Phân tích dữ liệu thông minh

Decision Tree

TS. Nguyễn Tiến Huy

ntienhuy@fit.hcmus.edu.vn

Nội dung

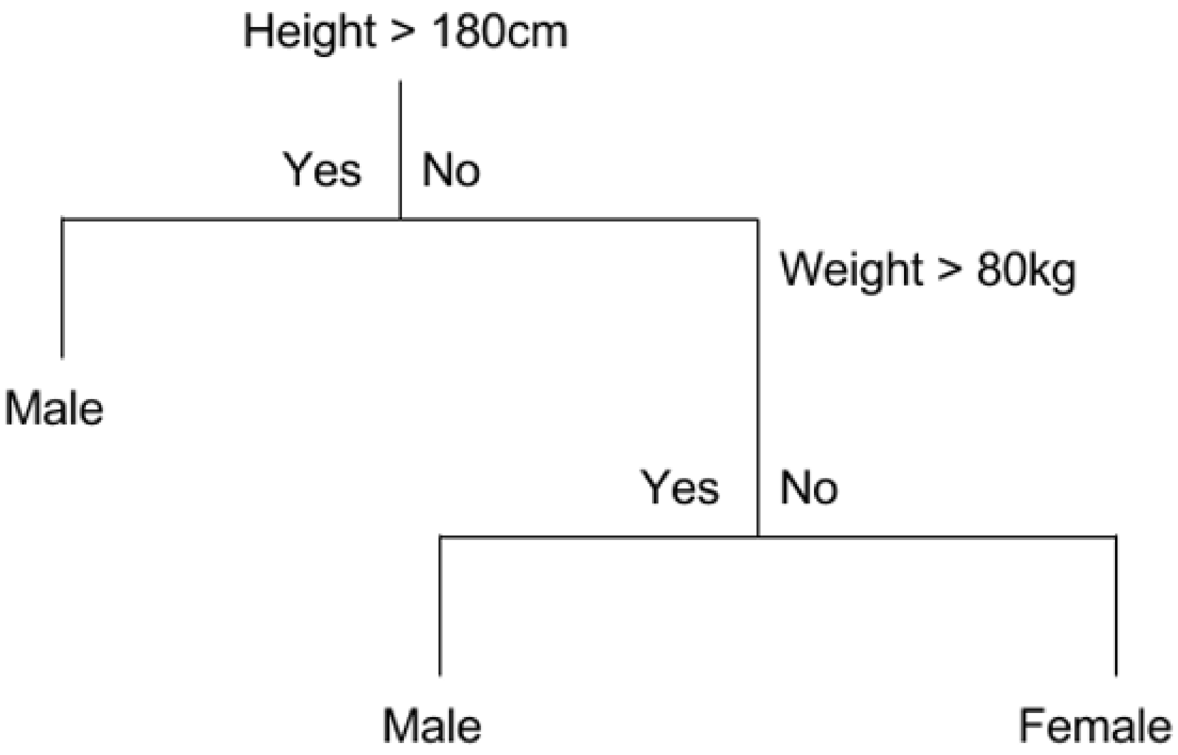
1. Problem introduction
2. Decision Tree for classification
3. Decision Tree for regression
4. Pruning Tree

Gender prediction

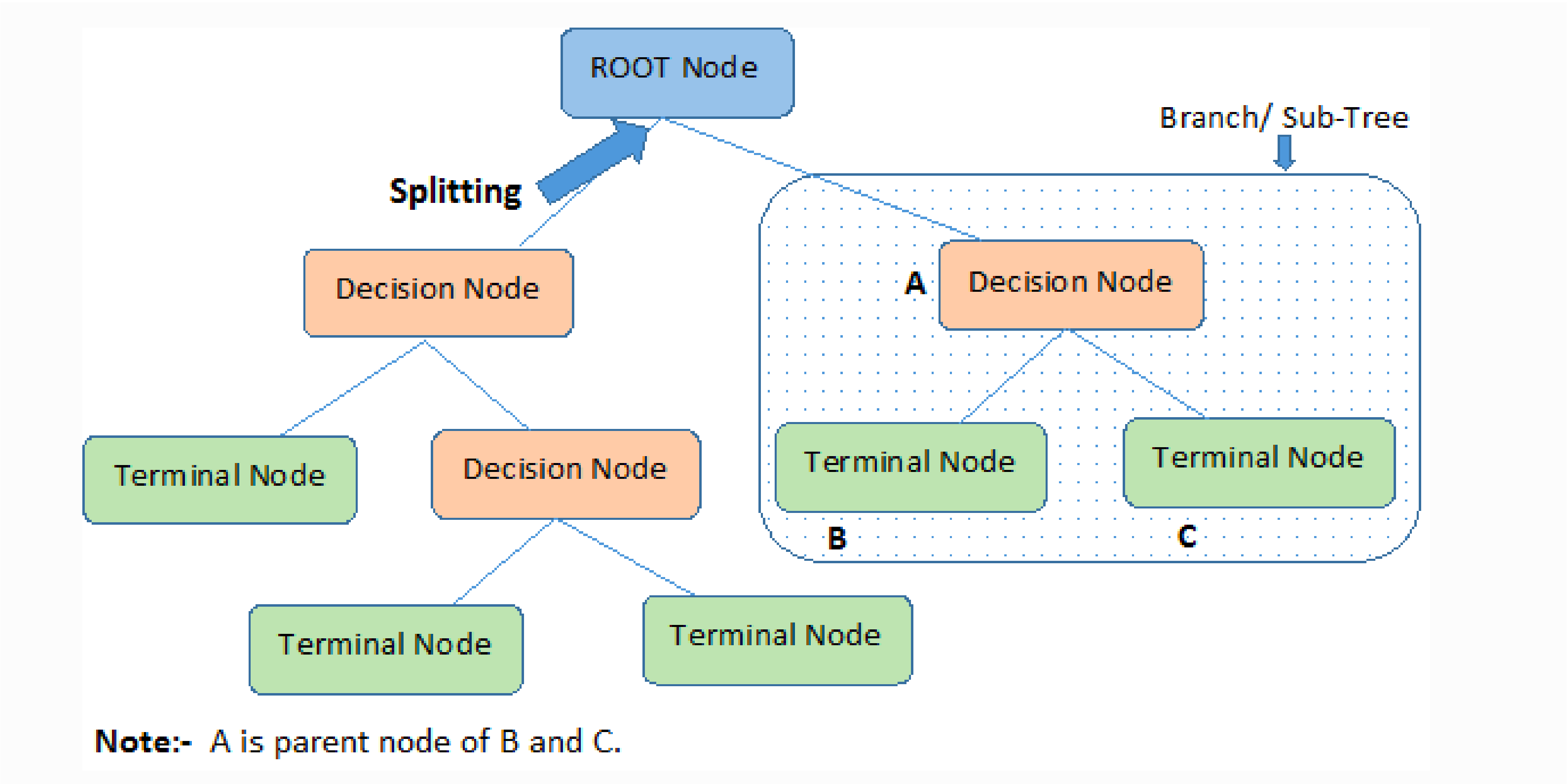
#No	Age	Weight	Height	Gender
1	54	65	181	Male
2	19	82	173	Male
3	24	54	165	Female
4	54	57	170	Female
5	18	81	185	Male
6	51	45	155	Female

Gender prediction

#No	Age	Weight	Height	Gender
1	54	65	181	Male
2	19	82	173	Male
3	24	54	165	Female
4	54	57	170	Female
5	18	81	185	Male
6	51	45	155	Female



