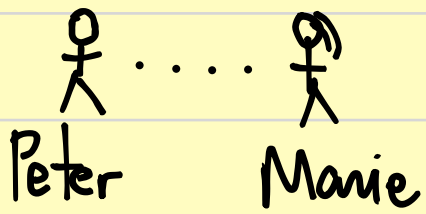
# Entscheidungsbäume (CART)

↖ Regression Trees




**Verunreinigung der Information.** Ein Dataset ist "sauber" (statistisch sauber), wenn der GINI-Index des Datasets NULL ist, und ist „schmutzig", wenn der GINI-Index EINS ist.

Dataset: {Apfel, Orange, Erdbeer, Apfel, Zitrone}

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

$p_i$ = Wahrscheinlichkeit dafür, dass „i" auftritt.

$$Gini = 1 - \left[\frac{2}{5}\right]^2 - \left[\frac{1}{5}\right]^2 - \left[\frac{1}{5}\right]^2 - \left[\frac{1}{5}\right]^2 = 0{,}72$$

Apfel Orange Erdbeer

Wir nutzen das Konzept um Entscheidungen zu treffen, auf basis von vorherigen Entscheidungen.

SEX Ja/Nein.

| | Sauberkeit | Emotionale Gespräche | Fitness | Mond | Sex |
|---|---|---|---|---|---|
| 1. | schmutzig | oft | mittel | voll | Ja |
| 2. | schmutzig | selten | intensiv | abnehmend | Nein |
| 3. | sauber | selten | kein | neu | Ja |
| 4. | mittel.sauber | oft | mittel | zunehmend | Ja |
| 5. | schmutzig | selten | intensiv | voll | Nein |
| 6. | schmutzig | oft | kein | abnehmend | Ja |
| 7. | sauber | selten | mittel | neu | Ja |
| 8. | schmutzig | oft | kein | zunehmend | Nein |
| 9. | sauber | selten | mittel | voll | Ja |
| 10. | mittel·sauber | oft | intensiv | abnehmend | Nein |

Wir gehen durch die Kriterien und suchen wir den mit dem kleinsten GINI·Index.

| Sauberkeit | Ja | Nein | # |
|---|---|---|---|
| schmutzig | 2 | 3 | 5 |
| mittel·sauber | 1 | 1 | 2 |
| sauber | 3 | 0 | 3 |
| | | | 10 |

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

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

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

$$Gini(\text{Sauberkeit}) = \frac{5}{10} \cdot 0'48 + \frac{2}{10} \cdot 0'5 + \frac{3}{10} \cdot 0 = 0'34$$

| Emotionale Gespräche (EG) | Ja | No | # |
|---|---|---|---|
| oft | 3 | 2 | 5 |
| selten | 3 | 2 | 5 |

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

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

$$Gini(EG) = \frac{5}{10} \cdot 0'48 + \frac{5}{10} \cdot 0'48 = 0'48$$

| Fitness | Ja | Nein | # |
|---|---|---|---|
| kein | 2 | 1 | 3 |
| mittel | 4 | 0 | 4 |
| intensiv | 0 | 3 | 3 |

$$Gini(F\ kein) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0'44$$

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

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

$$Gini(Fitness) = \frac{3}{10} \cdot 0'44 + 0 + 0 = 0'133$$

| Mond | Ja | Nein | # |
|---|---|---|---|
| voll | 2 | 1 | 3 |
| abnehm. | 1 | 2 | 3 |
| neu | 2 | 0 | 2 |
| Zunehm. | 1 | 1 | 2 |

