Entscheidungsbäume...

Wir messen die Verunreinigung der Daten um eine Entscheidung zu treffen.
Diese wird mit dem Gini-Index gemessen. $Gini = 1 - \sum_{i=1}^{n} p_i^2$

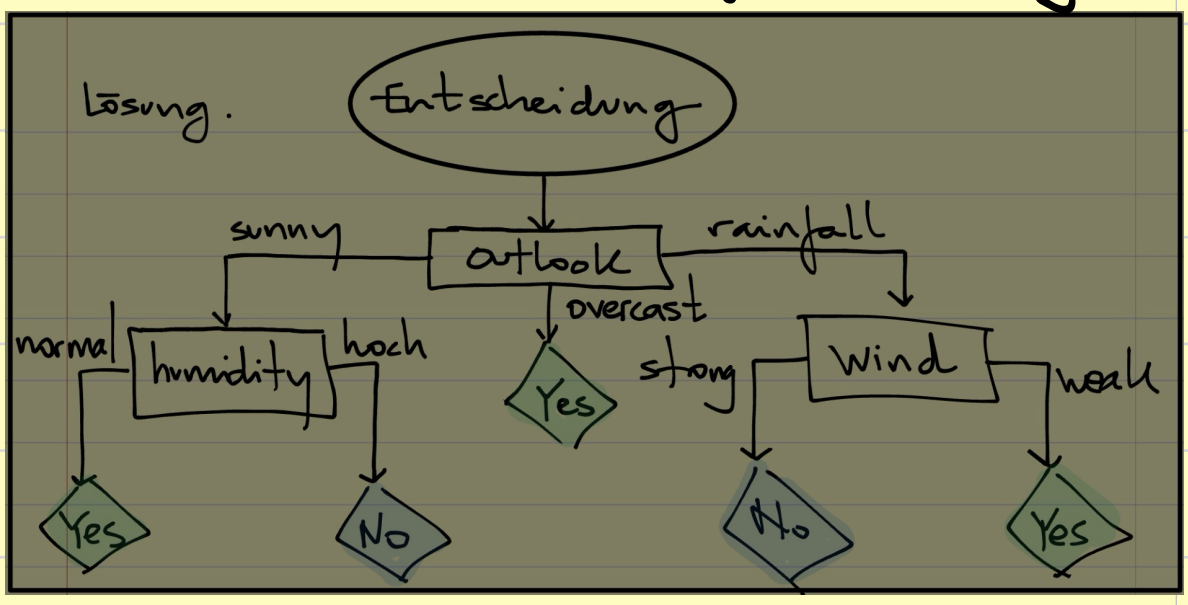
$p_i \in [0,1] \equiv$ W, dafür, dass die Probe zur Klasse gehört.

$Gini = 0 \equiv$ die Stichprobe ist homogen.

$Gini = 1 \equiv$ die Stichprobe ist maximal verunreinigt.

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.

OUTLOOK.	Ja	Nein	#
sunny	2	3	5
overcast	4	0	4
Rainfall	3	2	5
			14

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

$Gini(Outlook\,Overc.) = 1 - \left[\left(\frac{4}{4}\right)^2 + 0\right] = 0$

$Gini(Outlook\,Rain) = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0{,}48$

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

TEMPERATUR.	Ja	Nein	#
hot	2	2	4
mild	4	2	6
cool	3	1	4

$$Gini(temp\ hot) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0{,}5$$

$$Gini(Temp\ mild) = 1 - \left[\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right] = 0{,}44$$

$$Gini(Temp\ cool) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0{,}37$$

$$Gini(Temp) = \frac{4}{14} \cdot 0{,}5 + \frac{6}{14} \cdot 0{,}44 + \frac{4}{14} \cdot 0{,}37 = 0{,}43$$

HUMIDITY.	Ja	Nein	#
high	3	4	7
normal	6	1	7

$$Gini(Hum.high) = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0{,}49$$

$$Gini(Hum.high) = 1 - \left[\left(\frac{6}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right] = 0{,}24$$

$$Gini(humidity) = \frac{7}{14} \cdot 0{,}49 + \frac{7}{14} \cdot 0{,}24 = 0{,}367$$

