

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

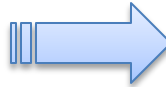
$$\begin{aligned} \text{Entropy (5,3,2)} &= \text{Entropy (0.5,0.3,0.2)} \\ &= - (0.5 * \log_2 0.5) - (0.3 * \log_2 0.3) - (0.2 * \log_2 0.2) \\ &= 1.49 \end{aligned}$$

# Entropy - Target

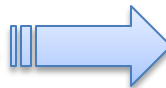
Play Golf
No
No
Yes
Yes
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
Yes
No



Play Golf
No
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes



$$5 / 14 = 0.36$$



$$9 / 14 = 0.64$$

$$\begin{aligned}\text{Entropy}(\text{PlayGolf}) &= \text{Entropy}(5,9) \\ &= \text{Entropy}(0.36, 0.64) \\ &= - (0.36 \log_2 0.36) - (0.64 \log_2 0.64) \\ &= 0.94\end{aligned}$$



# Frequency Tables

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

# Entropy – Frequency Table

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$


$$\begin{aligned} \mathbf{E}(\text{PlayGolf}, \text{Outlook}) &= \mathbf{P}(\text{Sunny}) * \mathbf{E}(3,2) + \mathbf{P}(\text{Overcast}) * \mathbf{E}(4,0) + \mathbf{P}(\text{Rainy}) * \mathbf{E}(2,3) \\ &= (5/14) * 0.971 + (4/14) * 0.0 + (5/14) * 0.971 \\ &= 0.693 \end{aligned}$$

# Information Gain

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

$$\begin{aligned} \mathbf{G}(\text{PlayGolf}, \text{Outlook}) &= \mathbf{E}(\text{PlayGolf}) - \mathbf{E}(\text{PlayGolf}, \text{Outlook}) \\ &= 0.940 - 0.693 = 0.247 \end{aligned}$$

# Information Gain – the best predictor?

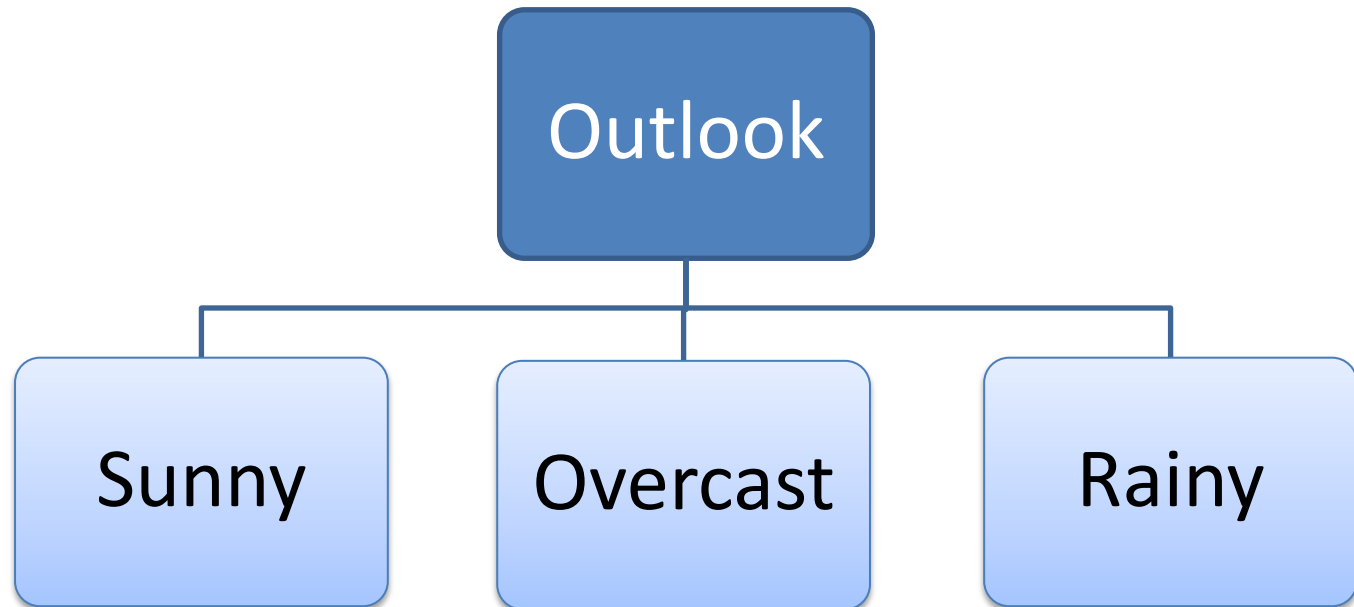
		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

# Decision Tree – Root Node



# Dataset – Sorted by Outlook

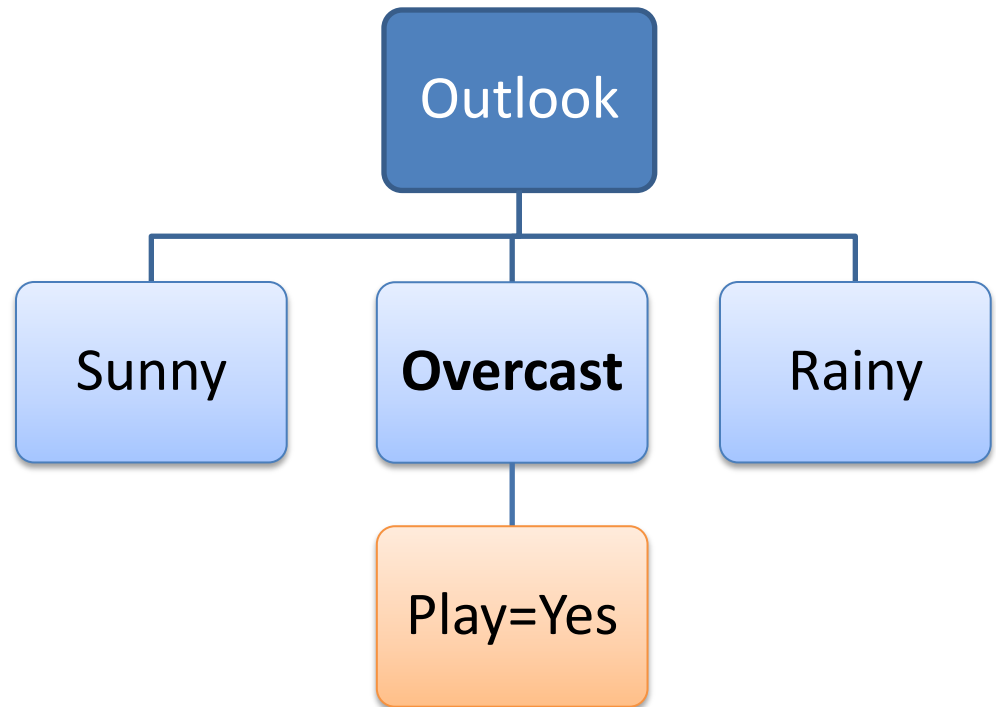
Outlook	Temp.	Humidity	Windy	Play Golf
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes

