5. Entscheidungsbäume (Teil 2) & Random Forest

Agenda für heute

- Kurze Wiederholung von Entscheidungsbäumen
- Nachteile von Entscheidungsbäumen
- Wie kann man den Nachteilen von Entscheidungsbäumen begegnen?
  → Random Forest
- Veranschaulichung anhand von Beispielen
- Interpretierbarkeit des Random Forest

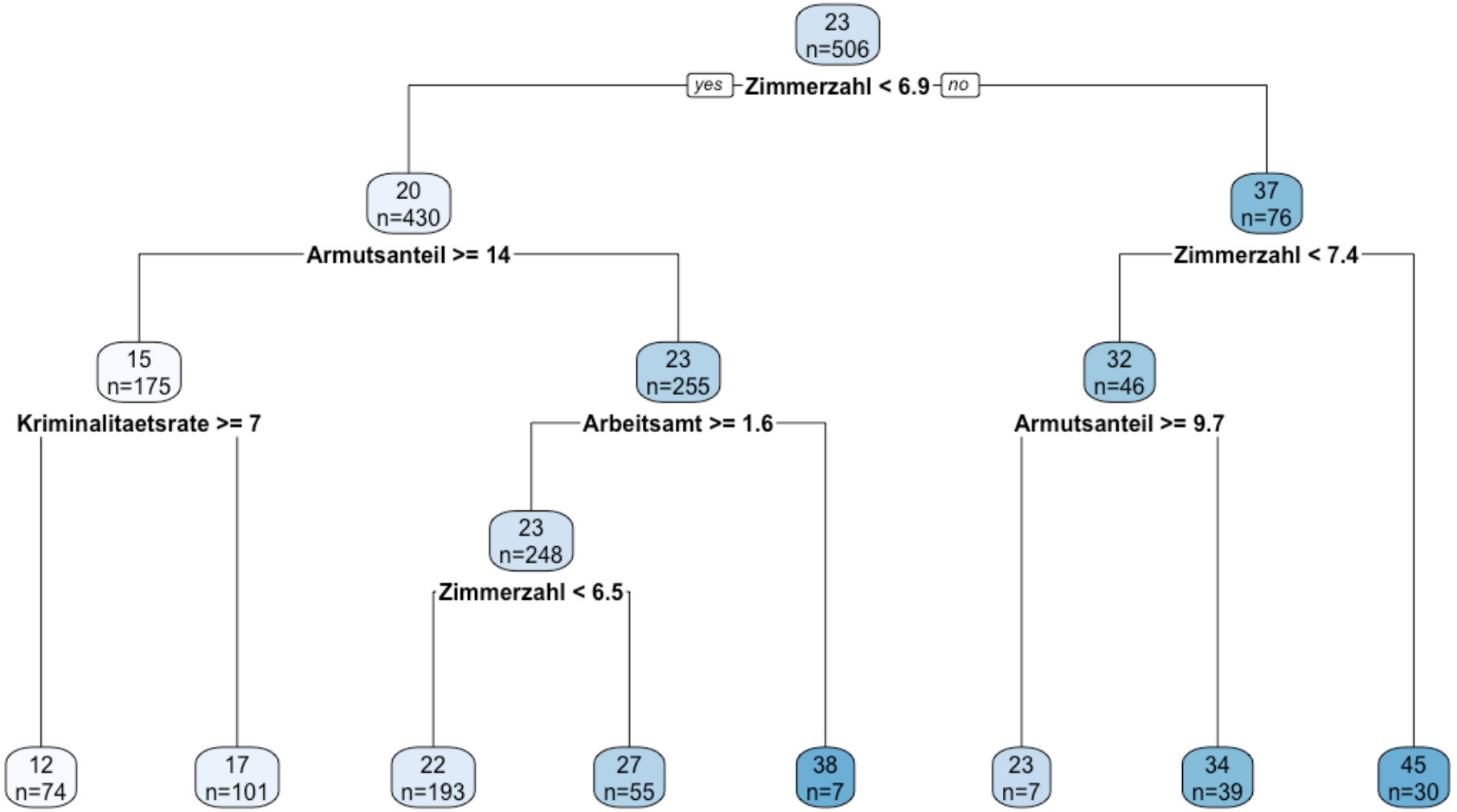
Wiederholung

- Entscheidungsbäume → CART („Classification and Regression Trees“)
- Rekursiver Split-Algorithmus
  - Wahl der Splitvariable und des Splitpunkts
  - Impurity-Maße um den optimalen Split zu bestimmen
- Vorteile von Entscheidungsbäumen
  - grafische Anschaulichkeit
  - tendenziell niedriger Bias
  - automatische Auswahl von Prädiktoren
  - auf fast alle Datentypen anwendbar
  - nicht anfällig für Ausreißer

Bias und Varianz von Entscheidungsbäumen

- Bei Verwendung tiefer Bäume haben Entscheidungsbäume einen **sehr geringen Bias**. Die Baumstruktur berücksichtigt ...
  - nonlineare Zusammenhänge zwischen den Prädiktorvariablen und der Kriteriumsvariable
  - (nonlineare) Interaktionen zwischen den Prädiktorvariablen
- Gleichzeitig weisen tiefe Bäume eine **sehr hohe Varianz** auf:
  - Bei einer neuen Stichprobe aus der gleichen Population kann sich die gesamte Baumstruktur und damit auch die Vorhersagen des Baumes teilweise stark ändern (Entscheidungsbäume sind „instabil“)
  - Aufgrund der hohen Varianz ist die intuitive Interpretierbarkeit der Entscheidungsbäume fragwürdig
    → mit der Baumstruktur ändert sich oft auch die Interpretation