Decision Tree



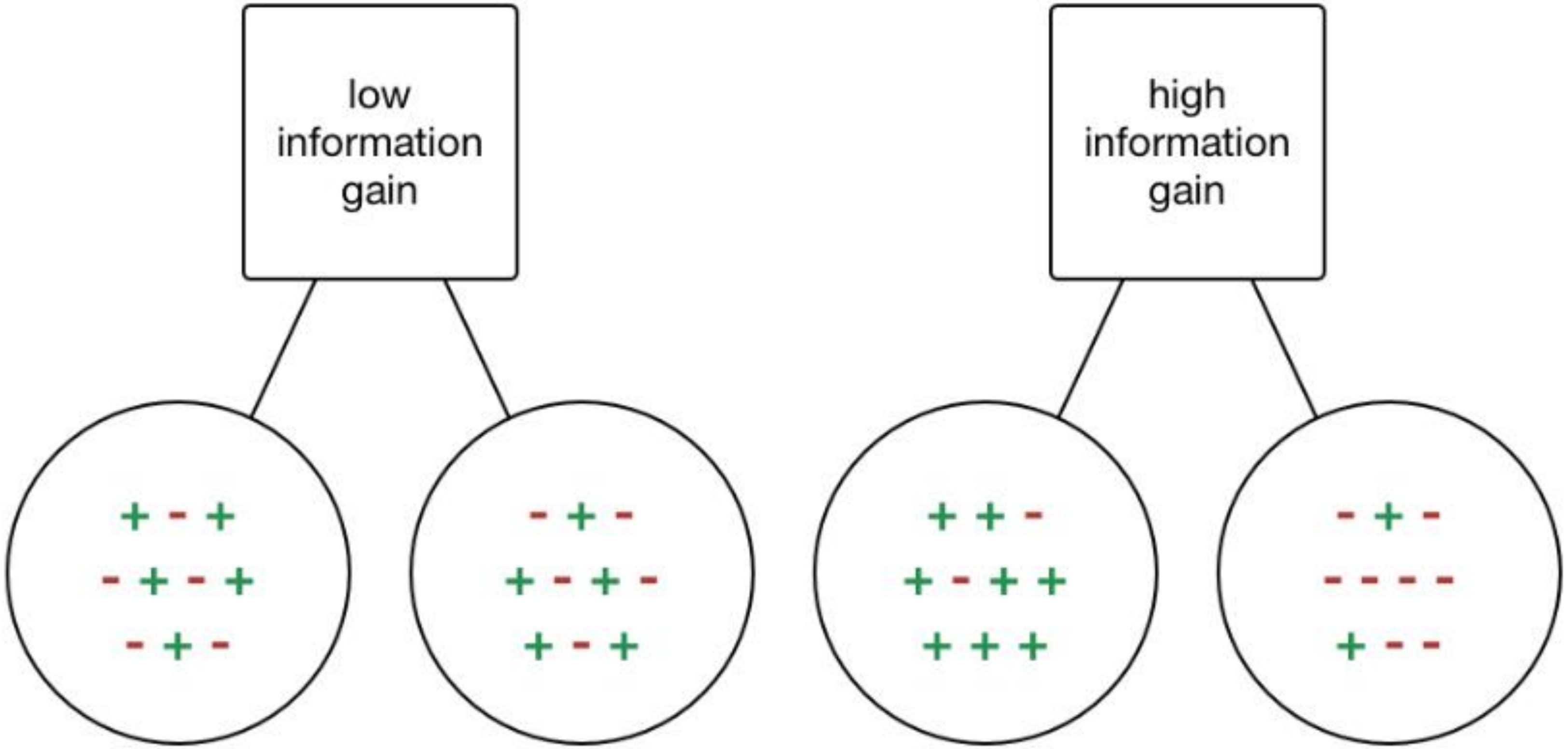
Note:- A is parent node of B and C.

How to build a decision tree

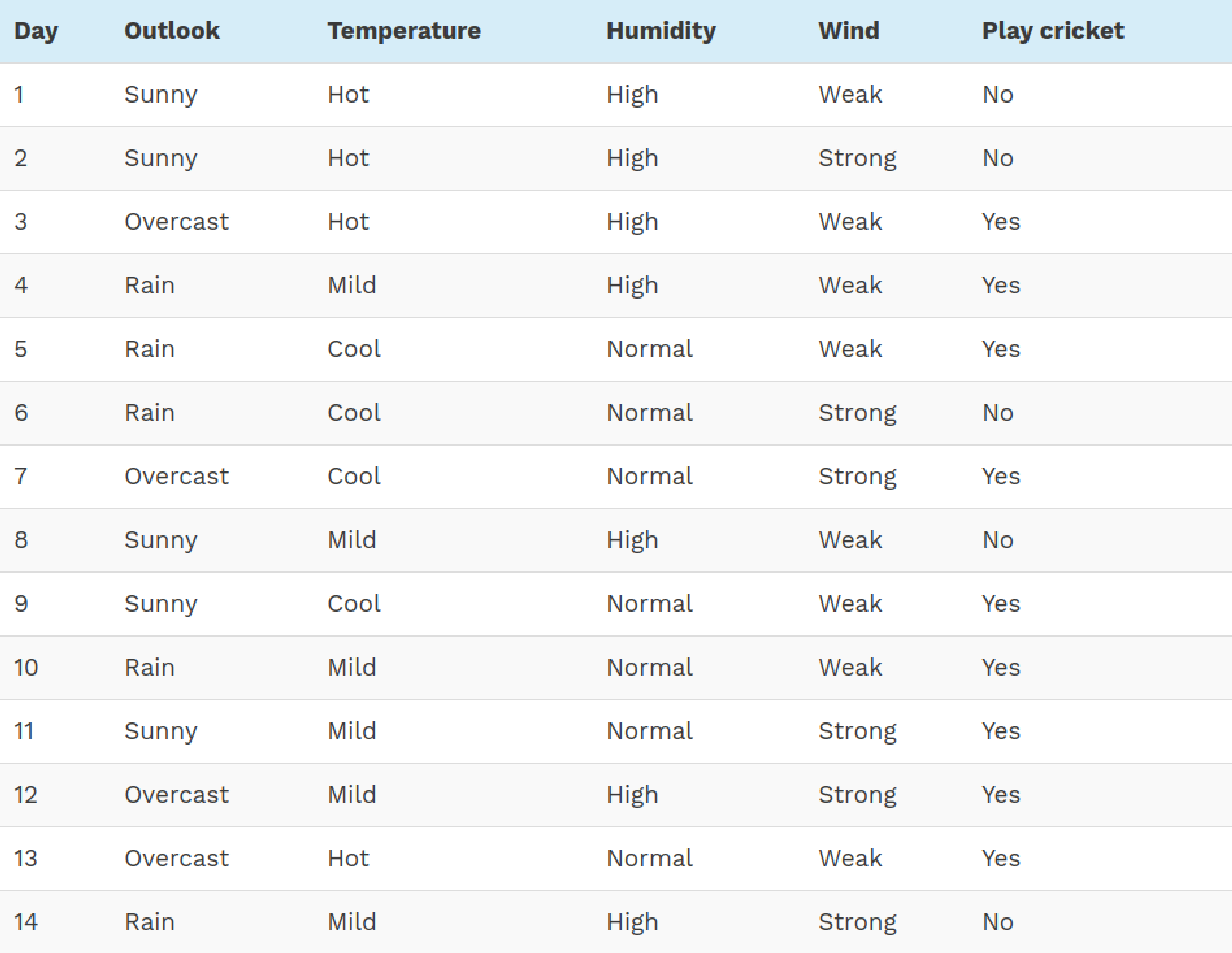
- CART (Classification and Regression Trees) → uses ***Gini Index(Classification)*** as metric.

- ID3 (Iterative Dichotomiser 3) → uses ***Entropy function*** and ***Information gain*** as metrics.

