4. Entscheidungsbäume (Teil 1)

Agenda für heute

- Einführung und Prinzip
- Modellschätzung
- Beispiel zur Veranschaulichung
- Vorteile von Entscheidungsbäumen

Motivation

- Kennenlernen einer prädiktiven Modellklasse die *Nonlinearität*, *Interaktionen* sowie eine *große Anzahl an Prädiktoren* berücksichtigen kann und damit in vielen praktischen Anwendungen eine höhere Vorhersagegüte erzielt als lineare Modelle
  → „Random Forest“

- Ein Random Forest (Breiman, 2001) entsteht durch die Kombination einer großen Zahl von „Entscheidungsbäumen“

- Für Entscheidungsbäume existieren viele verschiedene Algorithmen mit leicht unterschiedlichem Konstruktionsprinzip. Der bekannteste Vertreter, der auch im Random Forest verwendet wird ist der „Classification and Regression Trees“ Algorithmus (CART; Breiman 1983)

Depressionsklassifikation

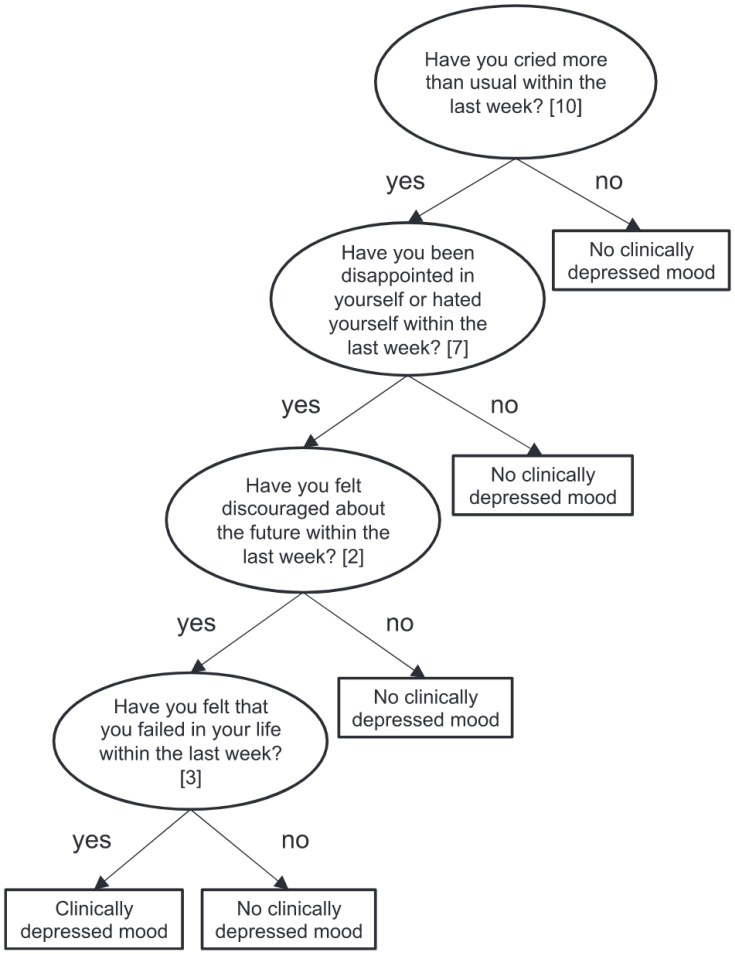


Fig. 2. Fast and frugal tree screening for depressed mood (according to the total BDI score). The numbers in brackets indicate the position of the respective item in the BDI. The full wording of the BDI cues is presented in Table 1; for this figure, it has been translated into binary questions.

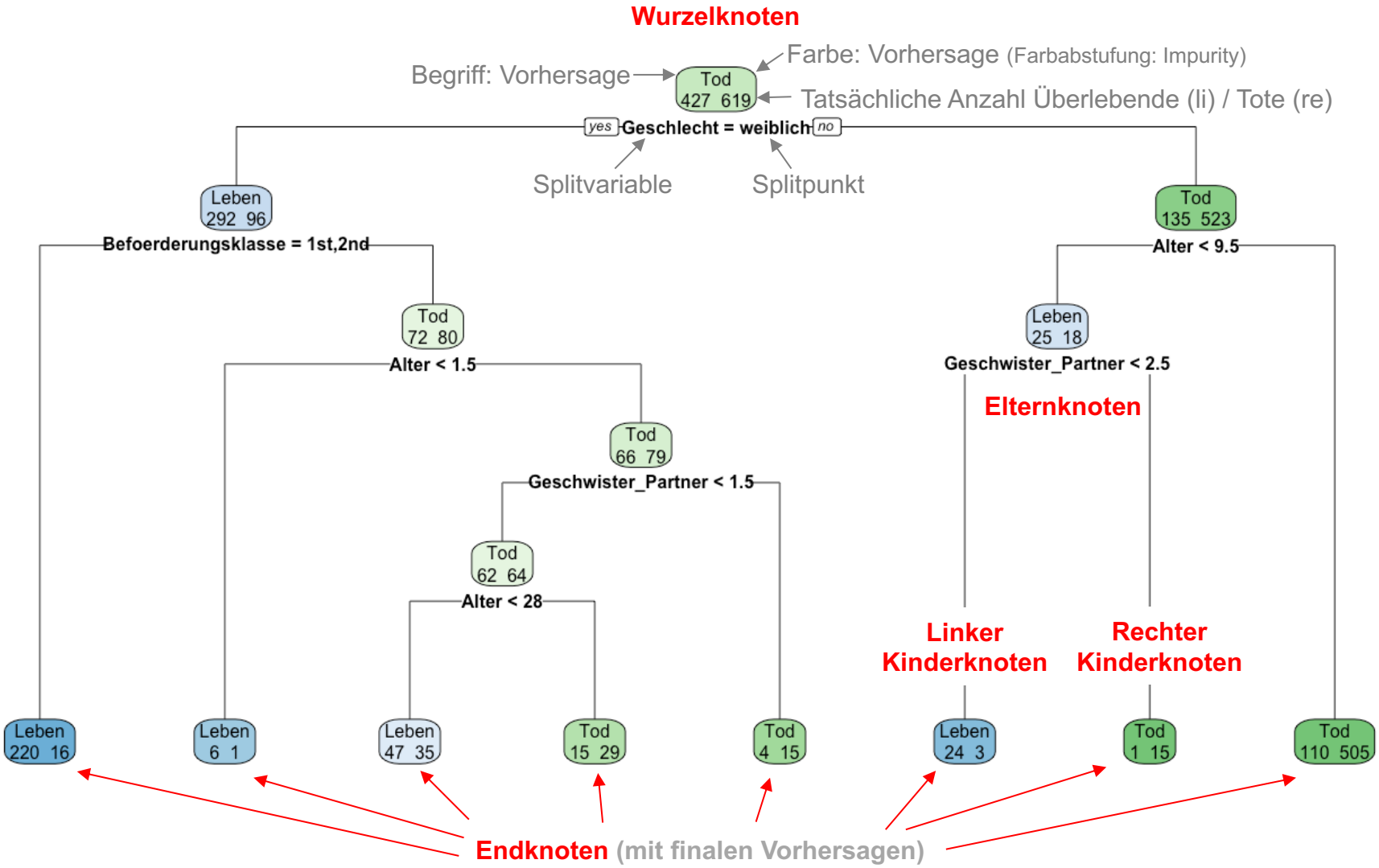
zitiert nach Jenny, M. A., Pachur, T., Williams, S. L., Becker, E., & Margraf, J. (2013). Simple rules for detecting depression. *Journal of Applied Research in Memory and Cognition*, *2*(3), 149–157. https://doi.org/10.1037/h0101797

