

## ENTSCHEIDUNGSBÄUME (CART)

## classification and Regression Trees.

Beispiel. SEX ♂ ♀ ♂ ♂ ♂ ♂ ♂ ♂ ♂ ♂ ...

	WOHNUNGS-BESCHÜTZUNG	SINNVOLLE GESPRÄCHE	FITNESS NIVEAU	MANN	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

KONZEPT. Verunreinigung der Information.

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

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

Gini-Index  $\in [0,1]$

Gini-Index = 0  $\rightarrow$  die Stichprobe ist vollkommen homogen und alle Elemente gehören zur Klasse.

Gini-Index = 1  $\rightarrow$  die Stichprobe ist maximal Verunreinigt.

---

ziel Entscheidungsbau um zu wissen ob das Paar sex hat, anhand von nichtnumerischen (nominalen) Variablen.

Wir suchen den ersten Entscheidungsvariable : das ist die Variable mit geringsten Verunreinigung.

W. BESCHMUTZUNG	Ja	Nein	#
stark	3	1	4
schwach	1	2	3
sauber	2	1	3

$$\text{Gini}(\text{W. Besch. stark}) = 1 - \sum p_i^2 = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0'375$$

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

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

$$\text{Gini(W. Besch.)} = \frac{4}{10} \cdot 0'375 + \frac{3}{10} \cdot 0'44 + \frac{3}{10} \cdot 0'44 = 0'414$$

<u>sinnvolle Gespräche</u>	Ja	Nein	#
oft	4	3	7
selten	2	1	3

$$\text{Gini(S.Gespr. oft)} = 1 - \left( \frac{4}{7} \right)^2 - \left( \frac{3}{7} \right)^2 = 0'49$$

$$\text{Gini(S.Gespr. selten)} = 1 - \left( \frac{2}{3} \right)^2 - \left( \frac{1}{3} \right)^2 = 0'44$$

$$\text{Gini(S.Gespr)} = \frac{7}{10} \cdot 0'49 + \frac{3}{10} \cdot 0'44 = 0'475$$

<u>FITNESS</u>	Ja	Nein	#
hoch	3	2	5
mittel	1	1	2
gering	2	1	3

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

$$\text{Gini(Fitness mittel)} = 1 - \left( \frac{1}{2} \right)^2 - \left( \frac{1}{2} \right)^2 = 0'5$$

$$\text{Gini(Fitness gering)} = 1 - \left( \frac{2}{3} \right)^2 - \left( \frac{1}{3} \right)^2 = 0'44$$

$$\text{Gini(Fitness)} = \frac{5}{10} \cdot 0'48 + \frac{2}{10} \cdot 0'5 + \frac{3}{10} \cdot 0'44 = 0'472$$

<u>MOND</u>	Ja	Nein	#
voll	3	2	5
wachsend	1	1	2
abnehmend	1	0	1
neu	1	1	2

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

$$\text{Gini(Mond wach)} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0'5$$

$$\text{Gini(Mond abn.)} = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0$$

$$\text{Gini(Mond neu)} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0'5$$

$$\text{Gini(Mond)} = \frac{5}{10} \cdot 0'48 + \frac{2}{10} \cdot 0'5 + \frac{1}{10} \cdot 0 + \frac{2}{10} \cdot 0'5 = 0'44$$