Entropy and Information Gain

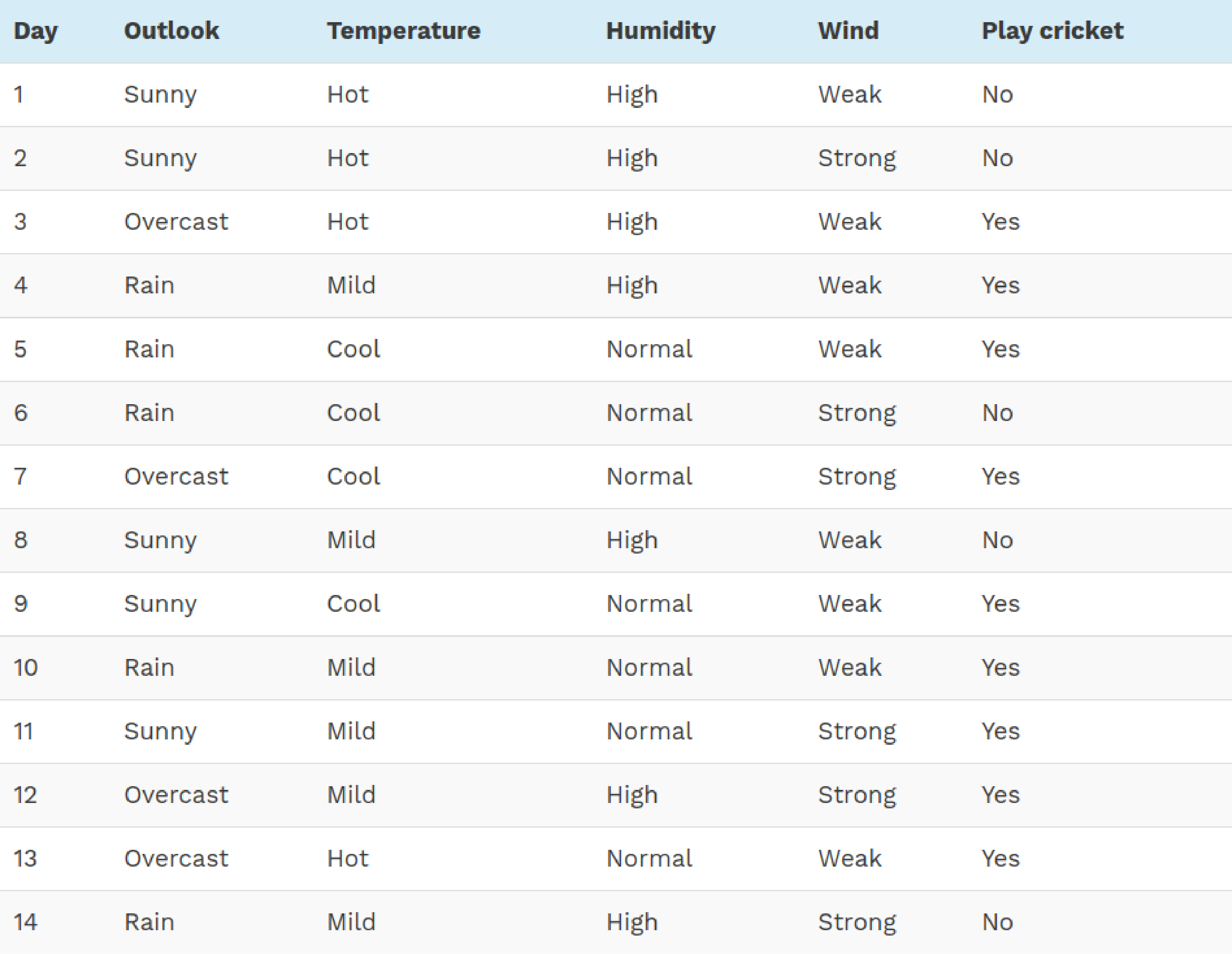


Weather prediction using Information Gain



Day	Outlook	Temperature	Humidity	Wind	Play cricket
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Weather prediction



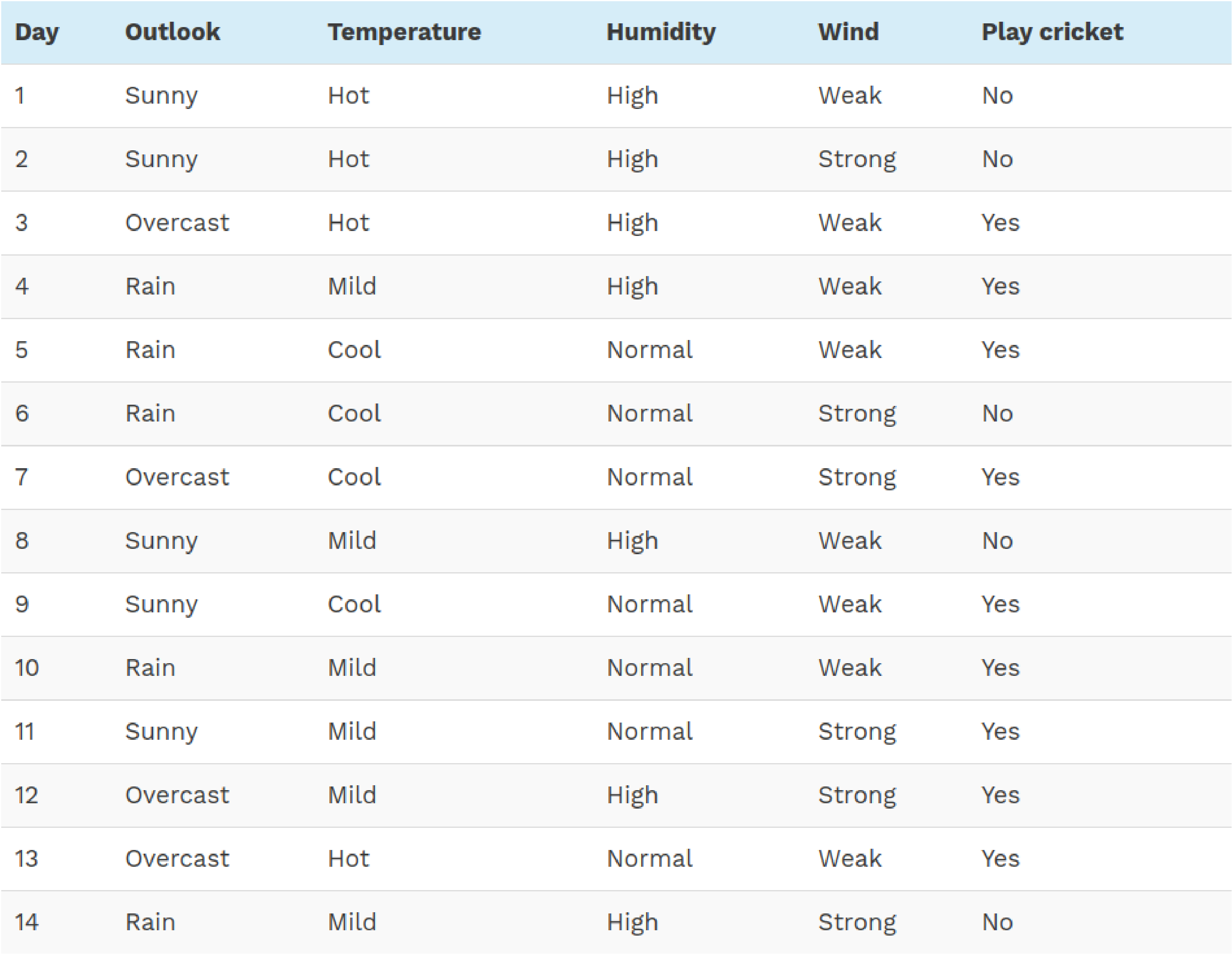
Day	Outlook	Temperature	Humidity	Wind	Play cricket
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$H(S) = \sum_{c \in C} -p(c) \log_2 p(c)$$

