# 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

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

z.B. [Apfel, Orange, Apfel, Orange, Banane]

$$\text{Gini-Index Apfel} = 1 - \left(\frac{2}{5}\right)^2$$

- Gini Index = 0 → die Stichprobe ist vollkommend homogen
  [Apfel, Apfel, Apfel, Apfel, Apfel]

$$\text{Gini-Index Apfel} = 1 - \left(\frac{5}{5}\right)^2 = 0$$

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 |

# 1. Ebene.

| W. verschmutzung. | Ja | Nein | # |
|---|---|---|---|
| stark | 3 | 1 | 4 |
| schwach | 1 | 2 | 3 |
| sauber | 2 | 1 | 3 |
| | | | 10 |

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

$$Gini(W.V\,schw) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0'474$$

$$Gini(W.V\,saub) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0'474$$

$$Gini(W.V) = \frac{4}{10}\cdot 0'375 + \frac{3}{10}\cdot 0'474 + \frac{3}{10}\cdot 0'474 = 0'4344$$

| Sinnvolle Gespräche | Ja | Nein | # |
|---|---|---|---|
| oft | 4 | 3 | 7 |
| selten | 2 | 1 | 3 |
| | | | 10 |

$$Gini(S.G.\ oft) = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0'489$$

$$Gini(S.G.\ selten) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0'474$$

$$Gini(S.G.) = \frac{7}{10} \cdot 0'489 + \frac{3}{10} \cdot 0'474 = 0'485$$

| Fitness. Niveau. | Ja | Nein | # |
|---|---|---|---|
| hoch | 3 | 2 | 5 |
| mittel | 1 | 1 | 2 |
| gering | 2 | 1 | 3 |

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

$$Gini(F.N.\ mittel) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0'5$$

$$Gini(F.N.\ gering) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0'474$$

$$Gini(F.N.) = \frac{5}{10} \cdot 0'48 + \frac{2}{10} \cdot 0'5 + \frac{3}{10} \cdot 0'474 = 0'4822$$

| Mond | Ja | Nein | # |
|---|---|---|---|
| voll | 3 | 2 | 5 |
| wachsend | 1 | 1 | 2 |
| abnehmend | 1 | 0 | 1 |
| neu | 1 | 1 | 2 |

