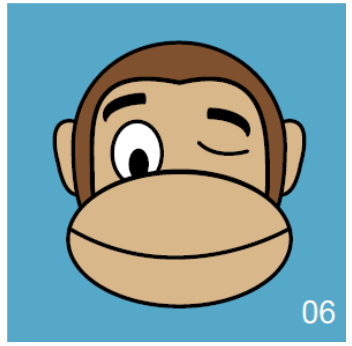
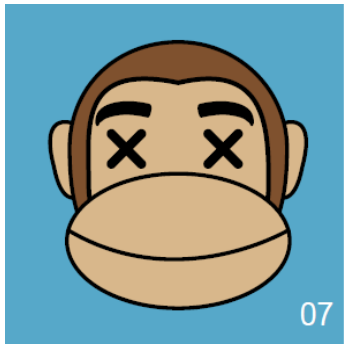
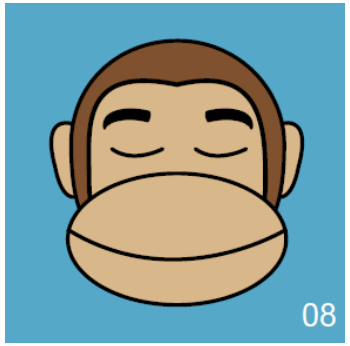
Gutes Äffchen – böses Äffchen

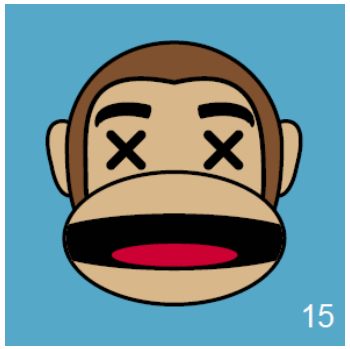
Überleben im Zoo
dank Entscheidungsbäumen