$$H(S) = -\left(\frac{9}{14}\right) log_2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) log_2 \left(\frac{5}{14}\right)$$

$$H(S) = 0.94$$

Weather prediction



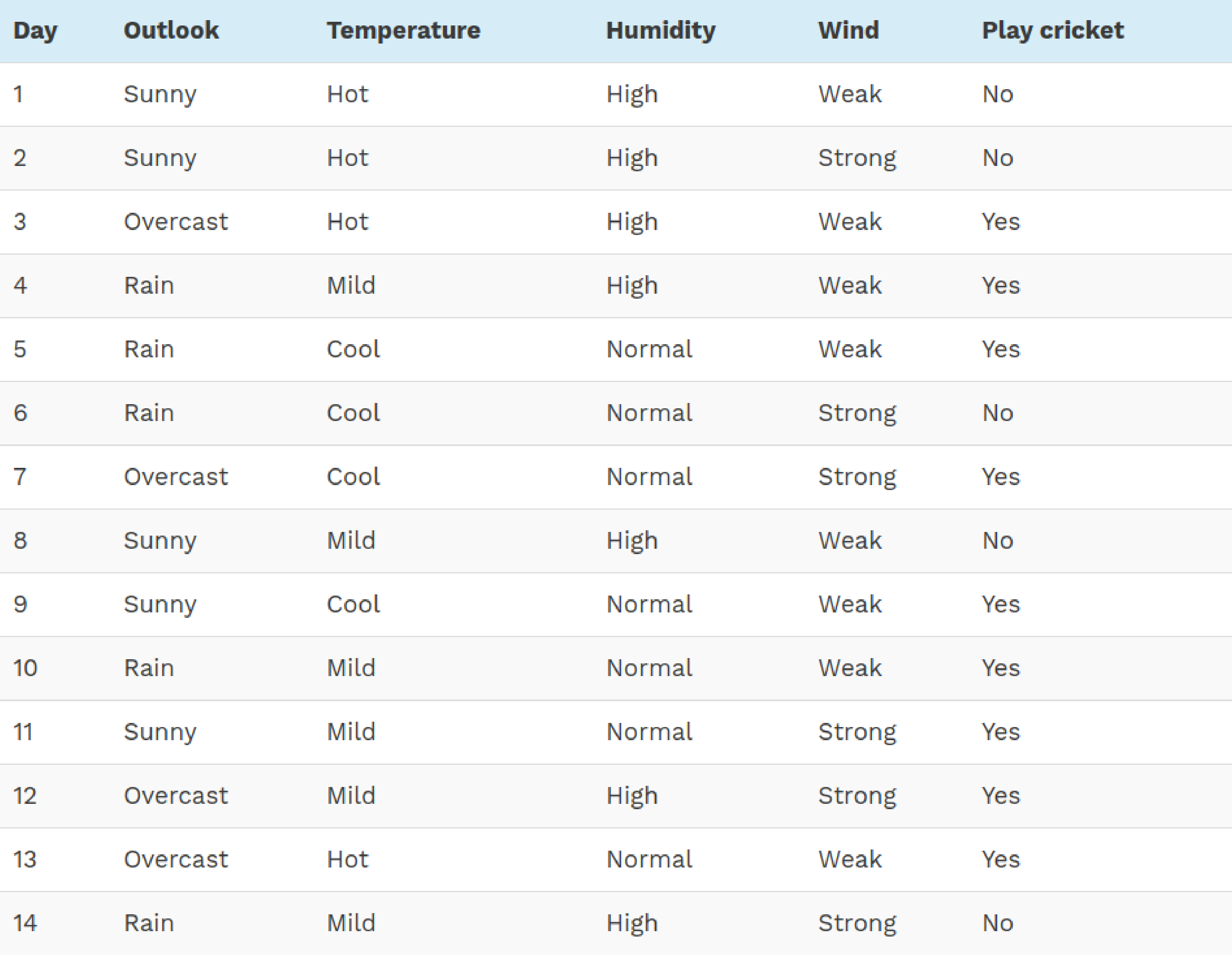
Day	Outlook	Temperature	Humidity	Wind	Play cricket
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$IG(S, Wind) = H(S) - \sum_{t \in T} P(t) * H(t)$$

$$P(S_{weak}) = \frac{\text{Number of weak}}{Total} = \frac{8}{14}$$

$$P(S_{strong}) = \frac{\text{Number of strong}}{Total} = \frac{6}{14}$$

Weather prediction



Day	Outlook	Temperature	Humidity	Wind	Play cricket
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$IG(S, Wind) = H(S) - \sum_{t\in T} P(t) * H(t)$$

$$\begin{aligned} H(S_{weak}) &= -\left(\frac{6}{8}\right) log_2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right) log_2\left(\frac{2}{8}\right) \\ &= 0.811 \end{aligned}$$

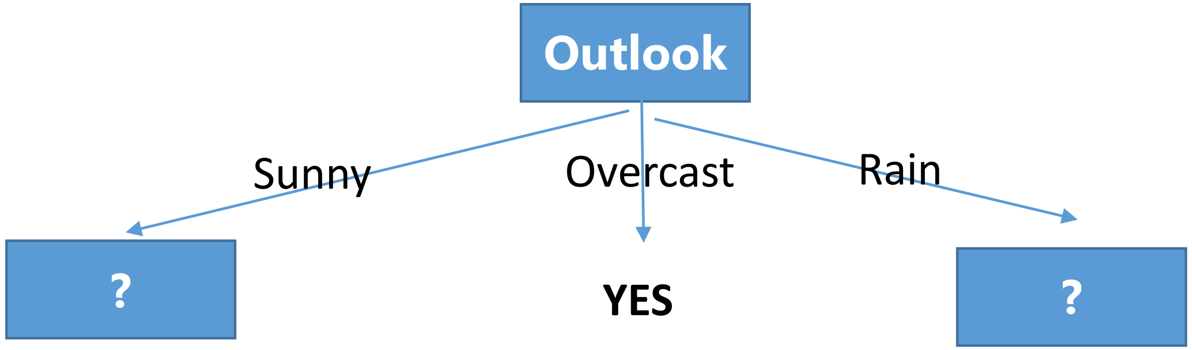
$$\begin{aligned} H(S_{strong}) &= -\left(\frac{3}{6}\right) log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right) log_2\left(\frac{3}{6}\right) \\ &= 1 \end{aligned}$$

