ENTSCHEIDUNGSBÄUME (CART)

KONZEPT. Verunreinigung der Information.

- Die Verunreinigung misst die Homogenität einer Datenprobe. Wenn die Daten in der Probe homogen sind, gehören die Stichproben zur gleichen Klasse und die Verunreinigung ist NULL ($\emptyset$).
- Wir messen die Verunreinigung mit dem Gini-Index:

$$\text{Gini-Index} = 1 - \sum_{i=1}^{n} p_i^2 \quad ; p_i \in [0,1] \text{ W. dafür dass die Probe zur Klasse gehört.}$$

Gini-Index ist ein Maß für die Verunreinigung einer Stichprobe und hat einen Wert zwischen 0 und 1.

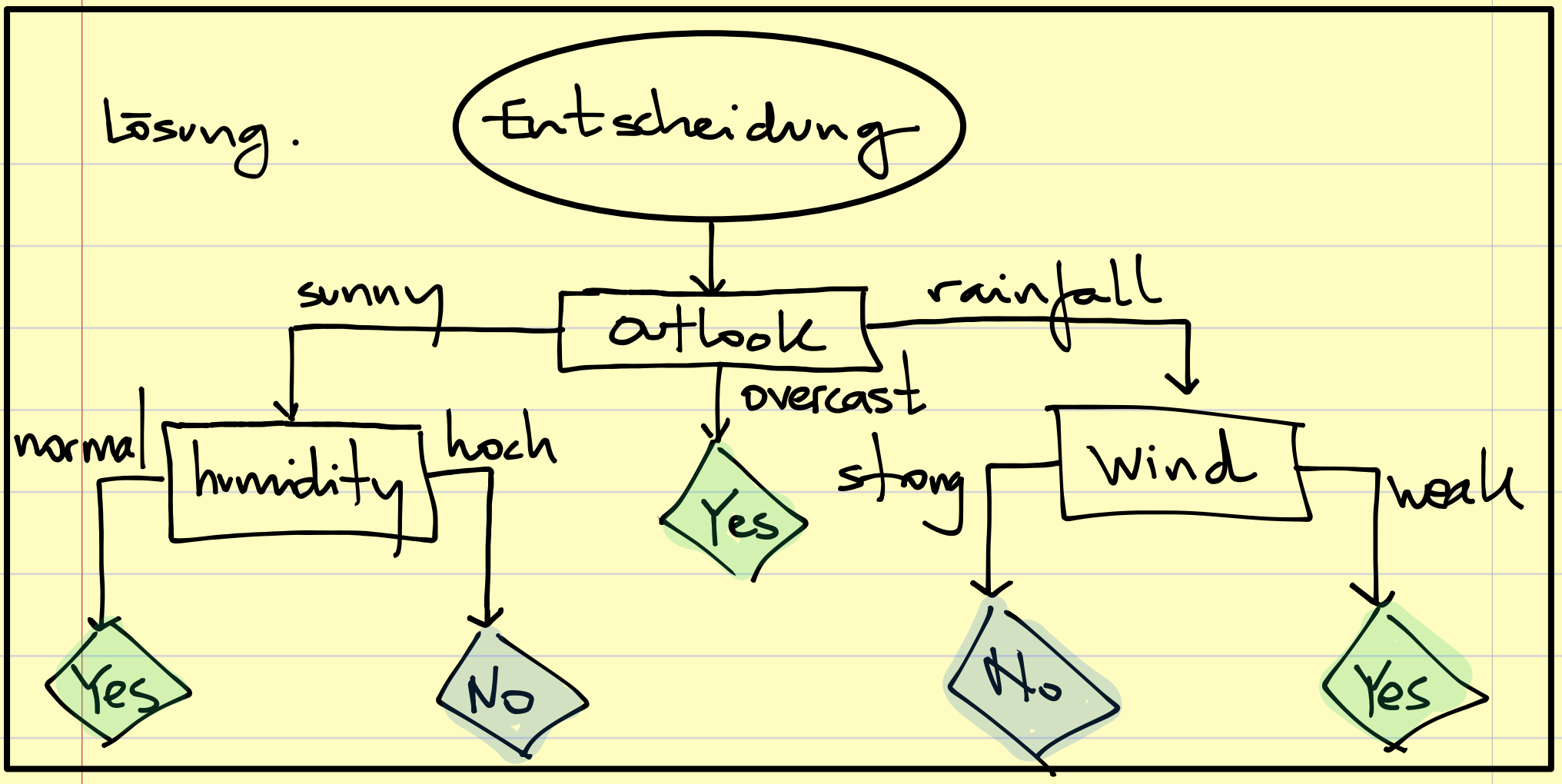
- Gini-Index = 0 bedeutet, dass die Stichprobe vollkommend homogen ist und alle Elemente zur Klasse gehören.
- Gini-Index = 1 bedeutet, maximale Verunreinigung bzw Ungleichheit zw. den Elementen.