Beispieldatensätze

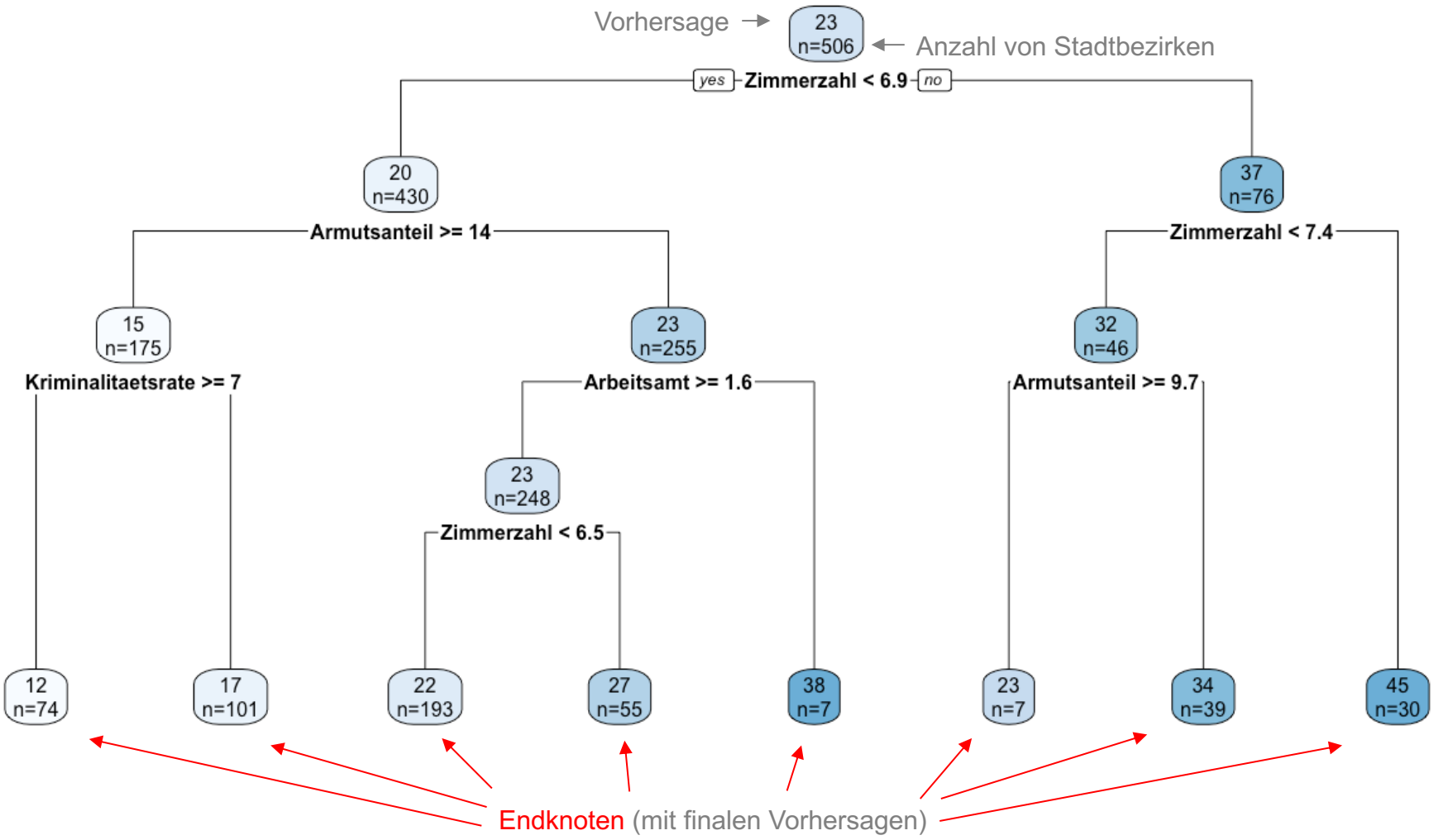
- **Klassifikation**: Titanic
  - Überleben von 1046 Passagieren auf der Titanic:
  - $y_i = 1$: Passagier hat überlebt
  - $y_i = 0$: Passagier hat nicht überlebt
  - 6 Prädiktoren: Beförderungsklasse, Geschlecht, Alter, Anzahl mitreisender Geschwister und Partner, Anzahl mitreisender Eltern und Kinder

- **Regression**: Boston Housing
  - Wohnungspreise aus 506 Stadtbezirken in Boston (im Jahr 1970)
  - y_i: Median des Wohnungspreises in Stadtbezirk i (Einheit: 1000 $)
  - 13 Prädiktoren: mittlere Anzahl von Zimmern pro Wohnung, Armutsanteil im Bezirk, Kriminalitätsrate im Bezirk, Nähe zum Arbeitsamt, Anteil an Altbauwohnungen, ...

Entscheidungsbaum: Titanic (Klassifikation)



Entscheidungsbaum: Boston Housing (Regression)



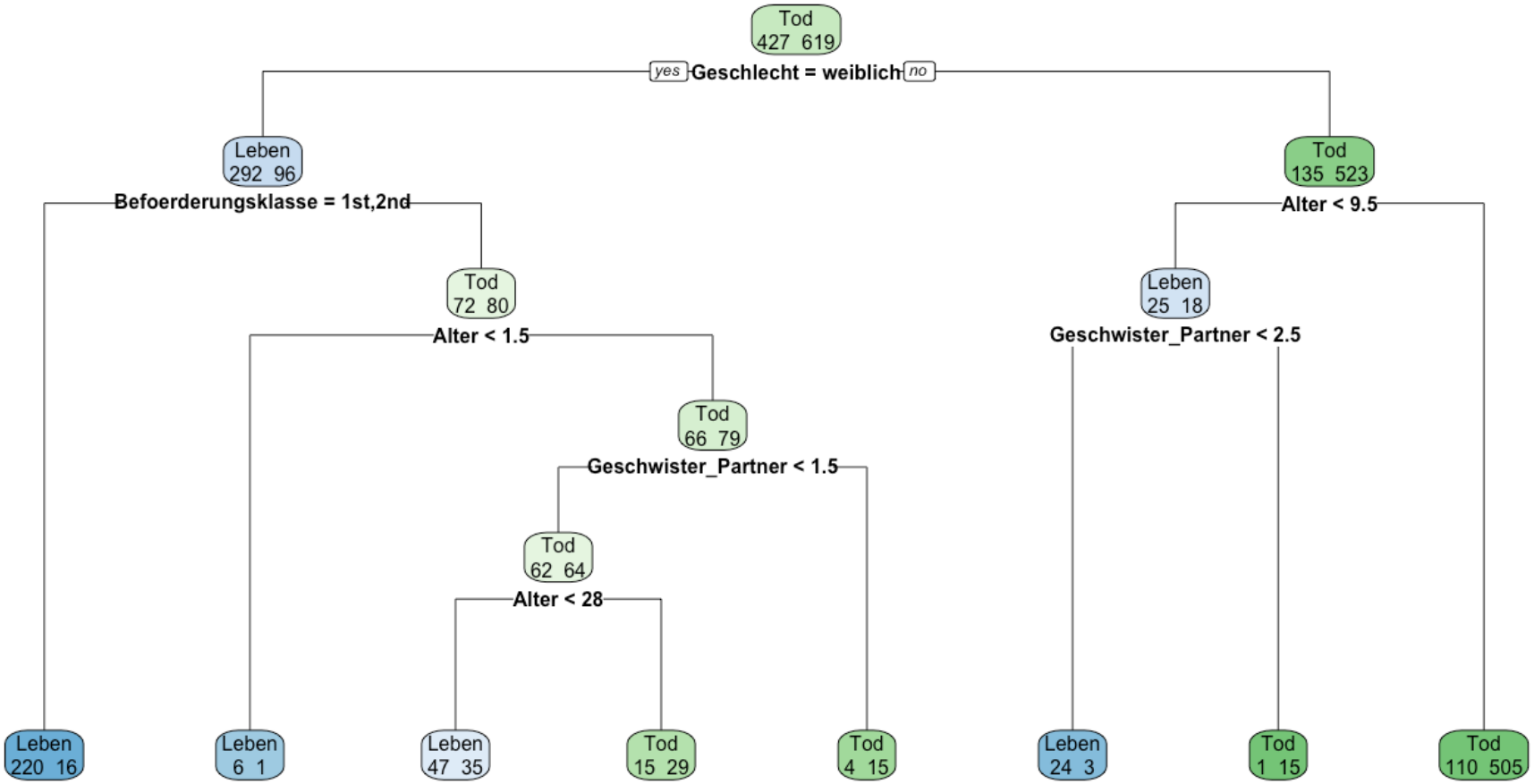
Berechnung von Vorhersagen

- Ein Entscheidungsbaum trifft für jede Beobachtung innerhalb eines Knotens die gleiche, konstante Vorhersage
  - Regression:
    Mittelwert der Kriteriumswerte aller Beobachtungen im Knoten
  - Klassifikation:
    Häufigster Kriteriumswert aller Beobachtungen im Knoten

- Berechnung der Vorhersage für eine **neue** Beobachtung:
  - Zuordnung der Beobachtung zu einem Endknoten anhand der Werte auf den Prädiktorvariablen („In welchem Endknoten landet die Beobachtung, wenn man sie oben in den Baum wirft“)
  - Die Vorhersage für die neue Beobachtung besteht aus der konstanten Vorhersage, die sich im Training des Modells für den entsprechenden Endknoten ergeben hat

Beispiel für die Berechnung einer Vorhersage

Rose
- ✓ Erste Klasse
- ✓ Weiblich
- ✓ 17 Jahre alt
- ✓ An Bord mit Mutter und Verlobtem



Modellschätzung: Rekursiver Split – Algorithmus

Wie entsteht die Struktur eines Entscheidungsbaumes, basierend auf einem vorliegenden Datensatz (mit bekannten Werten für alle Variablen)?