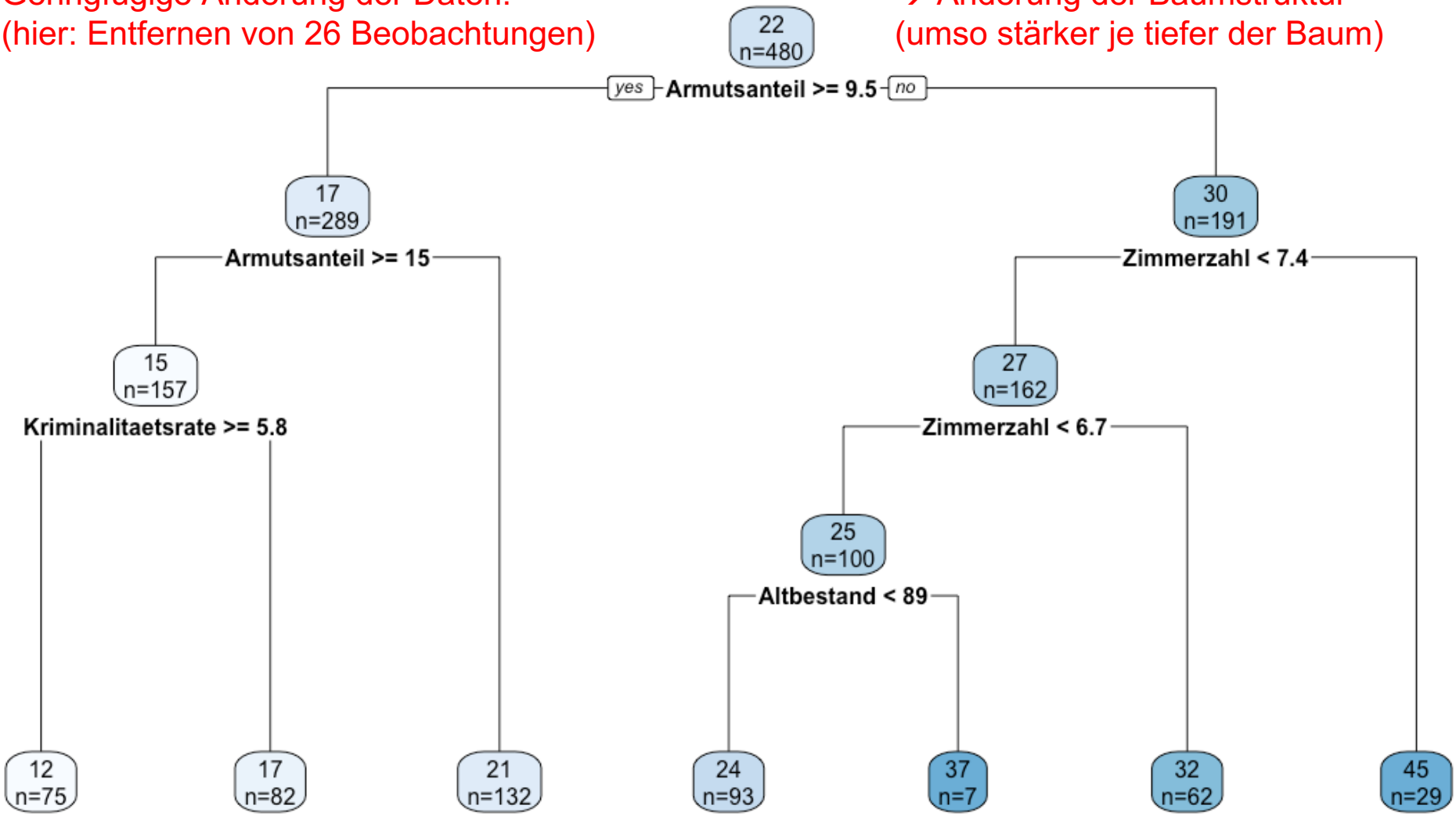
Veranschaulichung Instabilität: Boston Housing (**N = 506**)



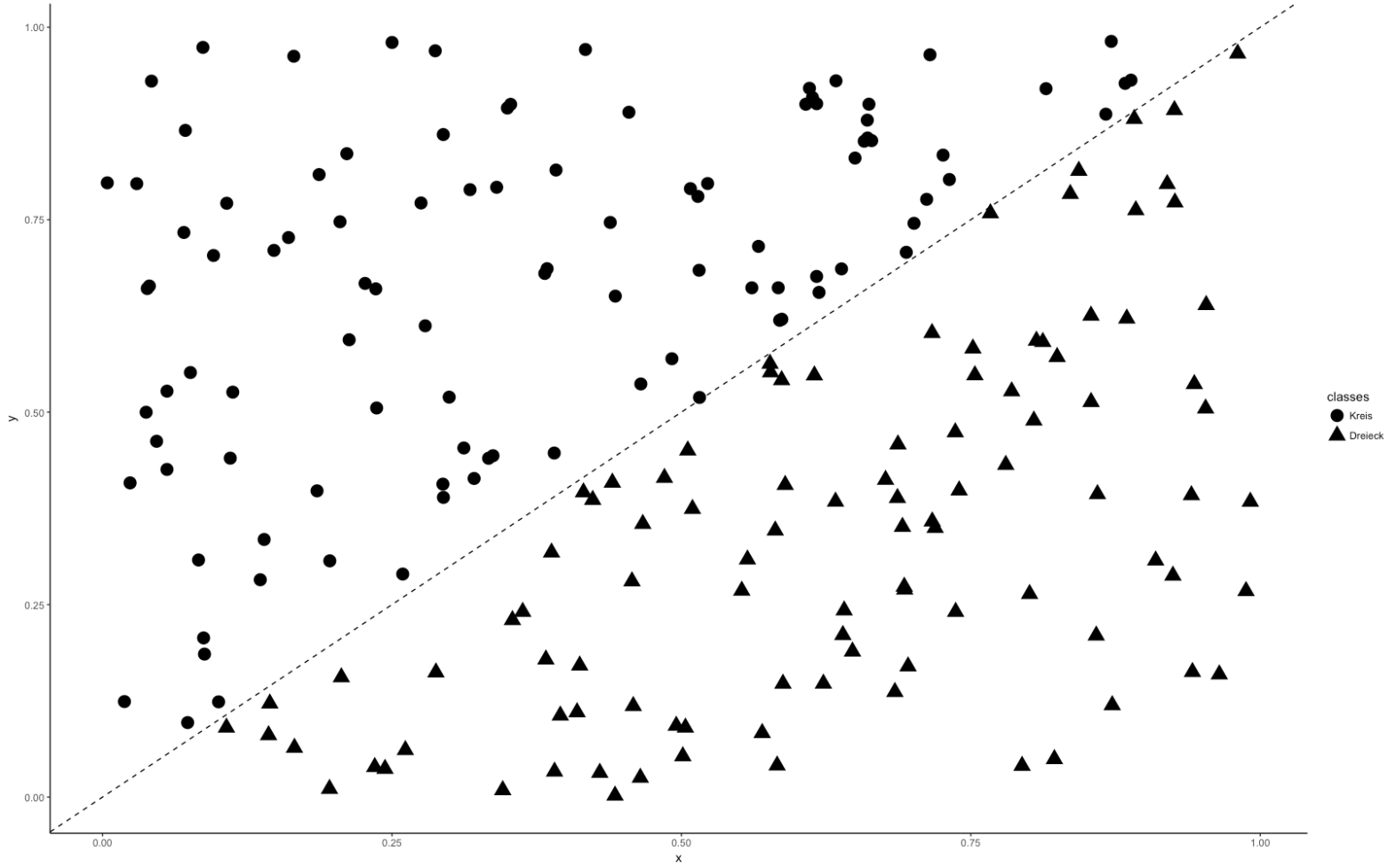
Veranschaulichung Instabilität: Boston Housing (**N = 480**)