# Subset (Outlook = Overcast)

Temp.	Humidity	Windy	Play Golf
Hot	High	FALSE	Yes
Cool	Normal	TRUE	Yes
Mild	High	TRUE	Yes
Hot	Normal	FALSE	Yes
Hot	High	FALSE	Yes



# Subset (Outlook = Sunny)

Temp.	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	Normal	FALSE	Yes
Mild	High	TRUE	No

		Play Golf	
		Yes	No
Temp.	Mild	2	1
	Cool	1	1
Gain = 0.02			

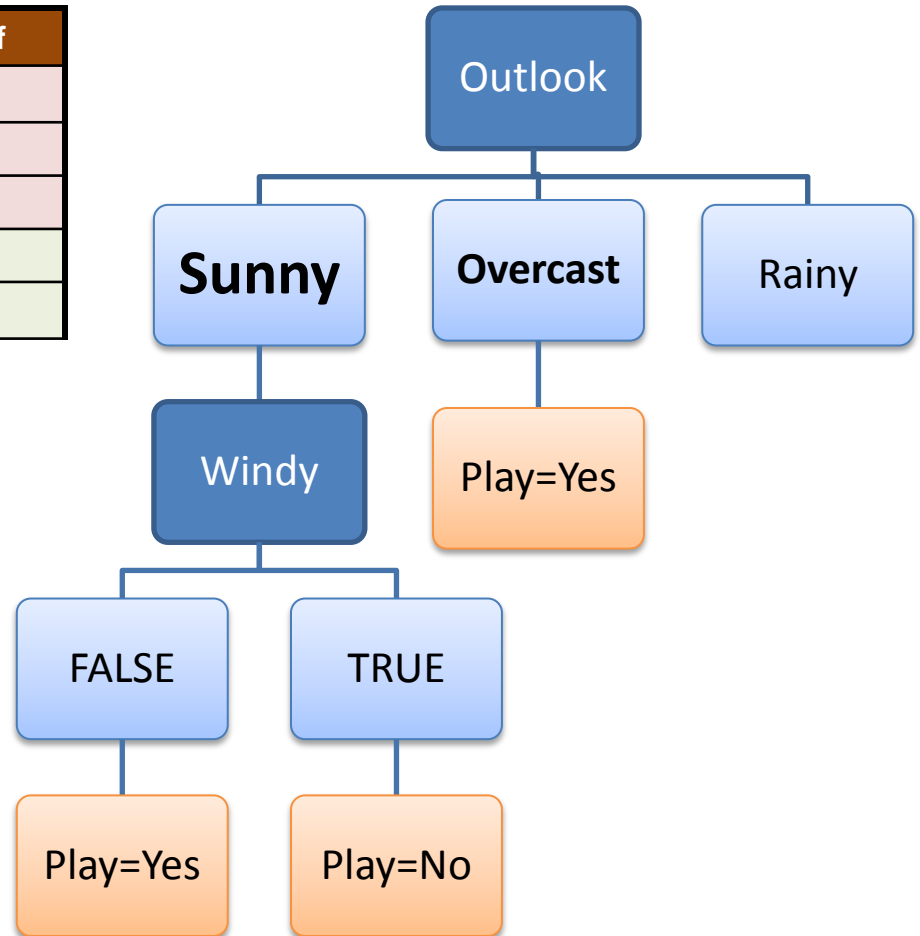
		Play Golf	
		Yes	No
Humidity	High	1	1
	Normal	2	1
Gain = 0.02			

		Play Golf	
		Yes	No
Windy	False	3	0
	True	0	2
Gain = 0.97			



# Subset (Outlook = Sunny)

Temp.	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No



# Subset (Outlook = Rainy)

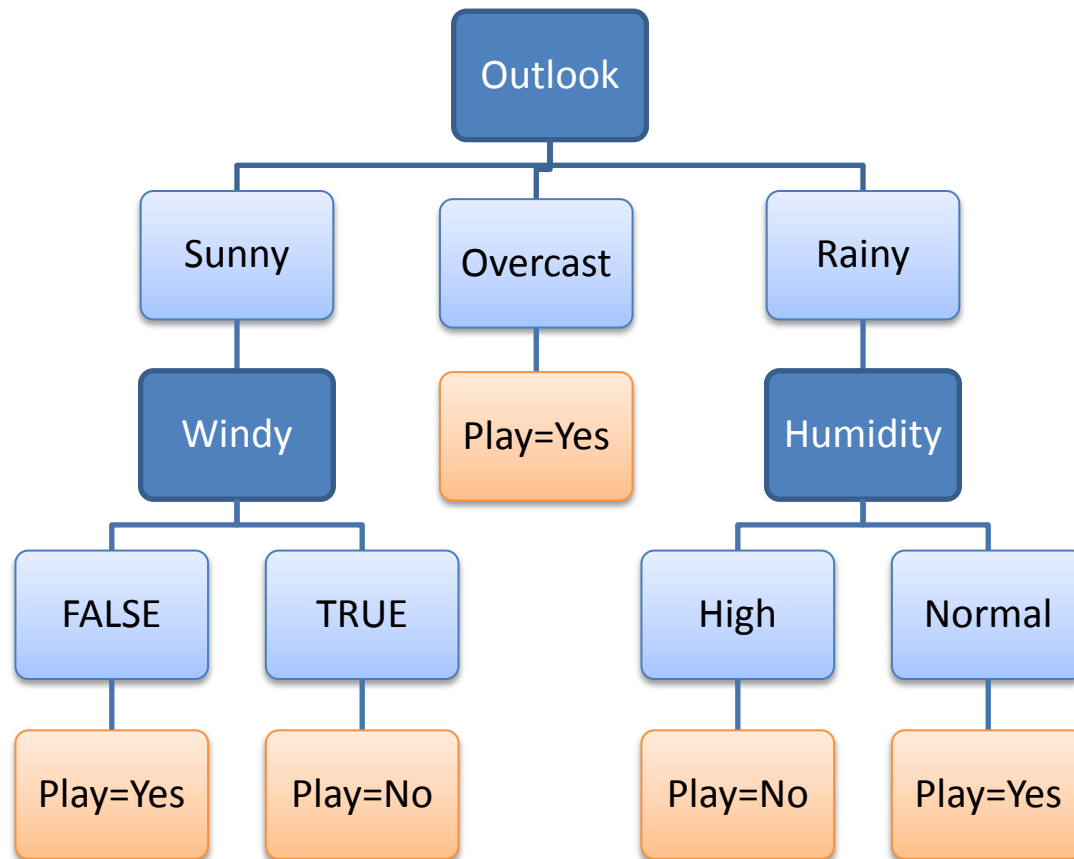
Temp.	Humidity	Windy	Play Golf
Hot	High	FALSE	No
Hot	High	TRUE	No
Mild	High	FALSE	No
Cool	Normal	FALSE	Yes
Mild	Normal	TRUE	Yes