- Wiederholte Teilung des Datensatzes anhand einzelner Prädiktorvariablen, bis ein bestimmtes Abbruchkriterium erreicht wird
- Splitvariable und Splitpunkt jeder Teilung des Datensatzes werden gleichzeitig mithilfe eines Optimalitätskriteriums bestimmt
  - Für jeden möglichen Split werden alle Kombinationen aus Splitvariable und Splitpunkt miteinander verglichen
  - **Ziel (Optimalitätskriterium):** Knoten sollen möglichst rein sein, d.h. die Beobachtungen sollen möglichst gleiche Werte auf der Kriteriumsvariable aufweisen
  - Bestimmung der Reinheit eines Knotens durch ein „**Impurity–Maß**“ (unterschiedliches Maß, je nachdem ob Regression oder Klassifikation).

Modellschätzung: Impurity–Maß bei Klassifikation

- Im Klassifikationsfall kann der MMCE als Impurity–Maß verwendet werden
- Unreinheit eines Elternknotens E:

$$Imp_E^{(Klas)} = \frac{1}{N_E} \sum_{i=1}^{N_E} I(y_i \neq \hat{y}_E)$$

- $\hat{y}_E$: der häufigste Kriteriumswert im Elternknoten (entweder 0 oder 1)

- Bei Verwendung des MMCE als Impurity–Maß entspricht die Unreinheit also dem Anteil der kleineren Klasse innerhalb des Knotens

- In der Praxis wird bei Klassifikation nicht der MMCE als Impurity–Maß verwendet, sondern der „Gini–Index“ oder die „Shannon–Entropie“. Beide Maße bevorzugen bei gleichem MMCE komplett reine Kinderknoten und führen damit in der Regel zu einer besseren Vorhersagegüte.

Beispiel

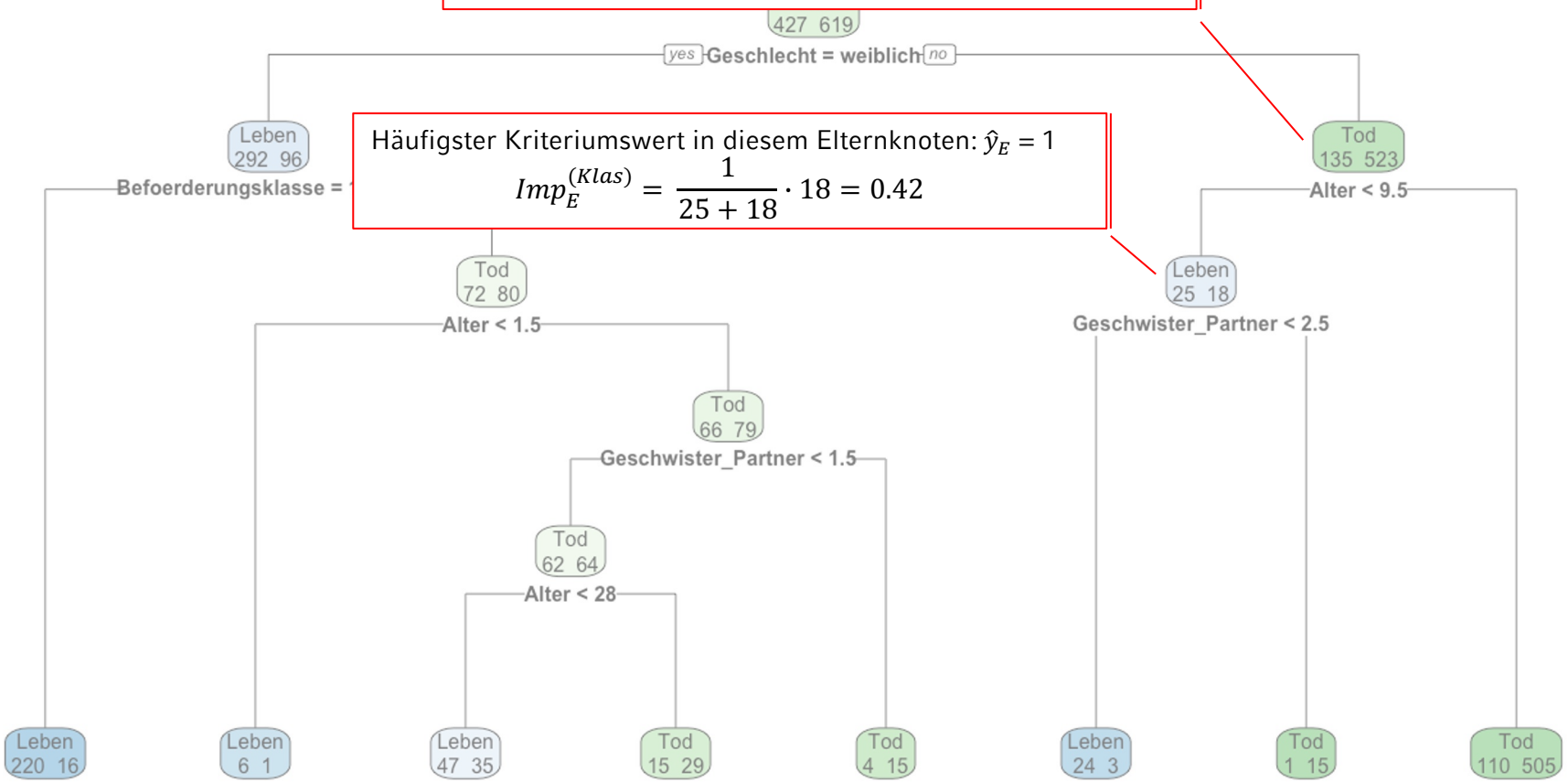
Zur Erinnerung: $Imp_E^{(Klas)} = \frac{1}{N_E}\sum_{i=1}^{N_E} I(y_i \neq \hat{y}_E)$

Häufigster Kriteriumswert in diesem Elternknoten: $\hat{y}_E = 0$

$$Imp_E^{(Klas)} = \frac{1}{135 + 523} \cdot 135 = 0.21$$

Häufigster Kriteriumswert in diesem Elternknoten: $\hat{y}_E = 1$

$$Imp_E^{(Klas)} = \frac{1}{25 + 18} \cdot 18 = 0.42$$



Modellschätzung: Impurity – Maß bei Regression

- Im Regressionsfall kann der MSE als Impurity–Maß verwendet werden
- Unreinheit eines Elternknotens E:

$$Imp_E^{(Regr)} = \frac{1}{N_E}\sum_{i=1}^{N_E}(y_i - \hat{y}_E)^2$$

$$\hat{y}_E = \bar{y}_E = \frac{1}{N_E}\sum_{i=1}^{N_E} y_i$$

- N_E: Anzahl von Beobachtungen im Elternknoten E

- Bei Verwendung des MSE als Impurity–Maß entspricht die Unreinheit also der Varianz der Kriteriumsvariable innerhalb des Knotens

Modellschätzung: **Split**kriterium

Wie wird bei der Schätzung des Baums jeweils ein optimaler Split bestimmt?

- Gewählt wird die Kombination aus Splitvariable und Splitpunkt mit der größten „Impurity–Reduction“:

$$Imp_E - \frac{N_L}{N_E} Imp_L - \frac{N_R}{N_E} Imp_R$$

- Imp_E, Imp_L, Imp_R: Unreinheit des Elternknotens, linken Kinderknotens, rechten Kinderknotens
- N_E, N_L, N_R: Anzahl von Beobachtungen im Elternknoten, linken Kinderknoten, rechten Kinderknoten