Beißt

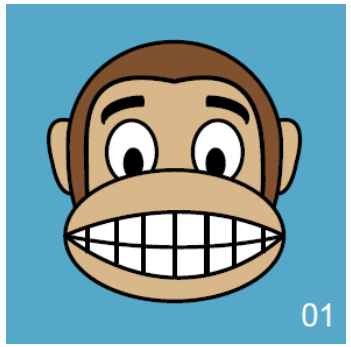


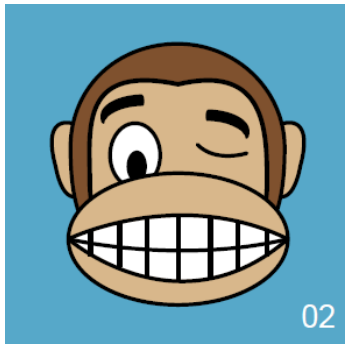


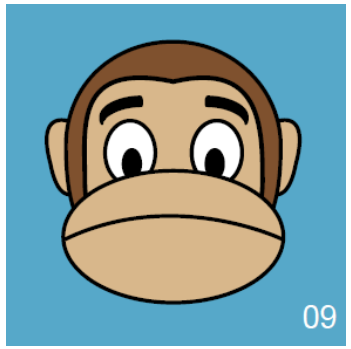


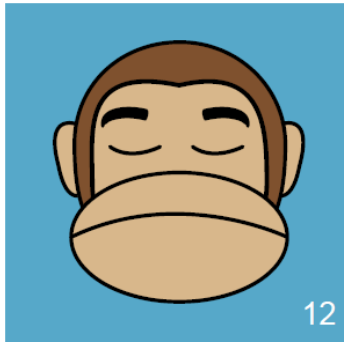


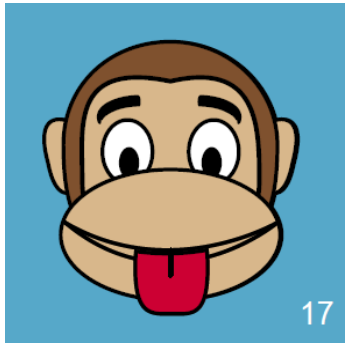
Beißt nicht

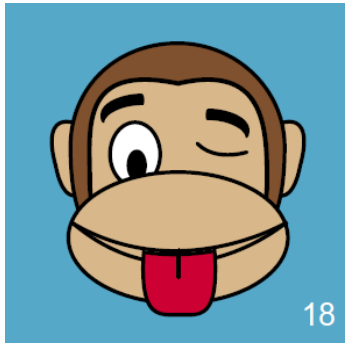






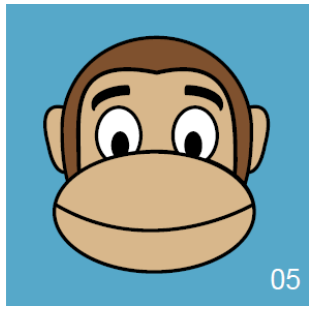




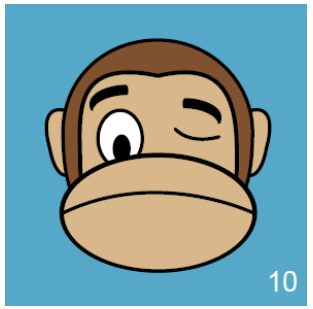


Haben Sie gelernt, wer beißt?

