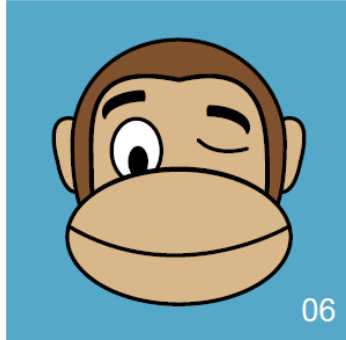


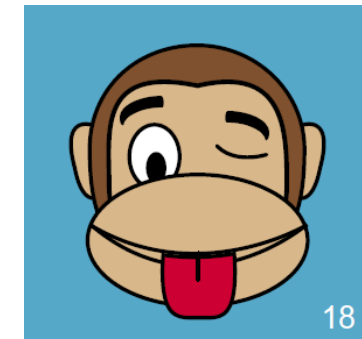
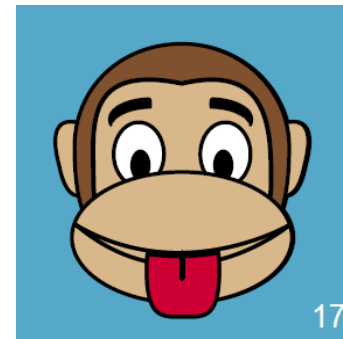
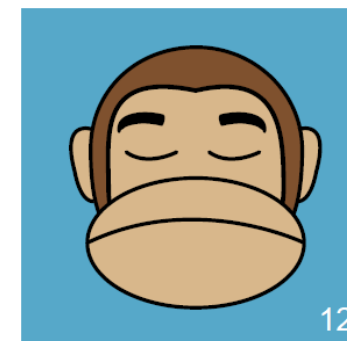
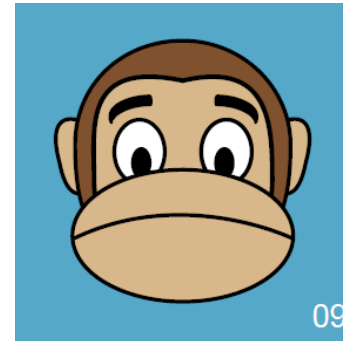
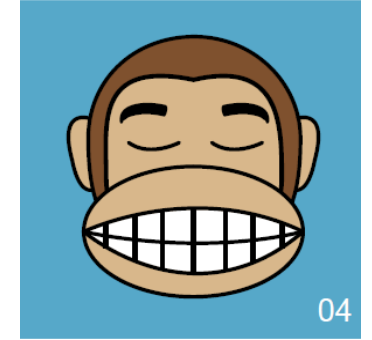
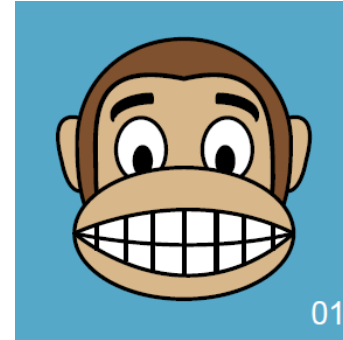
Überwachter Lernalgorithmus

Der Entscheidungsbaum

Beißt



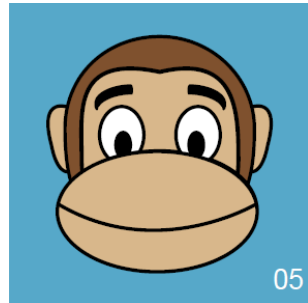
Beißt nicht



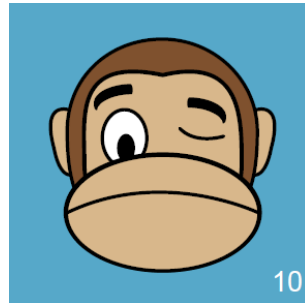
Haben Sie gelernt, wer beißt?



