Entscheidungsbäume (CART)

konzept. Verunreinigung der Information in einer Stichprobe

Die Verunreinigung misst die Homogeneität in der Datenprobe. Wenn die Daten in der Probe homogen sind, gehören die Stichproben zur gleichen Klasse und die Verunreinigung ist 0.

Wir messen die Verunreinigung mit dem Gini-Index:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

Gini-Index ist ein Maß für die Verunreinigung einer Stichprobe.
Hat einen Wert zw. $[0,1]$.

- Gini-Index = 0 bedeutet, dass die Stichprobe vollkommend homogen ist und alle Elemente ähnlich sind.
- Gini-Index = 1 bedeutet, maximale Verunreinigung bzw. Ungleichheit zw. den Elementen.

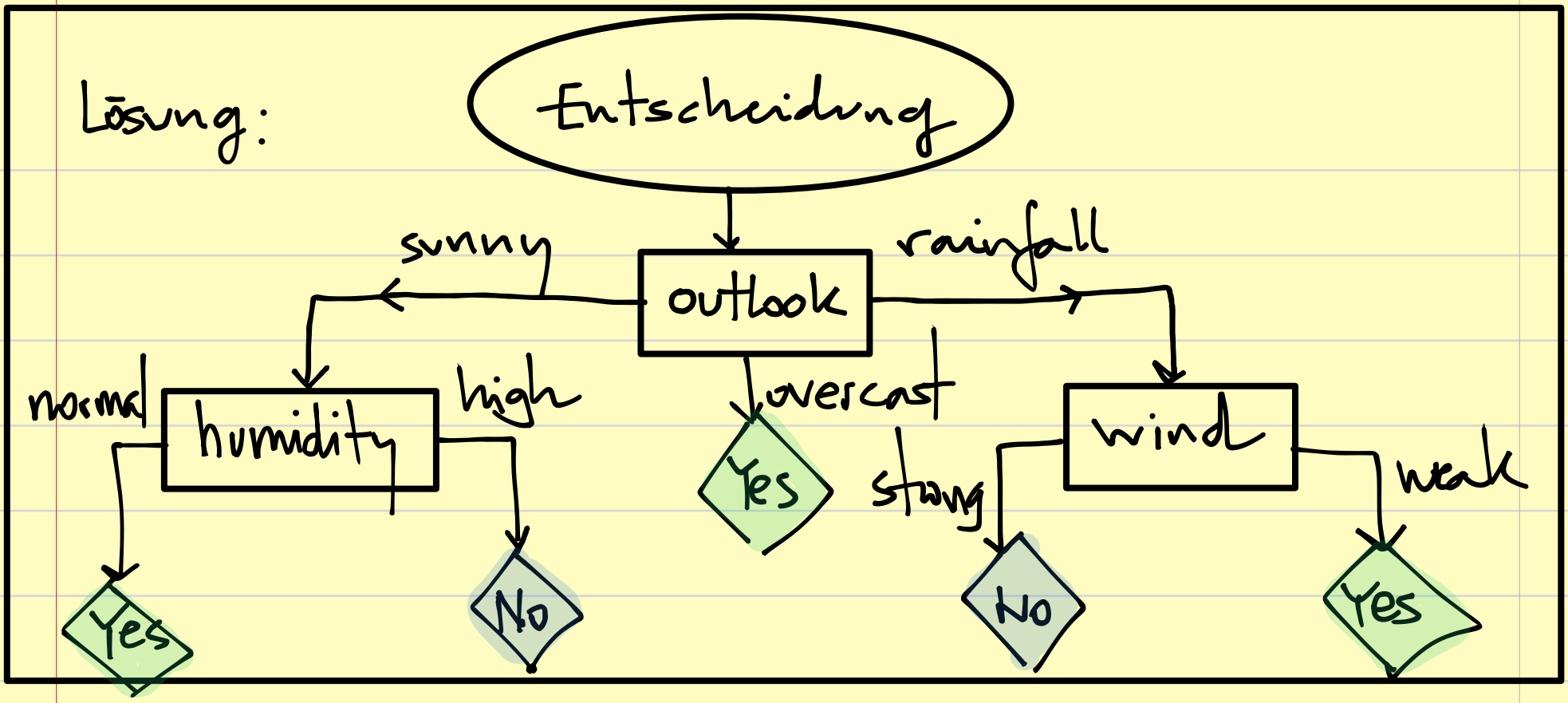
Ziel: Entscheiden ob ein Fußballspiel stattfindet anhand verschiedene nicht nummerische (nominale) Variabeln (Wetterbedingungen).