beißt

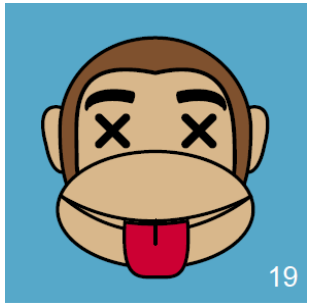


beißt

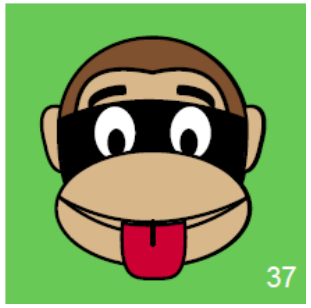


beißt nicht

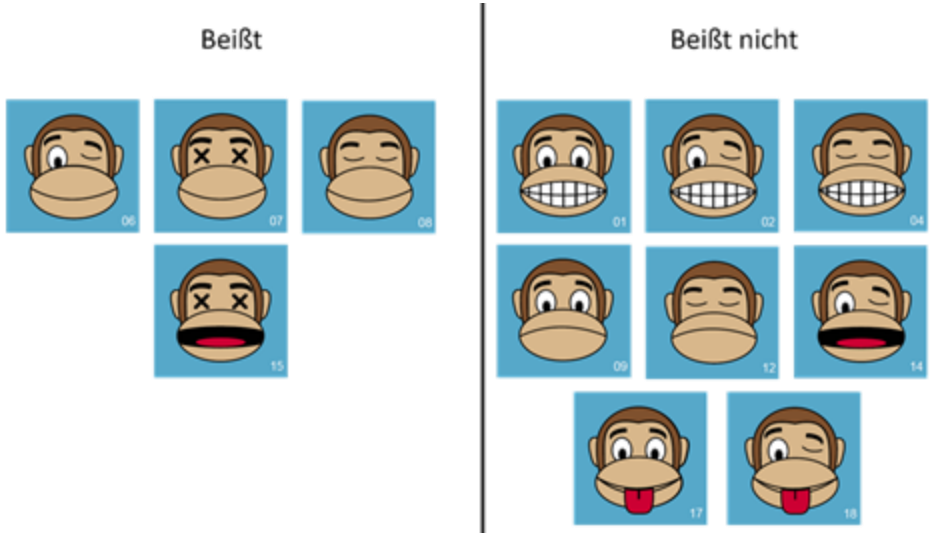
beißt nicht

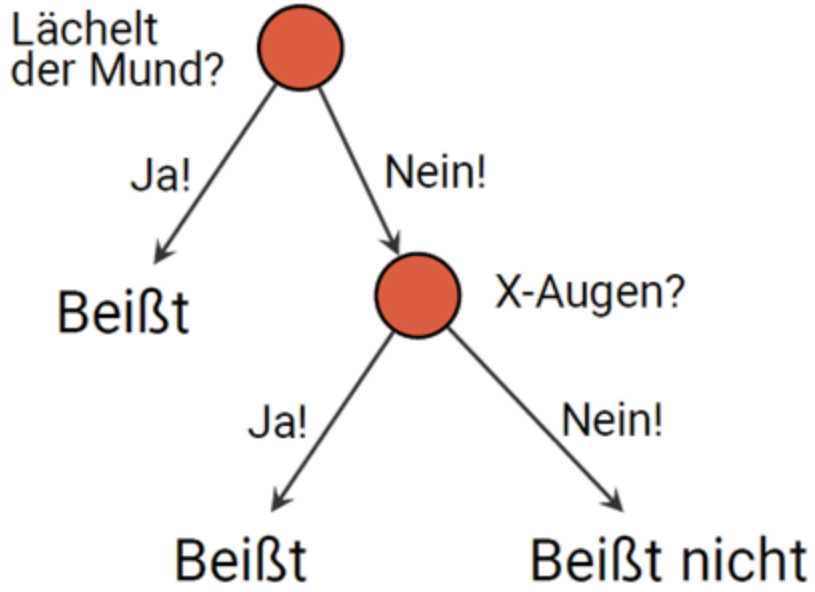


beißt

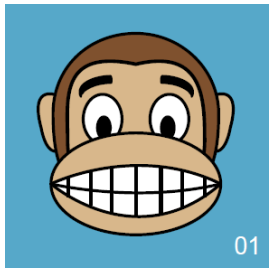


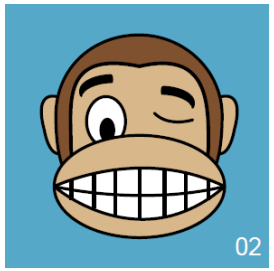
beißt nicht (?)