WIND	Ja	Nein	#
weak	6	2	8
strong	3	3	6

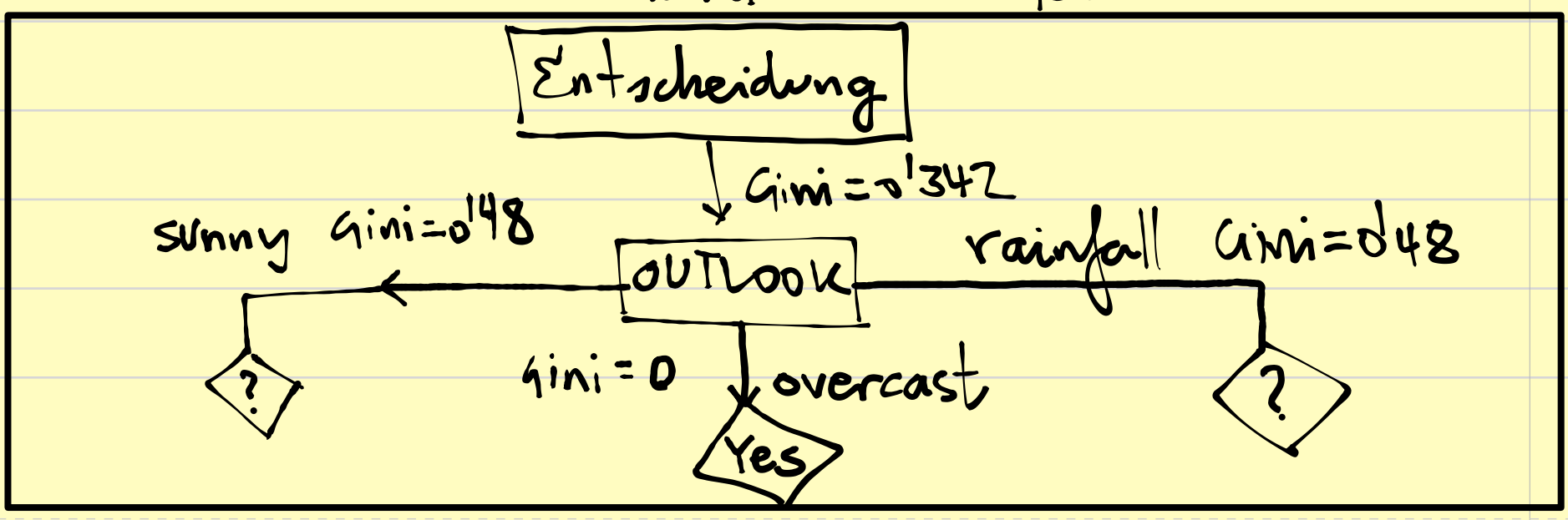
$$Gini(Wind\ weak) = 1 - \left[\left(\frac{6}{8}\right)^2 + \left(\frac{2}{8}\right)^2\right] = 0{,}37$$

$$Gini(Wind\ strong) = 0{,}5$$

$$Gini(Wind) = \frac{8}{14} \cdot 0{,}37 + \frac{6}{14} \cdot 0{,}5 = 0{,}428$$

Entscheidung:

Variablen	Gini	
Outlook	0,342	→ Outlook hat die geringste Verunreinigung
Temp	0,439	
Humidity	0,367	
Wind	0,428	



