

Entscheidungsbäume (CART)

KONZEPT. VERUNREINIGUNG der INFORMATION

Die Verunreinigung misst die Homogenität in der Datenprobe. Wenn die Datenprobe homogen ist, gehören die Stichproben zur gleichen Klasse und die Verunreinigung ist 0.

Wir messen die Verunreinigung mit dem GINI-Index:

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

Gini-Index ist ein Maß für die Verunreinigung einer Stichprobe. Hat einen Wert zw. [0,1].

- Gini-Index = 0 bedeutet, dass die Stichprobe vollkommen homogen ist und alle Elemente ähnlich sind.
- Gini-Index = 1 bedeutet maximale Verunreinigung bzw Ungleichheit zw. den Elementen.

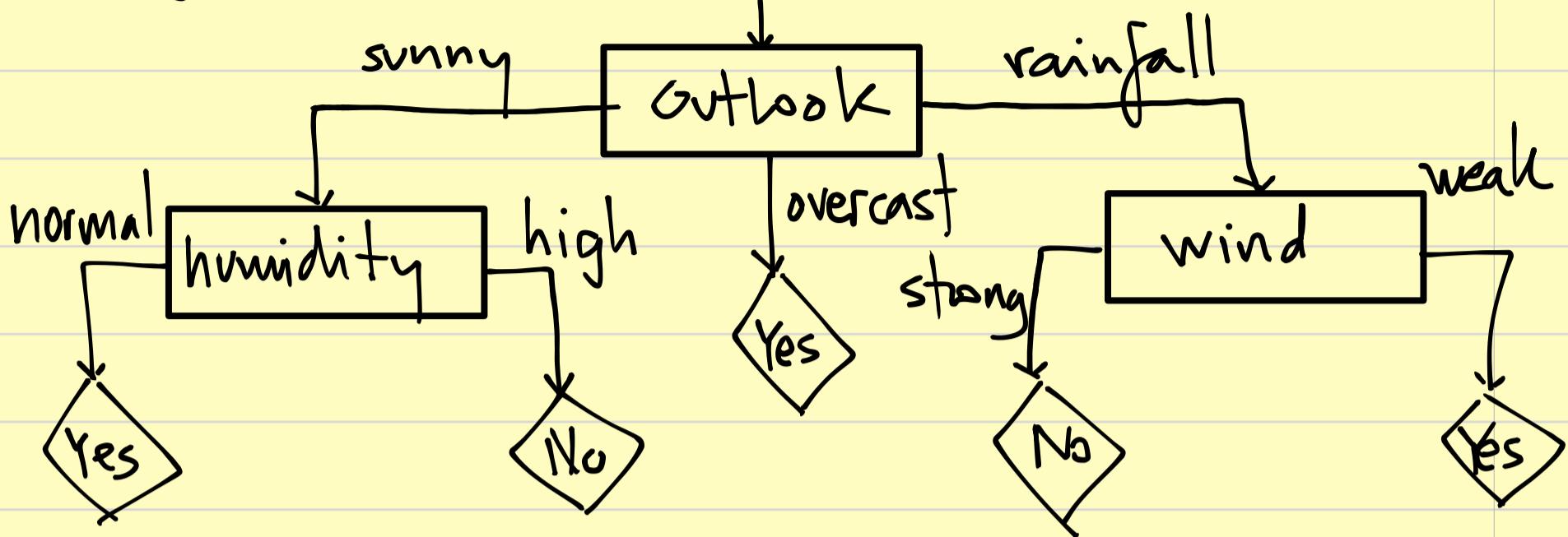
Beispiel

ZIEL. Findet ein Fußballspiel statt (Decision) basierend auf den Wetterbedingungen

Ergebnis: Entscheidungsbäum

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

Lösung:



Outlook: Outlook ist ein nominales Merkmal.
Es kann drei Werte annehmen (Sonnig, bewölkt, regen)

Outlook	Yes	No	#
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5

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

$$\sum = 14$$

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

$$Gini\text{-Index}(Outlook=overcast) = 1 - \left(\frac{4}{4}\right)^2 = 0 \quad (\bullet)$$