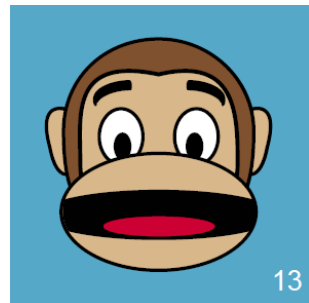
beißt



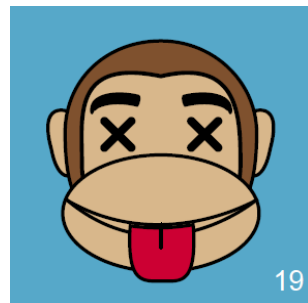
beißt



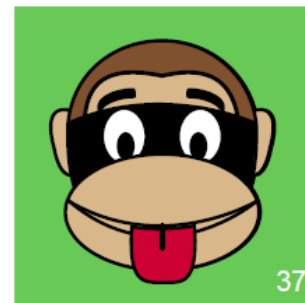
beißt nicht



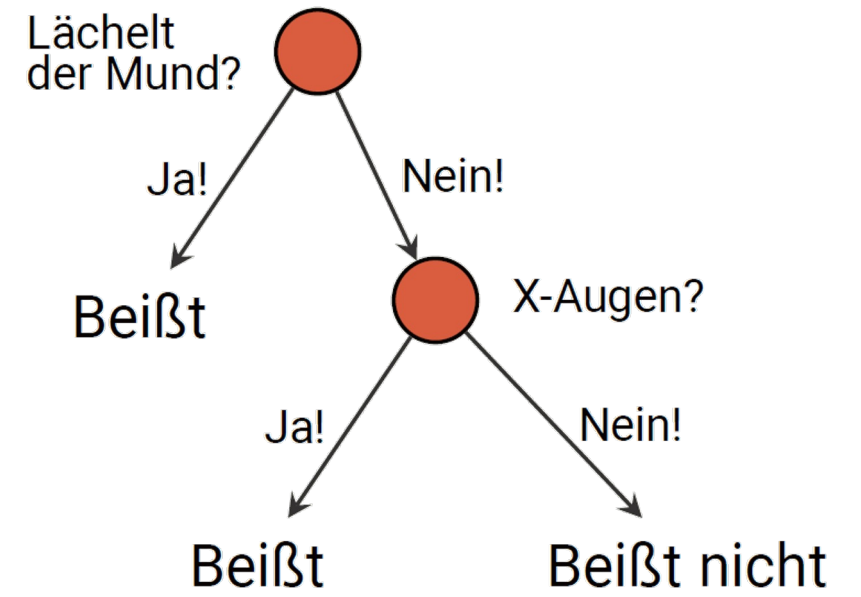
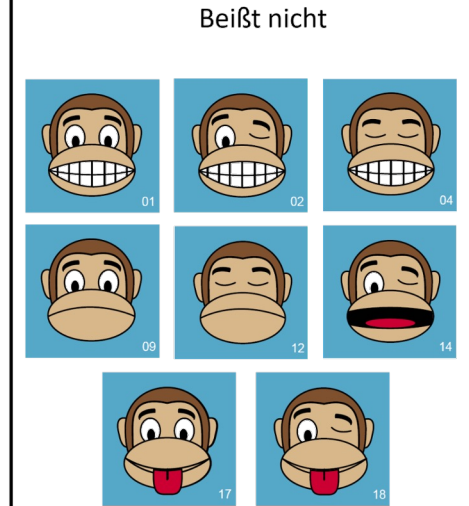
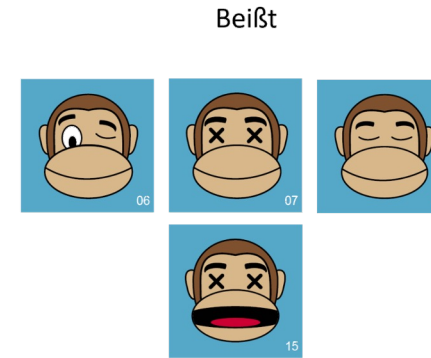
beißt nicht