Variablen: Gini

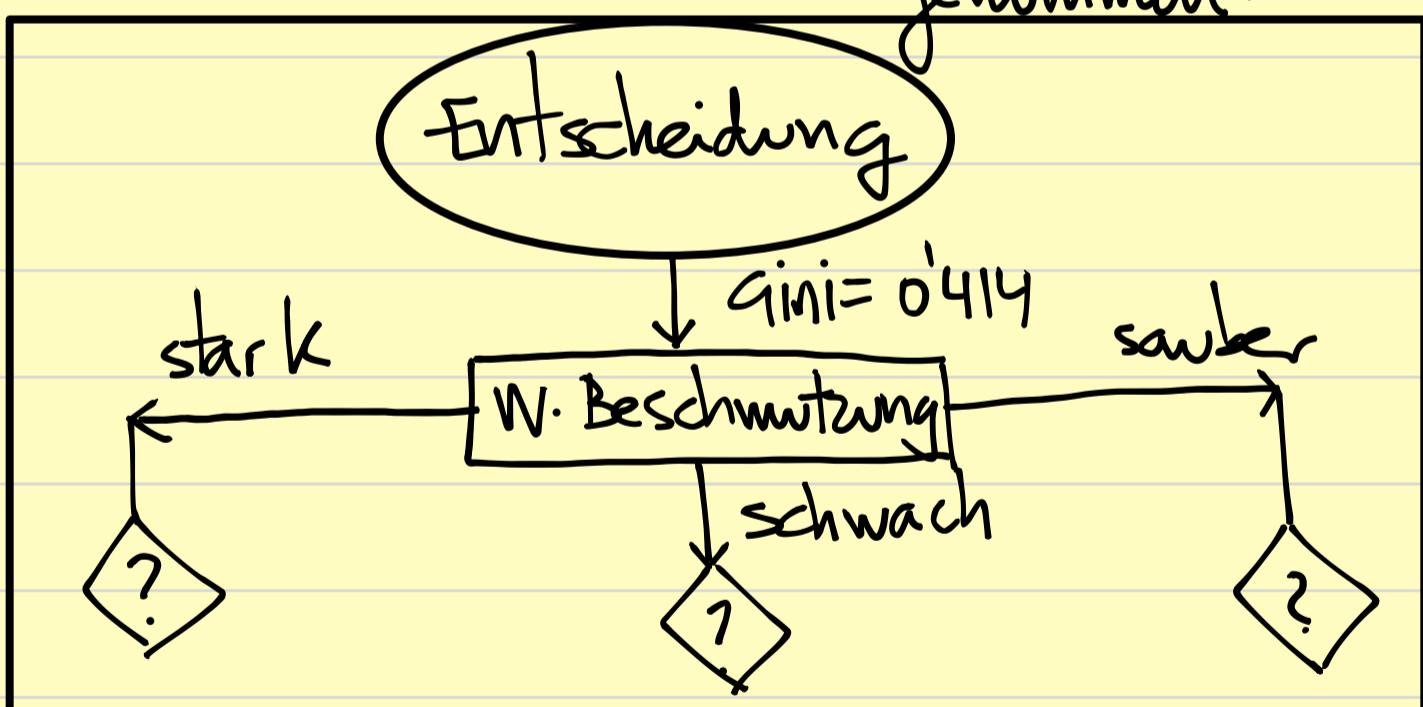
W. Besch. 0'414

Sinn.Gesp. 0'475

Fitness 0'472

Mond 0'44

Die Variable mit der geringsten Verunreinigung ist W. Beschmutzung und wird als 1. Knoten in dem Entscheidungsbau genommen.