Geringfügige Änderung der Daten:
(hier: Entfernen von 26 Beobachtungen)

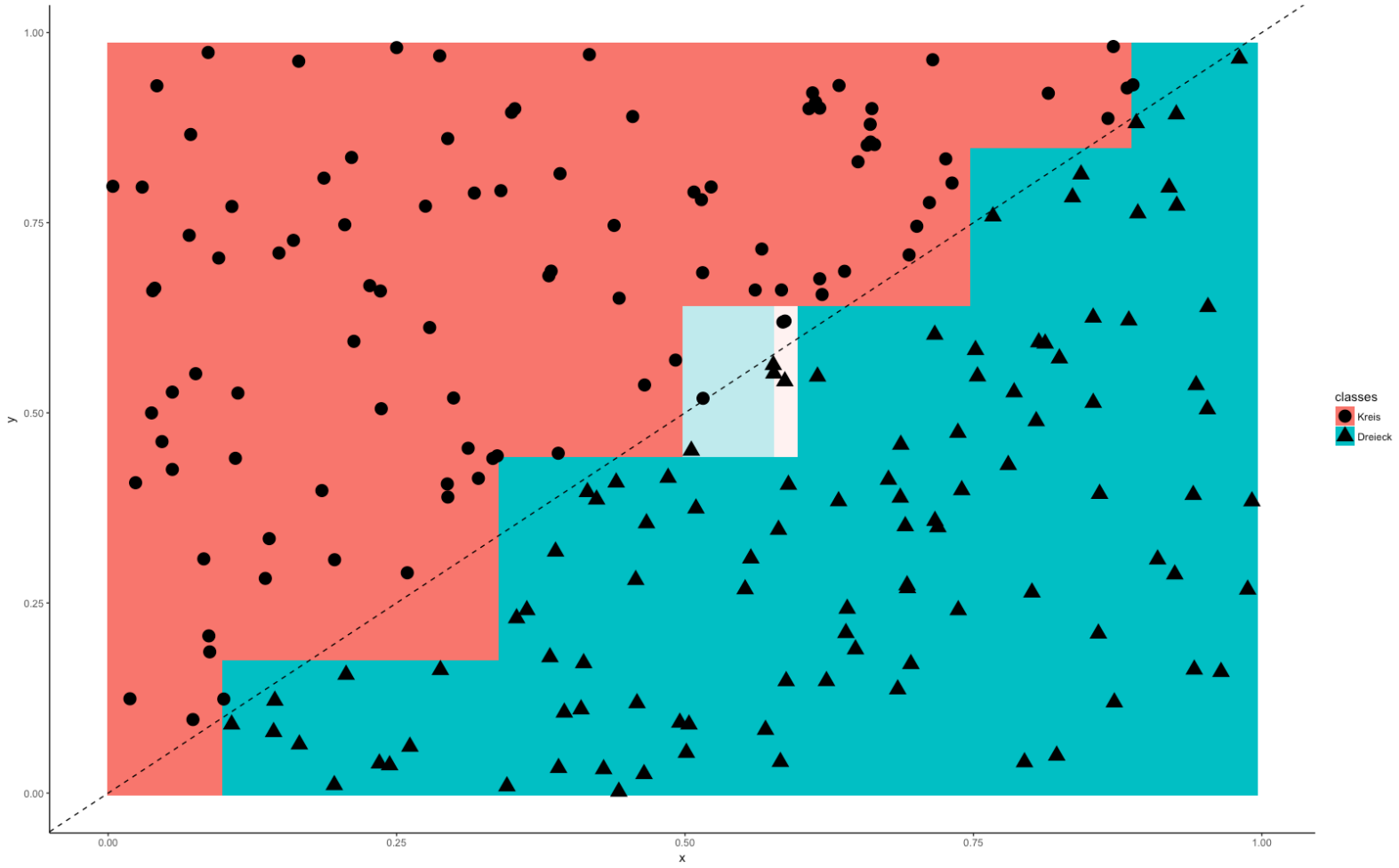
→ Änderung der Baumstruktur
(umso stärker je tiefer der Baum)



Veranschaulichung Linearität



Veranschaulichung Linearität



Lineare Zusammenhänge können durch Bäume nur approximiert werden

Nachteile von Entscheidungsbäumen

- Suboptimale Vorhersagegüte bei out-of-sample Daten aufgrund einer tendenziell hohen Varianz der Modellklasse
  → Hauptproblem: Instabilität der Bäume
- Um die bestmögliche Vorhersagegüte zu erzielen, müssen die Abbruchkriterien beim „Wachsen der Bäume“ datengesteuert gewählt werden.
  → Verhindern von Overfitting!