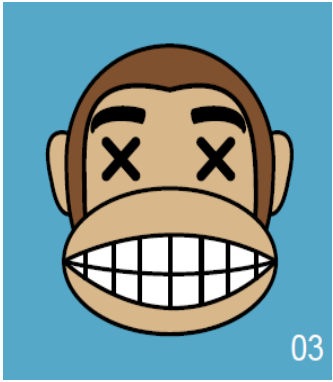


beißt



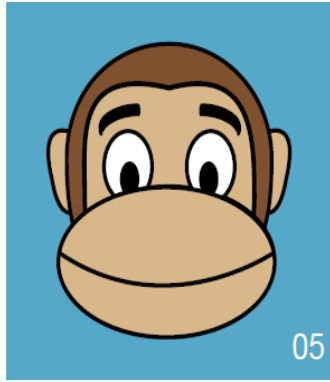
beißt nicht





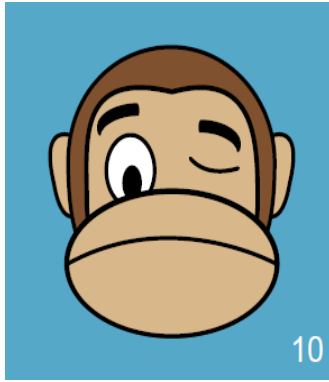
03

beißt



05

beißt



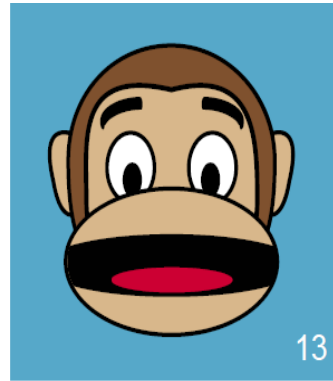
10

beißt nicht



11

beißt



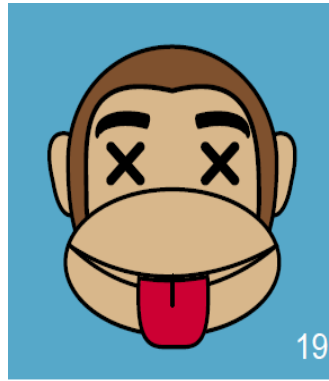
13

beißt nicht



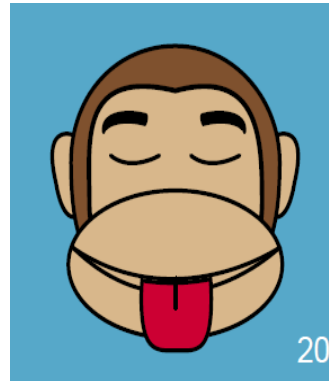
16

beißt nicht



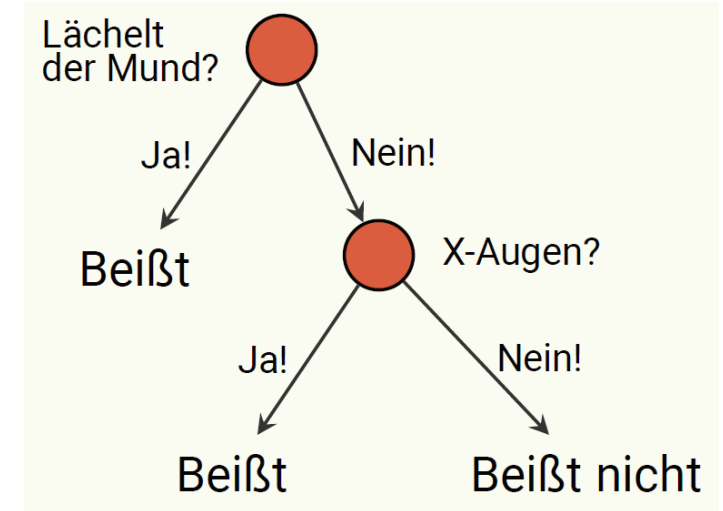
19

beißt



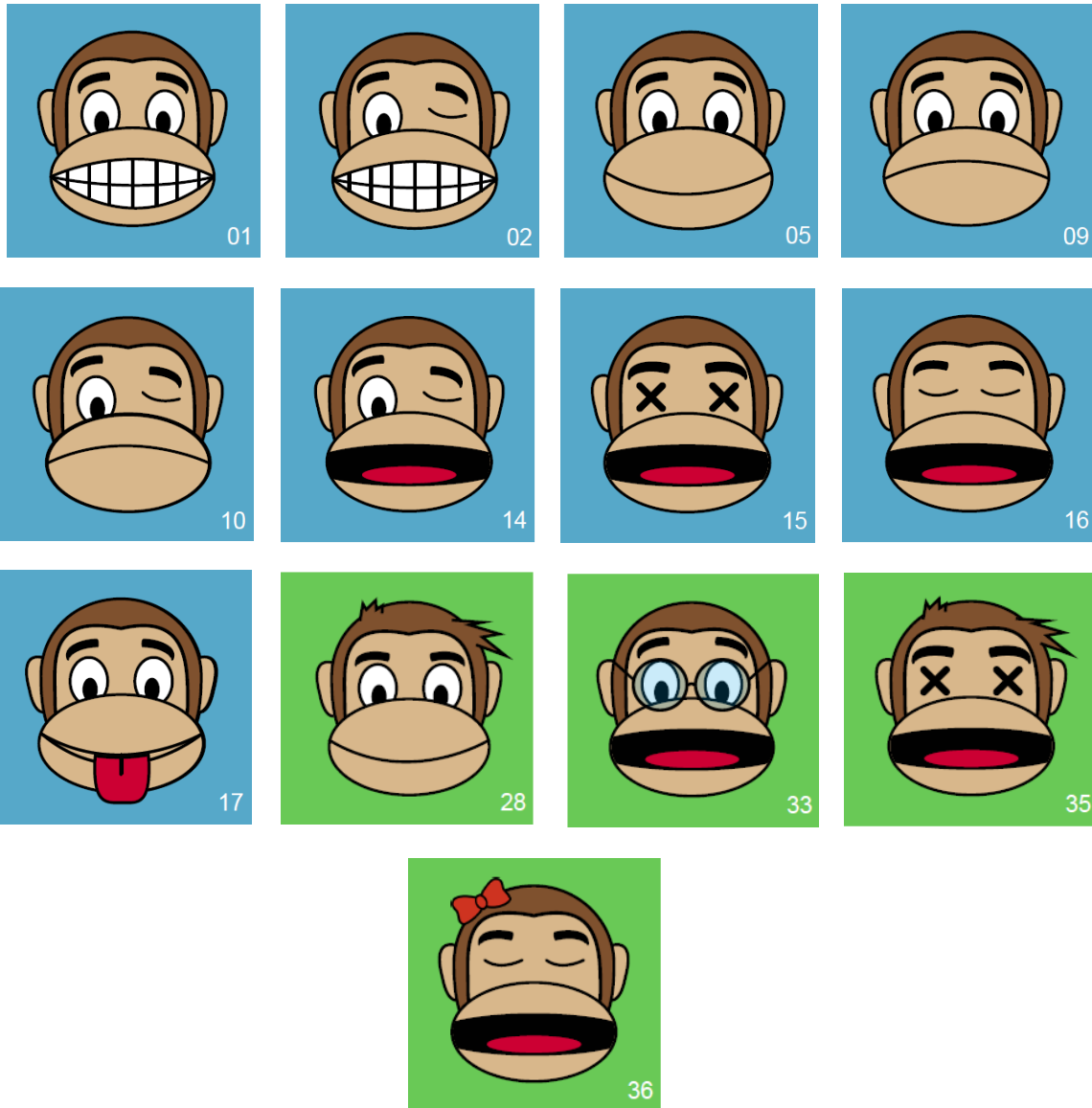