		Play Golf	
		Yes	No
Temp.	Hot	0	2
	Mild	1	1
	Cool	1	0
Gain = 0.57			

		Play Golf	
		Yes	No
Humidity	High	0	3
	Normal	2	0
Gain = 0.97			

		Play Golf	
		Yes	No
Windy	False	1	2
	True	1	1
Gain = 0.02			

# Subset (Outlook = Rainy)



# Decision Rules

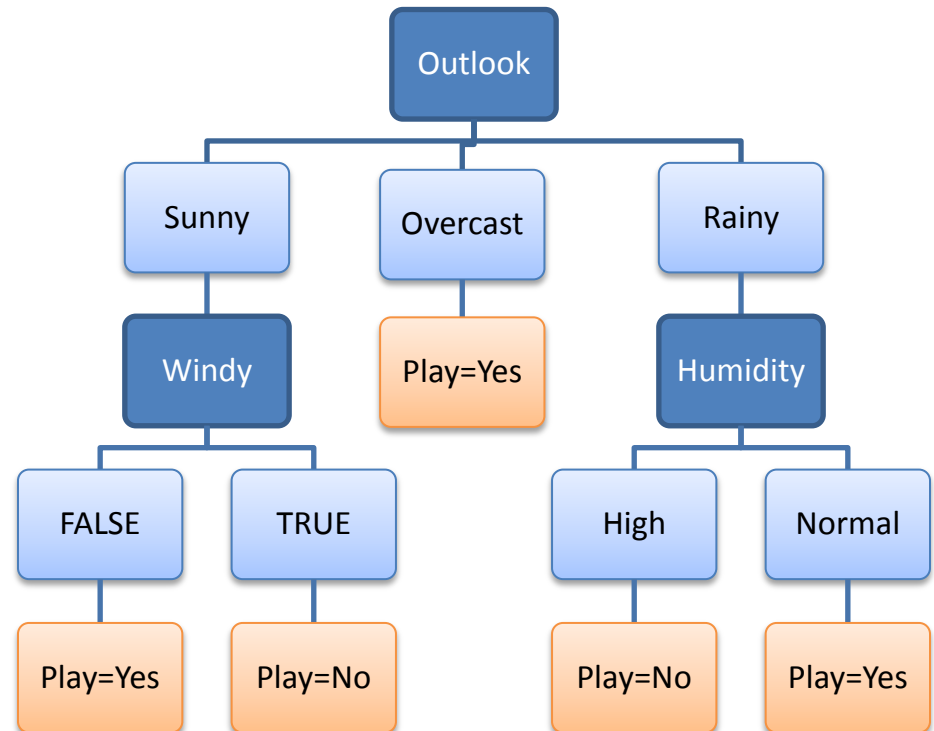
**R<sub>1</sub>:** IF (Outlook=Sunny) AND (Windy=FALSE) THEN Play=Yes

**R<sub>2</sub>:** IF (Outlook=Sunny) AND (Windy=TRUE) THEN Play=No

**R<sub>3</sub>:** IF (Outlook=Overcast) THEN Play=Yes

**R<sub>4</sub>:** IF (Outlook=Rainy) AND (Humidity=High) THEN Play=No

**R<sub>5</sub>:** IF (Outlook=Rain) AND (Humidity=Normal) THEN Play=Yes



# Decision Tree - Issues

- ❖ Working with Continuous Attributes
- ❖ Overfitting and Pruning
- ❖ Super Attributes (attributes with many values)
- ❖ Working with Missing Values
- ❖ Attributes with Different Costs

# Numeric Variables - Binning

Temp	B_Temp	Play Golf
85	80-90	No
80	80-90	No
83	80-90	Yes
70	70-80	Yes
68	60-70	Yes
65	60-70	No
64	60-70	Yes
72	70-80	No
69	60-70	Yes
75	70-80	Yes
75	70-80	Yes
72	70-80	Yes
81	80-90	Yes
71	70-80	No

		Play Golf	
		Yes	No
B_Temp	60-70	3	1
	70-80	4	2
	80-90	2	2

# Continuous Attributes - Discretization

- **Equal Frequency**

This strategy creates a set of  $N$  intervals with the same number of elements.

- **Equal width**

The original range of values is divided into  $N$  intervals with the same range.

- **Entropy based**

For each numeric attribute, instances are sorted and, for each possible threshold, a binary  $<$ ,  $\geq$  test is considered and evaluated in exactly the same way that a categorical attribute would be.

# Avoid Overfitting

**Overfitting** when our learning algorithm continues develop hypotheses that reduce training set error at the cost of an increased test set error.

- ◆ Stop growing when data split not statistically significant ( $\text{Chi}^2$  test)
- ◆ Grow full tree then post-prune
- ◆ Minimum description length (MDL):

*Minimize:*  $\text{size}(\text{tree}) + \text{size}(\text{misclassifications}(\text{tree}))$



# Avoid Overfitting - Post- Pruning

- First, build full tree then prune it.
  - Fully-grown tree shows all attribute interactions
  - Problem: some subtrees might be due to chance effects
- Two pruning operations:
  - *Subtree replacement*
  - *Subtree raising*
- Possible strategies:
  - error estimation
  - significance testing
  - MDL principle

# Error Estimation

- Transformed value for  $f$ : 
$$\frac{f - p}{\sqrt{p(1-p)/N}}$$
  
(i.e. subtract the mean and divide by the *standard deviation*)

- Resulting equation:

$$\Pr\left[-z \leq \frac{f - p}{\sqrt{p(1-p)/N}} \leq z\right] = c$$

- Solving for  $p$ :

$$p = \left( f + \frac{z^2}{2N} \pm z \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) / \left( 1 + \frac{z^2}{N} \right)$$