OUTLOOK SUNNY { Temp, Hum., Wind

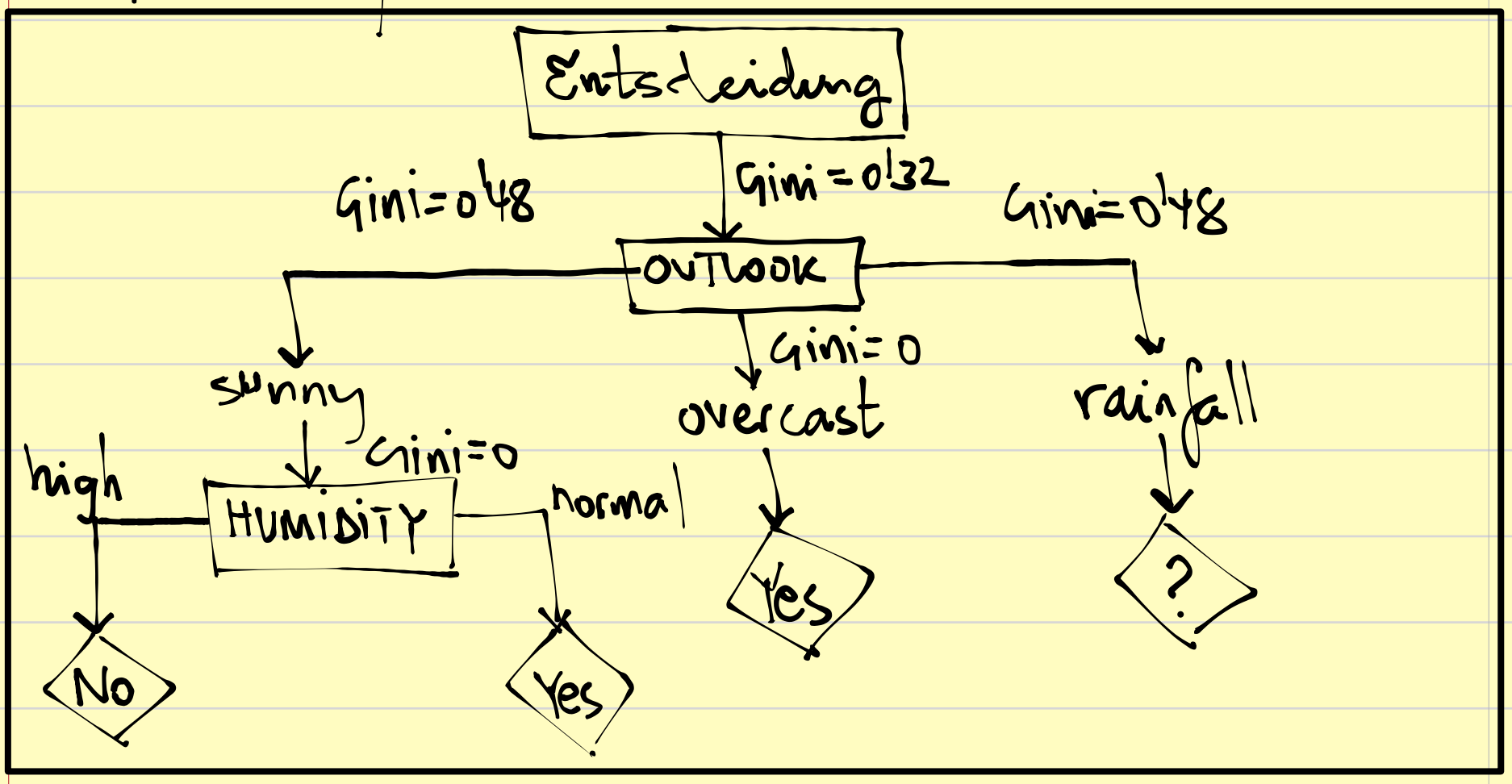
0. SUNNY + Temp	Ja	Nein	#
hot	0	2	2
mild	1	1	2
cool	1	0	1

$$Gini(Sunny+Temp) = \frac{2}{5}\cdot\left[1-\left[\left(\frac{0}{2}\right)^2+\left(\frac{2}{2}\right)^2\right]\right]+\frac{2}{5}\cdot\left[1-\left[\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right]\right]+\frac{1}{5}\left[1-\left[\left(\frac{1}{1}\right)^2\right]\right]$$