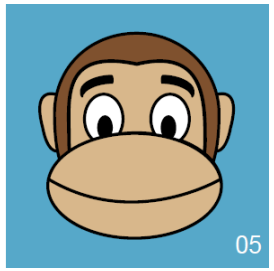


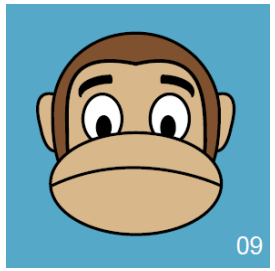


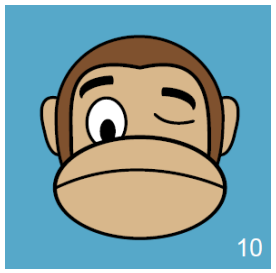
Beißt

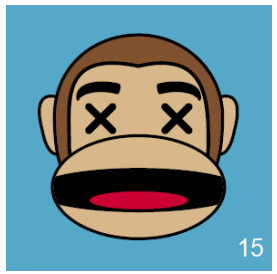


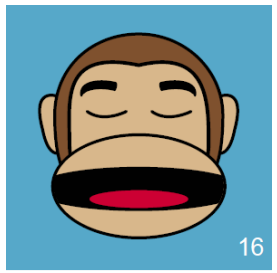


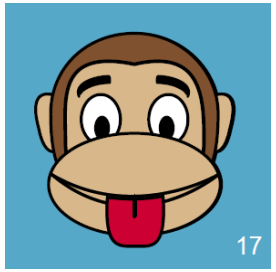


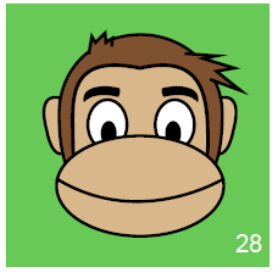


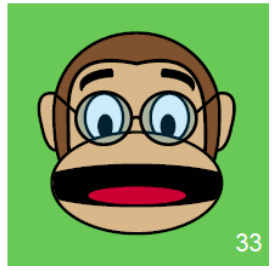


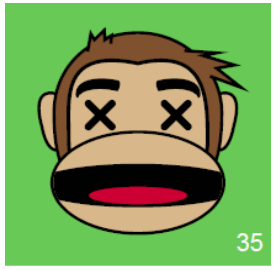


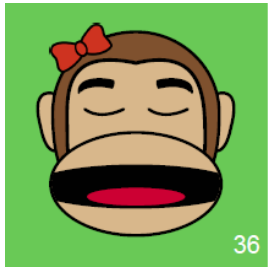




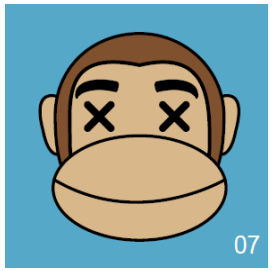


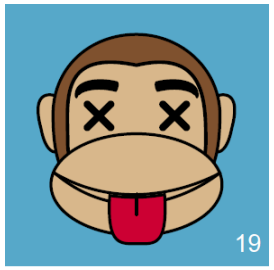


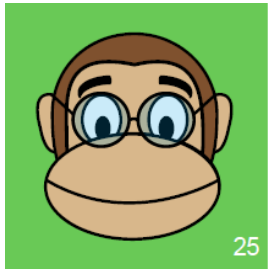


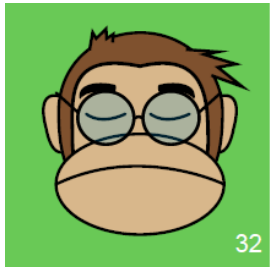


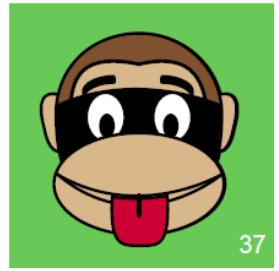
Beißt nicht

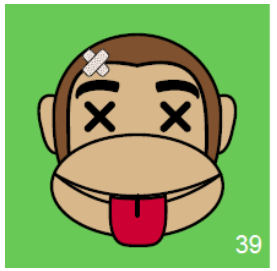


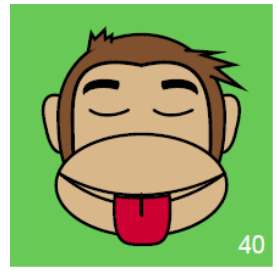


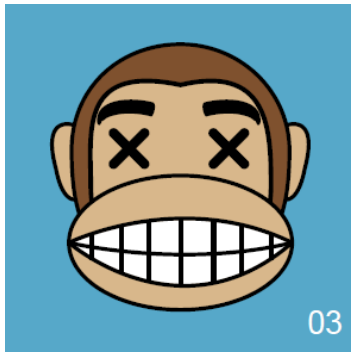




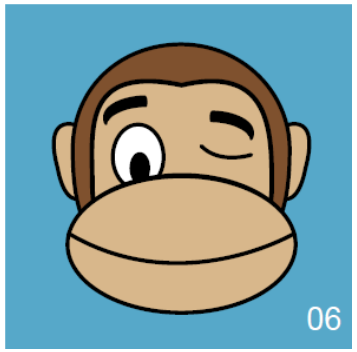




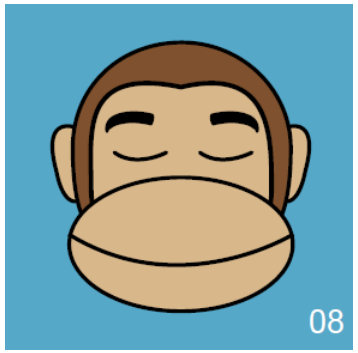