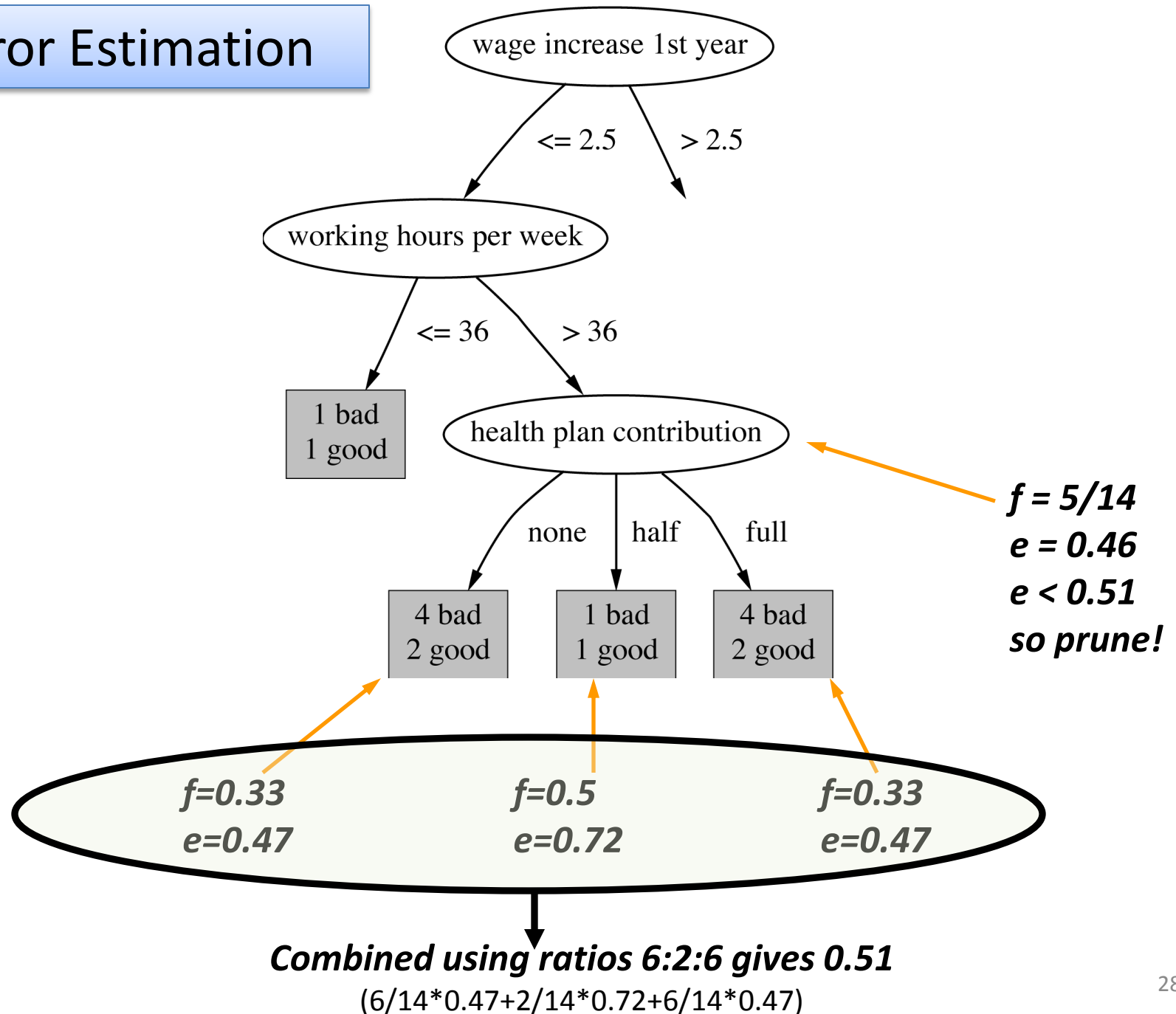
# Error Estimation

- Error estimate for subtree is weighted sum of error estimates for all its leaves
- Error estimate for a node (upper bound):

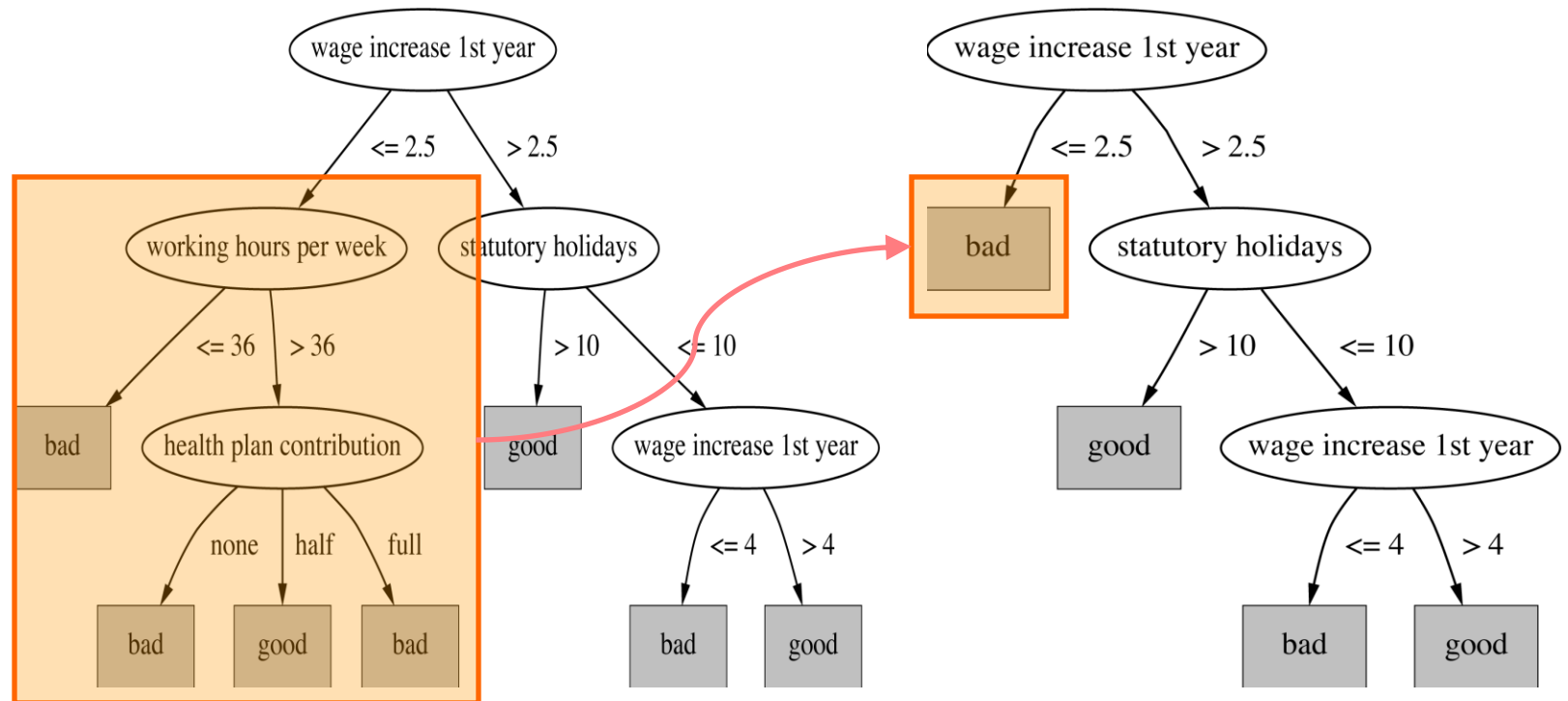
$$e = \left( f + \frac{z^2}{2N} + z \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right) / \left( 1 + \frac{z^2}{N} \right)$$