$$Gini(M\ voll) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0'44$$

$$Gini(M\ abn.) = 0'44$$

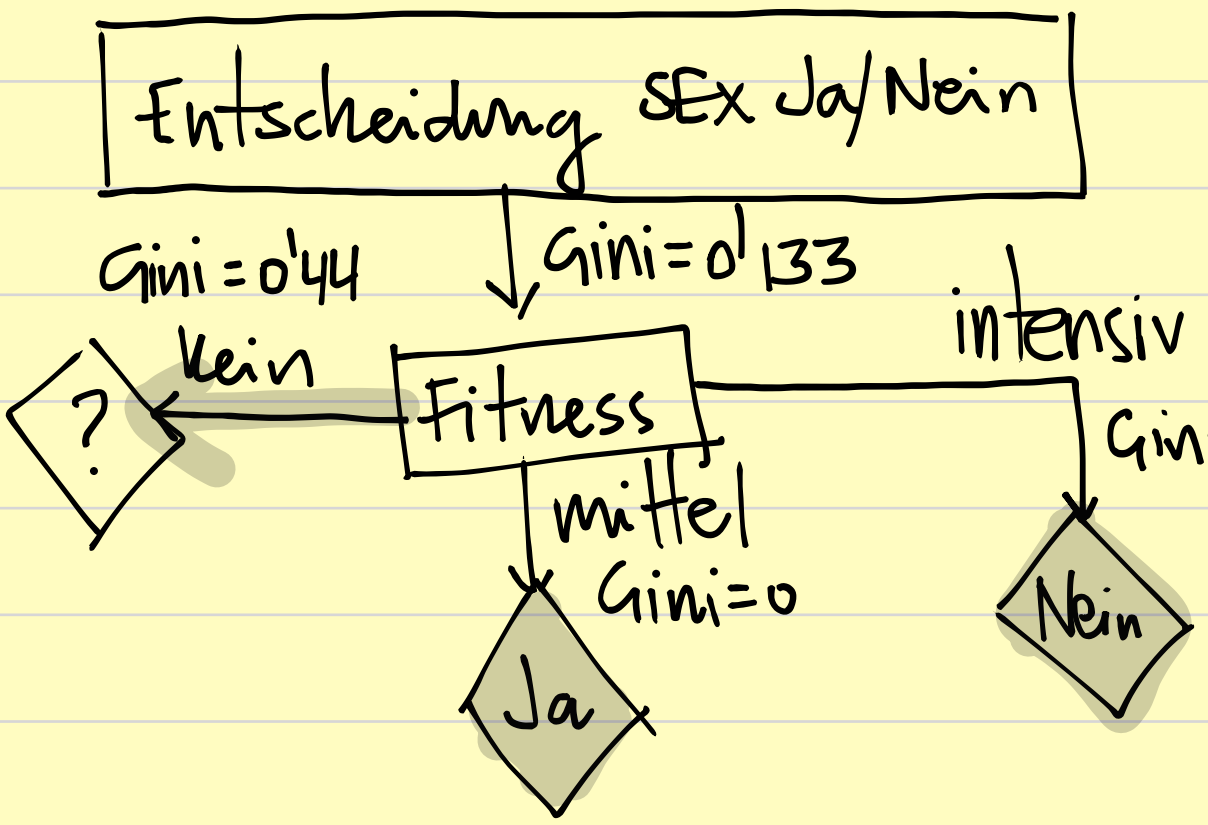
$$Gini(M\ neu) = 0$$

$$Gini(M\ zun.) = 0'5$$

$$Gini(Mond) = \frac{3}{10} \cdot 0'44 + \frac{3}{10} \cdot 0'44 + 0 + \frac{2}{10} \cdot 0'5 = 0'364$$

| | Gini |
|---|---|
| Sauberkeit | 0'34 |
| EG | 0'48 |
| F | 0'133 |
| M | 0'364 |

Der erste Knote in unserem Baum ist FITNESS mit dem kleinsten Gini




| Fitness kein + Sauberkeit | Ja | Nein | # |
|---|---|---|---|
| schmutzig | 1 | 1 | 2 |
| sauber | 1 | 0 | 1 |