Hinweis:

- Wie man optimale Abbruchkriterien in der Praxis datengesteuert wählt wurde nicht weiter erklärt, da Entscheidungsbäume in dieser Vorlesung nur als die nötigen Bausteine des leistungsfähigeren Random Forests behandelt werden. Wie wir später noch sehen werden, spielt im Random Forest die Wahl der Abbruchkriterien keine große praktische Rolle.

Verbesserung der Vorhersagegüte von Entscheidungsbäumen

… oder der „Weg“ der Entscheidungsbäume zum Random Forest:

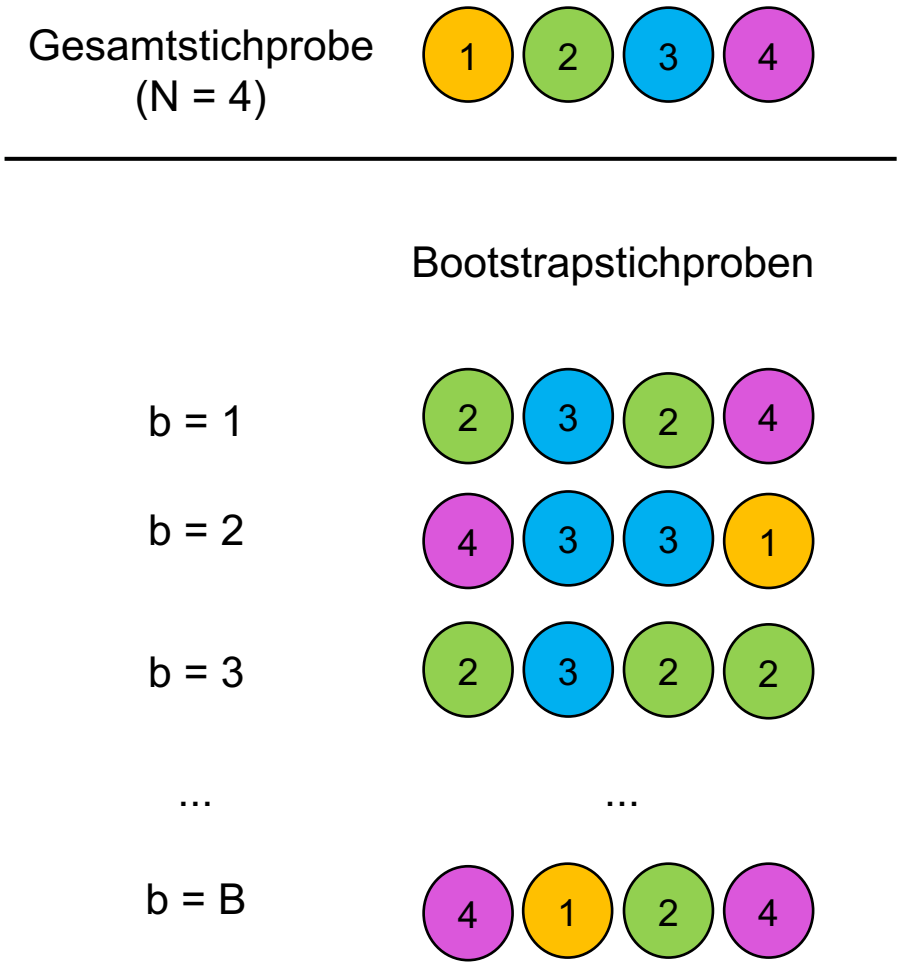
- Schritt 1: Tree Bagging (#11-14)
- Schritt 2: Verbesserung von Tree Bagging durch Modifikation der einzelnen Bäume (#15-16)

Schritt 1: Tree Bagging

- Ziel: Verbesserung der Vorhersagegüte von Entscheidungsbäumen durch Reduktion der Varianz (ohne eine gleichzeitig starke Erhöhung des Bias)
- **Naive Idee:**
  - Ziehe mehrere Stichproben aus der Population
  - Trainiere mit jeder Stichprobe einen eigenen Entscheidungsbaum (Bäume sind aufgrund der hohen Varianz leicht unterschiedlich)
  - Zur Vorhersage einer neuen Beobachtung, berechne den Mittelwert der Vorhersagen aus den einzelnen Bäumen/Stichproben
  - Bei einer großen Anzahl von Bäumen/Stichproben sollten die kombinierten Vorhersagen relativ stabil sein
- **Problem:** Mehrere Stichproben zu erheben ist praktisch nicht umsetzbar
- **Lösung:** Nutze die vorliegenden Daten zur Erzeugung mehrerer „Pseudostichproben“ mit der gleichen Größe wie der Ursprungsdatensatz

Schritt 1: Tree Bagging - Bootstrap

Idee: Simuliere das Ziehen echter neuer Stichproben aus der Population



Bootstrapstichprobe:
Ziehe N mal eine Beobachtung aus der Gesamtstichprobe **mit** Zurücklegen

- Bootstrapstichprobe hat die gleiche Größe wie die Gesamtstichprobe
- Manche Beobachtungen aus der Gesamtstichprobe können in der Bootstrapstichprobe mehrfach enthalten sein, andere gar nicht
- Es existieren theoretisch $\binom{2N-1}{N}$ mögliche Bootstrapstichproben: in der Regel zu viele
- Betrachte daher nur *B* zufällig ausgewählte Bootstrapstichproben

Schritt 1: Tree Bagging

Modellschätzung:

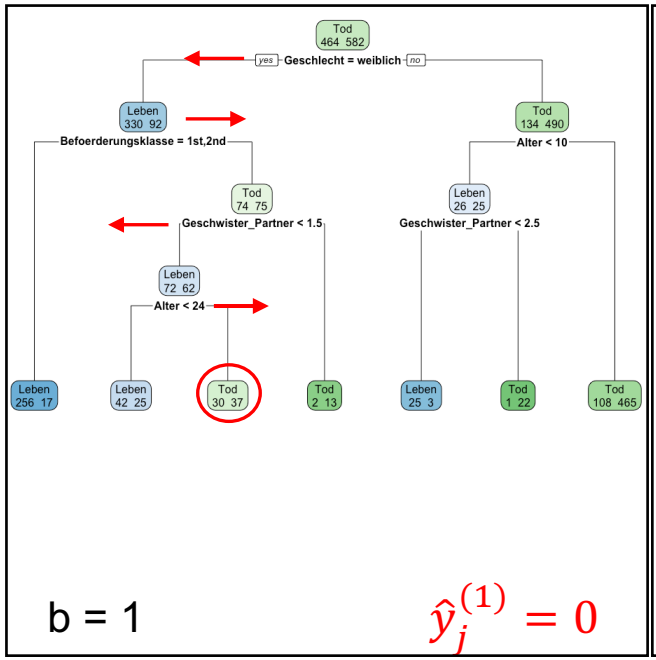
- Ziehe *B* Bootstrapstichproben aus der Gesamtstichprobe
- Für jede Bootstrapstichprobe: trainiere einen Entscheidungsbaum

Berechnung von Vorhersagen:

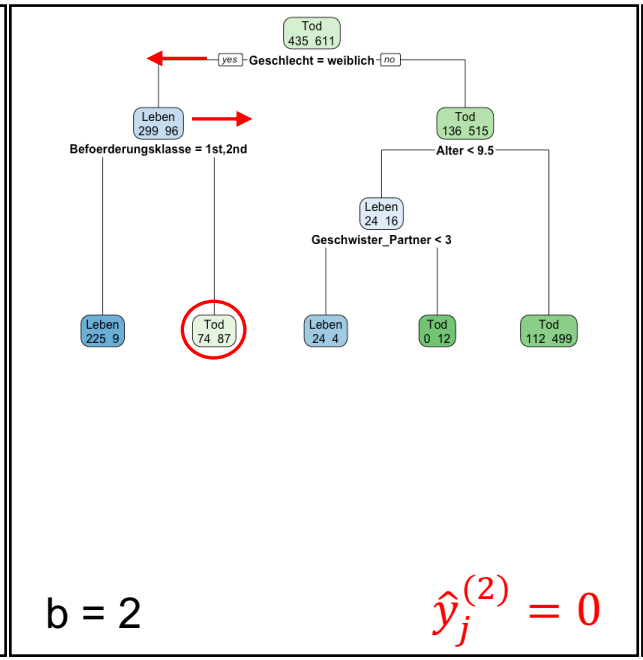
- Für eine neue Beobachtung, berechne Vorhersagen aus allen *B* Bäumen
- Kombiniere die *B* Vorhersagen der Bäume zu einer Gesamtvorhersage
  - Regression:
    Gesamtvorhersage ist der Mittelwert der *B* einzelnen Vorhersagen
  - Klassifikation:
    Gesamtvorhersage ist der häufigste Kriteriumswert der *B* einzelnen Vorhersagen („Mehrheitsentscheid“)

→ Das Mitteln von Vorhersagen mehrerer prädiktiver Modelle der gleichen Modellklasse, die auf Bootstrapstichproben trainiert wurden, bezeichnet man allgemein als „**Bagging**“ (steht für „Bootstrap Aggregation“).

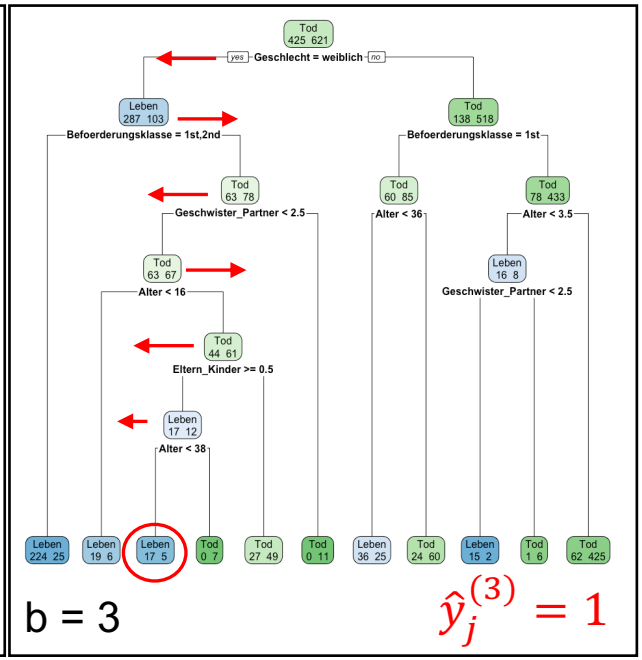
Veranschaulichung Bagging: 3 Bäume (Titanic)



b = 1 $\hat{y}_j^{(1)} = 0$



b = 2 $\hat{y}_j^{(2)} = 0$



b = 3 $\hat{y}_j^{(3)} = 1$

$$\hat{y}_j^{(Bagging)} = 0$$

Neue Passagier:in j:

- Befoerderungsklasse = 3
- Geschlecht = weiblich
- Alter = 30
- Geschwister_Partner = 1
- Eltern_Kinder = 2