$$= \frac{2}{5}\cdot 0 + \frac{2}{5}\cdot 0{,}5 + \frac{1}{5}\cdot 0 = 0{,}2$$

0. SUNNY + Humidity	Ja	Nein	#
high	0	3	3
normal	2	0	2

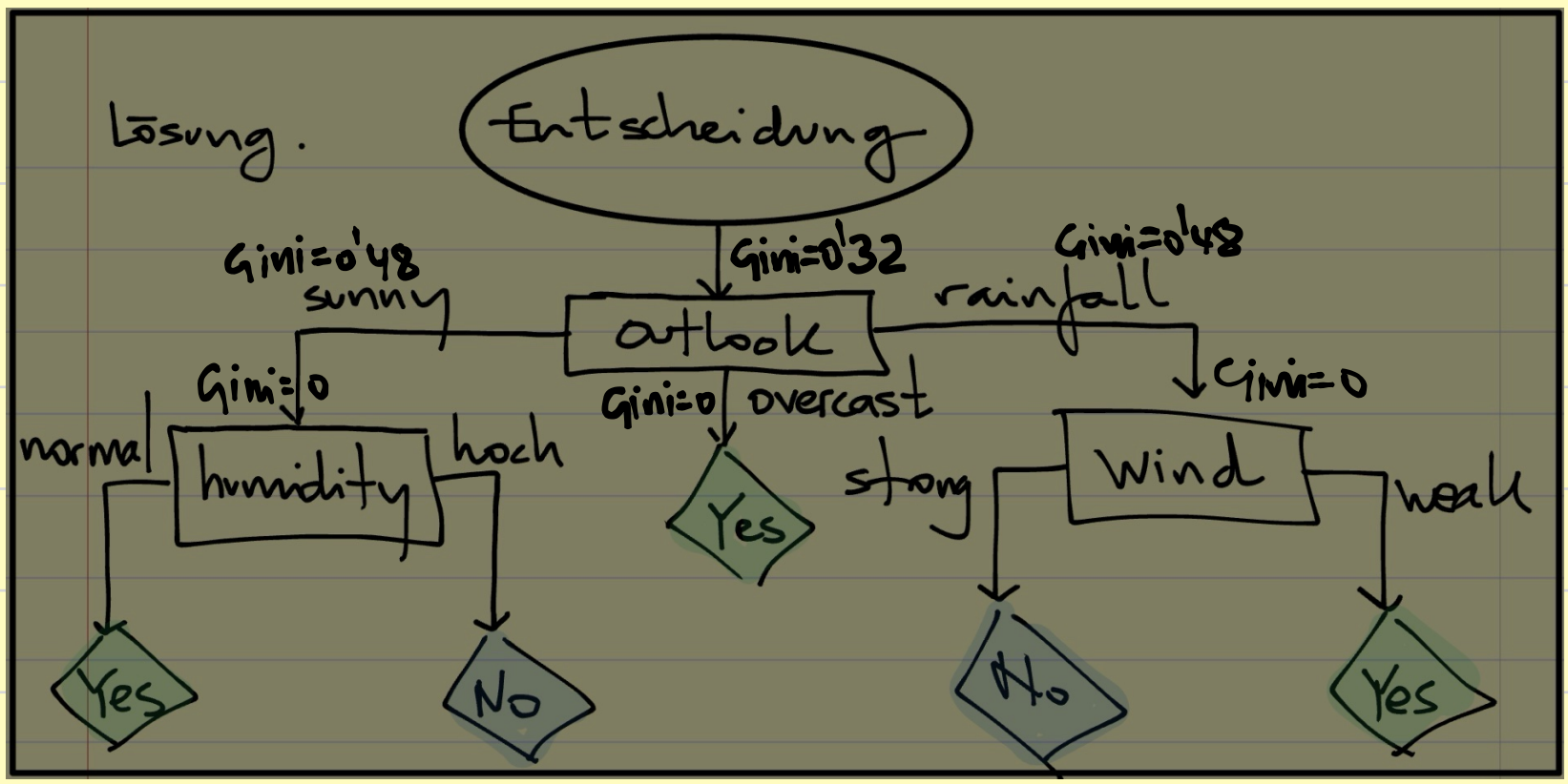
$$Gini(Sunny+Hum.) = 0$$



OUTLOOK RAINFALL+ { Temp, Wind

0. RAINFALL + WIND	Ja	Nein	#
Weak	3	0	3
strong	0	2	2

Gini(0. Rainfall + Wind) = 0



Entscheidungsbaum ≡ CART

Übung.

Beispiel. SEX Ja/Nein.

	Wohnungs. Verschmutzung	Sinnvolle Gespräche	Fitness Niveau	Mond	Sex
1.	stark	oft	hoch	voll	Ja
2.	schwach	oft	gering	wachsend	Nein
3.	sauber	selten	hoch	voll	Ja
4.	stark	oft	mittel	abnehmend	Ja
5.	stark	selten	hoch	voll	Nein
6.	sauber	oft	hoch	wachsend	Ja
7.	schwach	oft	mittel	voll	Nein
8.	stark	oft	gering	voll	Ja
9.	schwach	selten	gering	neu	Ja
10.	sauber	oft	hoch	neu	Nein