20

beißt nicht

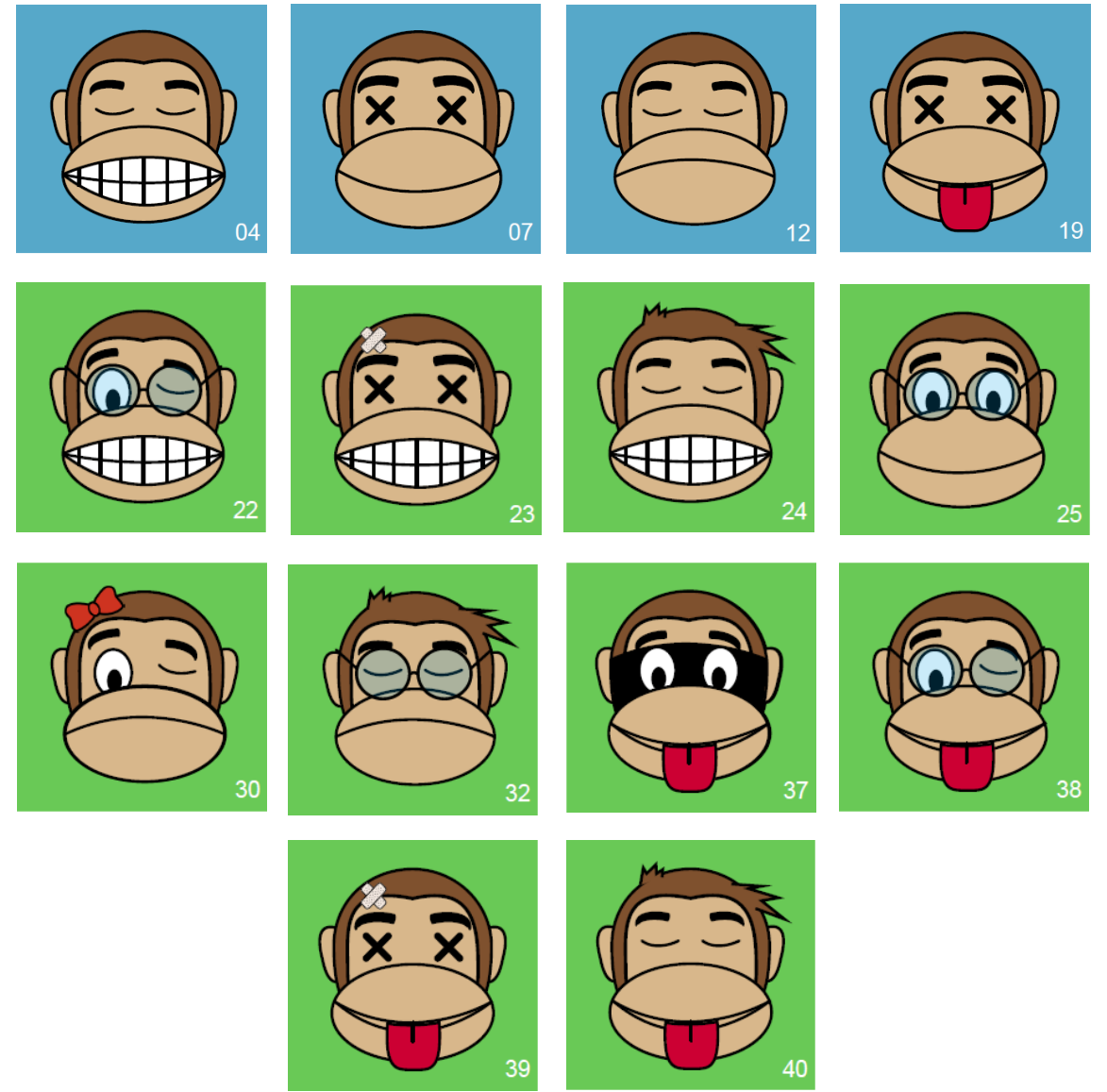


Spielvariante 2:
Erstellen Sie einen
Entscheidungsbaum
für einen anderen Zoo

Beißt

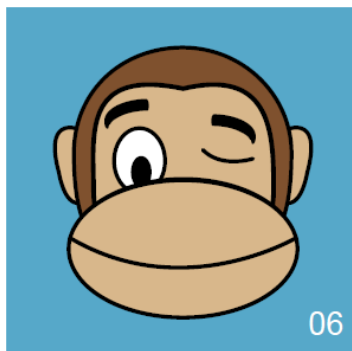


Beißt nicht





beißt nicht



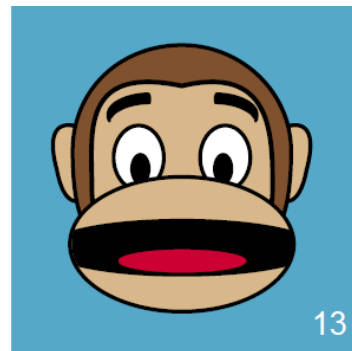
beißt



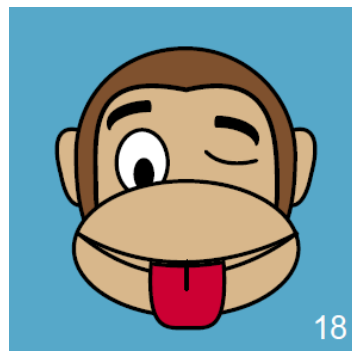
beißt nicht



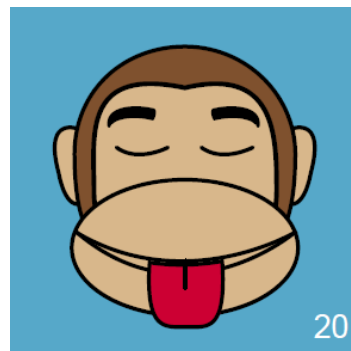
beißt nicht



beißt



beißt



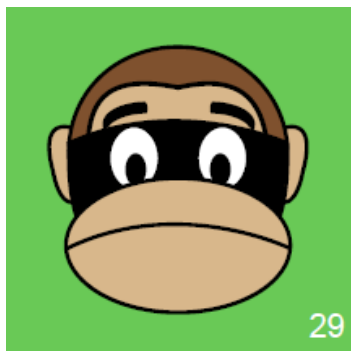