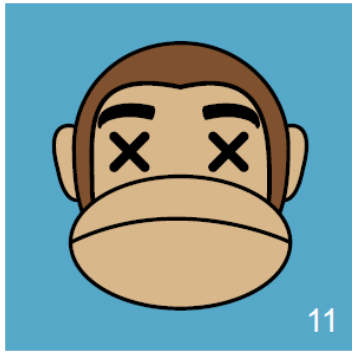
beißt nicht



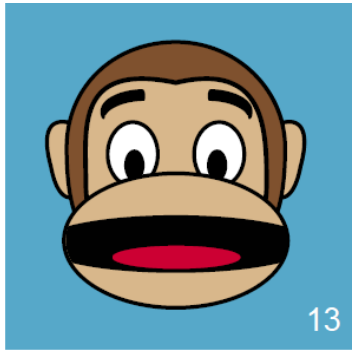
beißt



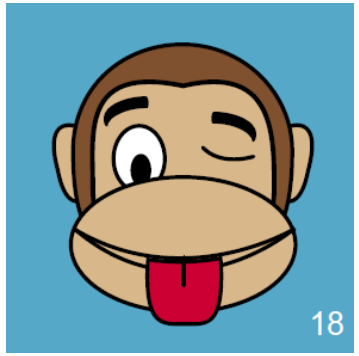
beißt nicht



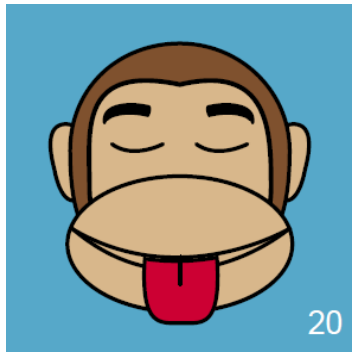
beißt nicht



beißt

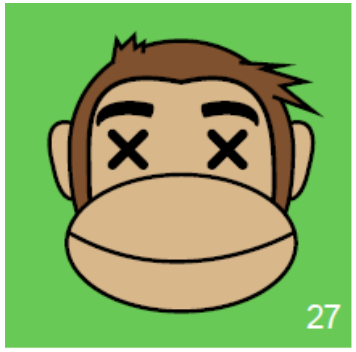


beißt



beißt nicht

beißt nicht

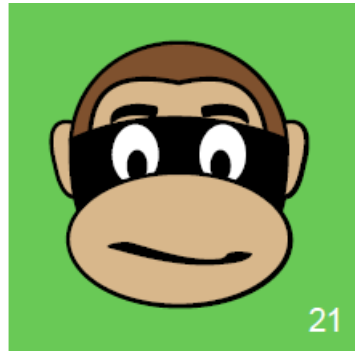


beißt nicht

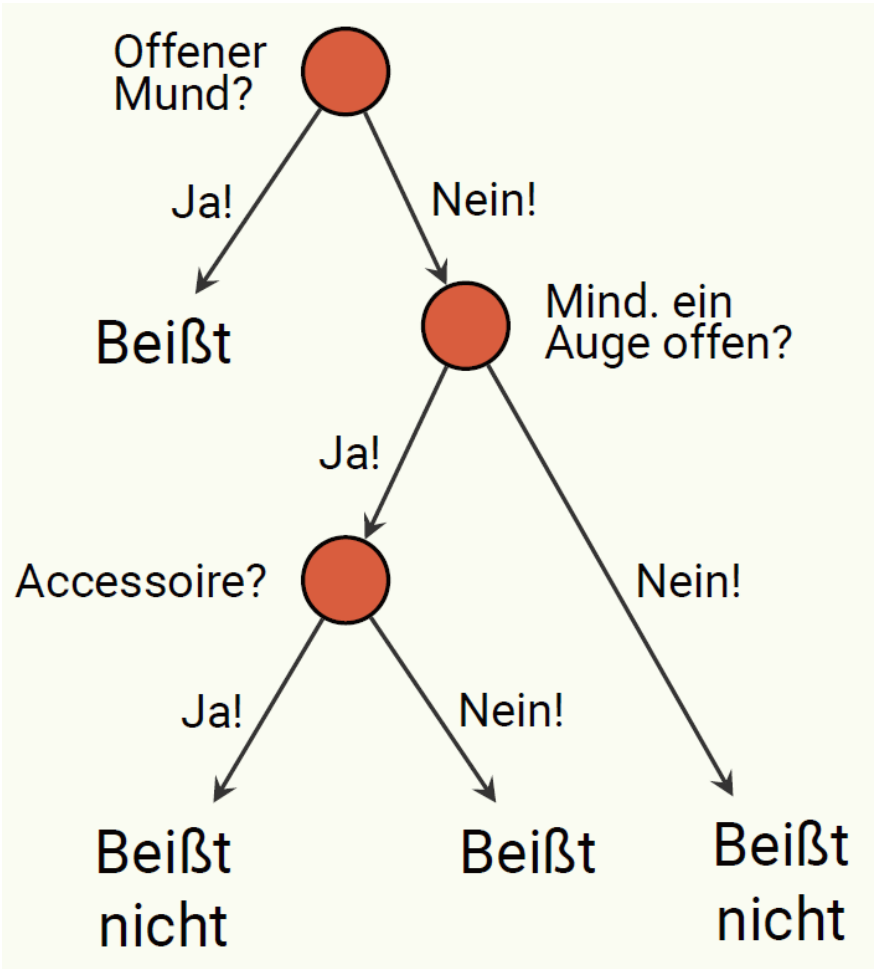
beißt nicht

beißt nicht

beißt



beißt nicht


Offener Mund?
Ja!
Nein!
Beißt
Mind. ein Auge offen?
Ja!
Accessoire?
Nein!
Ja!
Nein!
Beißt nicht
Beißt
Beißt nicht

Was soll ich am Wochenende unternehmen? Entwerfen Sie einen Entscheidungsbaum!

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.: *Money* als Wurzel. Argument: Wenn Money=Poor, geht man auf jeden Fall ins Kino
- Bsp.: *Parents* als Wurzel. Argument: Wenn Parents=Yes, geht man auf jeden Fall ins Kino

Der Gini-Koeffizient

- Der *Gini-Koeffizient* ist eine Zahl zwischen 0 und 1
- misst, wie *heterogen* („unordentlich“) die Daten sind, d.h. wie ungleichmäßig die Werte des Ziel-Features verteilt sind
- Gini = 0: Alle Daten haben beim Ziel-Feature denselben Wert → perfekte Ordnung
- Bsp. *Weather=Windy* → Werte bei *Decision:* [Cinema, Cinema, Shopping, Cinema]

  → ziemlich einheitlich → niedriger Gini-Koeffizient
- Bsp. *Parents=No* → [Tennis, Stay-In, Cinema, Shopping, Tennis]

  → eher uneinheitlich → hoher 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 (die Gini-Unreinheit)

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

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

Maximalwert 1, aber in der Regel deutlich niedriger

- Bei 2 möglichen Werten: Maximalwert Gini = 0,5
- Bei n möglichen Werten: Maximalwert Gini = $1 - \frac{1}{n}$ (nähert sich 1)