3 Einzelvorhersagen:
2 mal „Tod“, 1 mal „Leben“
→ Mehrheitsentscheid: „Tod“

Schritt 2: Zufällige Variabilität der Bäume

- Verbesserung von Tree Bagging durch Modifikation der einzelnen Bäume:
  - Verwende tiefe Bäume ohne vorzeitige Abbruchkriterien (bzw. mit einem sehr milden Abbruchkriterium, wie z.B. bei Regression mindestens 5 Beobachtungen in jedem Endknoten beizubehalten)
  - Berücksichtige für jeden Split nicht alle Prädiktorvariablen, sondern nur eine zufällig ausgewählte Teilmenge aller Prädiktorvariablen (bei jedem neuen Split stehen dann wieder alle Prädiktorvariablen prinzipiell zur Verfügung)
- **Idee:**
  - Je niedriger der Bias der einzelnen Bäume, desto niedriger ist auch der Bias der kombinierten Vorhersagen (Bias bleibt nahezu gleich) → daher tiefe Bäume
  - Je unterschiedlicher die Struktur der einzelnen Bäume (d.h. niedrige Korrelation zwischen den Vorhersagen der einzelnen Bäume), desto besser kann die Varianz der Random-Forest-Vorhersage durch die Verwendung einer großen Anzahl an Bäumen reduziert werden.
    („Ausmitteln ist am effektivsten bei unkorrelierten Zufallsvariablen“, siehe Standardfehler des Mittelwerts)

Random Forest

Die target-Variable wird immer benötigt

Wähle zufällig ein Subset an features (d.h., Prädiktoren) aus

ID	Y	X_1	X_2	X_3	...	X_{10}
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
...						
N						

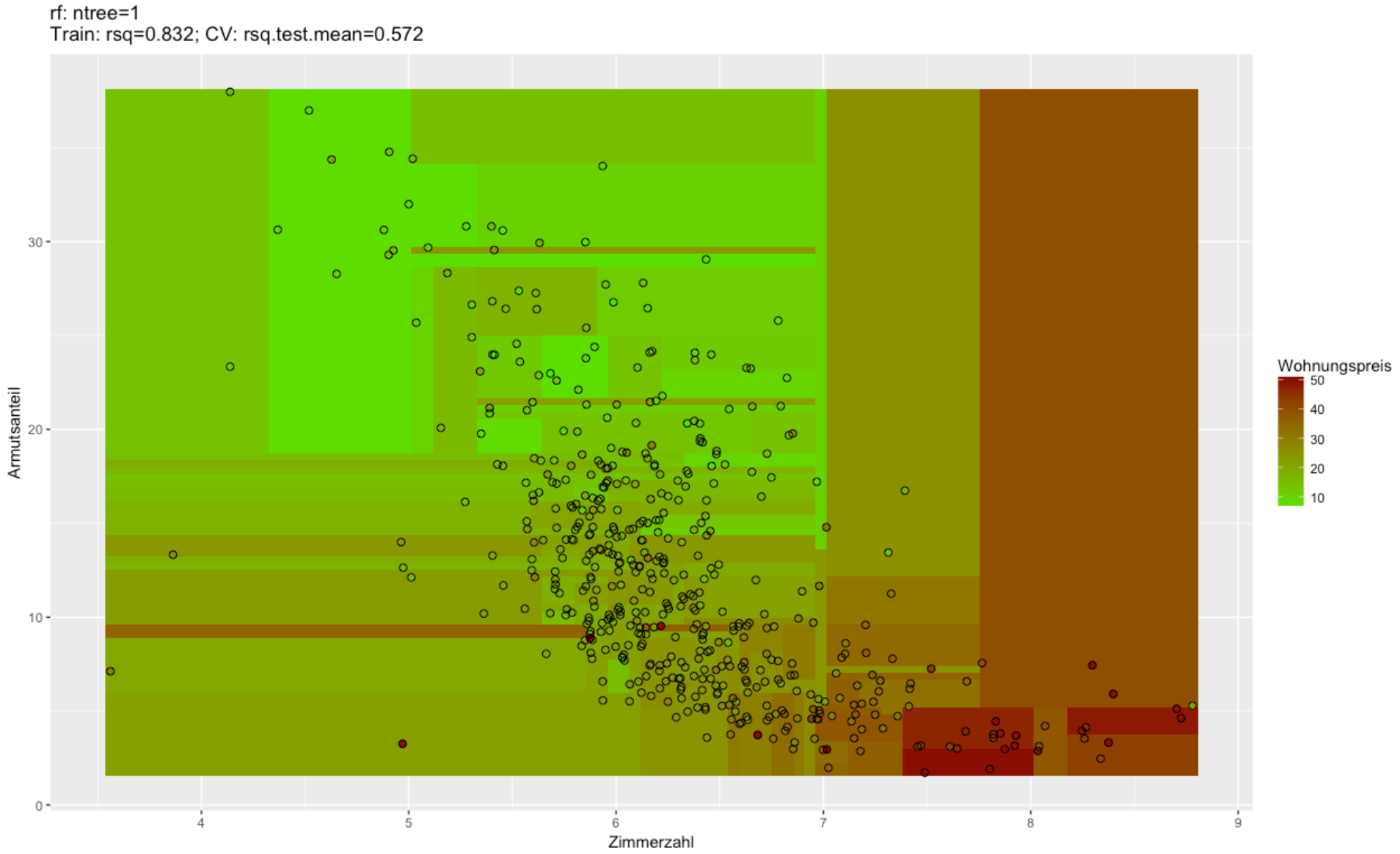
Wähle zufällig (*mit* Zurücklegen) ein Subset an Fällen aus → Bootstrap