beißt nicht



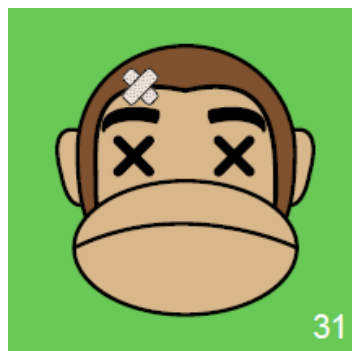
beißt nicht



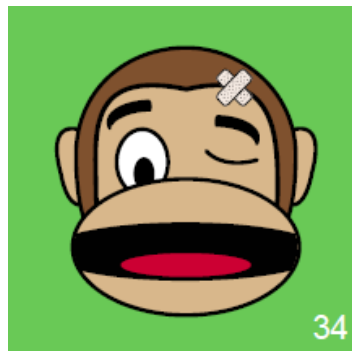
beißt nicht



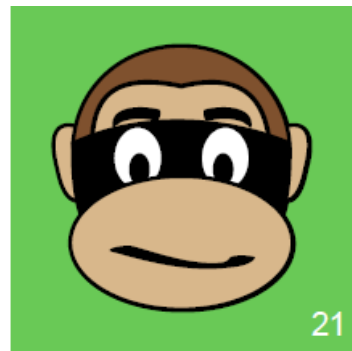
beißt nicht



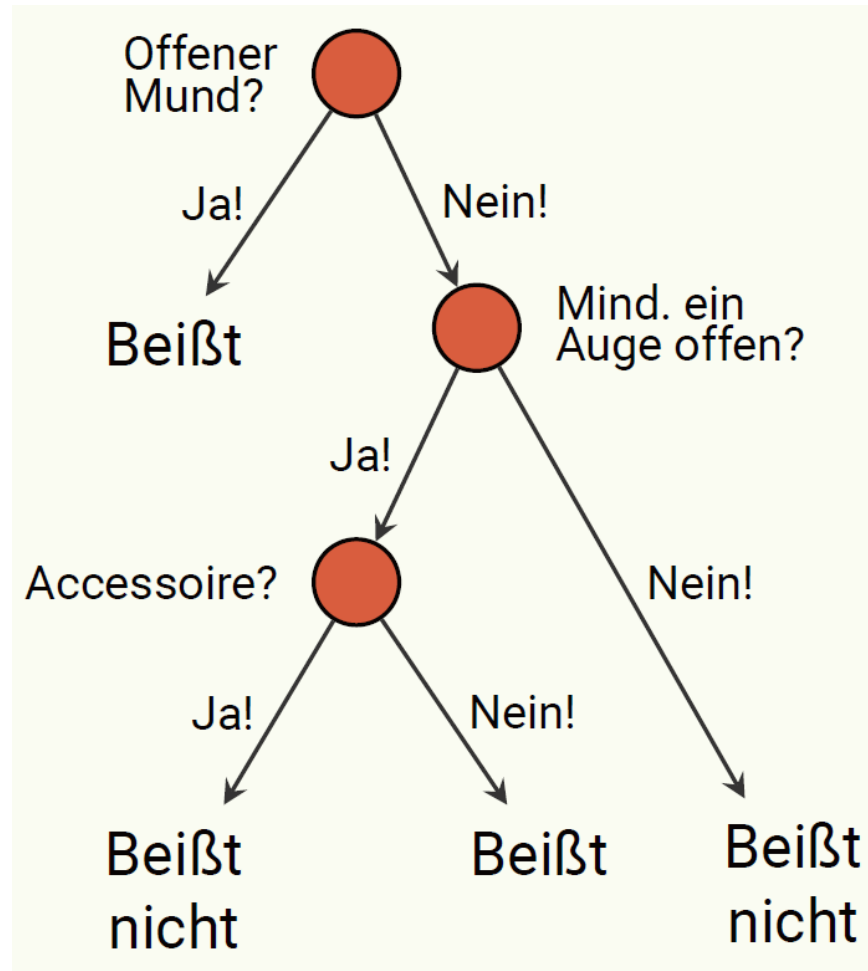
beißt nicht



beißt



beißt nicht



Aufgabe: Entwerfen Sie einen Entscheidungsbaum für dieses Szenario

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Welches Attribut steht im Wurzelknoten Ihres Baums? Warum?