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

Die gewichtete Summe des Gini-Index für die Merkmale des Ausblicks ist

$$Gini\text{-Index(Ausblick)} = \frac{5}{14} \cdot 0'48 + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0'48 = 0'342$$

Temperature : Nominales Merkmal (heiß, kalt, mild)

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

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

$$Gini(\text{Temperature}=\text{cool}) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0'375$$

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

Gewichtete Summe Gini · Temperature:

$$Gini(\text{Temperature}) = \frac{4}{14} \cdot 0'5 + \frac{4}{14} \cdot 0'375 + \frac{6}{14} \cdot 0'445 = 0'439$$

Humidity : Nominal (high, normal)

Humidity	Yes	No	#
high	3	4	7
normal	6	1	7

$$Gini(\text{humidity}=\text{high}) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0'489$$

$$Gini(\text{humidity}=\text{normal}) = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0'244$$

Gewichtete Summe Humidity

$$Gini(\text{Humidity}) = \frac{7}{14} \cdot 0'489 + \frac{7}{14} \cdot 0'244 = 0'367$$

Wind: Nominal (strong, weak)

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

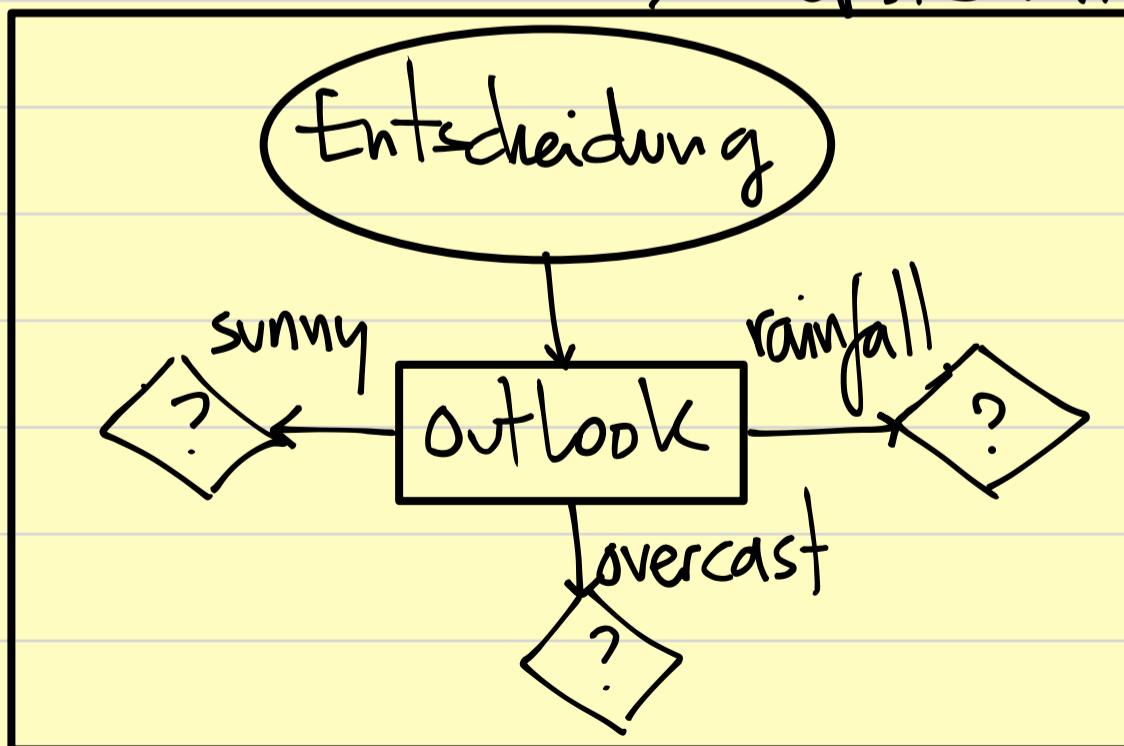
$$\text{Gini}(\text{Wind}=\text{strong}) = 1 - \left(\frac{6}{8} \right)^2 - \left(\frac{2}{8} \right)^2 = 0'375$$

$$\text{Gini}(\text{Wind}=\text{weak}) = 1 - \left(\frac{3}{6} \right)^2 - \left(\frac{3}{6} \right)^2 = 0'5$$

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

feature	Gini
outlook	0'342
Temperat.	0'439
Humidity	0'367
Wind	0'428

Der Gini Index zeigt, dass Outlook die sauberste Entscheidung anbietet. Wir wählen Outlook als erste Knoote.