- 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) =

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

Trennschärfe bestimmen

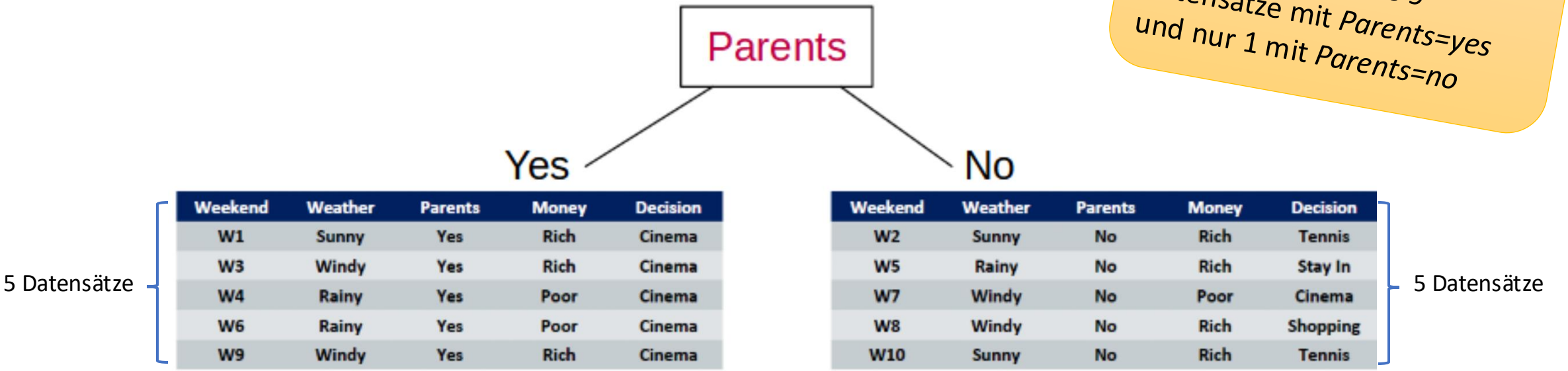
- Wir wollen herausfinden, wie gut wir das *Zielmerkmal* (hier: *Decision)* anhand der Werte eines anderen Merkmals vorhersagen können
- Bsp.: Es wäre ideal, wenn wir bei *Parents=no* immer Tennis spielen und bei *Parents=yes* immer ins Kino gehen würden
- Dann wäre *Parents* der perfekte Wurzelknoten für unseren Entscheidungsbaum, weil es ausreicht, für einen unbekannten Datensatz dessen Wert bei *Parent* zu betrachten, um eine Entscheidung zu treffen

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

Wie schätzt du die Trennschärfe von *Parents* ein?

Aufteilungen bewerten: Der *gewichtete* Gini-Koeffizient

Warum ist die Gewichtung notwendig?
Stell dir vor, es gäbe 9 Datensätze mit Parents=yes und nur 1 mit Parents=no



5 Datensätze

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

5 Datensätze

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

- 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

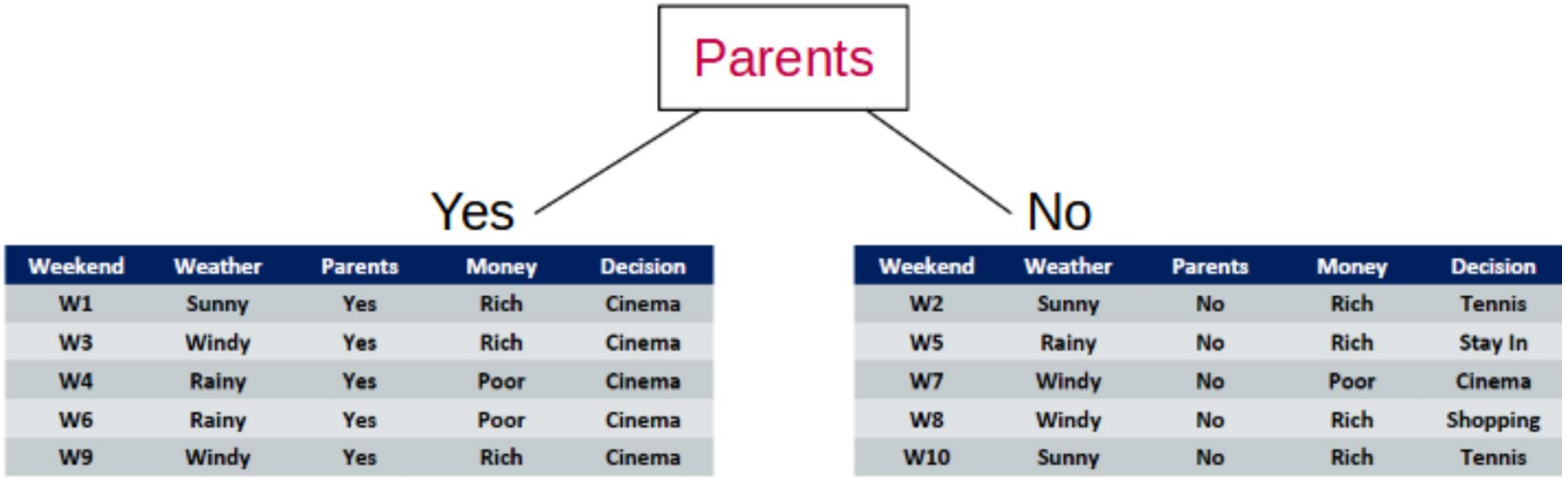
- $p_{Cinema} = \frac{5}{5}$
- $p_{Tennis} = \frac{0}{5}$
- $p_{Stay-In} = \frac{0}{5}$
- $p_{Shopping} = \frac{0}{5}$