W. Beschm.	Sinnvolle Gespräche	Ja	Nein	#
stark	oft	3	0	3
selten	selten	0	1	1

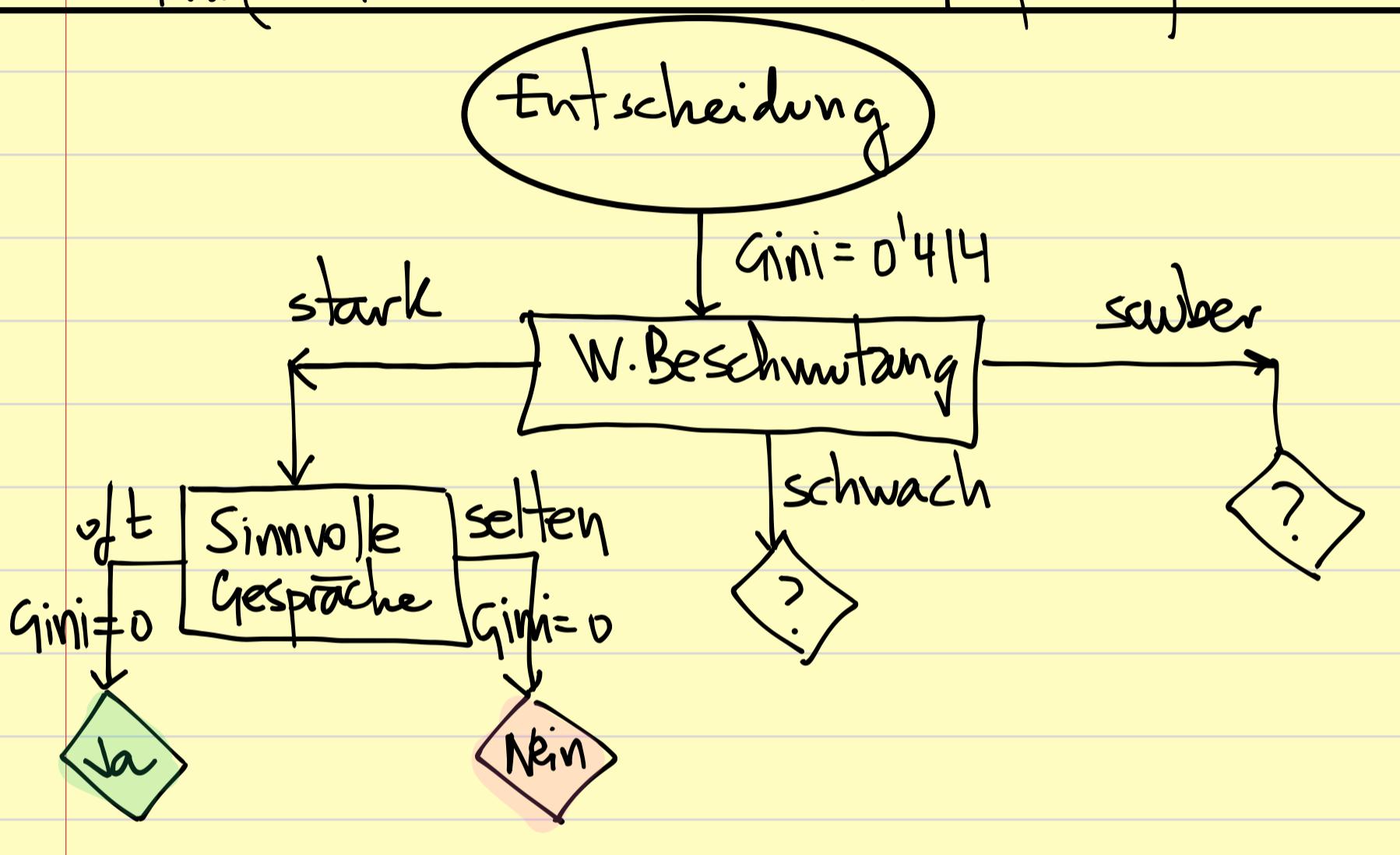
$Gini(W.\text{Besch starker + Sinnvolle Gespräche oft}) =$

$$= 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$Gini(W.\text{Besch. stark + Sinnvolle Gespräche selten}) =$

$$= 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2 = 0$$

$Gini(W.\text{Besch stark + Sinnvolle Gespräche}) = 0$



W.Beschwirung fitness Ja Nein #  
Schwach

"	hoch	0	0	0
"	mittel	0	1	1
"	gering	1	1	2

$Gini(W.\text{Beschm. schwach + Fitness}) =$

$$= \frac{1}{3} \cdot \left[ 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2 \right] + \frac{2}{3} \cdot \left[ 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 \right] = 0'33$$