/3

$$Gini(F\ kein + Saub.\ schm.) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0'5$$

$$Gini(F\ kein + Saub.\ saub) = 0$$

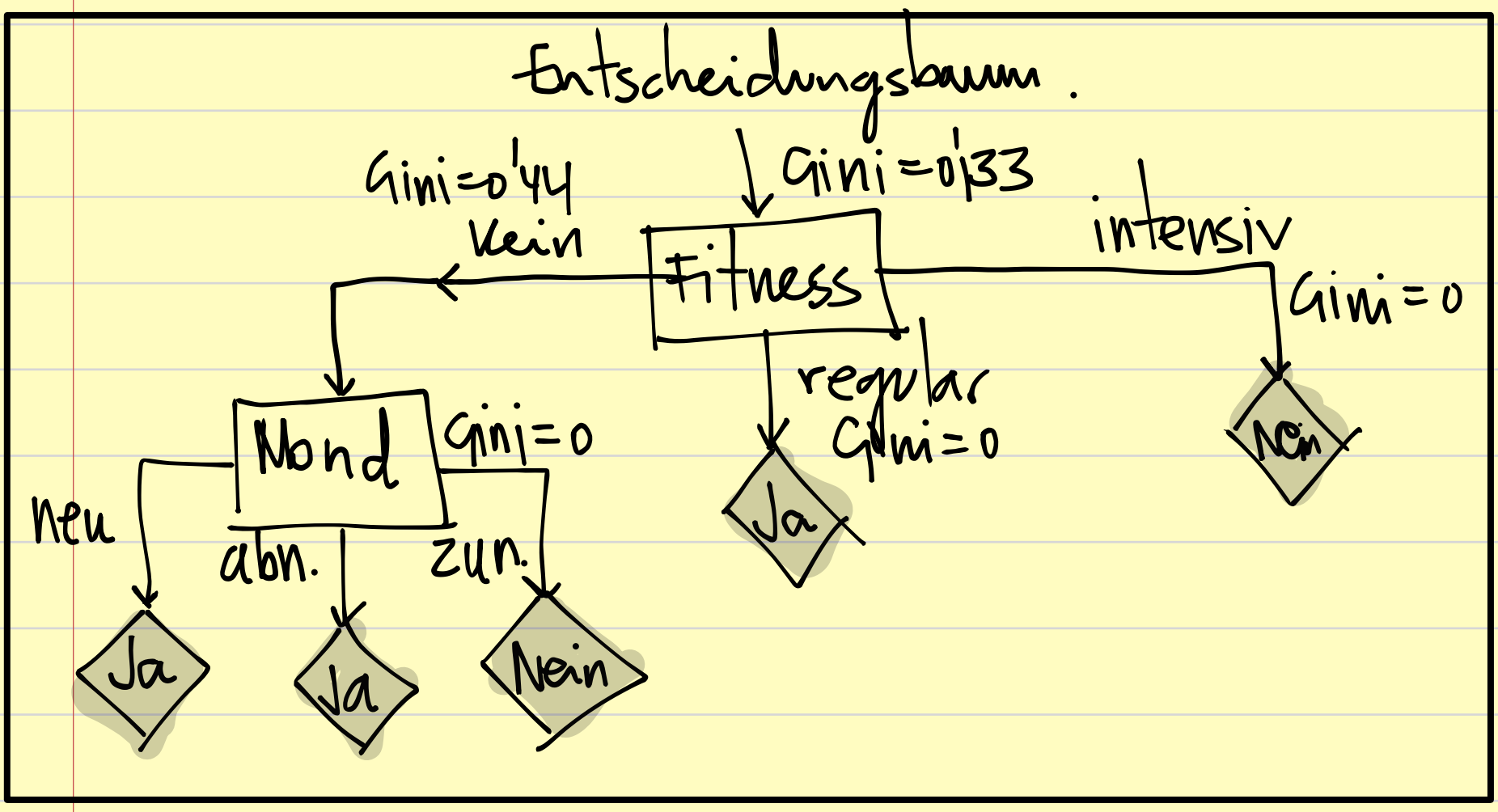
$$Gini(F\ kein + Sauberkeit) = \frac{2}{3} \cdot 0'5 + \frac{1}{3} \cdot 0 = 0'33$$

| Fitness kein + EG. | Ja | Nein | # |
|---|---|---|---|
| oft | 1 | 1 | 2 |
| selten | 1 | 0 | 1 |