Bitte einen Entscheidungsbaum erstellen.

Daten:

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Lösung:



Wir suchen den ersten Knoten mit dem geringsten Gini.Index (minimale Verunreinigung):

OUTLOOK: outlook ist ein nominales Merkmal. Es kann drei Wert annehmen: (sunny, overcast, rainfall)

outlook	Yes	No	#
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5
			$\Sigma = 14$

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

$$Gini(Outlook=Sunny) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0'48$$

$$Gini(outlook=overcast) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

$$Gini(Outlook=Rainfall) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0'48$$

Die gewichtete Summe von Outlook:

$$Gini(Outlook) = \frac{5}{14} \cdot 0'48 + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0'48 = 0'342$$

Temperature: (hot, cold, mild)

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Temperature	Yes	No	#
hot	2	2	4
cold	3	1	4
mild	4	2	6
			$\Sigma = 14$

$$Gini(Temp.=hot) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0'5$$

$$Gini(Temp.=cold) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0'375$$

$$Gini(Temp=mild) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0'445$$

$$\text{Gewichtete Summe } Gini(Temp.) = \frac{4}{14}\cdot 0'5 + \frac{4}{14}\cdot 0'375 + \frac{6}{14}\cdot 0'445 = 0'43$$

Humidity (hoch, normal)

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Humidity	Yes	No	#
high	3	4	7
normal	6	1	7
			$\Sigma = 14$

$$Gini(Humidity=high) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0'489$$

$$Gini(Humidity=normal) = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0'244$$

$$\text{Gewichtete } Gini(Humidity) = \frac{7}{14}\cdot 0'489 + \frac{7}{14}\cdot 0'244 = 0'367$$

Wind (strong, weak)