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

$$Gini(Mond\ wach) = 0{,}5$$

$$Gini(Mond\ ab) = 0$$

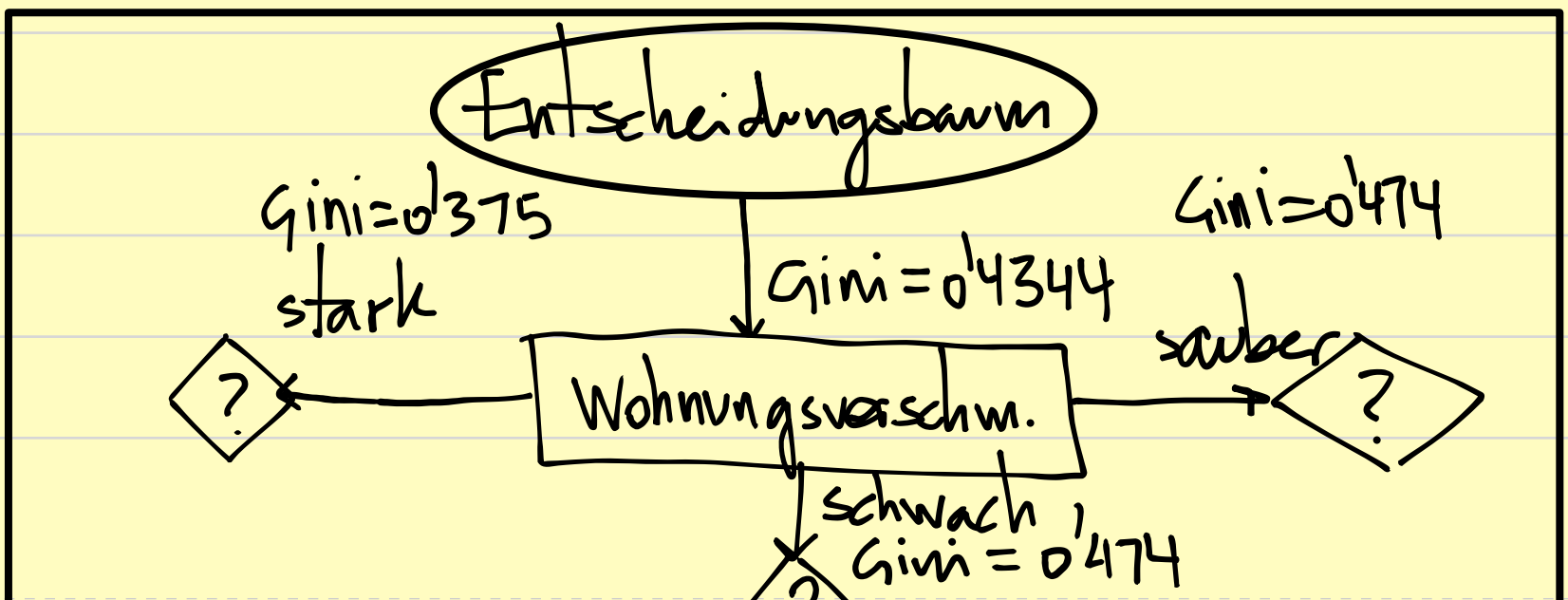
$$Gini(Mond\ neu) = 0{,}5$$

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

Gini Index 1. Ebene.

| | |
|---|---|
| W.V. | 0,4344 |
| S.G. | 0,485 |
| F.N. | 0,4822 |
| Mond | 0,44 |

→ 1. Entscheidungskriterium ist die W.V. weil die Informationen am saubersten liegt.



?