Welches Attribut ist am „hilfreichsten“ für die Entscheidungsfindung?

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Ein Attribut sollte möglichst weit oben im Baum stehen, wenn seine verschiedenen Attributwerte zu möglichst klaren Entscheidungen führen.

- Bsp.: Wenn Money=Poor, dann ist die Entscheidung immer Cinema
- Bsp.: Wenn Parents=Yes, dann ist die Entscheidung immer Cinema

Der Gini-Koeffizient

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Der Gini-Koeffizient misst, wie „unordentlich“ Daten sind, d.h. wie ungleichmäßig die Werte eines Ziel-Features verteilt sind
- Bsp. Weather=Windy → [Cinema, Cinema, Shopping, Cinema]
→ recht ordentlich → niedriger Gini-Koeffizient
- Bsp. Parents=No → [Tennis, Stay-In, Cinema, Shopping, Tennis]
→ eher unordentlich → hoher Gini-Koeffizient
- Gini = 0: Alle Daten haben beim Ziel-Feature denselben Wert → perfekte Ordnung
- Maximalwert 1, aber in der Regel deutlich niedriger
 - Bei 2 möglichen Werten: Maximalwert Gini = 0,5
 - Bei n möglichen Werten: Maximalwert Gini = $1 - \frac{1}{n}$ (näht sich 1)

Der Gini-Koeffizient

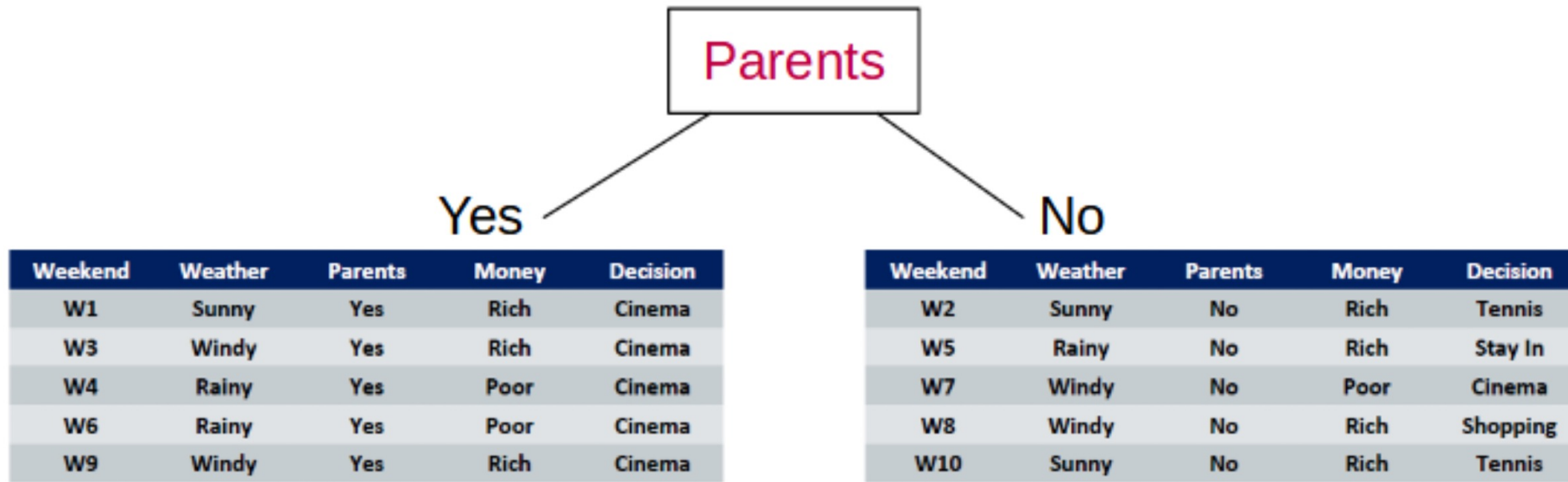
Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

$$Gini(D) = 1 - \sum_{i=1}^K (p_i)^2$$

- D: (Ausgewählte) Datensätze
- K: Anzahl möglicher Werte für das Ziel-Feature
 - hier: K=4 (4 Entscheidungen möglich)
- p_i : relative Häufigkeit des i-ten Werts in D
 - Bsp: In der Gesamtliste ist $p_{cinema} = \frac{6}{10}$ (an 6 von 10 Wochenenden gehe ich ins Kino)
- Gini(Gesamt) =

$$1 - \left(\left(\frac{6}{10} \right)^2 + \left(\frac{2}{10} \right)^2 + \left(\frac{1}{10} \right)^2 + \left(\frac{1}{10} \right)^2 \right) = 0,58$$