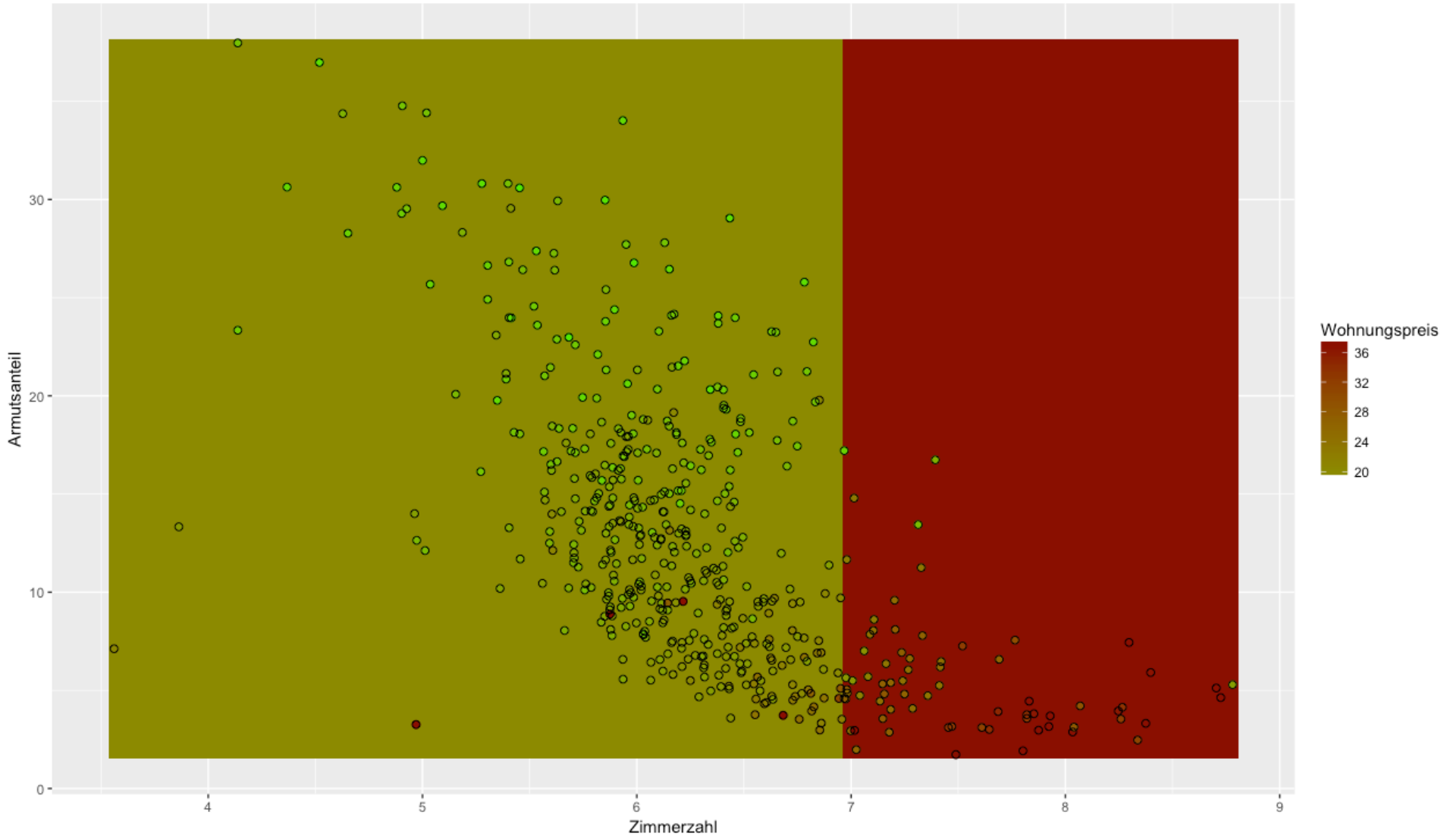
- Dabei wird die Unreinheit der resultierenden Kinderknoten mit deren relativer Größe gewichtet.

  → Dies bevorzugt Kinderknoten die möglichst rein und gleichzeitig nicht zu klein sind.

Modellschätzung: **Abbruch**kriterien

- Prinzipiell kann der Split–Algorithmus solange fortgeführt werden, bis sich in jedem Endknoten ausschließlich Beobachtungen mit dem gleichen Kriteriumswert bzw. nur noch eine einzige Beobachtung befindet.
- In beiden Fällen werden alle Beobachtungen mit denen das Modell trainiert wurde perfekt vorhergesagt. Die Vorhersage bei neuen Daten ist dann jedoch üblicherweise sehr schlecht (**Overfitting**). Um eine bessere Vorhersagegüte bei neuen Daten zu erzielen, ist es daher notwendig, den Algorithmus vorzeitig abzubrechen.
  → Reduziere die „Tiefe“ des Baumes
- Mögliche Abbruchkriterien (auch in Kombination möglich):
  - Minimale Anzahl von Beobachtungen im Elternknoten
  - Minimale Anzahl von Beobachtungen im Kinderknoten
  - Minimale Impurity–Reduction
  - Maximale Anzahl von Ebenen im Entscheidungsbaum

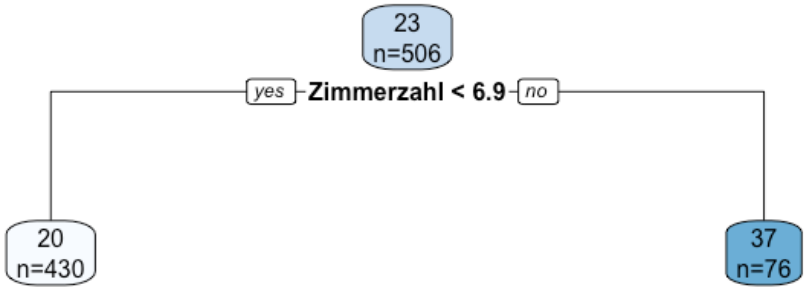
Veranschaulichung: 1 Split (Boston Housing mit nur 2 Prädiktoren)



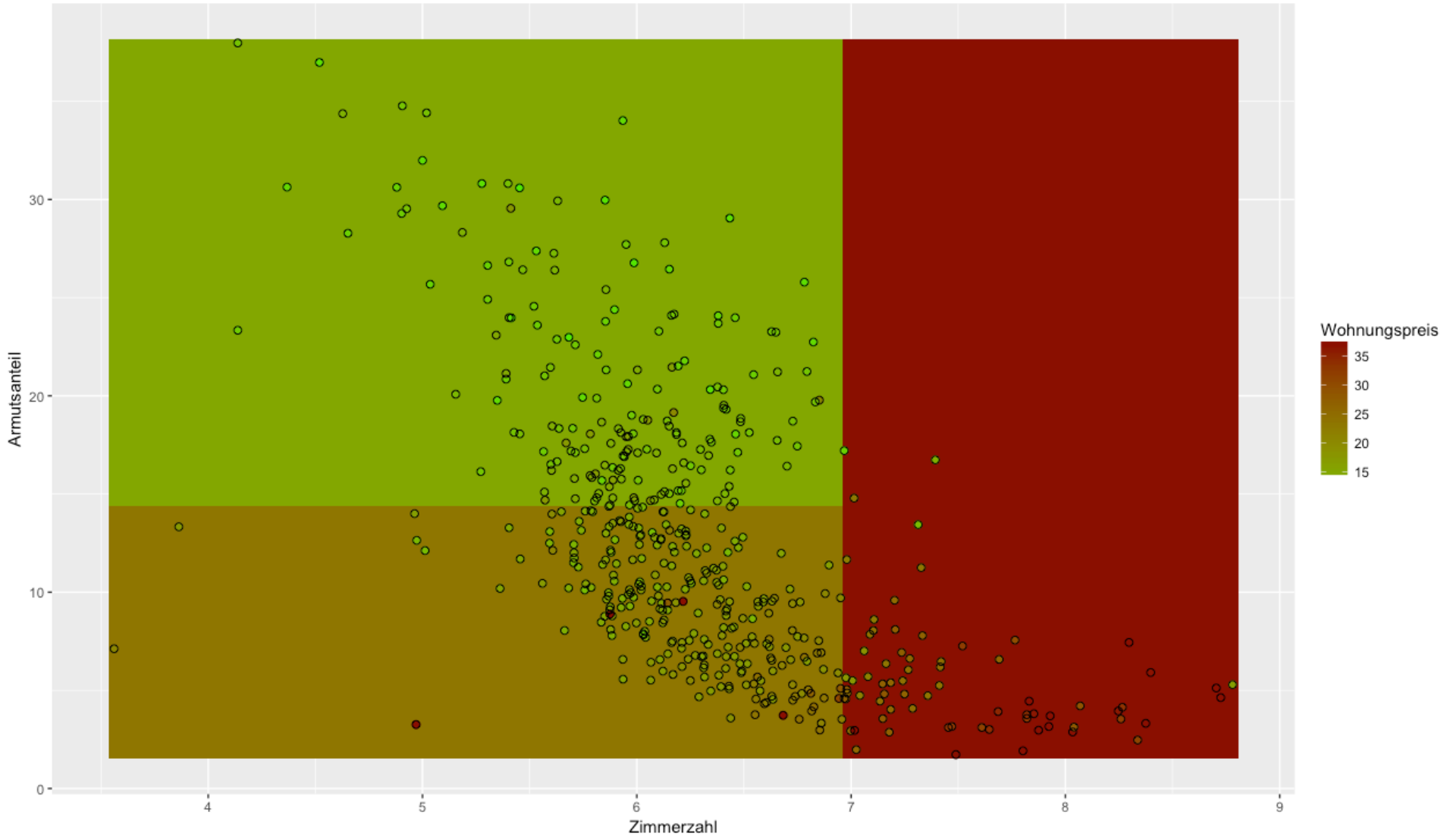
Veranschaulichung: 1 Split (Boston Housing mit nur 2 Prädiktoren)