Weather prediction

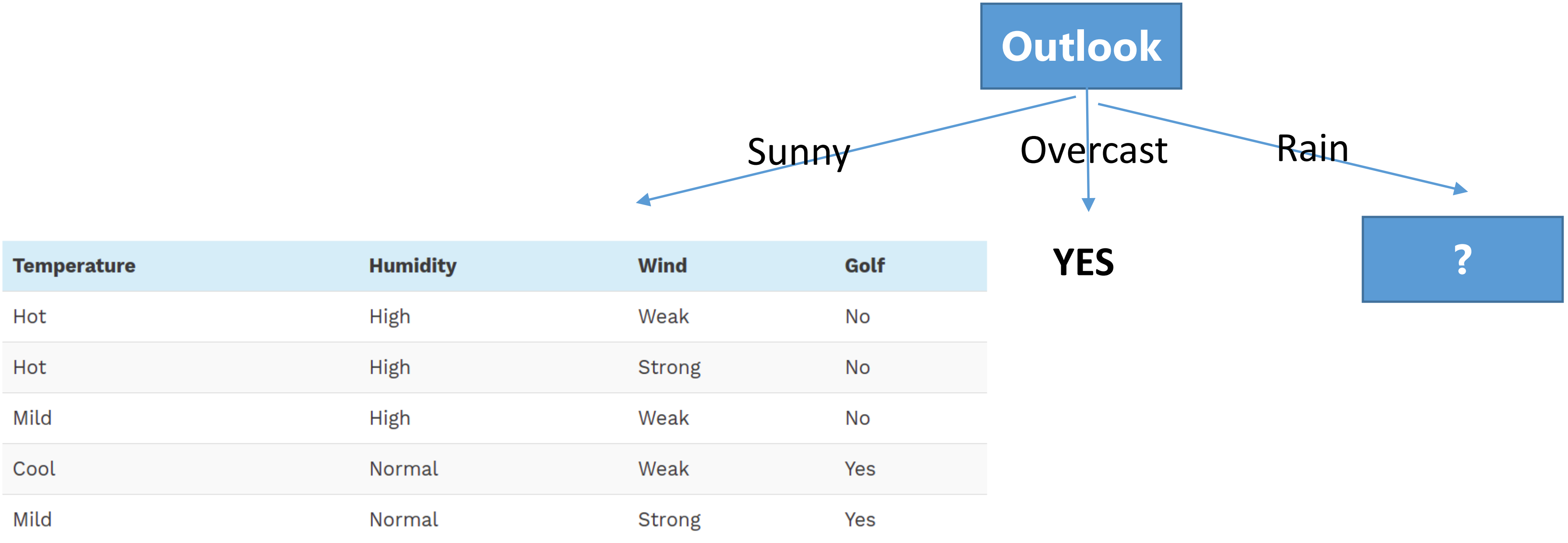
Day	Outlook	Temperature	Humidity	Wind	Play cricket
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$
\begin{aligned}
IG(S, Wind) &= H(S) - \sum_{t \in T} P(t) * H(t) \\
&= H(S) - P(S_{weak}) * H(S_{weak}) - P(S_{strong}) * H(S_{strong}) \\
&= 0.94 - \left(\frac{8}{14}\right)(0.811) - \left(\frac{6}{14}\right)(1.0) \\
&= 0.048
\end{aligned}
$$

Weather prediction



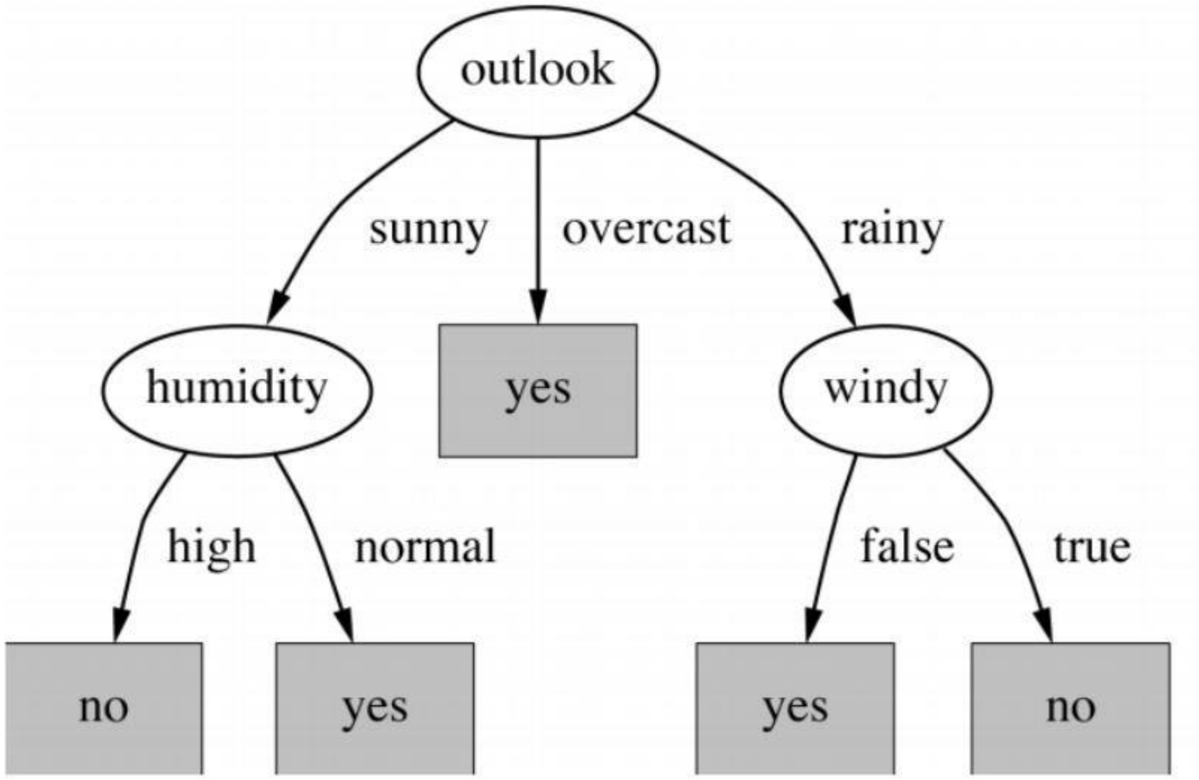
Weather prediction



Temperature	Humidity	Wind	Golf
Hot	High	Weak	No
Hot	High	Strong	No
Mild	High	Weak	No
Cool	Normal	Weak	Yes
Mild	Normal	Strong	Yes