- If  $c = 25\%$  then  $z = 0.69$  (from normal distribution)
- $f$  is the error on the training data
- $N$  is the number of instances covered by the leaf

# Error Estimation



# Subtree Replacement



# Super Attributes

- ◆ The information gain equation,  $G(T,X)$  is biased toward attributes that have a large number of values over attributes that have a smaller number of values.
- ◆ These 'Super Attributes' will easily be selected as the root, result in a broad tree that classifies perfectly but performs poorly on unseen instances.
- ◆ We can penalize attributes with large numbers of values by using an alternative method for attribute selection, referred to as GainRatio.

$$\text{GainRatio}(T,X) = \text{Gain}(T,X) / \text{SplitInformation}(T,X)$$

# Super Attributes

		Play Golf		
		Yes	No	<i>total</i>
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
Gain = 0.247				

$$Split(T, X) = - \sum_{c \in A} P(c) \log_2 P(c)$$

$$\text{Split (Play, Outlook)} = - (5/14 * \log_2(5/14) + 4/14 * \log_2(4/15) + 5/14 * \log_2(5/14)) \\ = 1.577$$

$$\text{Gain Ratio (Play, Outlook)} = 0.247/1.577 = \mathbf{0.156}$$

# Super Attributes

		Play Golf		
		Yes	No	<i>total</i>
ID	id1	1	0	1
	id2	0	1	1
	id3	1	0	1
	id4	1	0	1
	id5	0	1	1
	id6	0	1	1
	id7	1	0	1
	id8	1	0	1
	id9	0	1	1
	id10	1	0	1
	id11	1	0	1
	id12	0	1	1
	id13	1	0	1
	id14	1	0	1

$$\text{Entropy}(\text{Play}, \text{ID}) = 0$$

$$\text{Gain}(\text{Play}, \text{ID}) = 0.94$$

$$\text{Split}(\text{Play}, \text{ID}) = - (1/14 * \log_2(1/14) * 14 = 3.81$$

$$\text{Gain Ratio}(\text{Play}, \text{ID}) = 0.94/3.81 = 0.247$$








# Attributes with Different Costs

- ◆ Sometimes the best attribute for splitting the training elements is very costly. In order to make the overall decision process more cost effective we may wish to penalize the information gain of an attribute by its cost.

$$G'(T, X) = \frac{G(T, X)}{Cost(X)}$$

# Numeric Variables and Missing Values

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	85	High	False	No
Rainy	80	High	True	No
Overcast	? 	High	False	Yes
Sunny	70	High	False	Yes
Sunny	68	? 	False	Yes
Sunny	65	Normal	True	No
Overcast	64	Normal	True	Yes
Rainy	72	High	? 	No
Rainy	69	Normal	False	Yes
Sunny	? 	Normal	False	Yes
Rainy	75	Normal	True	Yes
? 	72	High	True	Yes
Overcast	81	Normal	False	Yes
Sunny	71	High	True	No

# Missing Values

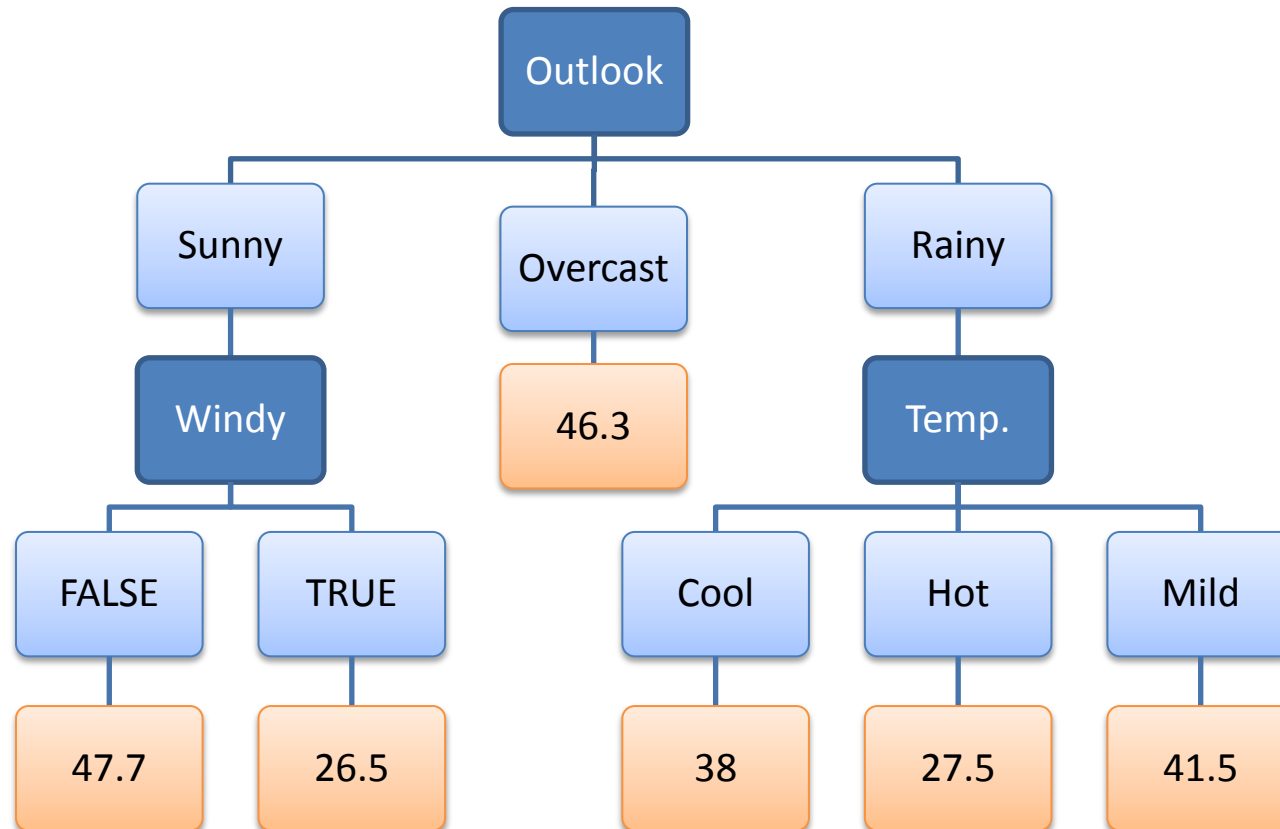
- For the numerical variables replace the missing value with the average or median.
- For the categorical variables replace the missing with.
  - Most common value
  - Most common value at node K
- K Nearest Neighbors (KNN)