✓ Ausprobieren aller möglichen Splitpunkte für alle möglichen Splitvariablen

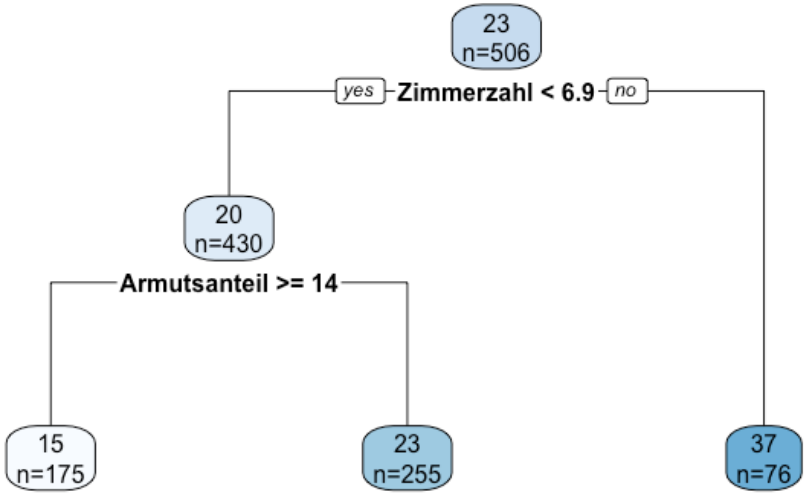
→ Wahl des Splits mit der höchsten Impurity – Reduction



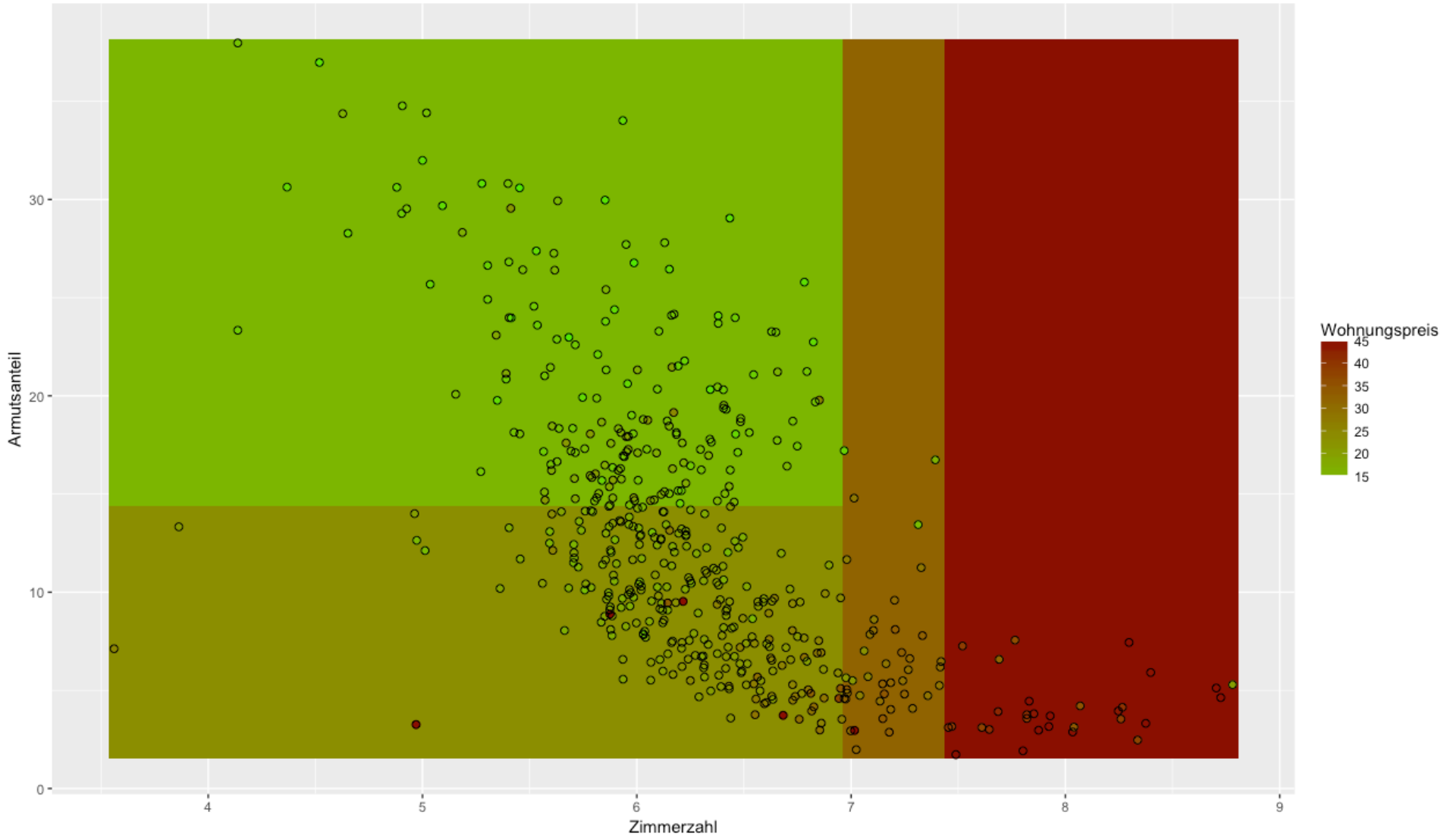
Veranschaulichung: 2 Splits (Boston Housing mit nur 2 Prädiktoren)



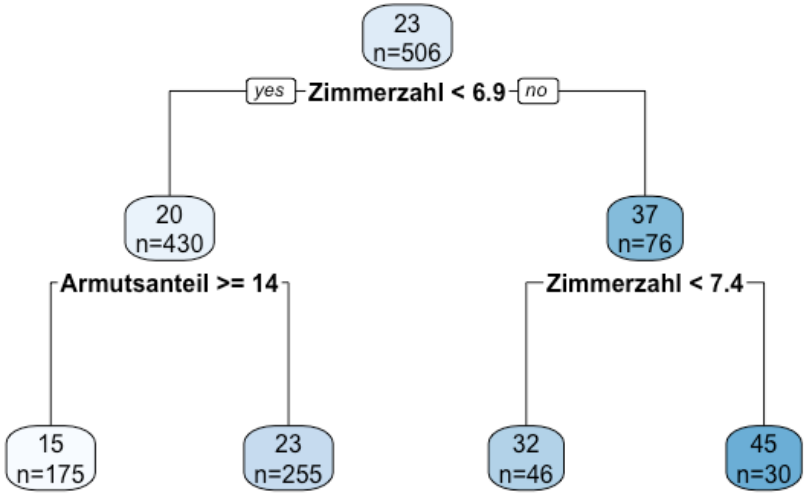
Veranschaulichung: 2 Splits (Boston Housing mit nur 2 Prädiktoren)



Veranschaulichung: 3 Splits (Boston Housing mit nur 2 Prädiktoren)