Weather prediction

Final decision tree



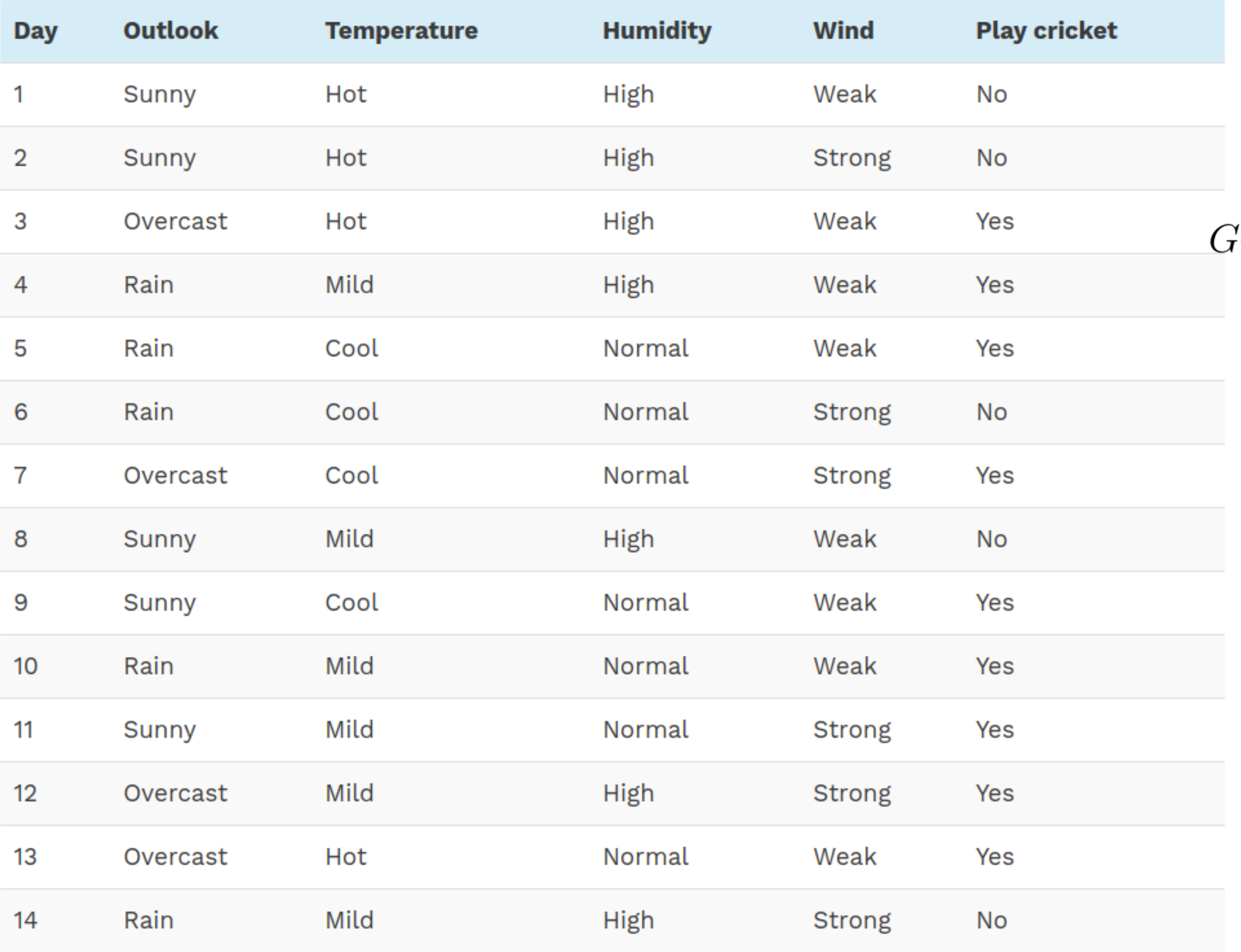
Weather prediction using Gini index

Day	Outlook	Temperature	Humidity	Wind	Play cricket
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$Gini(S, Outlook) = \sum_{i \in T} P(i) * Gini(i)$$

$$Gini(i) = 1 - \sum_{c \in C} P(c)^2$$

Weather prediction using Gini index



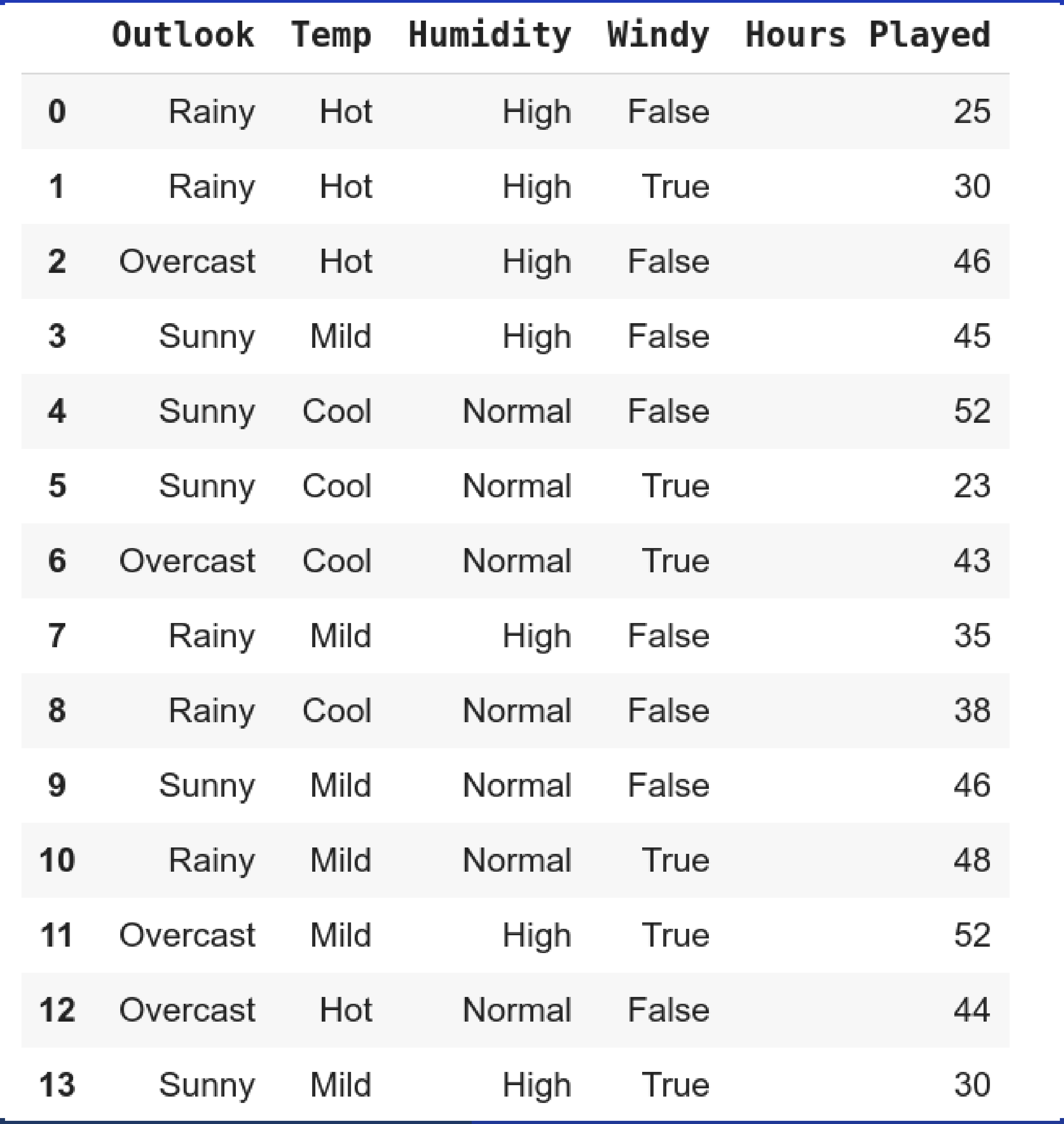
Day	Outlook	Temperature	Humidity	Wind	Play cricket
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$\begin{aligned}
Gini(Outlook = Sunny) &= 1 - P(Yes)^2 - P(No)^2 \\
&= 1 - (2/5)^2 - (3/5)^2 \\
&= 0.48 \\
Gini(Outlook = Overcast) &= 1 - (4/4)^2 - (0/4)^2 \\
&= 0 \\
Gini(Outlook = Rain) &= 1 - (3/5)^2 - (2/5)^2 \\
&= 0.48 \\
Gini(S, Outlook) &= (5/14) * 0.48 + (4/14) * 0 + (5/14) * 0.48 \\
&= 0.342
\end{aligned}$$