# Decision Tree - Regression

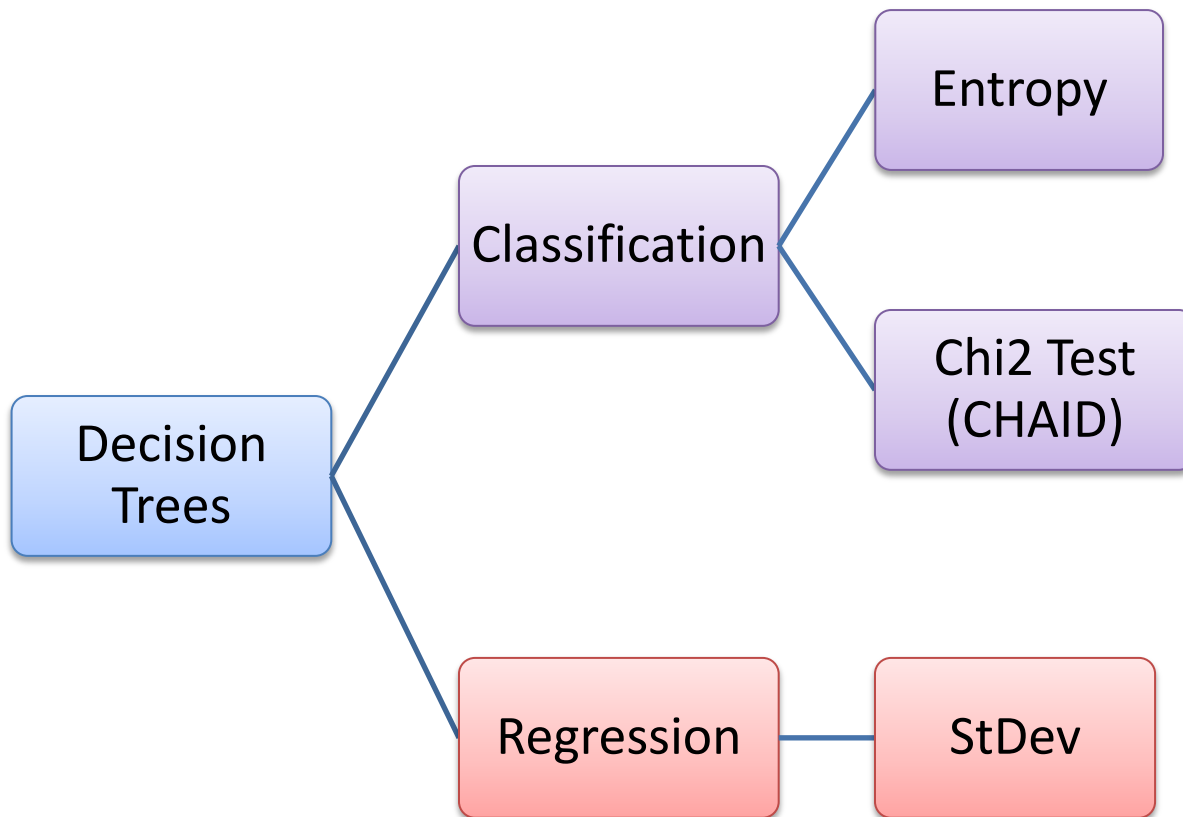
# Dataset

Predictors				Target
Outlook	Temp.	Humidity	Windy	Golf Players
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	46
Sunny	Mild	High	False	45
Sunny	Cool	Normal	False	52
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	46
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

# Decision Tree - Regression



# Entropy versus Standard Deviation



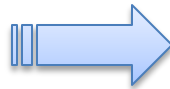
$$E = \sum_{i=1}^c -p_i \log_2 p_i$$

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$S = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

# Target – Standard Deviation & Average

Golf Players
25
30
46
45
52
23
43
35
38
46
48
52
44
30



StDev = 9.32  
Avg = 39.79



# Standard Deviation Tables

		Golf Players (StDev)
Outlook	Overcast	3.49
	Rainy	7.78
	Sunny	10.87

		Golf Players (StDev)
Temp.	Cool	10.51
	Hot	8.95
	Mild	7.65

		Golf Players (StDev)
Humidity	High	9.36
	Normal	8.37

		Golf Players (StDev)
Windy	False	7.87
	True	10.59

# Standard Deviation

		Golf Players (StDev)	Count
Outlook	Overcast	3.49	4
	Rainy	7.78	5
	Sunny	10.87	5
			14

$$S(T, X) = \sum_{c \in X} P(c)S(c)$$


$$\begin{aligned} \mathbf{S}(\text{Players}, \text{Outlook}) &= \mathbf{P}(\text{Sunny}) * \mathbf{S}(\text{Sunny}) + \mathbf{P}(\text{Overcast}) * \mathbf{S}(\text{Overcast}) + \mathbf{P}(\text{Rainy}) * \mathbf{S}(\text{Rainy}) \\ &= (4/14) * 3.49 + (5/14) * 7.78 + (5/14) * 10.87 \\ &= 7.66 \end{aligned}$$

# Standard Deviation Reduction (SDR)

$$SDR(T, X) = S(T) - S(T, X)$$

$$\begin{aligned} \mathbf{G}(\text{Players}, \text{Outlook}) &= \mathbf{S}(\text{Players}) - \mathbf{S}(\text{Players}, \text{Outlook}) \\ &= 9.32 - 7.66 = 1.66 \end{aligned}$$