Ziel. Entscheiden ob ein Fußballspiel stattfindet anhand verschiedene nichtnummerische (nominale) Variablen (Wetterbedingungen).

? Bitte einen Entscheidungsbaum erstellen.

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 Werte annehmen: sunny, overcast, rainfall.

OUTLOOK	Yes	NO	#
SUNNY	2	3	5
OVERCAST	4	0	4
RAINFALL	3	2	5
			$\sum 14$

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

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

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

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

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)

Temperature	Yes	No	#
hot	2	2	4
cold	3	1	4
mild	4	2	6

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

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

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

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

$$Gini(Temperature) = \frac{4}{14} \cdot 0'5 + \frac{4}{14} \cdot 0'375 + \frac{6}{14} \cdot 0'445 = 0'43$$

HUMIDITY : high, normal

HUMIDITY	Yes	No	#
high	3	4	7
Normal	6	1	7

$$Gini(Humidity \cdot High) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0'49$$

$$Gini(Humidity \cdot Normal) = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0'24$$

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

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

WIND : (strong, weak)

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

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

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

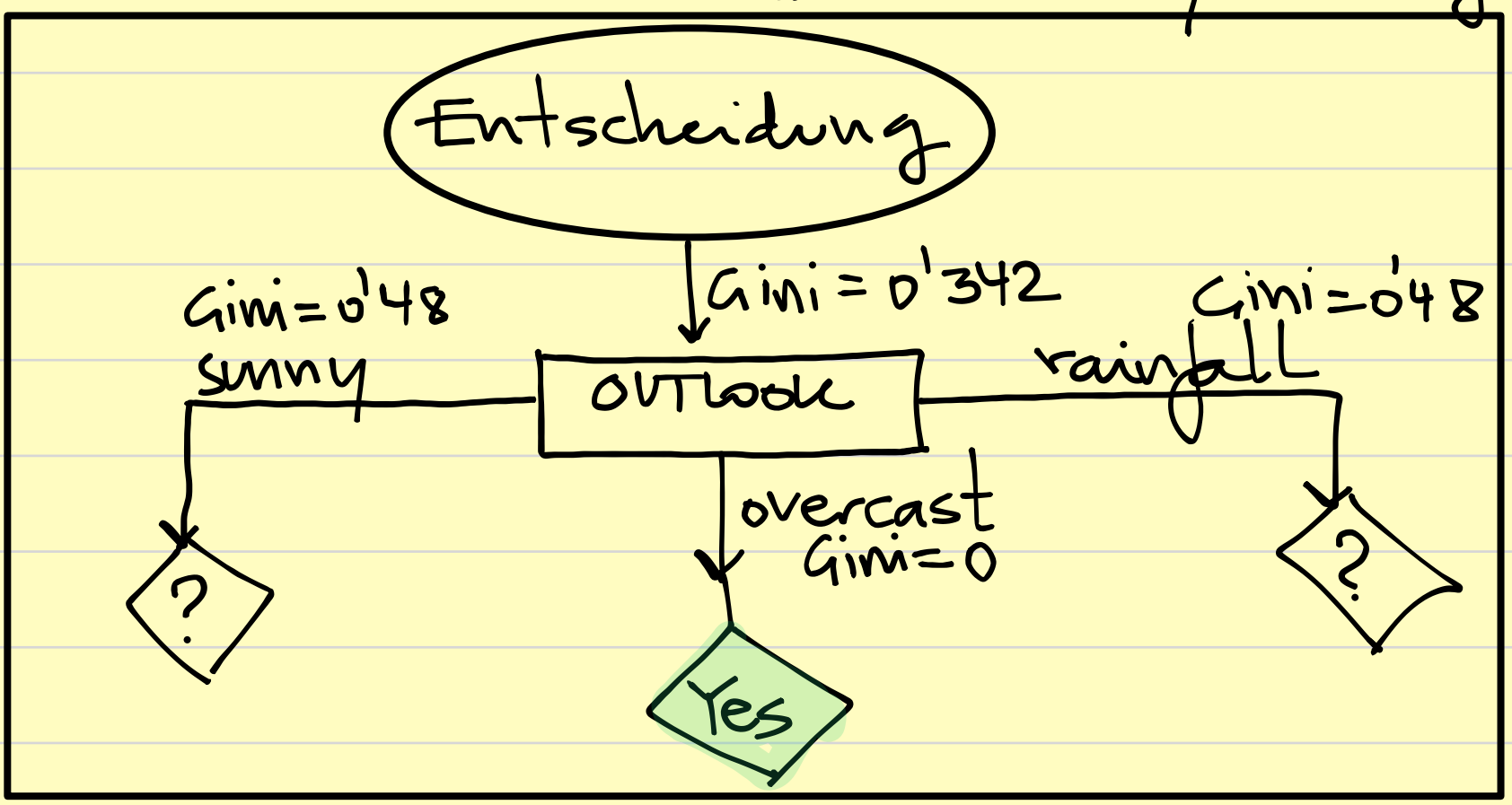
$$Gini(wind) = \frac{8}{14} \cdot 0'375 + \frac{6}{14} \cdot 0'5 = 0'428$$

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

1. Entscheidung

Variabeln	Gini
outlook	0'342
Temperature	0'439
Humidity	0'367
Wind	0'428

} outlook hat die geringste Verunreinigung und wird als Knote gewählt



OUTLOOK SUNNY . Temp / Humidity / Wind.

· Outlook Sunny + Temp.