Gini (Fitness kein + EG) = 0'33

| Fitness kein + Mond | Ja | Nein | # | |
|---|---|---|---|---|
| neu | 1 | 0 | 1 | Gini=0 |
| abnehmend | 1 | 0 | 1 | Gini=0 |
| voll | 0 | 0 | 0 | |
| zunehmend | 0 | 1 | 1 | Gini=0 |

Gini (F kein + Mond) = 0



Verein. Fußball spielen Ja/Nein

| 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 | Ja | Nein | # |
|---|---|---|---|
| sunny | 2 | 3 | 5 |
| overcast | 4 | 0 | 4 |
| rainfall | 3 | 2 | 5 |
| | | | 14 |

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

$$Gini(Outlook\,Overcast) = 0$$

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

$$Gini(Outlook) = \left(\frac{5}{14}\cdot 0'48+\frac{5}{14}\cdot 0'48\right) = 0'343$$

| TEMP. | Ja | Nein | # |
|---|---|---|---|
| hot | 2 | 2 | 4 |
| mild | 4 | 2 | 6 |
| cool | 3 | 1 | 4 |
| | | | 14 |

$$Gini(T\,hot) = 0'5$$

$$Gini(T\,mild) = 1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2 = 0'44$$

$$Gini(T\,cool) = 1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2 = 0'375$$

$$Gini(Temp) = \frac{4}{14}\cdot 0'5+\frac{6}{14}\cdot 0'44+\frac{4}{14}\cdot 0'375 = 0'438$$

| LUFTFEUCHTIGKEIT | Ja | Nein | # |
|---|---|---|---|
| high | 3 | 4 | 7 |
| normal | 6 | 1 | 7 |
| | | | 14 |

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

$$Gini(LF\,n) = 1-\left(\frac{6}{7}\right)^2-\left(\frac{1}{7}\right)^2 = 0'245$$