Aufteilungen bewerten: Der *gewichtete* Gini-Koeffizient



$$Gini(F) = \sum_{v \in V_F} p_v \cdot Gini(F = v)$$

- F : ein Feature (Bsp. *Parents*)
- V_F : mögliche Werte dieses Features (Bsp. $\{yes, no\}$)
- v : ein bestimmter Wert (Bsp. *no*)
- p_v : relative Häufigkeit von $F=v$ in den Daten (Bsp. $p_{no} = \frac{5}{10}$)

Gini(parents=no)

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- $p_{Tennis} = \frac{2}{5}$

- $p_{Stay-In} = \frac{1}{5}$

- $p_{Shopping} = \frac{1}{5}$

- $p_{Cinema} = \frac{1}{5}$

- $Gini(Parents=No) =$

$$1 - \left(\left(\frac{2}{5} \right)^2 + \left(\frac{1}{5} \right)^2 + \left(\frac{1}{5} \right)^2 + \left(\frac{1}{5} \right)^2 \right) = 0,72$$

Gini(parents=yes)

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- $p_{Cinema} = \frac{5}{5}$

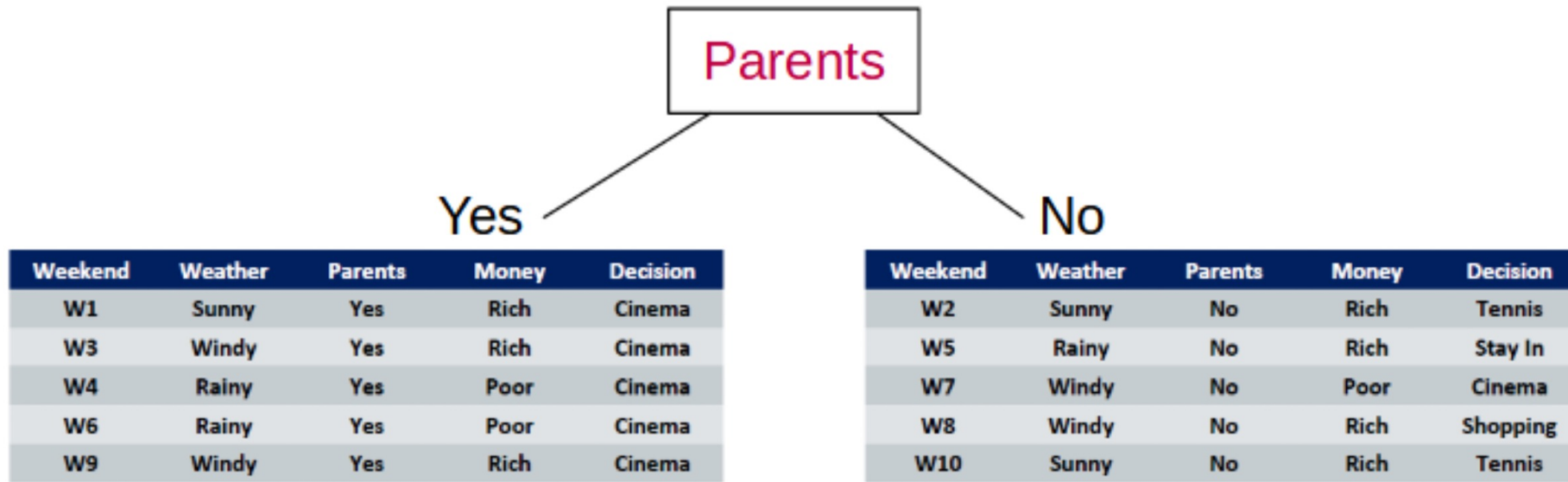
- $p_{Tennis} = \frac{0}{5}$

- $p_{Stay-In} = \frac{0}{5}$

- $p_{Shopping} = \frac{0}{5}$

- $$\begin{aligned} \text{Gini(Parents=Yes)} &= \\ &1 - \left(\left(\frac{5}{5}\right)^2 + \left(\frac{0}{5}\right)^2 + \left(\frac{0}{5}\right)^2 + \left(\frac{0}{5}\right)^2 \right) \\ &= 1 - 1 = 0 \end{aligned}$$

Aufteilungen bewerten: Der *gewichtete* Gini-Koeffizient



$$Gini(F) = \sum_{v \in V_F} p_v \cdot Gini(F = v)$$

$$Gini(Parents) = \frac{5}{10} \cdot 0 + \frac{5}{10} \cdot 0,72 = 0,36$$

Aufgabe

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Berechne die gewichteten Gini-Koeffizienten für die Features
 - Money und
 - Weather