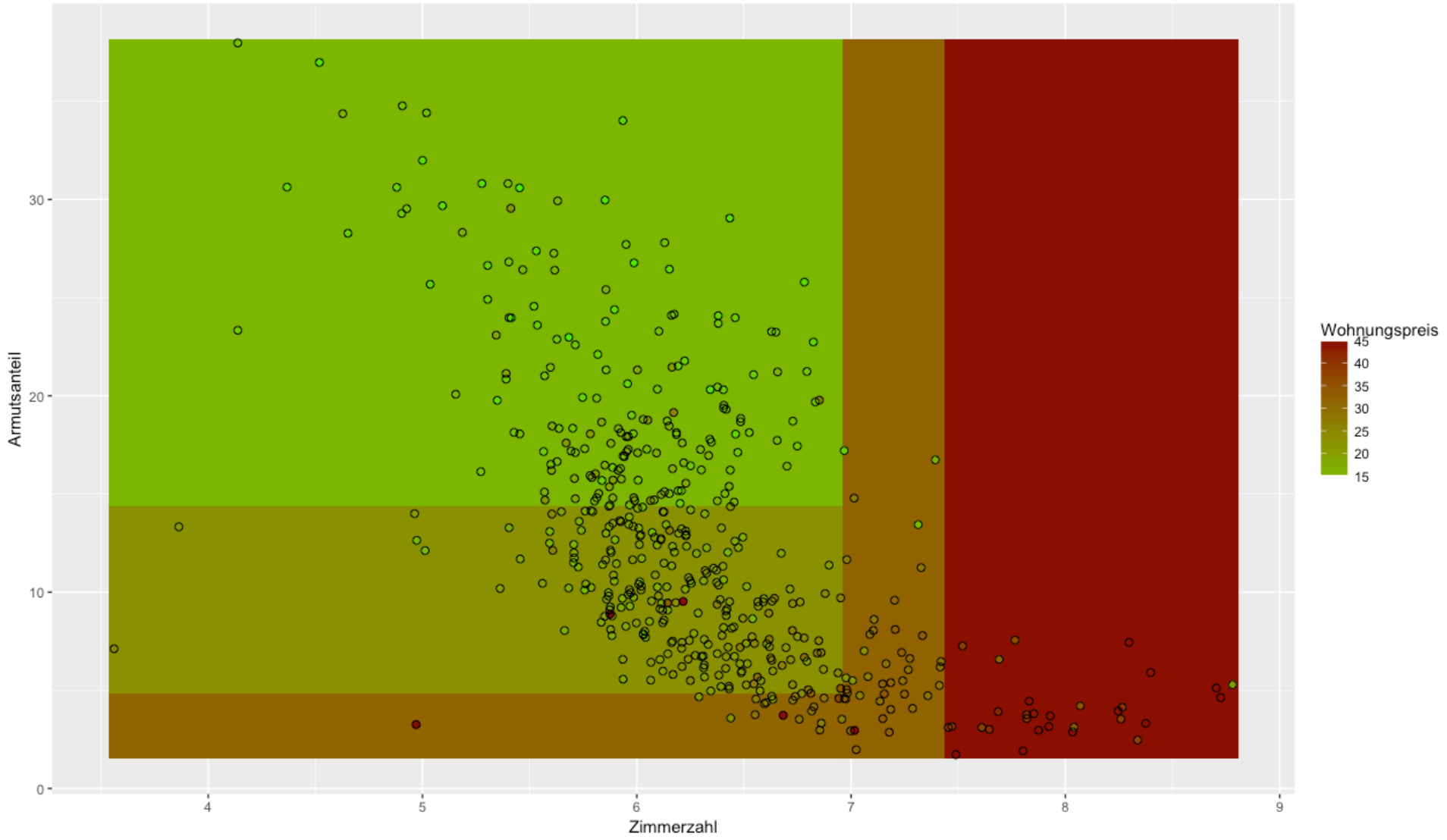


Veranschaulichung: 3 Splits (Boston Housing mit nur 2 Prädiktoren)

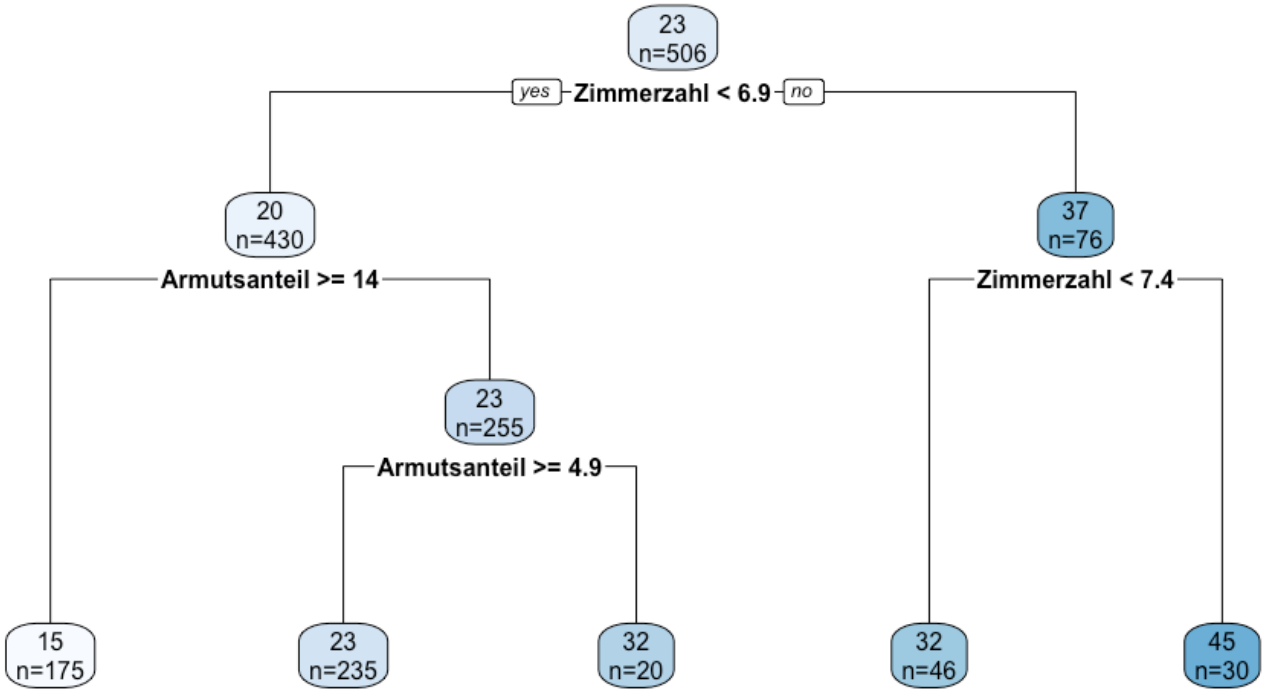


Für jeden Split werden immer alle Prädiktoren berücksichtigt
→ Prädiktoren können mehrfach ausgewählt werden

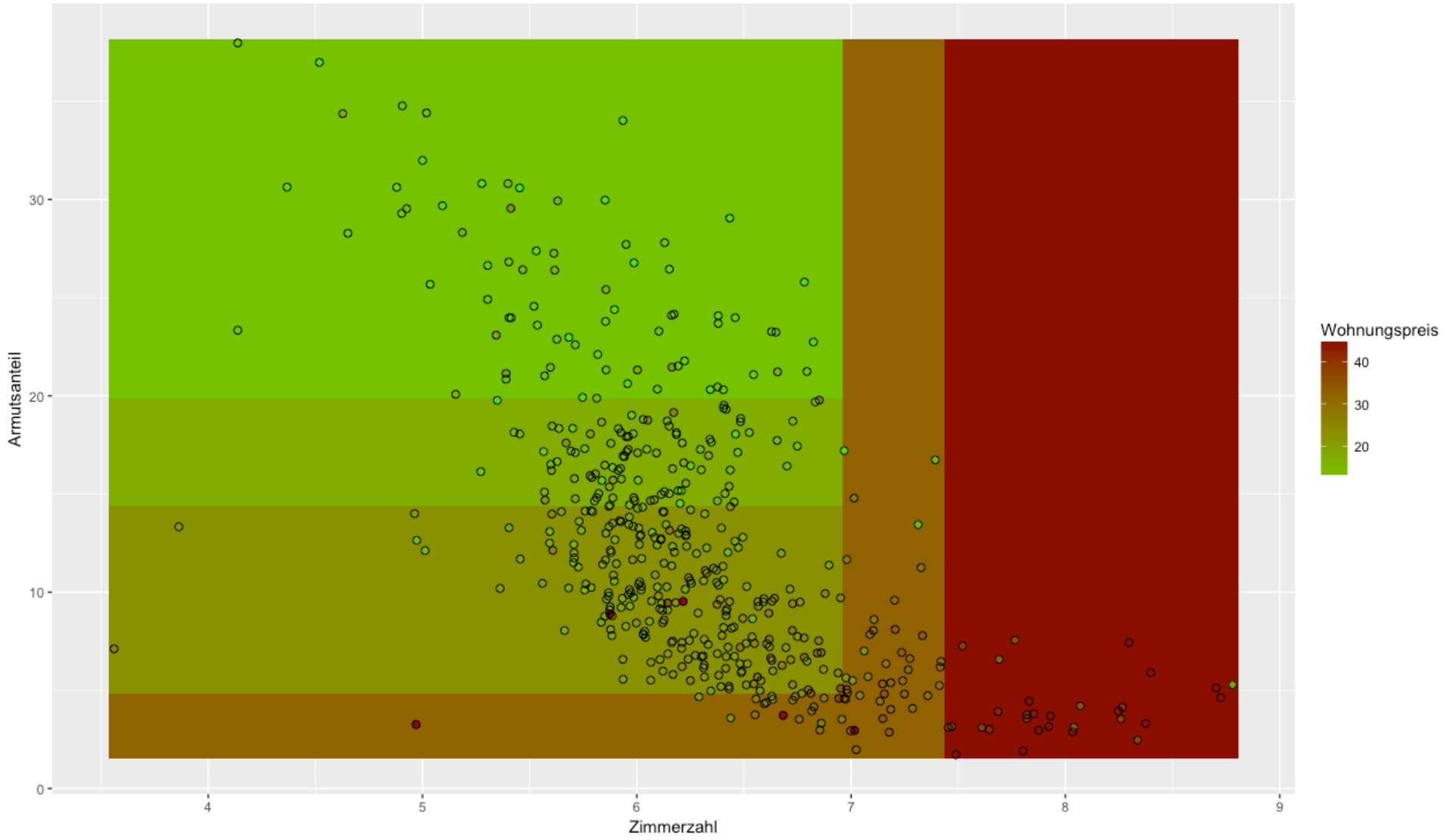
Veranschaulichung: 4 Splits (Boston Housing mit nur 2 Prädiktoren)