# W. Beschmutzung Mond Ja Nein #

---

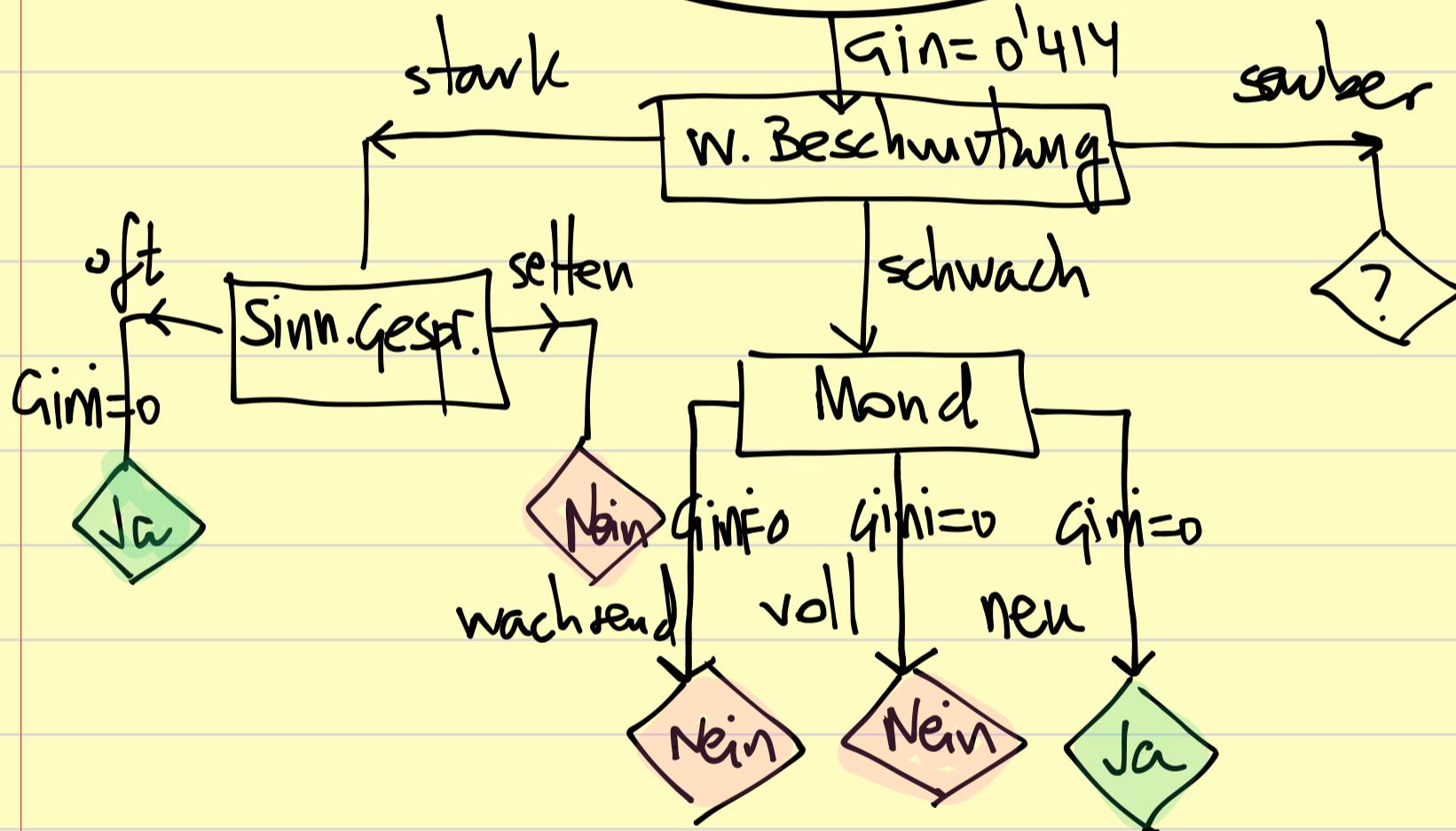
Schwach

wachsend	0	1	1
voll	0	1	1
neu	1	0	1

	SEX	JA / NEIN	WOHNUNGS- BESCHMUTZUNG	SINNLICHE GESPRÄCHE	FITNESS NIVEAU	MOND	SEX
1.	stark	oft	hoch	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	new	ja		
10.	sauber	oft	hoch	new	nein		