Sunny+Temp	Yes	No	#
hot	0	2	2
mild	1	1	2
cold	1	0	1
			Σ5

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

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

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

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

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

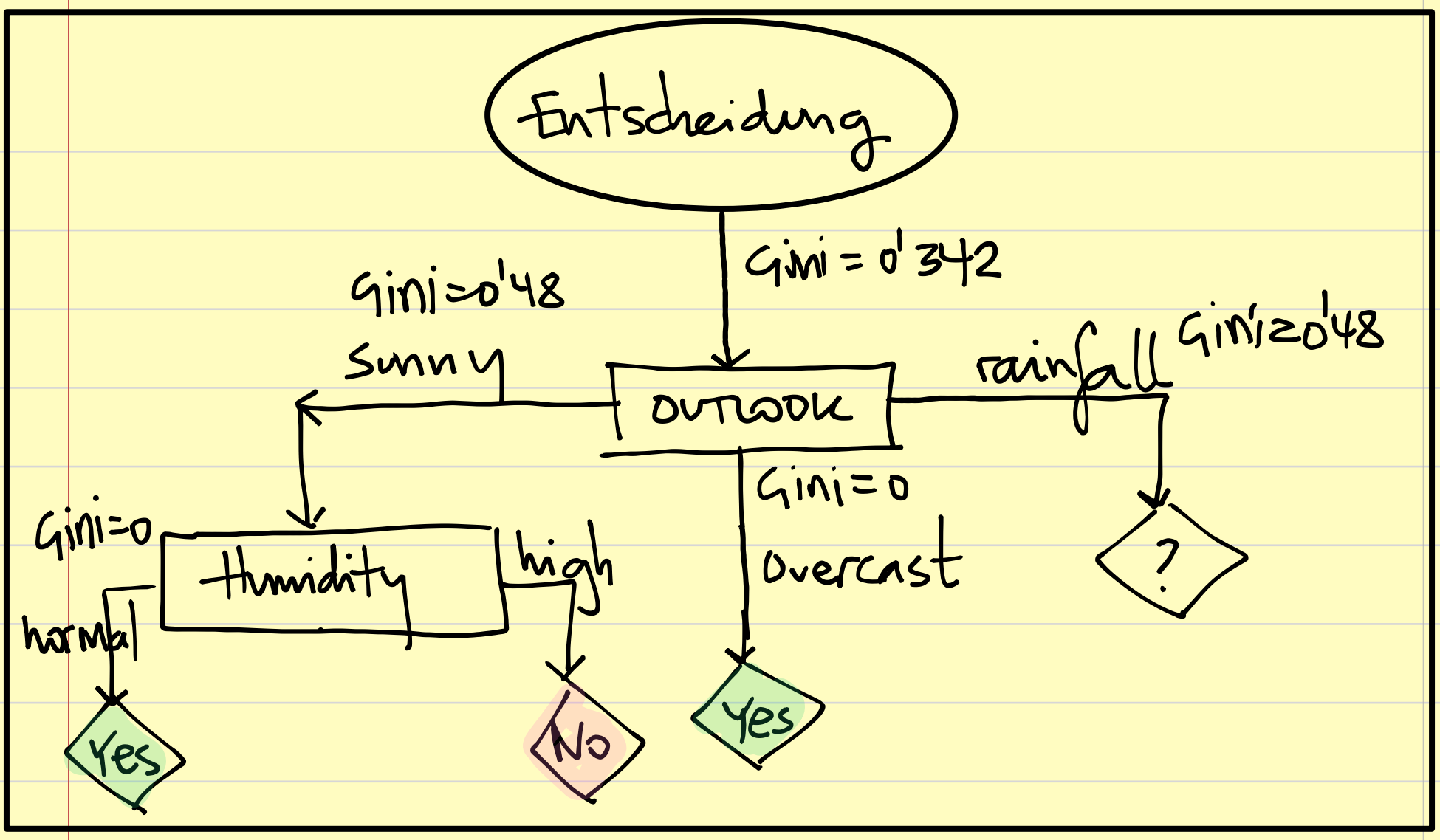
· Sunny + Humidity

Sunny+Humidity		Yes	No	#
Sunny	high	0	3	3
"	normal	2	0	2

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

$$Gini(Sunny+Humidity) = 0$$

Wir können somit die Suche nach dem Knote mit geringsten Verunreinigung beenden. Humidity + Sunny hat einen Gini = 0.