$$Gini(LF) = 0'367$$

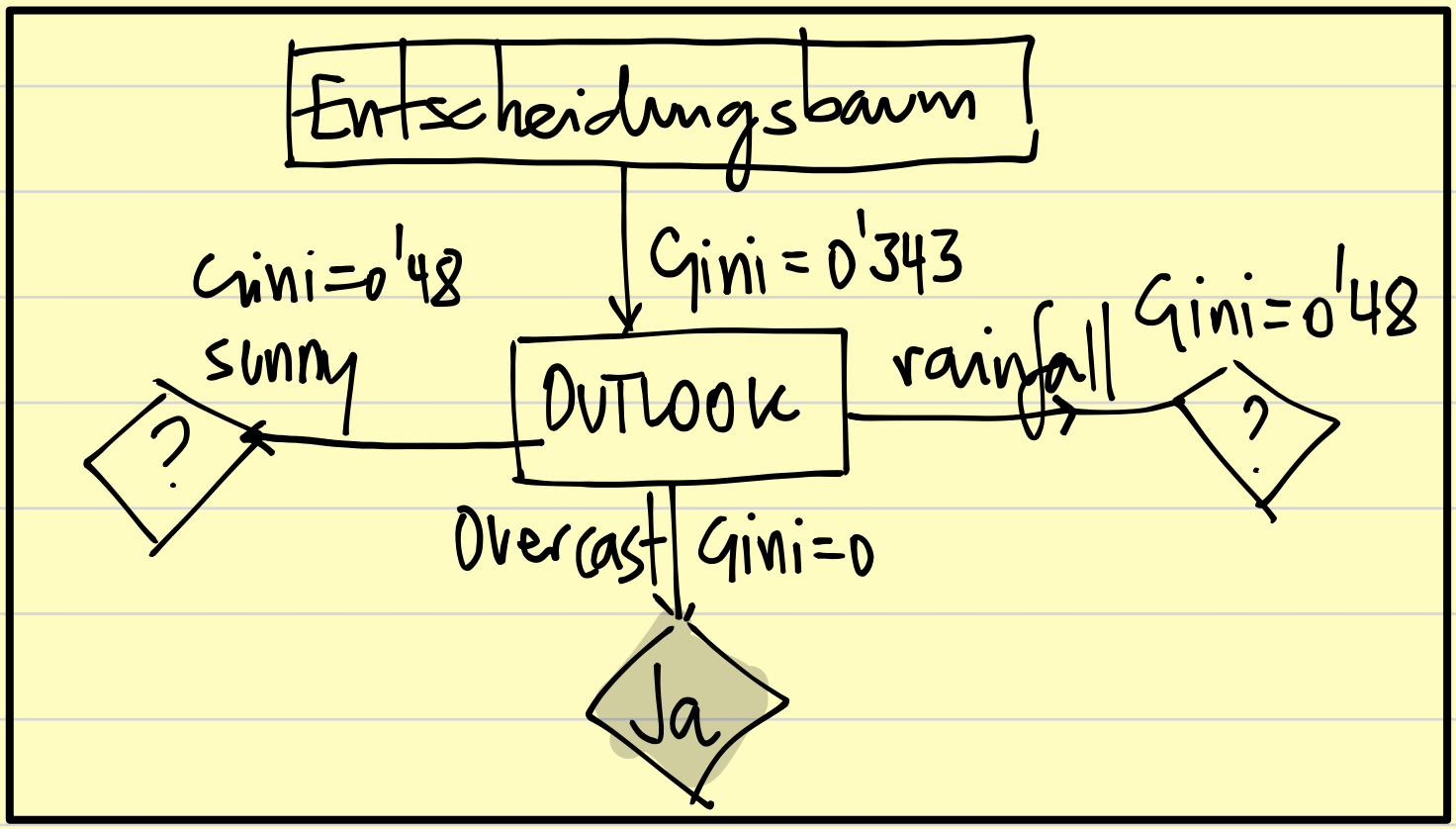
| WIND | Ja | Nein | # |
|---|---|---|---|
| weak | 6 | 2 | 8 |
| strong | 3 | 3 | 6 |

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

$$Gini(W\,strong) = 0'5$$

$$Gini(wind) = 0'428$$

| | Gini | |
|---|---|---|
| Outlook | 0'343 | → höchste Sauberkeit. |
| Temp. | 0'438 | |
| LF | 0'367 | |
| Wind | 0'428 | |