# Standard Deviation Reduction the best predictor?

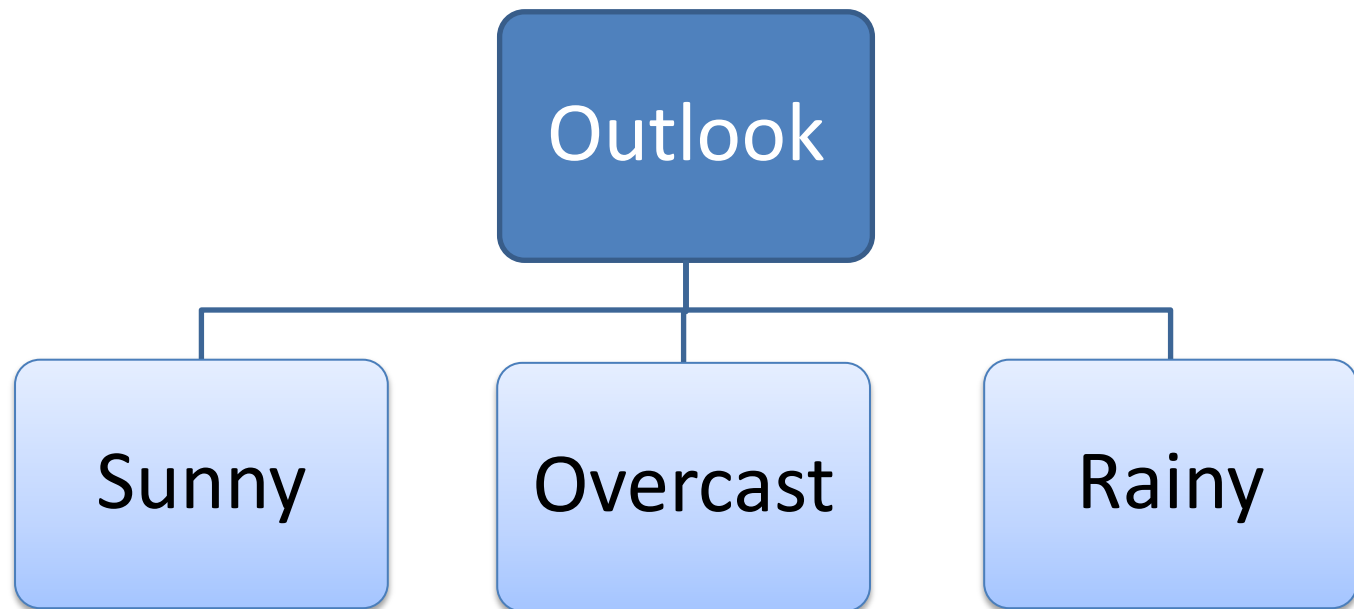
		Golf Players (StDev)
Outlook	Overcast	3.49
	Rainy	7.78
	Sunny	10.87
SDR=1.66		

		Golf Players (StDev)
Temp.	Cool	10.51
	Hot	8.95
	Mild	7.65
SDR=0.17		

		Golf Players (StDev)
Humidity	High	9.36
	Normal	8.37
SDR=0.28		

		Golf Players (StDev)
Windy	False	7.87
	True	10.59
SDR=0.29		

# Decision Tree – Root Node



# Dataset – Sorted by Outlook

Outlook	Temp.	Humidity	Windy	Golf Players
Sunny	Mild	High	FALSE	45
Sunny	Cool	Normal	FALSE	52
Sunny	Cool	Normal	TRUE	23
Sunny	Mild	Normal	FALSE	46
Sunny	Mild	High	TRUE	30
Rainy	Hot	High	FALSE	25
Rainy	Hot	High	TRUE	30
Rainy	Mild	High	FALSE	35
Rainy	Cool	Normal	FALSE	38
Rainy	Mild	Normal	TRUE	48
Overcast	Hot	High	FALSE	46
Overcast	Cool	Normal	TRUE	43
Overcast	Mild	High	TRUE	52
Overcast	Hot	Normal	FALSE	44

# Subset (Outlook = Sunny)

Temp.	Humidity	Windy	Golf Players
Mild	High	FALSE	45
Cool	Normal	FALSE	52
Cool	Normal	TRUE	23
Mild	Normal	FALSE	46
Mild	High	TRUE	30
			SD=10.87

		Golf Players (StDev)
Temp.	Cool	14.50
	Mild	7.32
SDR= 0.678		

		Golf Players (StDev)
Humidity	High	7.50
	Normal	12.50
SDR= 0.370		

		Golf Players (StDev)
Windy	False	3.09
	True	3.50
SDR= 7.62		

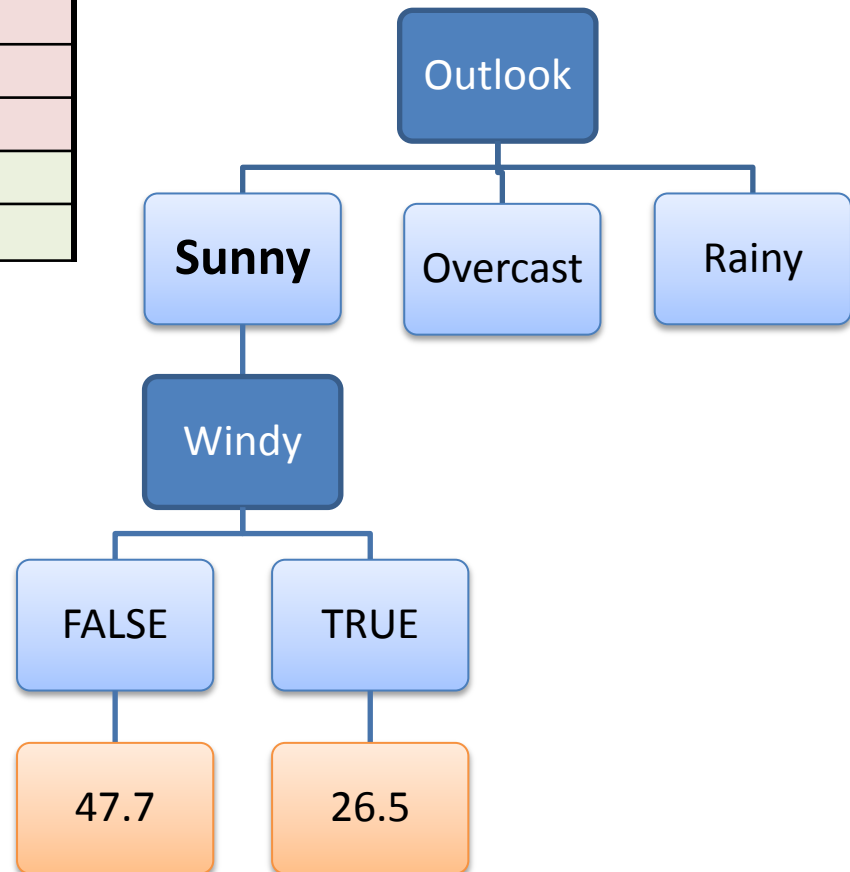
$$\text{SDR} = 10.87 - ((2/5) * 14.5 + (3/5) * 7.32)$$