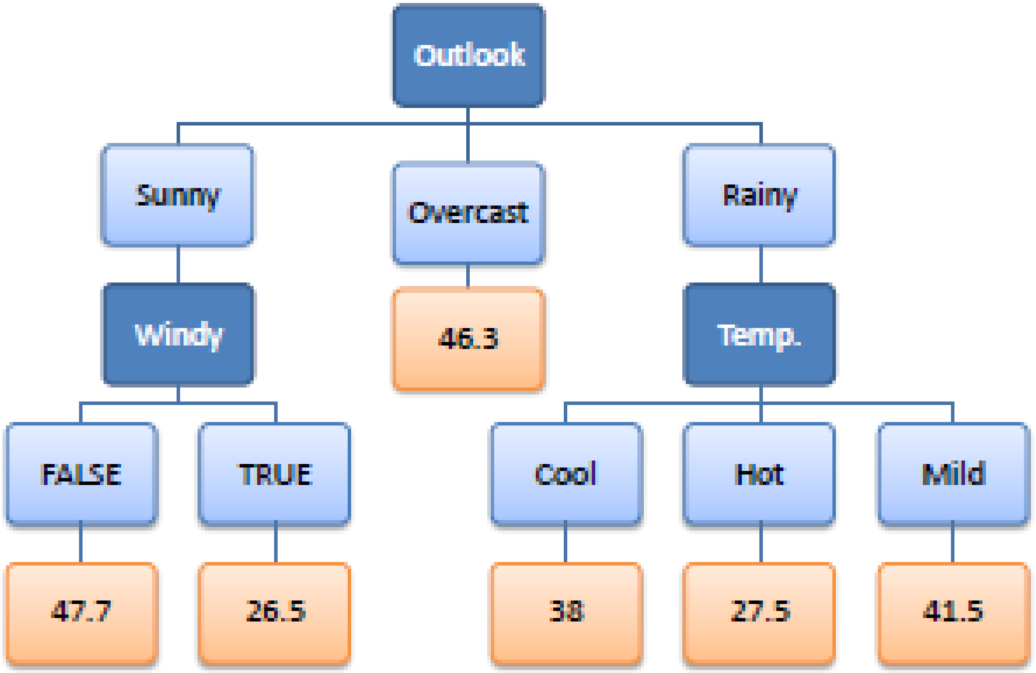
Categorial vs Numerical data

- **Numerical data** has meaning as a measurement, such as a person's height, weight, IQ, or blood pressure
- **Categorical data** represents characteristics such as a person's gender, marital status, hometown, or the types of movies they like.

Decision tree for regression

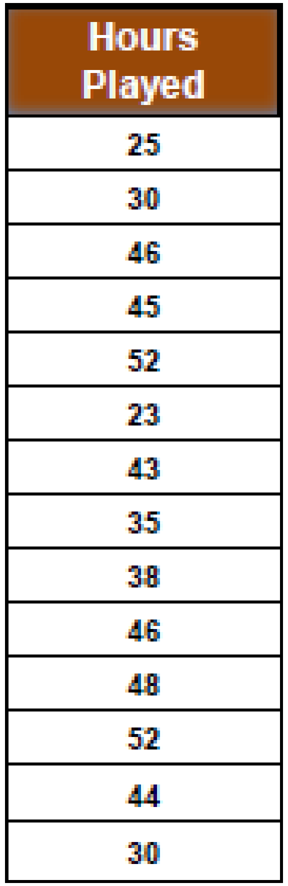


	Outlook	Temp	Humidity	Windy	Hours Played
0	Rainy	Hot	High	False	25
1	Rainy	Hot	High	True	30
2	Overcast	Hot	High	False	46
3	Sunny	Mild	High	False	45
4	Sunny	Cool	Normal	False	52
5	Sunny	Cool	Normal	True	23
6	Overcast	Cool	Normal	True	43
7	Rainy	Mild	High	False	35
8	Rainy	Cool	Normal	False	38
9	Sunny	Mild	Normal	False	46
10	Rainy	Mild	Normal	True	48
11	Overcast	Mild	High	True	52
12	Overcast	Hot	Normal	False	44
13	Sunny	Mild	High	True	30



Decision tree for regression

- Standart deviation



Hours Played
25
30
46
45
52
23
43
35
38
46
48
52
44
30

$$Count = n = 14$$

$$Average = \bar{x} = \frac{\sum x}{n} = 39.8$$

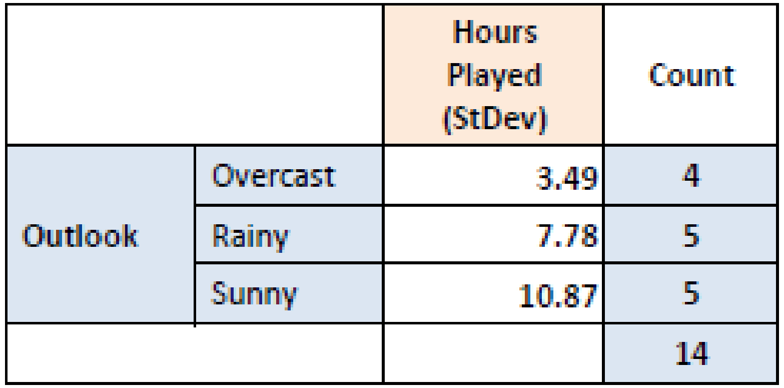
$$Standard\ Deviation = S = \sqrt{\frac{\sum(x - \bar{x})^2}{n}} = 9.32$$

$$Coeffeicient\ of\ Variation = CV = \frac{S}{\bar{x}} * 100\% = 23\%$$

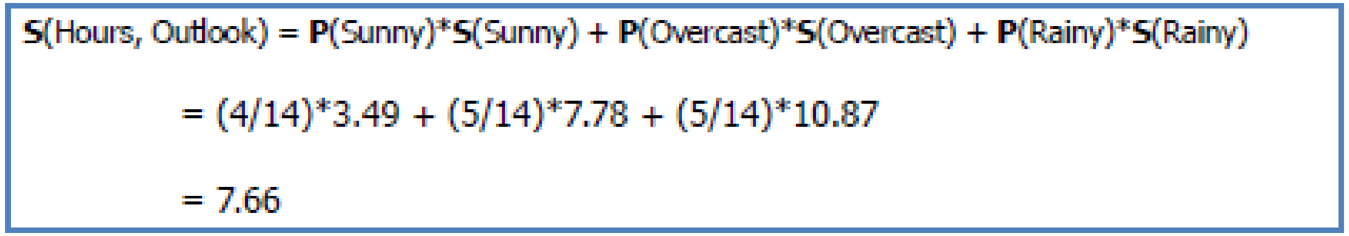
Decision tree for regression

- Standart deviation for a split

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



		Hours Played (StDev)	Count
Outlook	Overcast	3.49	4
	Rainy	7.78	5
	Sunny	10.87	5
			14

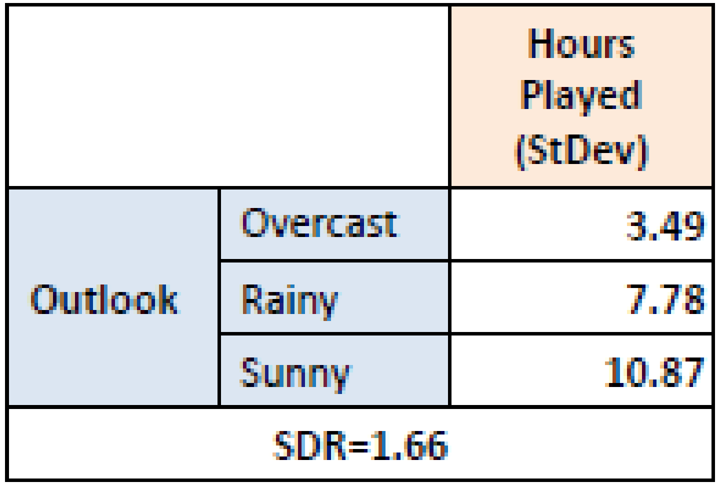


S(Hours, Outlook) = **P**(Sunny)***S**(Sunny) + **P**(Overcast)***S**(Overcast) + **P**(Rainy)***S**(Rainy)

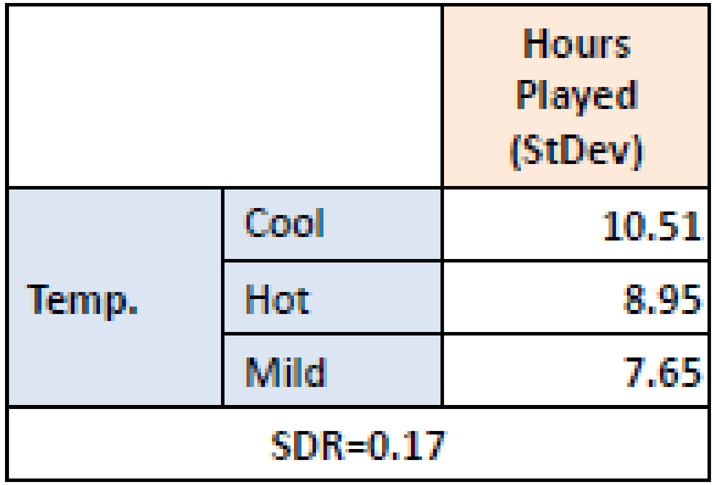
= (4/14)*3.49 + (5/14)*7.78 + (5/14)*10.87