wind	Yes	No	#
weak	6	2	8
strong	3	3	6
			≤ 14

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

$$Gini(Wind=weak) = 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0'375$$

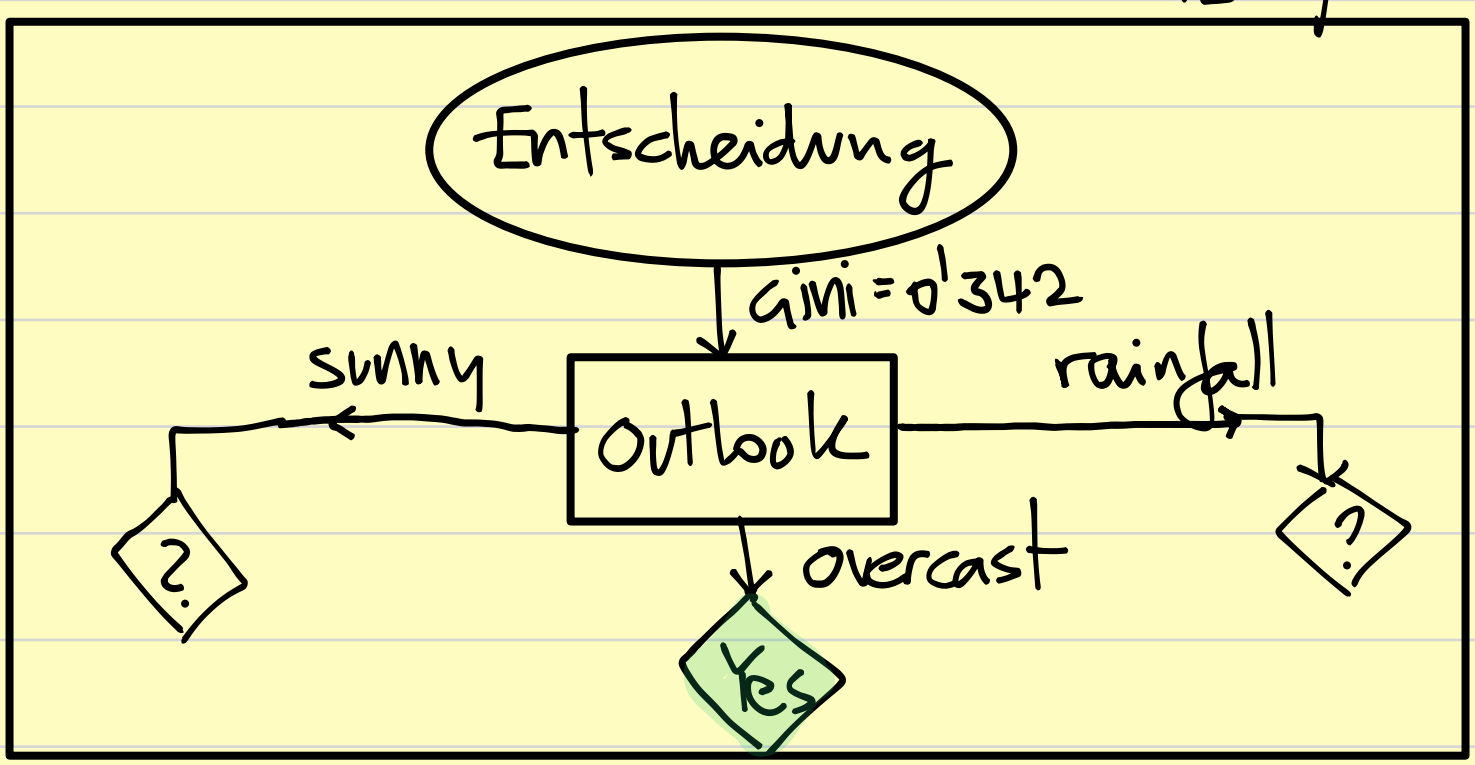
$$Gini(Wind=strong) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0'5$$

$$\text{Gewichtete Gini (Wind)} = \frac{8}{14} \cdot 0'375 + \frac{6}{14} \cdot 0'5 = 0'428$$

1. Entscheidung:

Feature	Gini
Outlook	0'342
Temp.	0'439
Humidity	0'367
Wind	0'428

Outlook hat die geringste Verunreinigung und wird als Knote gewählt.



Gini : Sunny + Temp.

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

outlook	Temp	Yes	No	#
Sunny	hot	0	2	2
"	cold	1	0	1
"	mild	1	1	2
				Σ: 5

$$Gini(sunny+hot) = 1-\left(\frac{0}{2}\right)^2-\left(\frac{2}{2}\right)^2 = 0$$

$$Gini(sunny+cold) = 1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2 = 0$$

$$Gini(sunny+mild) = 1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2 = 0{,}5$$

$$\text{Gewichtete } Gini(sunny+Temp) = \frac{2}{5}\cdot 0+\frac{1}{5}\cdot 0+\frac{2}{5}\cdot 0{,}5 = 0{,}25$$