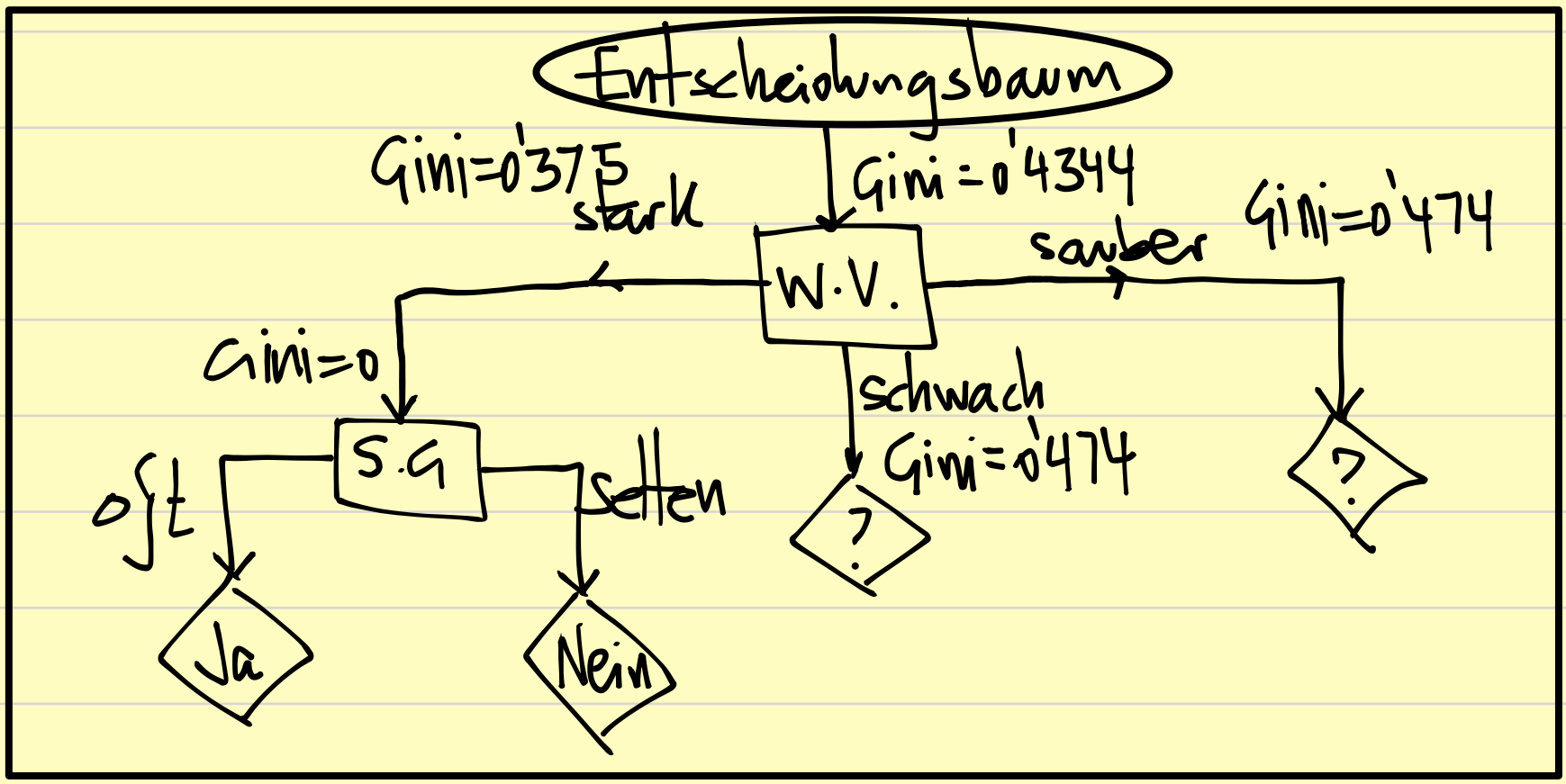
- W.V. STARK + { S.G. / F.N. / M

| W.V. STARK + S.G. | Ja | Nein | # |
|---|---|---|---|
| oft | 3 | 0 | 3 |
| selten | 0 | 1 | 1 |