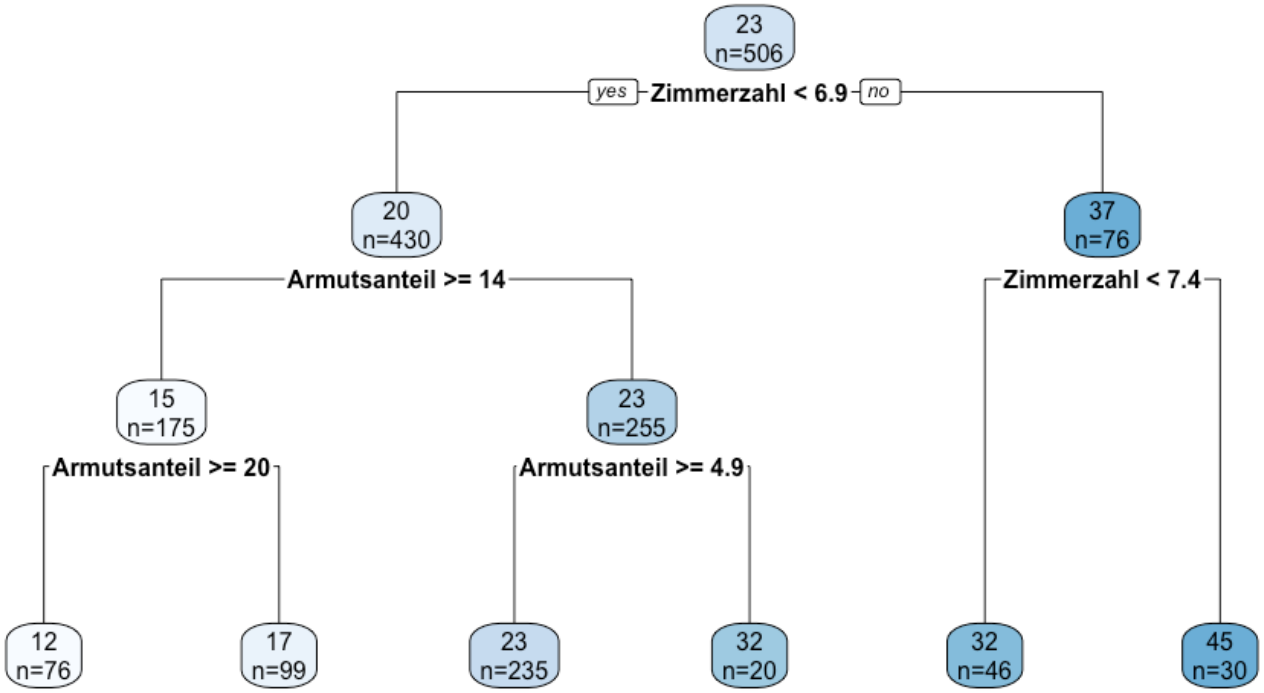
Veranschaulichung: 4 Splits (Boston Housing mit nur 2 Prädiktoren)



Veranschaulichung: 5 Splits (Boston Housing mit nur 2 Prädiktoren)



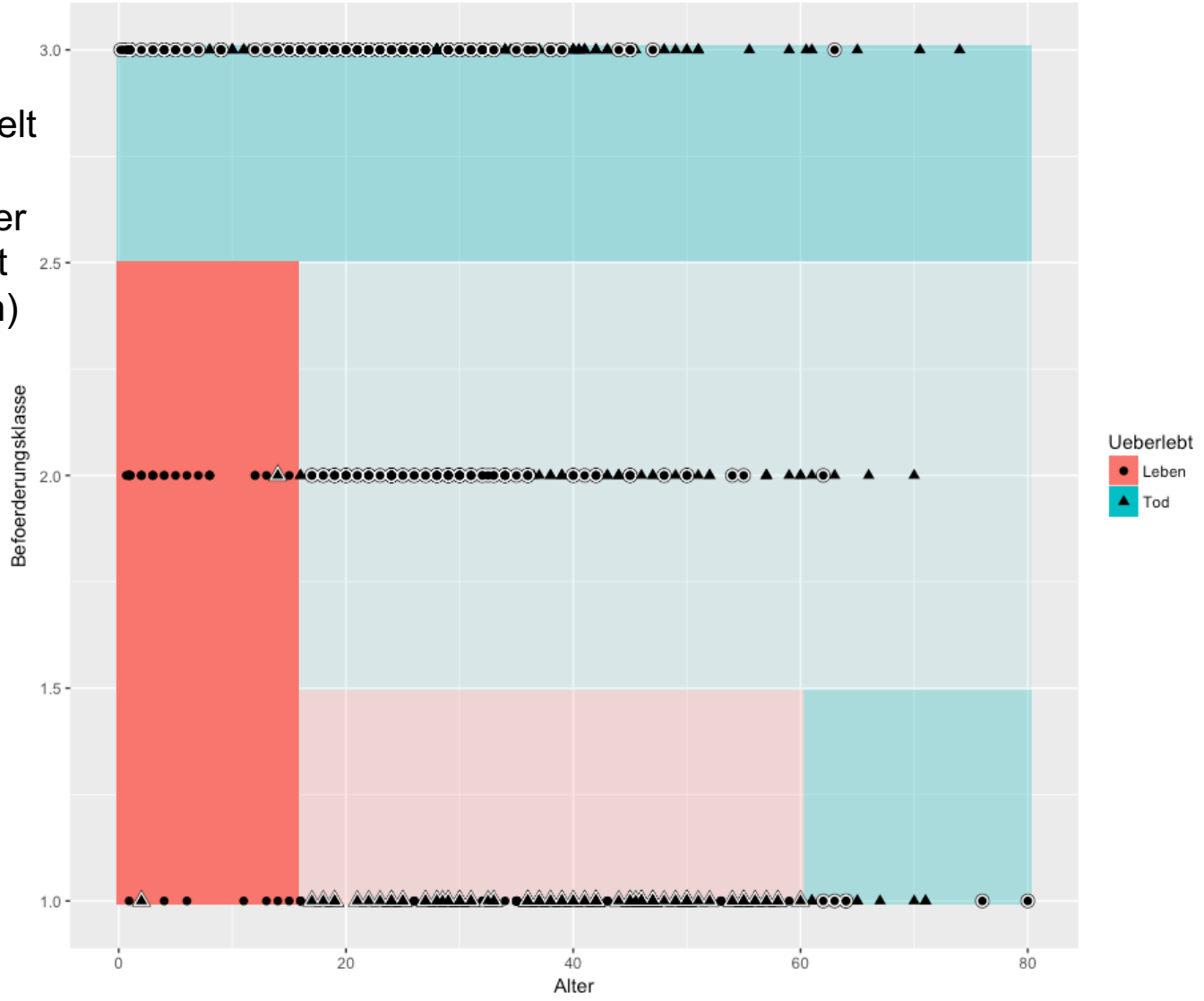
Veranschaulichung: 5 Splits (Boston Housing mit nur 2 Prädiktoren)



Umgang mit **diskreten Prädiktoren** (Titanic mit nur 2 Prädiktoren)

Hinweise:

- Beförderungsklasse wird hier als numerische Prädiktorvariable behandelt
- viele Beobachtungen liegen genau übereinander (häufigster Kriteriumswert optisch nicht zu erkennen)