Gini : Sunny + Humidity

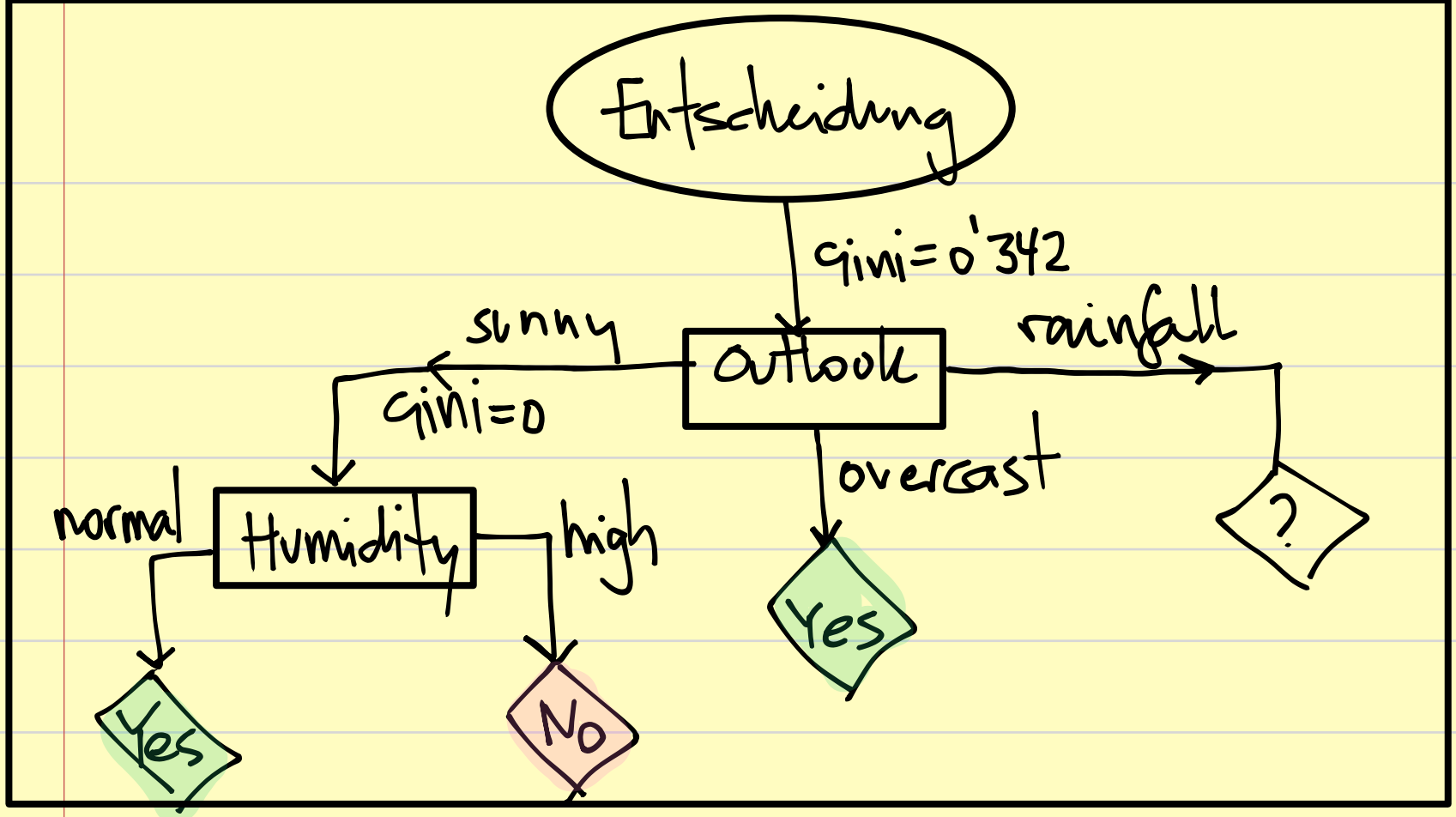
Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Outlook	Humidity	Yes	No	#
sunny	high	0	3	3
"	normal	2	0	2

$$Gini(sunny+Hum.high) = 1-\left(\frac{0}{3}\right)^2-\left(\frac{3}{3}\right)^2 = 0$$

$$Gini(sunny+Hum.norm.) = 1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2 = 0$$

$$\text{Gewichtete } Gini(sunny+Humidity) = \frac{3}{5}\cdot 0+\frac{2}{5}\cdot 0 = 0$$



Gini (Rainfall + Temperature)

Outlook	Temp.	Yes	No	#
Rainfall	cold	1	1	2
"	mild	2	1	3
				≤5

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

$$Gini(Rainfall + Temp.\,Cold) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0'5$$

$$Gini(Rainfall + Temp.\,Mild) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0'44$$

$$\text{Gewichtete } Gini(Temp + Rainfall) = \frac{2}{5}\cdot 0'5 + \frac{3}{5}\cdot 0'44 =$$

$$= 0'467$$

Gini (Rainfall + Wind)

Outlook	Wind	Yes	No	#
Rainfall	weak	3	0	3
"	strong	0	2	2

Gini (Rainfall + weak Wind) = 0

Gini (Rainfall + strong wind) = 0

Gewichtete Gini (Rainfall, Wind) = 0