Berechne einen tiefen Baum; wiederhole *ntree* mal

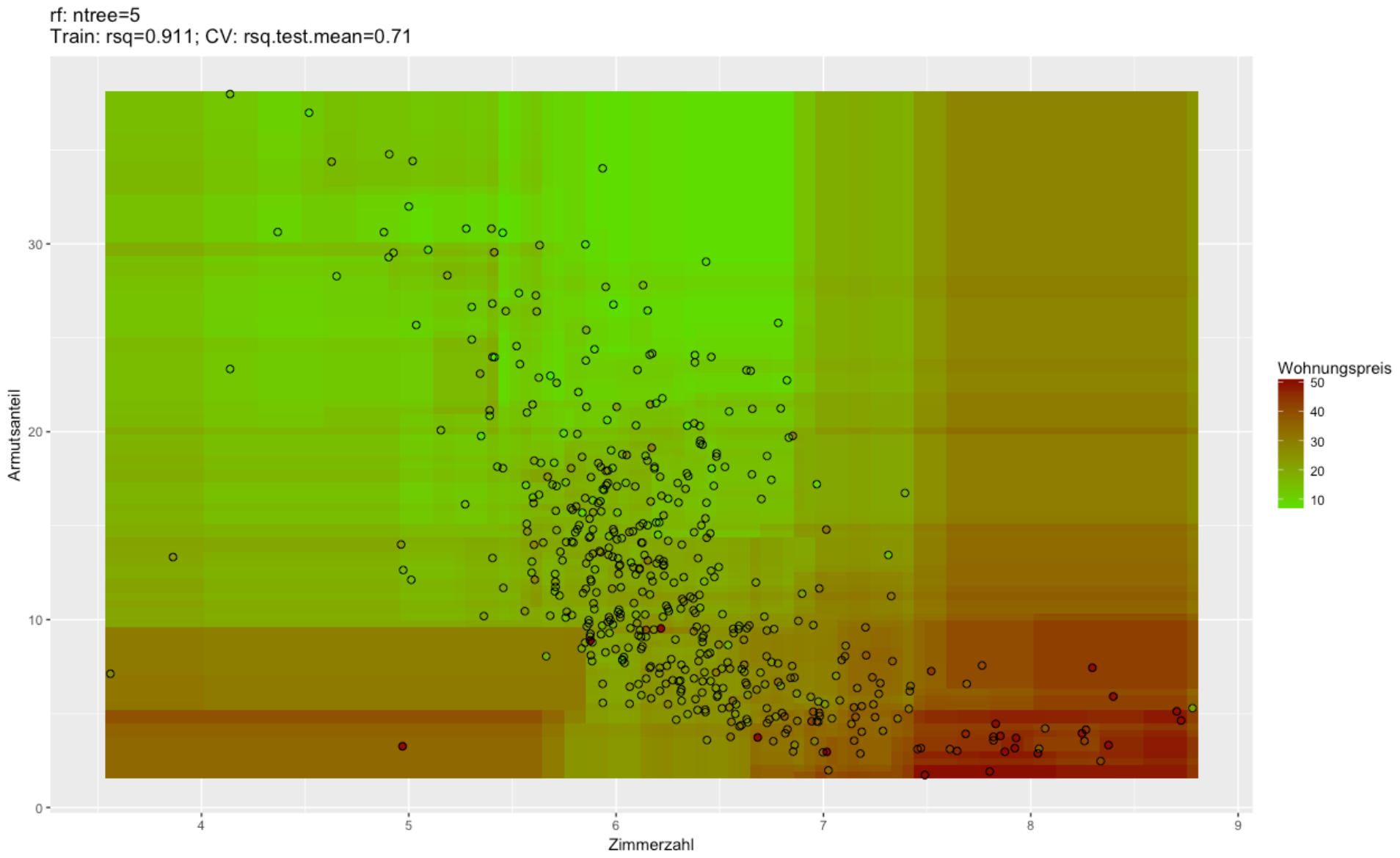
Hyperparameter

- Als „Hyperparameter“ bezeichnet man Einstellungsmöglichkeiten einer prädiktiven Modellklasse, die vor dem Training auf einen bestimmten Wert festgelegt werden müssen (können also nicht mitgeschätzt werden).
- Die wichtigsten Hyperparameter beim Random Forest sind:
  - Anzahl der Bäume (*ntree*)
    - Standardeinstellung: *ntree* = 500 (je mehr Bäume desto besser)
  - Anzahl zufällig ausgewählter Prädiktorvariablen pro Split (*mtry*). Sinnvolle default-Werte *(p* = Anzahl der Prädiktorvariablen*):*
    - Regression: $\frac{p}{3}$ (abrunden)
    - Klassifikation: $\sqrt{p}$ (abrunden)
  - Mindestanzahl an Beobachtungen im Endknoten (*min.node.size*). Sinnvolle default-Werte:
    - Regression: 5
    - Klassifikation: 1

