

3. Trainings- und Testdaten

Die verfügbaren gelabelten Daten werden getrennt verwendet:

- **Trainingsdaten** sind gelabelt bzw. beschriftet und werden zur Erzeugung des Modells verwendet.
- **Testdaten** sind gelabelt bzw. beschriftet und werden zur Beurteilung der Vorhersagequalität verwendet.

Beispiel: gelabelte Trainingsdaten

Die bekannten Äffchen sind mit den Labels *beißt* oder *beißt nicht* versehen. Aus diesen Beispielen kann ein Entscheidungsbaum Regeln für die Klassifikation neuer Äffchen lernen.

TRAININGSDATEN

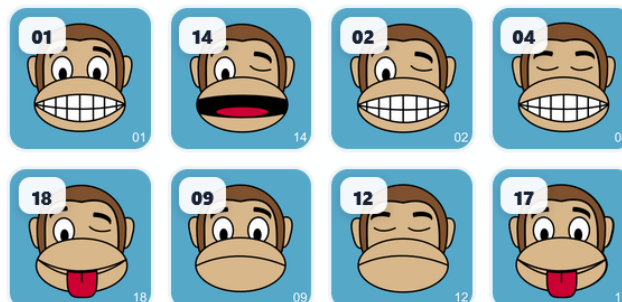
Bekannte Äffchen

Vergleiche die Bilder jederzeit mit deinen Regeln.

● Beißt · 4



● Beißt nicht · 8



Beispielgrafik: vom Nutzer bereitgestellt; Äffchen-Illustrationen im Ausgangsmaterial von Annabel Lindner und Stefan Seegerer, CC BY-NC.

Merke

Die Trennung verhindert, dass ein Modell nur an bereits bekannten Daten gut wirkt. Erst neue Testdaten zeigen, wie zuverlässig seine **Vorhersagen** sind.

Zwei Entscheidungsbäume können dieselben Trainingsdaten vollständig richtig klassifizieren und bei unbekannten Testdaten trotzdem unterschiedlich gut funktionieren.

Die Trainingsdaten müssen nicht jede mögliche Merkmalskombination enthalten.