$$\text{Gini}(\text{W. Besch. schwach + Mond}) = 0$$

Entscheidung



# W. Beschmutzung Fitness Ja Nein #

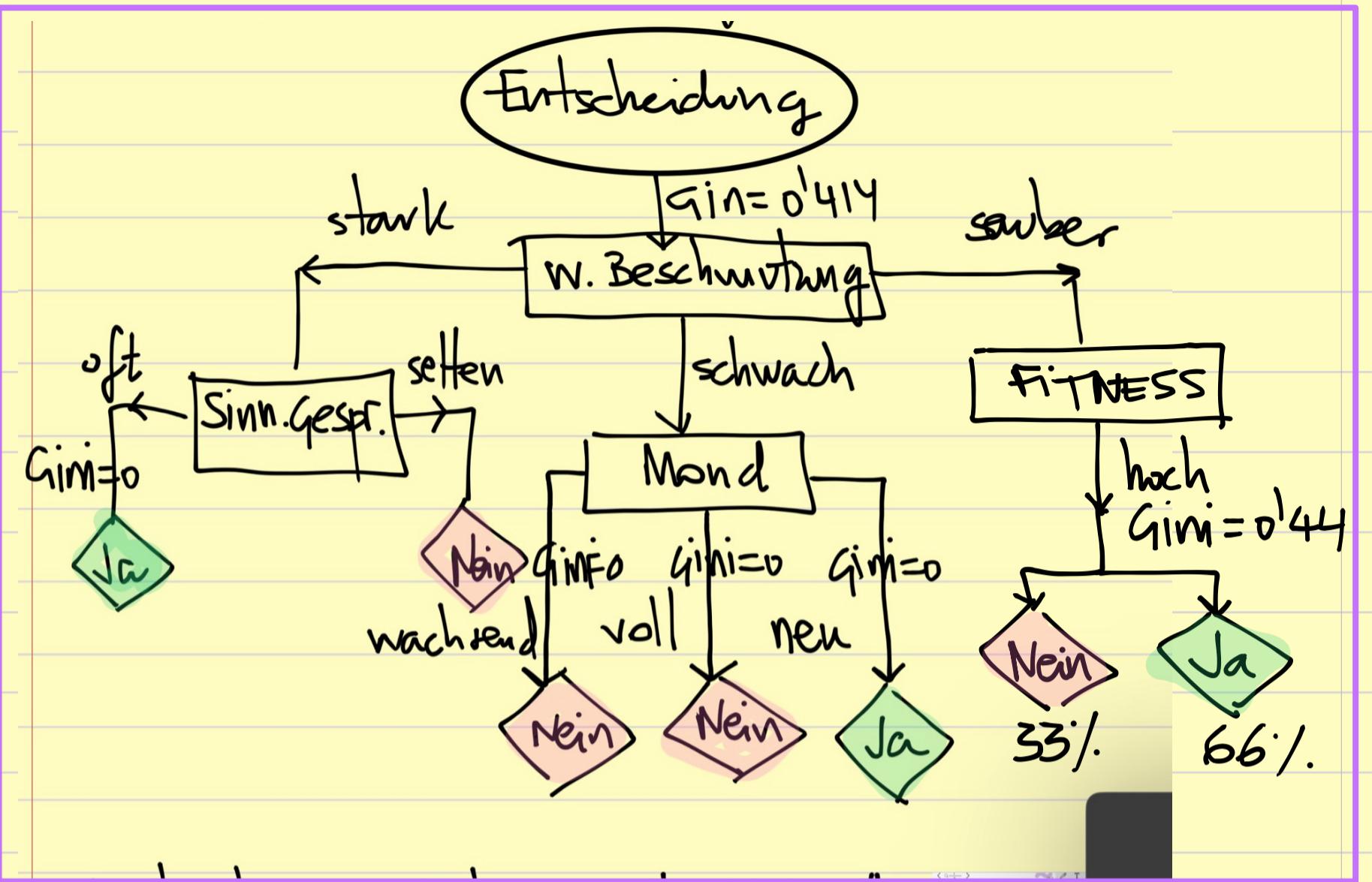
---

sauber

„	hoch	2	1	3
---	------	---	---	---

	SEX	JA / NEIN	WOHNUNGS- BESCHMUTZUNG	SINNLICHE GESPRÄCHE	FITNESS NIVEAU	MOND	SEX
1.	stark	oft	hoch	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	new	ja		
10.	sauber	oft	hoch	new	nein		

$$\text{Gini}(\text{W. Besch sauber + Fitness}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.44$$



## Übung: Datensatz für Entscheidungsbaum CART.

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	strong	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

Ziel: Entscheidung ob ein Fußballspiel stattfindet anhand Wetterbedingungen