$$Gini(WV\,stark + SG\,oft) = 1 - \left(\frac{3}{3}\right)^2 = 0$$

$$Gini(WV\,stark + SG\,selten) = 1 - \left(\frac{1}{1}\right)^2 = 0$$

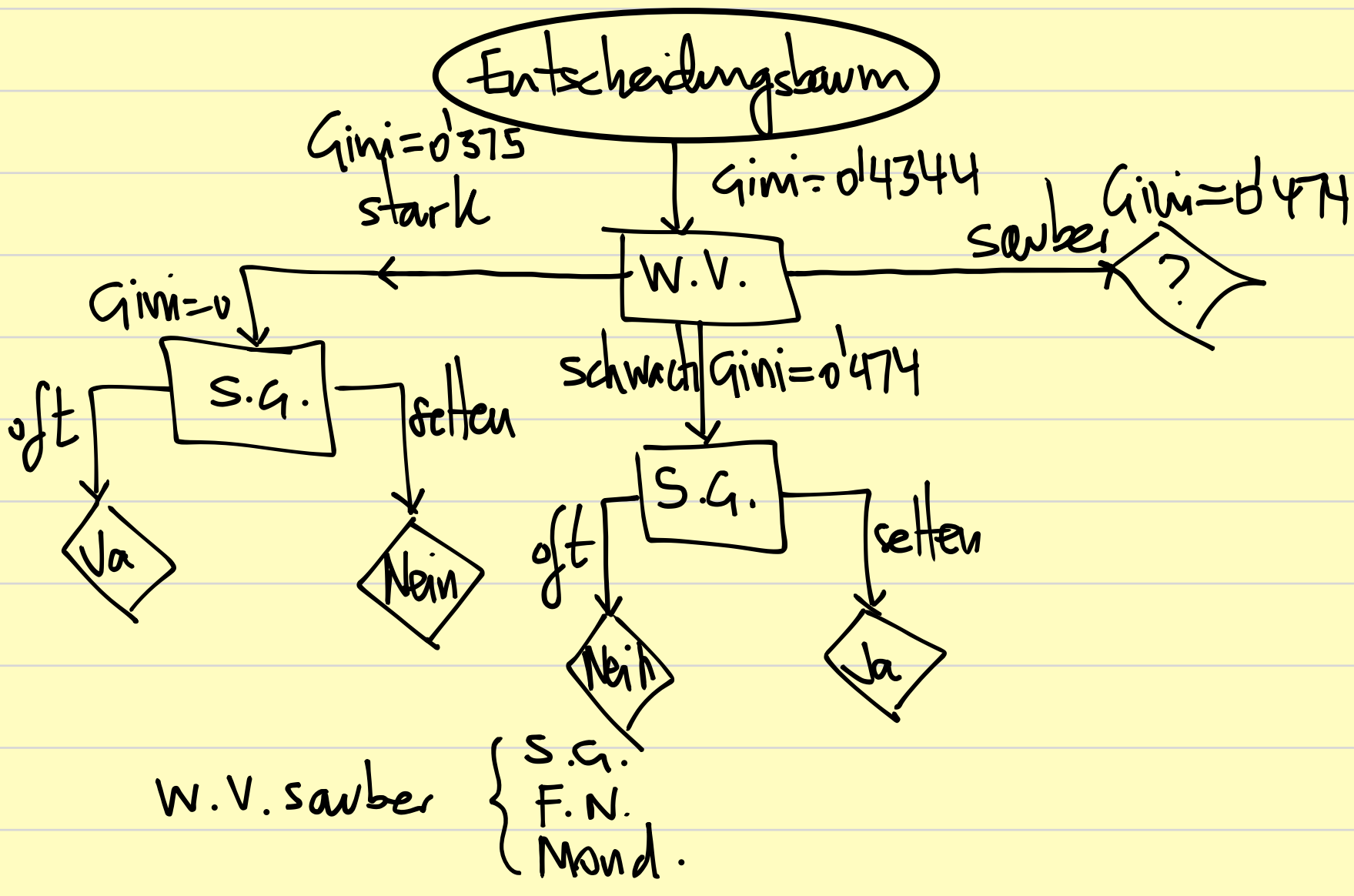
$$Gini(WV\,stark + SG) = \frac{3}{4} \cdot 0 + \frac{1}{4} \cdot 0 = 0$$




W.V. Schwach + { S.G. / F.N. / Mond.

| W.V. Schwach. | S.G. | Ja | Nein | # |
|---|---|---|---|---|
| oft | | 0 | 2 | 2 |
| selten | | 1 | 0 | 1 |

Gini (W.V. schwach + S.G.) = 0




W.V. sauber { S.G. / F.N. / Mond.

| W.V. Sauber + S.G. | Ja | Nein | # |
|---|---|---|---|
| oft | 1 | 1 | 2 |
| selten | 1 | 0 | 1 |