= 7.66

Decision tree for regression



		Hours Played (StDev)
Outlook	Overcast	3.49
	Rainy	7.78
	Sunny	10.87
SDR=1.66		

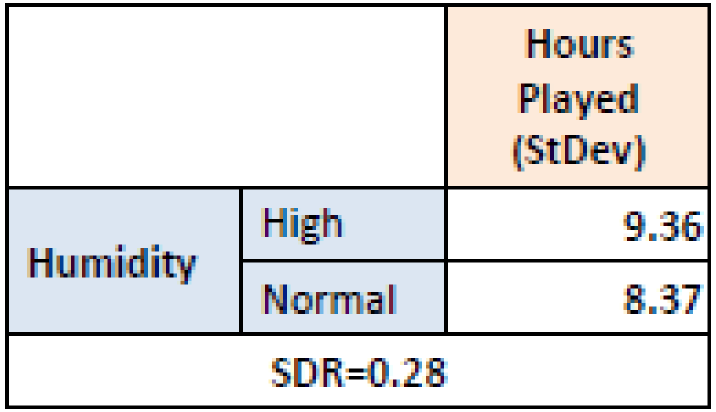


		Hours Played (StDev)
Temp.	Cool	10.51
	Hot	8.95
	Mild	7.65
SDR=0.17		

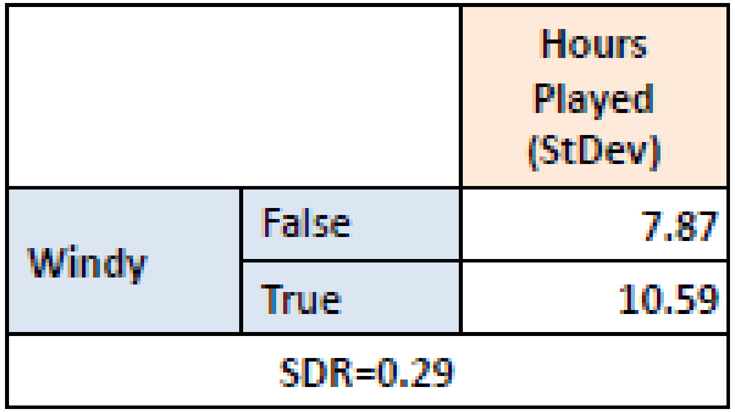
$$SDR(T,X) = S(T) - S(T,X)$$

SDR(Hours , Outlook) = **S**(Hours) – **S**(Hours, Outlook)

= 9.32 – 7.66 = 1.66

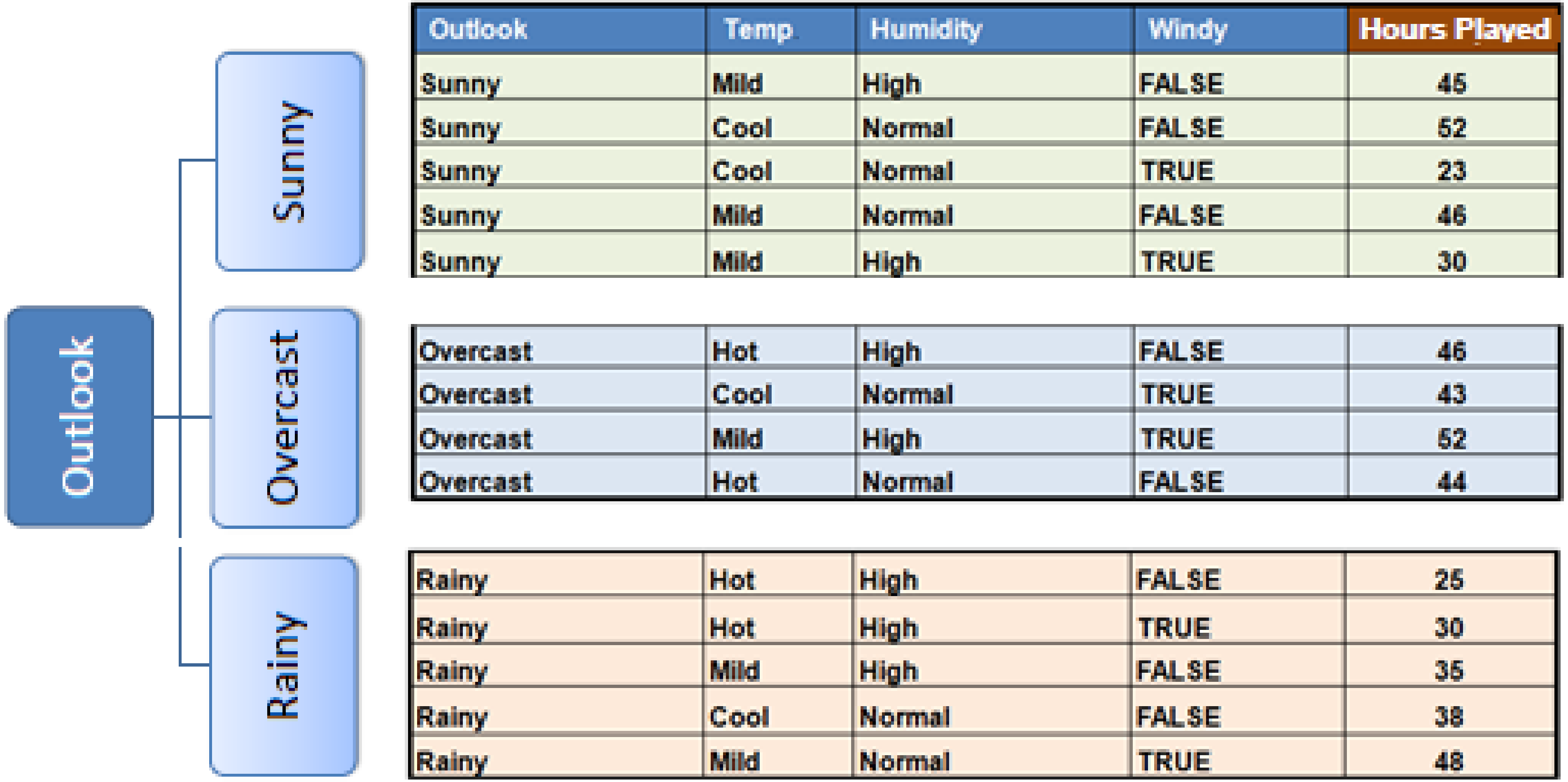


		Hours Played (StDev)
Humidity	High	9.36
	Normal	8.37
SDR=0.28		



		Hours Played (StDev)
Windy	False	7.87
	True	10.59
SDR=0.29		

Decision tree for regression



Outlook	Temp.	Humidity	Windy	Hours Played
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
Overcast	Hot	High	FALSE	46
Overcast	Cool	Normal	TRUE	43
Overcast	Mild	High	TRUE	52
Overcast	Hot	Normal	FALSE	44
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

How to stop

- Coefficient of variant.
- The number of samples in a branch.

Pruning Tree

- **Pruning Tree: to avoid overfitting.**
- How:
  - Control tree's depth
  - Control the number of samples in each node.

Cảm ơn đã theo dõi!