$$\text{SDR} = 10.87 - ((2/5) * 7.5 + (3/5) * 12.5)$$

$$\text{SDR} = 10.87 - ((3/5) * 3.09 + (2/5) * 3.5)$$

# Subset (Outlook = Sunny)

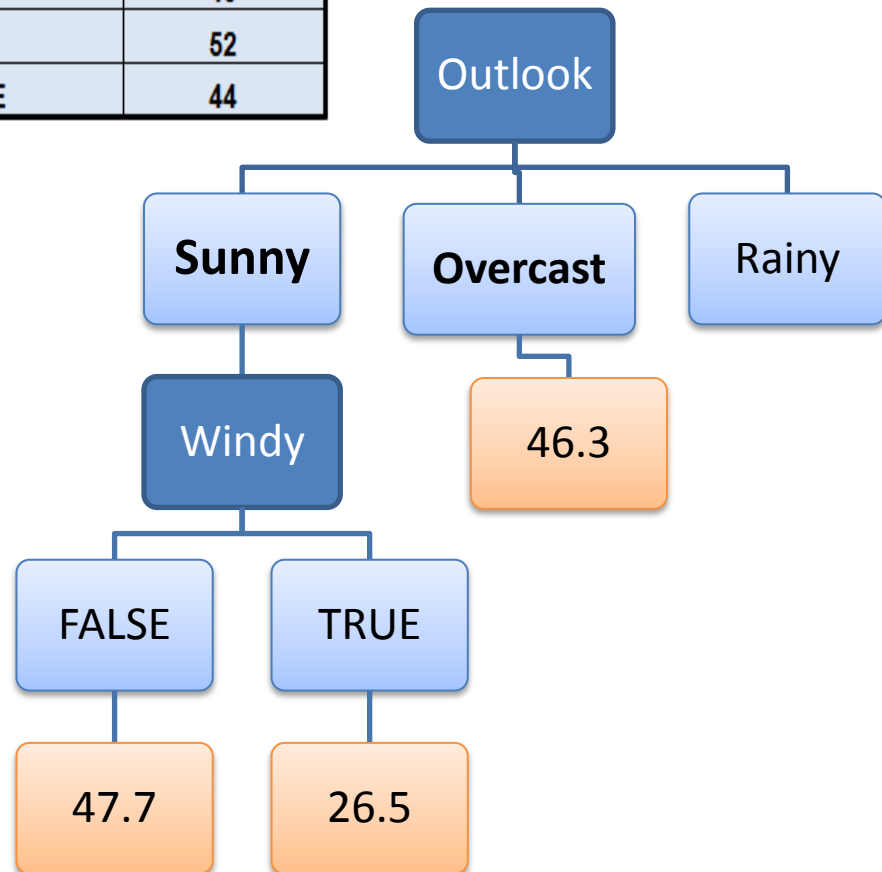
Temp.	Humidity	Windy	Golf Players
Mild	High	FALSE	45
Cool	Normal	FALSE	52
Mild	Normal	FALSE	46
Cool	Normal	TRUE	23
Mild	High	TRUE	30





# Subset (Outlook = Overcast)

Overcast	Hot	High	FALSE	46
Overcast	Cool	Normal	TRUE	43
Overcast	Mild	High	TRUE	52
Overcast	Hot	Normal	FALSE	44



# Subset (Outlook = Rainy)

Temp.	Humidity	Windy	Golf Players
Hot	High	FALSE	25
Hot	High	TRUE	30
Mild	High	FALSE	35
Cool	Normal	FALSE	38
Mild	Normal	TRUE	48
			StDev=7.78

★		Golf Players (StDev)
Temp.	Cool	0
	Hot	2.5
	Mild	6.5
SDR= 4.18		

$$\text{SDR} = 7.78 - ((2/5)*2.5 + (2/5)*6.5)$$

		Golf Players (StDev)
Humidity	High	4.1
	Normal	5.0
SDR= 3.32		

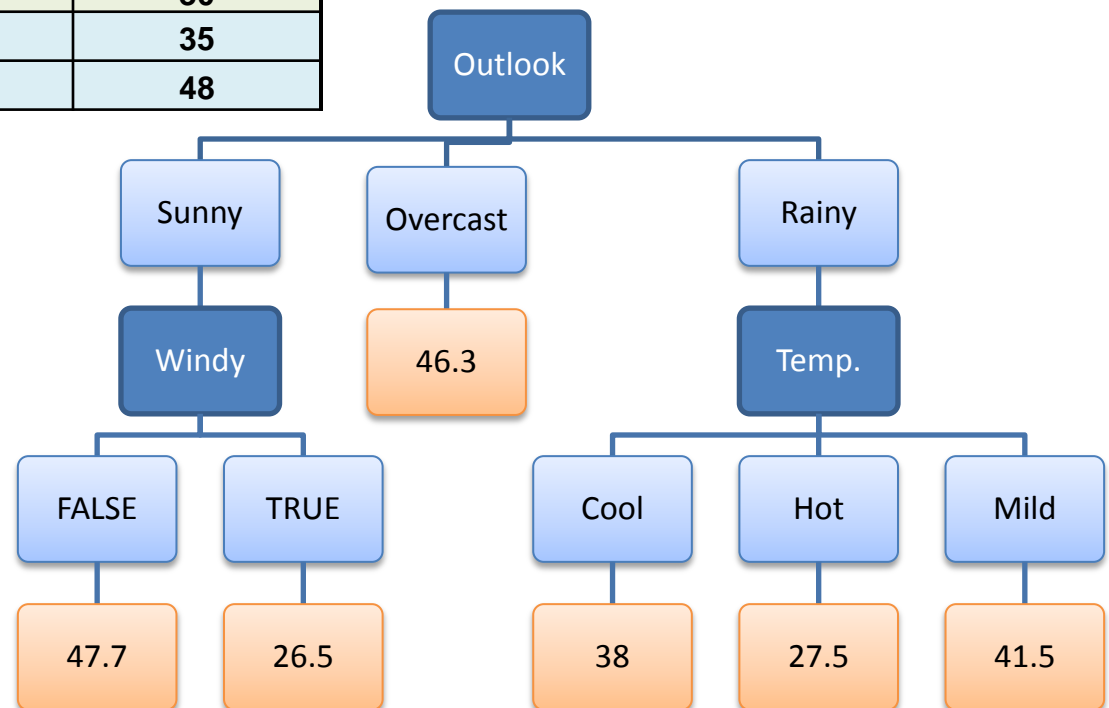
$$\text{SDR} = 7.78 - ((3/5)*4.1 + (2/5)*5.0)$$

		Golf Players (StDev)
Windy	False	5.6
	True	9.0
SDR= 0.82		

$$\text{SDR} = 7.78 - ((3/5)*5.6 + (2/5)*9.0)$$

# Subset (Outlook = Rainy)

Temp.	Humidity	Windy	Golf Players
Cool	Normal	FALSE	38
Hot	High	FALSE	25
Hot	High	TRUE	30
Mild	High	FALSE	35
Mild	Normal	TRUE	48



# Questions?