- Gini(Parents=Yes) =

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

$$= 1 - 1 = 0$$

Aufteilungen bewerten:
Der *gewichtete* Gini-Koeffizient



Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

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

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

Formelsammlung TGI

6.2 Gini-Unreinheit

Für eine (ausgewählte) Menge von Datensätzen D und einem Ziel-Feature mit k möglichen Ausprägungen ist die **Gini-Unreinheit** (auch: Gini-Koeffizient, Gini-Index, Gini Impurity) wie folgt definiert:

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

wobei p_i die relative Häufigkeit der i-ten Ausprägung des Ziel-Merkmals ist.

Mit $Gini(F = v)$ bezeichnen wir die Gini-Unreinheit der Auswahl von Datensätzen, bei denen das Merkmal/Feature F den Wert v hat.

Ein Feature F kann verschiedene Werte $v \in V_f$ annehmen. Tritt ein bestimmter Wert v mit der relativen Häufigkeit p_v auf, dann berechnet sich die **gewichtete Gini-Unreinheit** für das Feature F folgendermaßen:

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

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

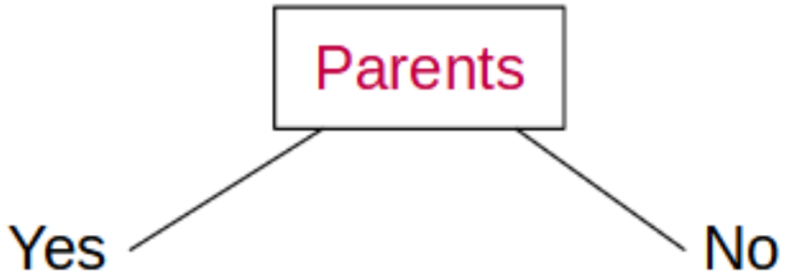
Lösung

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

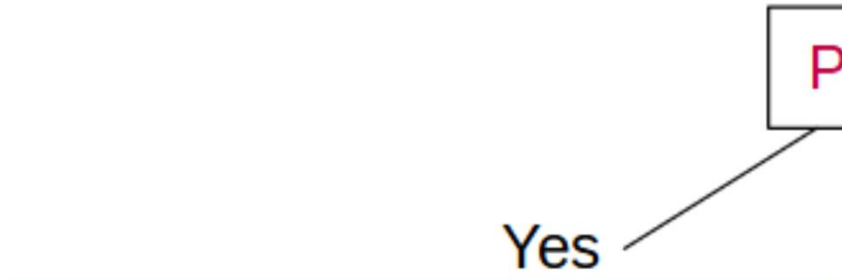
So sieht unser Entscheidungsbaum jetzt aus



Wie geht es jetzt weiter?

- Wir ergänzen jeden Teilbaum nach derselben Methode wie eben
  - gewichtete Gini-Koeffizienten für alle Features (außer *Parents*) berechnen. Feature mit bestem (= niedrigsten) Wert auswählen
- Rekursion!

So sieht unser Entscheidungsbaum jetzt aus



Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Wichtig: In jedem Teilbaum nur noch diejenigen Datensätze verwenden, die zu den getroffenen Entscheidung passen!

- Bsp.: lm Ast *Parents=Yes* müssen alle Daten beim Feature *Parents* den Wert *Yes* haben.

Entscheidungsbaum der Tiefe 2