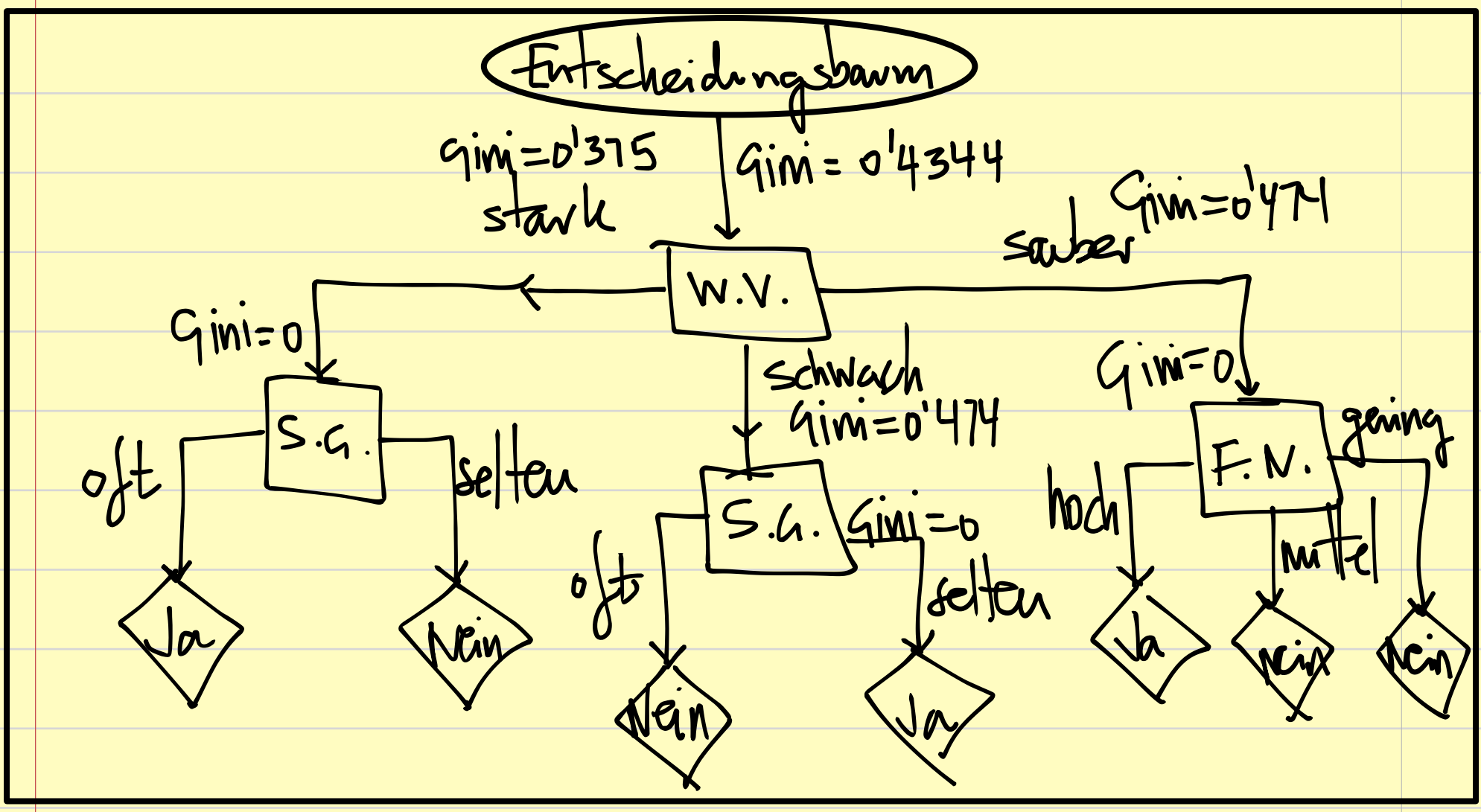
Gini(W.V. sauber + S.G. oft) = 0'5

Gini(W.V. sauber + S.G. selten) = 0

$Gini(W.V.\ sauber + S.G.) = \frac{2}{3} \cdot 0'5 + 0 = 0'33$

| W.V. Sauber + F.N. | Ja | Nein | # |
|---|---|---|---|
| hoch | 3 | 0 | 3 |
| mittel | 0 | 0 | 0 |
| gering | 0 | 0 | 0 |

Gini(W.V. sauber + F.N) = 0


Entscheidungsbaum
Gini=0'375
stark
Gini=0'4344
W.V.
sauber
Gini=0'474
Gini=0
S.G.
oft
selten
Ja
Nein
schwach
Gini=0'474
S.G.
Gini=0
oft
selten
Nein
Ja
Gini=0
F.V.
hoch
mittel
gering
Ja
nein
nein