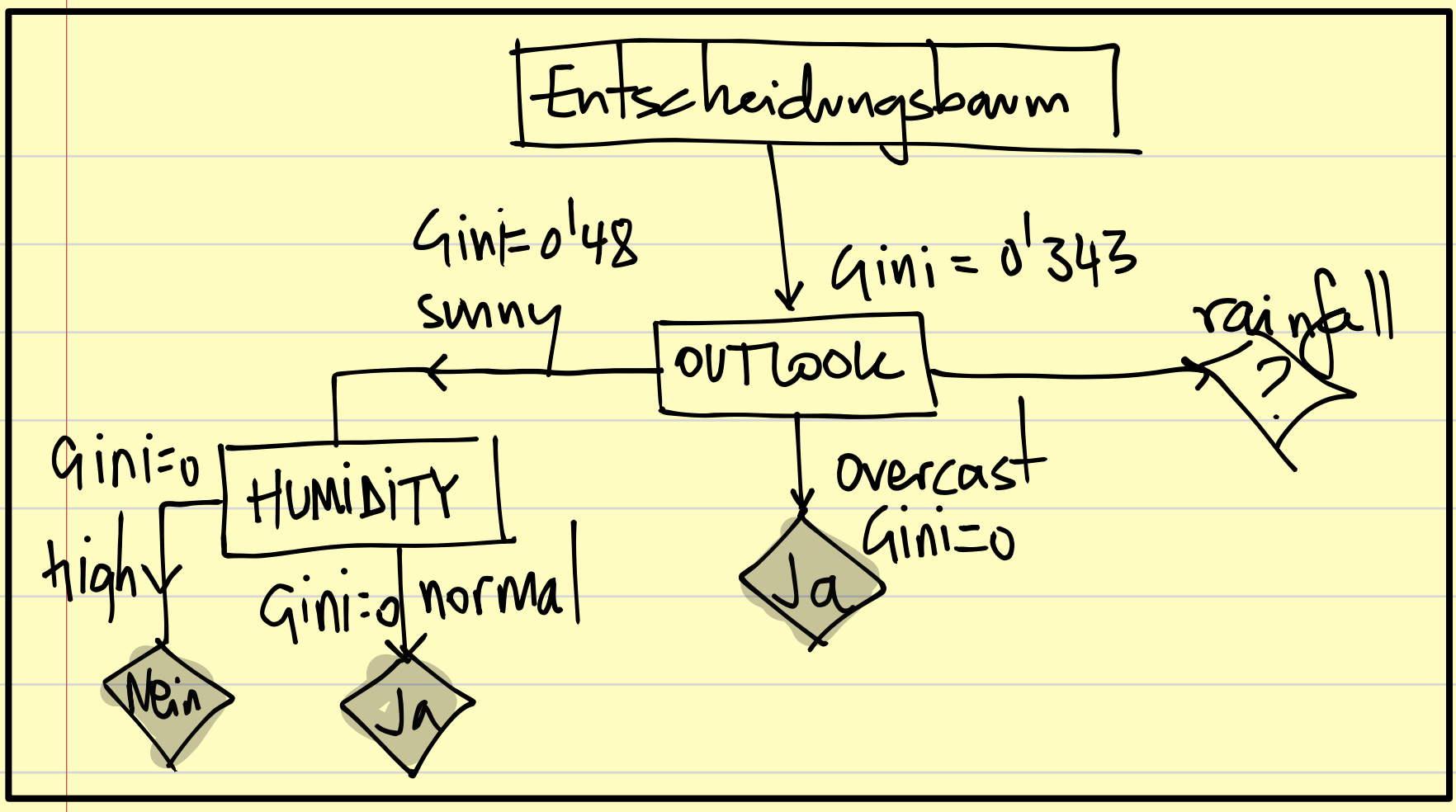

| Outlook S + Temp | Ja | Nein | # | |
|---|---|---|---|---|
| hot | 0 | 2 | 2 | → Gini = 0 |
| mild | 1 | 1 | 2 | → Gini = 0'5 |
| cool | 1 | 0 | 1/5 | → Gini = 0 |

$$Gini = \frac{2}{5} \cdot 0'5 = 0'2$$

(Outlook S + Temp)

| Outlook S + LF | Ja | Nein | # | |
|---|---|---|---|---|
| high | 0 | 3 | 3 | → Gini = 0 |
| normal | 2 | 0 | 2 | → Gini = 0 |

Gini(Outlook S + LF) = 0




| Overlook R+T | Ja | Nein | # | |
|---|---|---|---|---|
| mild | 2 | 1 | 3 | $\rightarrow Gini = 1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2 = 0'44$ |
| cool | 1 | 1 | 2 | $\rightarrow Gini = 0'5$ |

$$Gini(Overlook\ R+T) = \frac{3}{5}\cdot 0'44 + \frac{2}{5}\cdot 0'5 = 0'364$$

| Overlook R+W | Ja | Nein | # | |
|---|---|---|---|---|
| weak | 3 | 0 | 3 | $\rightarrow Gini(Over.R+W) = 0$ |
| strong | 0 | 2 | 2 | |