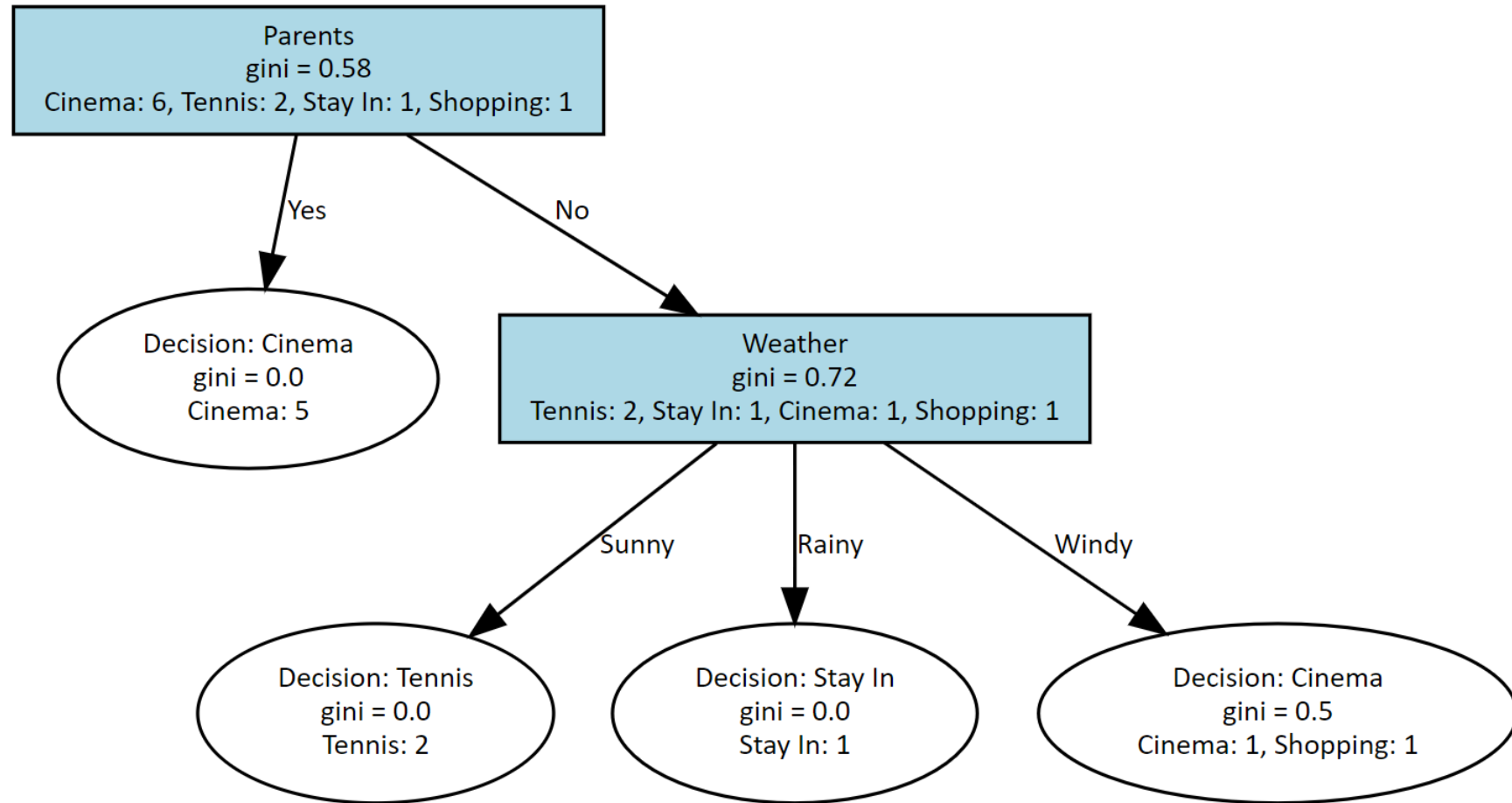
Lösung

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- $Gini(Money) = \frac{3}{10} \cdot 0 + \frac{7}{10} \cdot 0,694 = 0,486$
- $Gini(Weather) = \frac{3}{10} \cdot 0,444 + \frac{3}{10} \cdot 0,444 + \frac{4}{10} \cdot 0,375 = 0,416$
- $Gini(Parents) = \frac{5}{10} \cdot 0 + \frac{5}{10} \cdot 0,72 = 0,36$

Parents hat den niedrigsten Gini-Wert, d.h. sorgt für am meisten Ordnung
→ bester Wurzelknoten des Entscheidungsbaums

Entscheidungsbaum der Tiefe 2



Entscheidungsbaumlernen: Richtig oder falsch?

1. Der vorgestellte Algorithmus zum Lernen von Entscheidungsbäumen ist ein *unüberwachtes* Lernverfahren, weil er aus den Trainingsdaten selbständig einen Entscheidungsbaum erstellt, ohne dass ein Mensch eingreifen muss.
2. An der Wurzel eines Entscheidungsbaums kann man ablesen, welches Label einem Datensatz zugeordnet werden soll.
3. Ein Knoten wird nicht weiter aufgeteilt, wenn alle Datensätze dasselbe Label haben.
4. Ein erlernter Entscheidungsbaum ist immer eindeutig, d.h. wird der Algorithmus mit denselben Daten wiederholt, entsteht derselbe Entscheidungsbaum.