Gini . Index : Sunny + Temperature .

outlook	Temperature	Yes	No	#
Sunny	hot	0	2	2
Sunny	cold	1	0	1
Sunny	mild	1	1	2

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

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

$$\text{Gini}(\text{sunny+mild}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$\text{Gini}(\text{sunny+Temperature}) = \frac{2}{5} \cdot 0 + \frac{1}{5} \cdot 0 + \frac{2}{5} \cdot 0.5 = 0.2$$

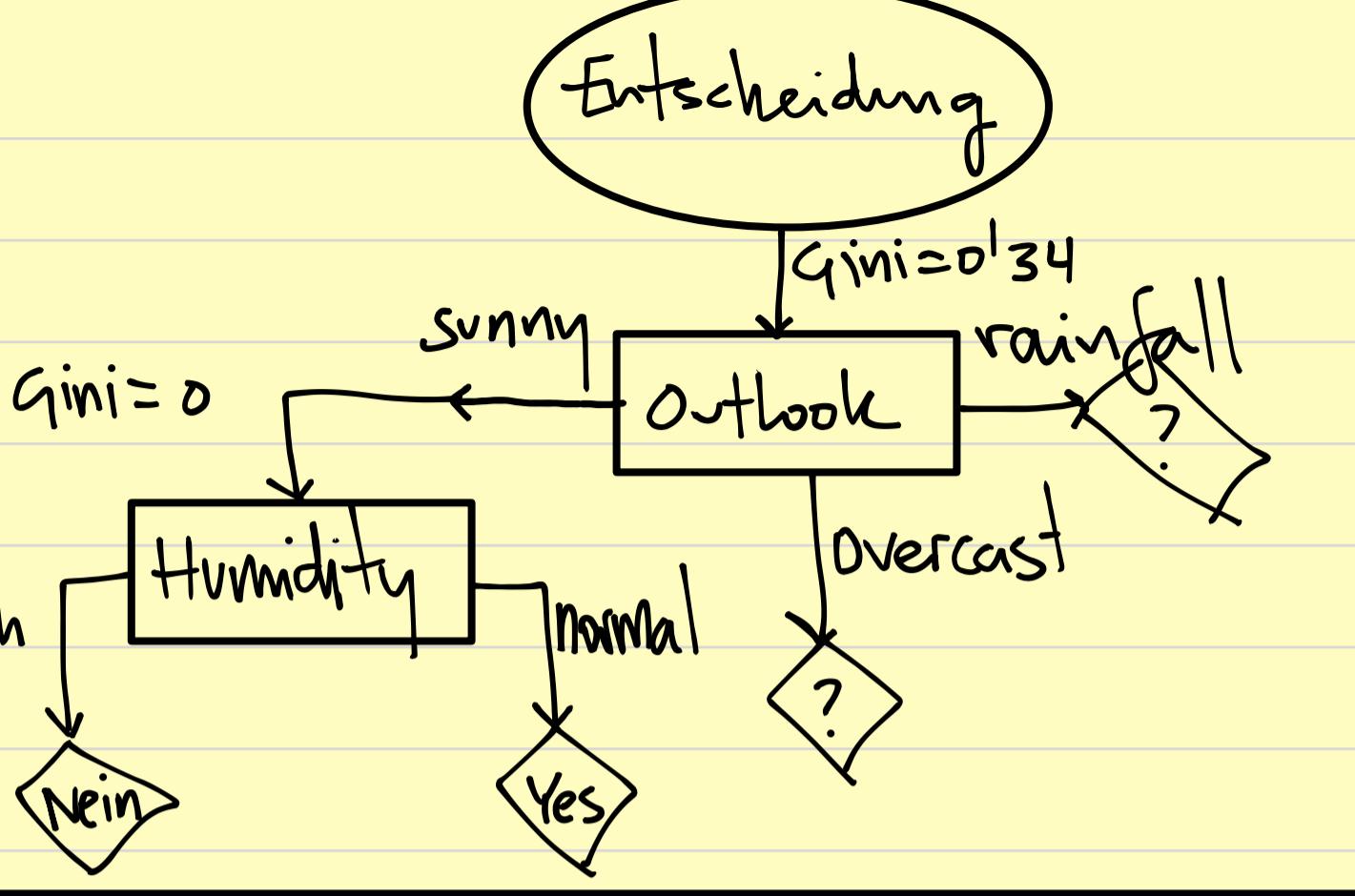
Gini Sunny + Humidity

outlook	humidity	Yes	No	#
sunny	high	0	3	3
sunny	normal	2	0	2

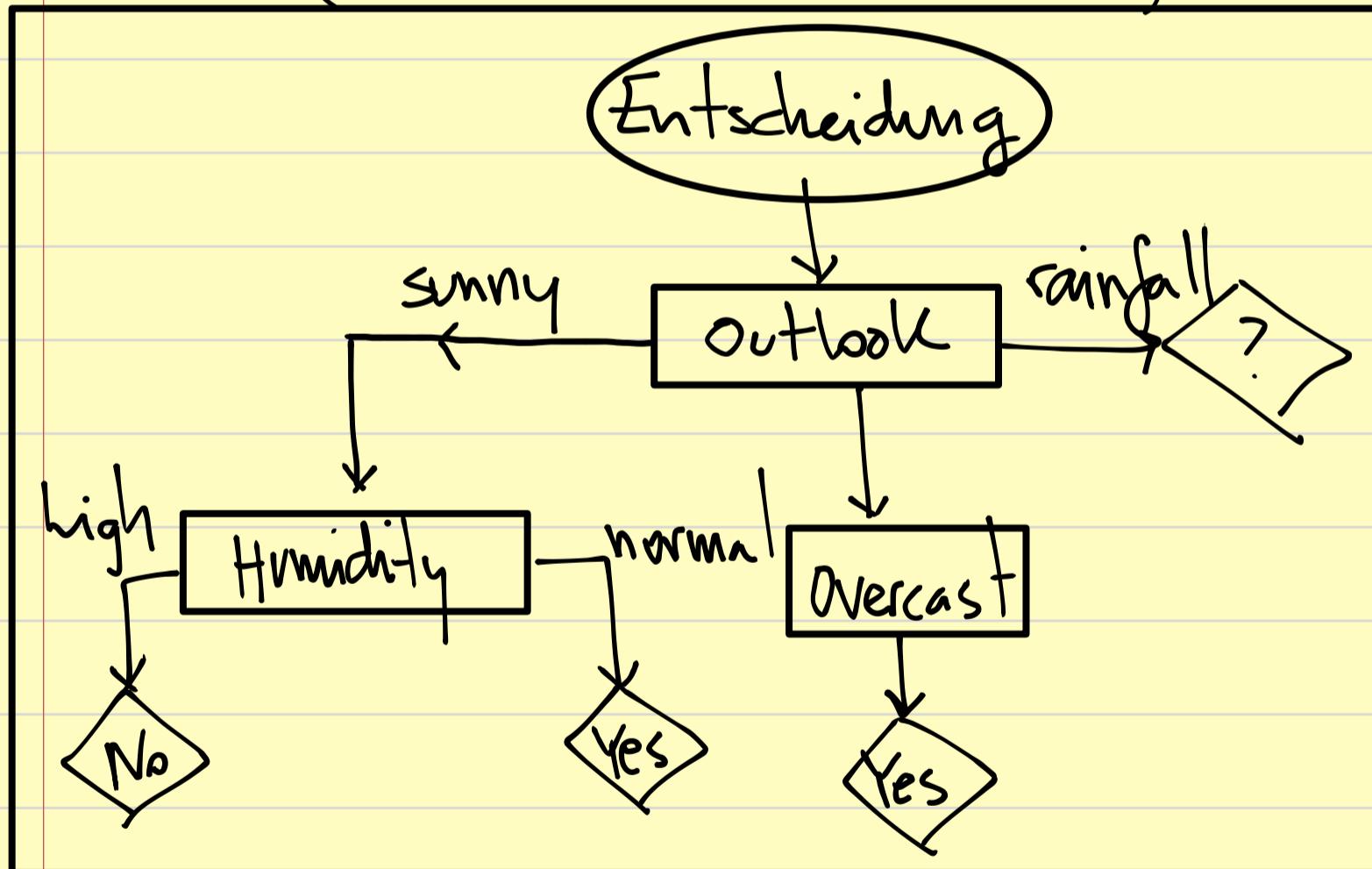
$$\text{Gini}(\text{sunny+high}) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$\text{Gini}(\text{sunny+normal}) = 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = 0$$