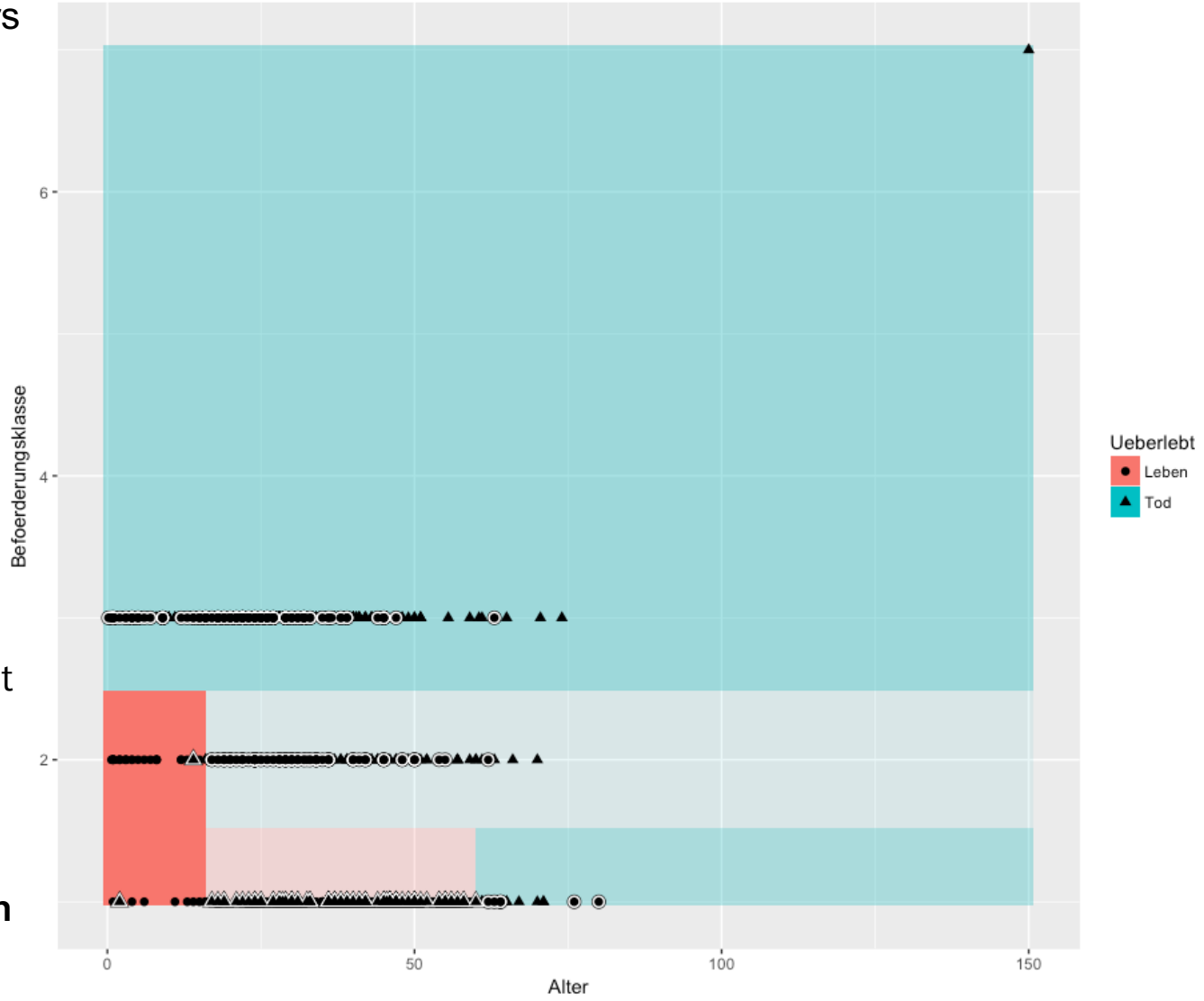


Bei numerischen Prädiktoren findet der Split in der Mitte zwischen den nächsten im Datensatz beobachteten Werten statt

→ **Diskrete Prädiktoren sind kein Problem**

Umgang mit **Ausreißerwerten** (Titanic mit nur 2 Prädiktoren)

Hinzufügen eines Ausreißers mit den Prädiktorwerten:
Alter = 150
Beförderungsklasse = 7



Auf Vorhersagen für Beobachtungen im normalen Wertebereich hat der Ausreißer so gut wie keinen Einfluss

→ **Robust gegenüber Ausreißerwerten in den Prädiktorvariablen**

Depressionsklassifikation

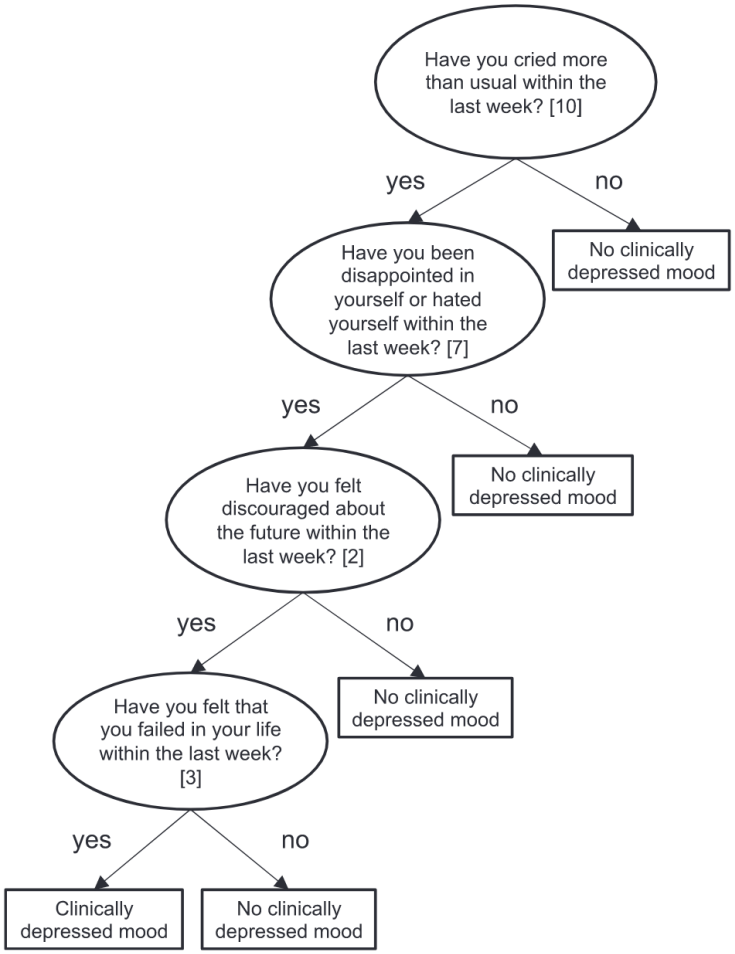


Fig. 2. Fast and frugal tree screening for depressed mood (according to the total BDI score). The numbers in brackets indicate the position of the respective item in the BDI. The full wording of the BDI cues is presented in Table 1; for this figure, it has been translated into binary questions.

zitiert nach Jenny, M. A., Pachur, T., Williams, S. L., Becker, E., & Margraf, J. (2013). Simple rules for detecting depression. *Journal of Applied Research in Memory and Cognition*, *2*(3), 149–157. https://doi.org/10.1037/h0101797

Vorteile von Entscheidungsbäumen

- Grafische Veranschaulichung des Modells
  → anschauliche Darstellung von Interaktionen
- Tendenziell niedriger Bias:
  - Interaktionen werden automatisch berücksichtigt
  - Nonlineare Zusammenhänge werden automatisch berücksichtigt
- Automatische Auswahl der wichtigsten Prädiktorvariablen
  → unwichtige Prädiktoren werden nie als Splitvariable ausgewählt
- Auf fast alle Datentypen anwendbar
  - Klassifikation (auch mehr als 2 Kategorien) und Regression
  - einfacher Umgang mit diskreten Prädiktorvariablen
- Nicht anfällig gegenüber Ausreißerwerten in den Prädiktorvariablen
- Standardisierung (und alle anderen monotonen Transformationen) der Prädiktorvariablen haben keinen Einfluss auf die Vorhersagen

Eine visuelle Wiederholung …

R2
D3

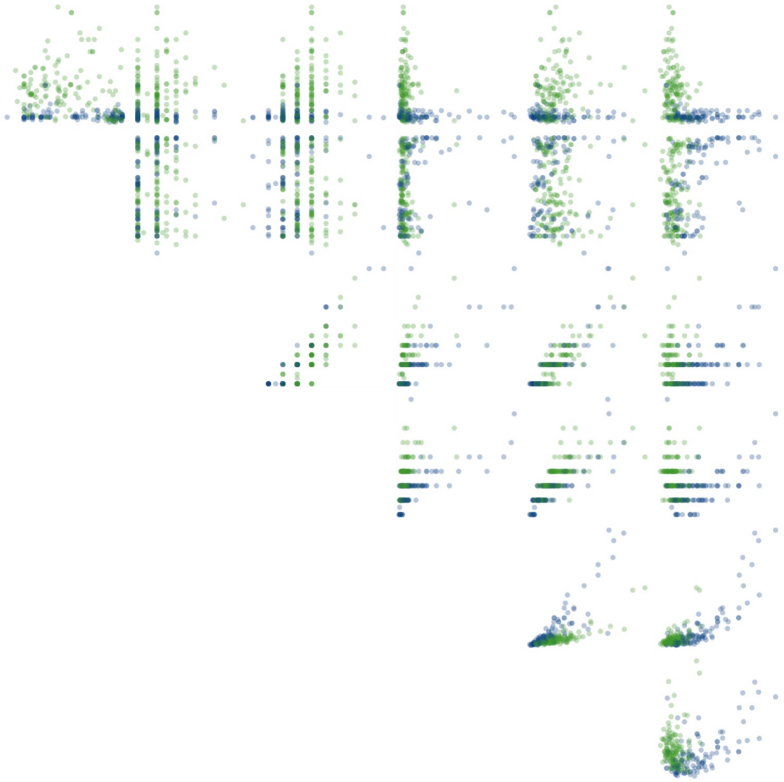
A visual introduction to machine learning

English

In machine learning, computers apply **statistical learning** techniques to automatically identify patterns in data. These techniques can be used to make highly accurate predictions.

Keep scrolling. Using a data set about homes, we will create a machine learning model to distinguish homes in New York from homes in San Francisco.

SCROLL



http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
http://www.r2d3.us/visual-intro-to-machine-learning-part-2/

Nächste Woche

- Bias und Varianz von Entscheidungsbäumen
  → Nachteile von Entscheidungsbäumen
- Möglichkeiten zur Verbesserung der Vorhersagegüte
  → **Random Forest**