- Outlook + Temp

Rainfall + Temp		Yes	No	#
Rainfall	cold	1	1	2
"	mild	2	1	3
"	hot	0	0	0
				Σ5

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

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

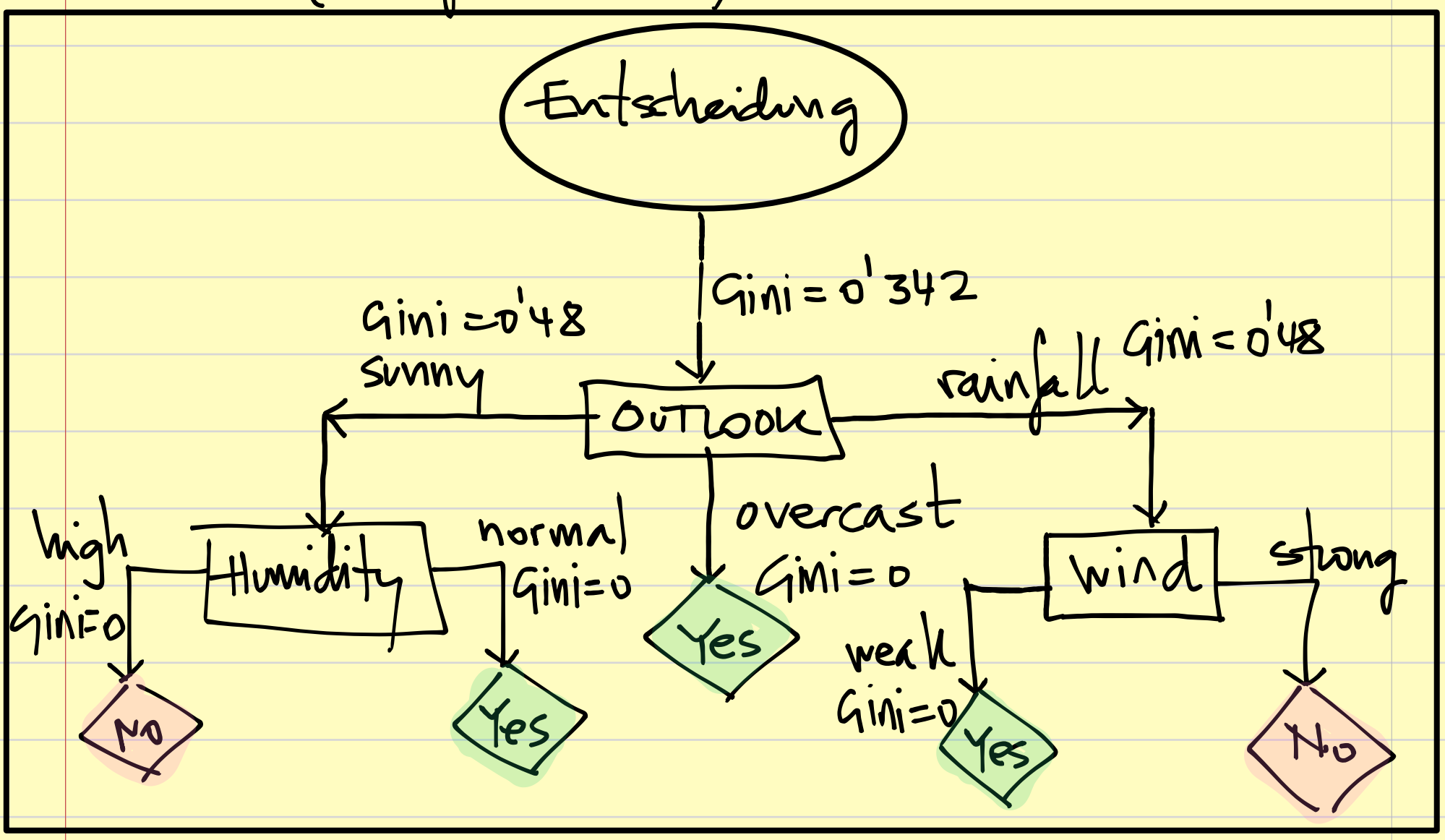
$$Gini(Rainfall+Temp.mild) = 1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2 = 0'44$$

$$Gini(Rainfall+Temp) = \frac{2}{5}\cdot 0'5+\frac{3}{5}\cdot 0'44 = 0'467$$

- Outlook + Wind

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

Gini(Rainfall + Wind) = 0



Beispiel. SEX . Ja / Nein .

	WOHNUNGS-BESCHMUTZUNG	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

$$Gini(\text{Wohnung.Beschmutzung}) =$$

$$= \frac{4}{10} \cdot 0{,}375 + \frac{3}{10} \cdot 0{,}44 + \frac{3}{10} \cdot 0{,}44$$

Beschmutzung	Ja	Nein	#
stark	3	1	4
schwach	1	2	3
sauber	2	1	3

$$Gini(\text{Beschm. stark}) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0{,}375$$

$$Gini(\text{Besch. schw.}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0{,}44$$

$$Gini(\text{Besch. sauber}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0{,}44$$

...

Beispiel. SEX. Ja / Nein.

	WOHNUNGS-BESCHMUTZUNG	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