$$\text{Gini}(\text{sunny+humidity}) = \frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 0 = 0$$



$Gini(\text{Overcast}) = 0$ (sich oben) (*)



Gini (rainfall + Temperature)

outlook	Temperature	Yes	No	#
Rainfall	cool	1	1	2
Rainfall	mild	2	1	3

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

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

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

Gini (Rainfall + Wind)

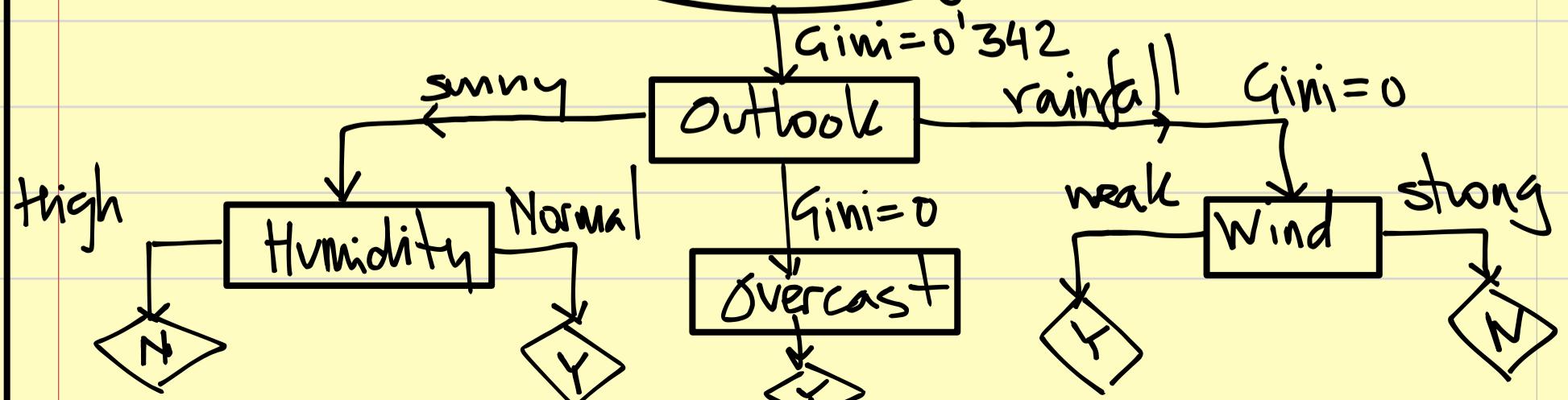
Outlook	Wind	Yes	No	#
Rainfall	weak	3	0	3
Rainfall	strong	0	2	2

$$\text{Gini}(\text{Rainfall} + \text{weak}) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$\text{Gini}(\text{Rainfall} + \text{strong}) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

$$\boxed{\text{Gini}(\text{Rainfall} + \text{wind}) = 0}$$

Entscheidung



$\text{Gini}(\text{Rainfall} + \text{Wind} = \text{Strong}) = 0$ (No)

$\text{Gini}(\text{Rainfall} + \text{Wind} = \text{Weak}) = 0$ (Yes)