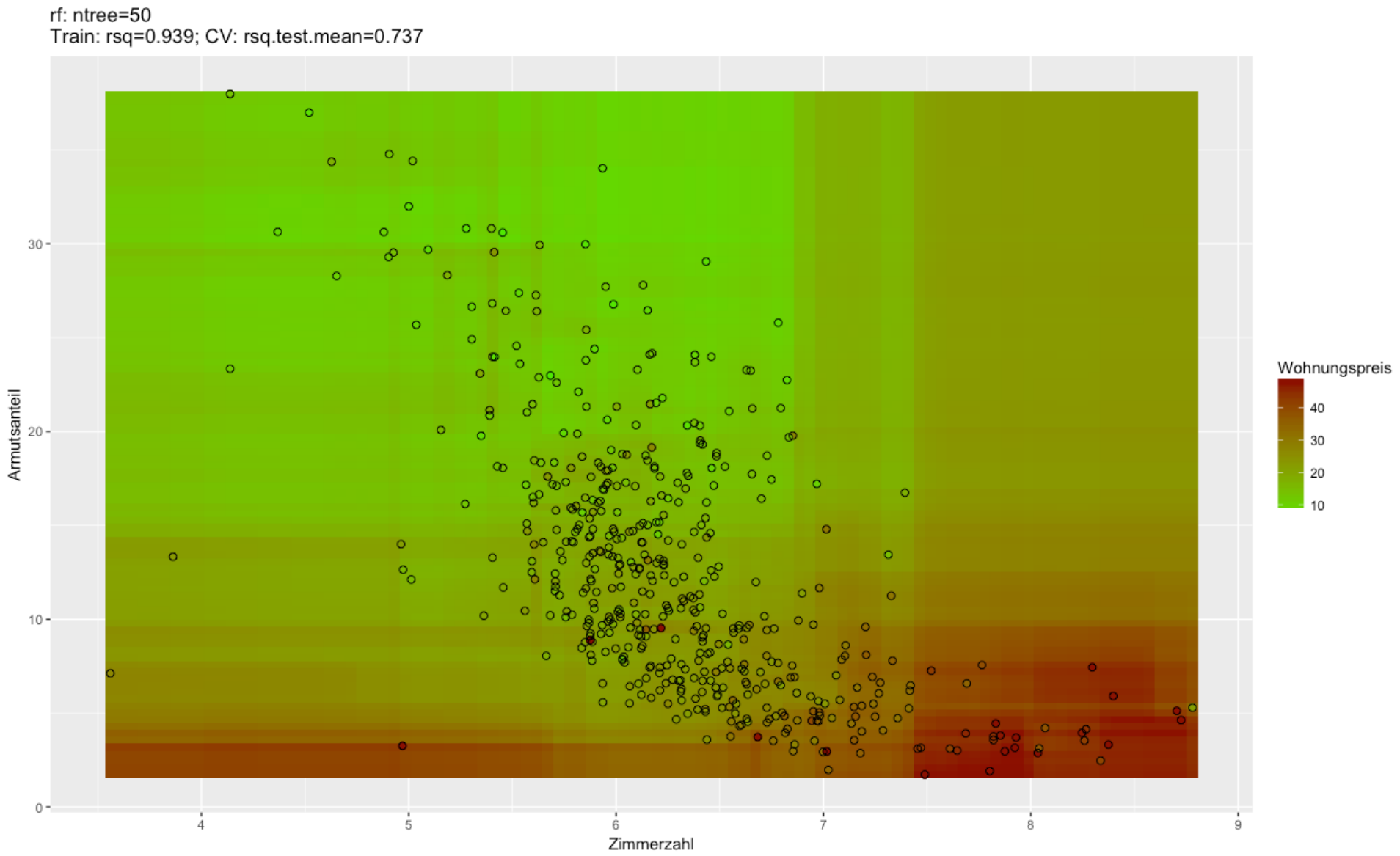
Boston Housing: *ntree* = 1 (sehr tiefer) Baum



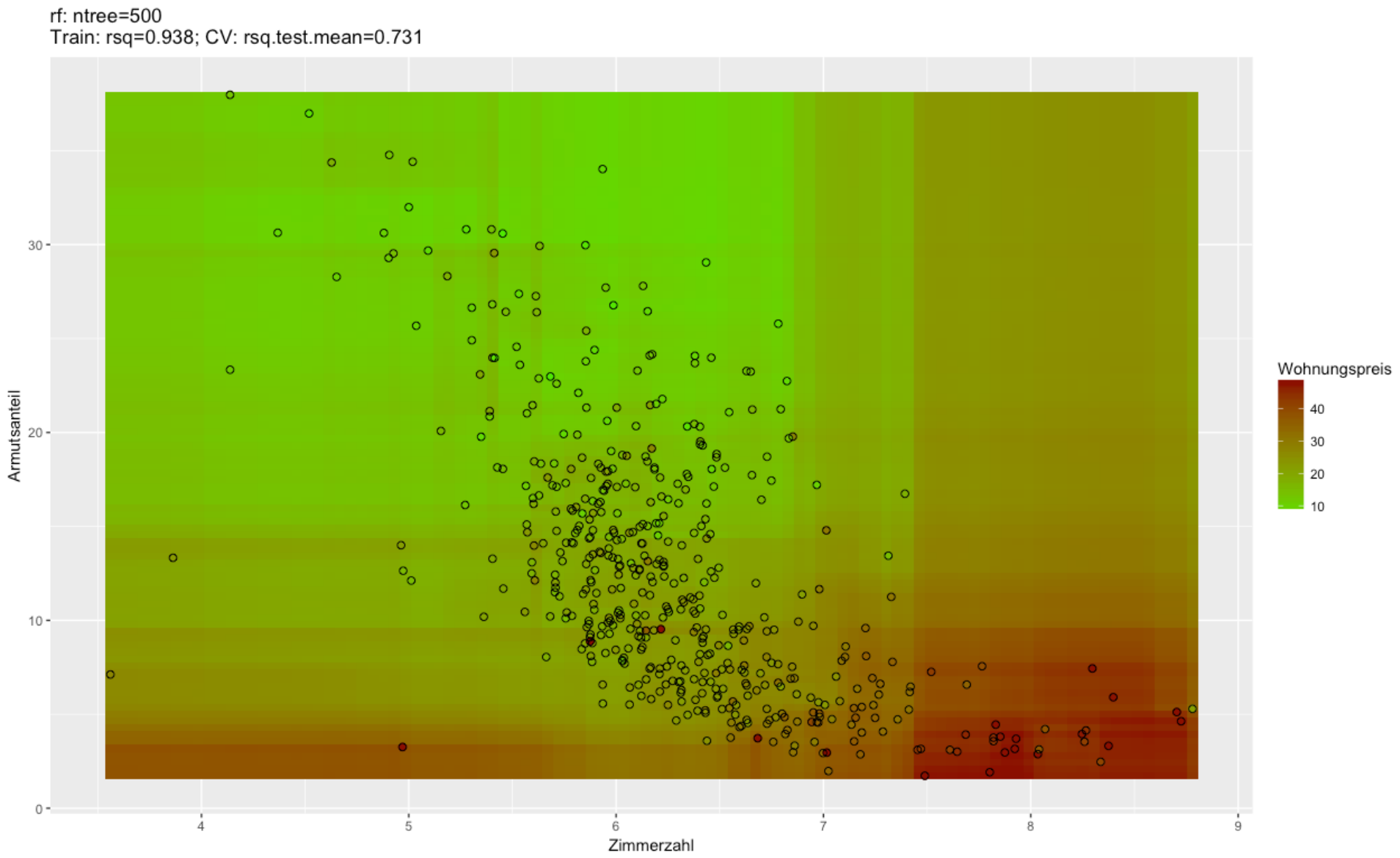
Boston Housing: *ntree* = 5 Bäume



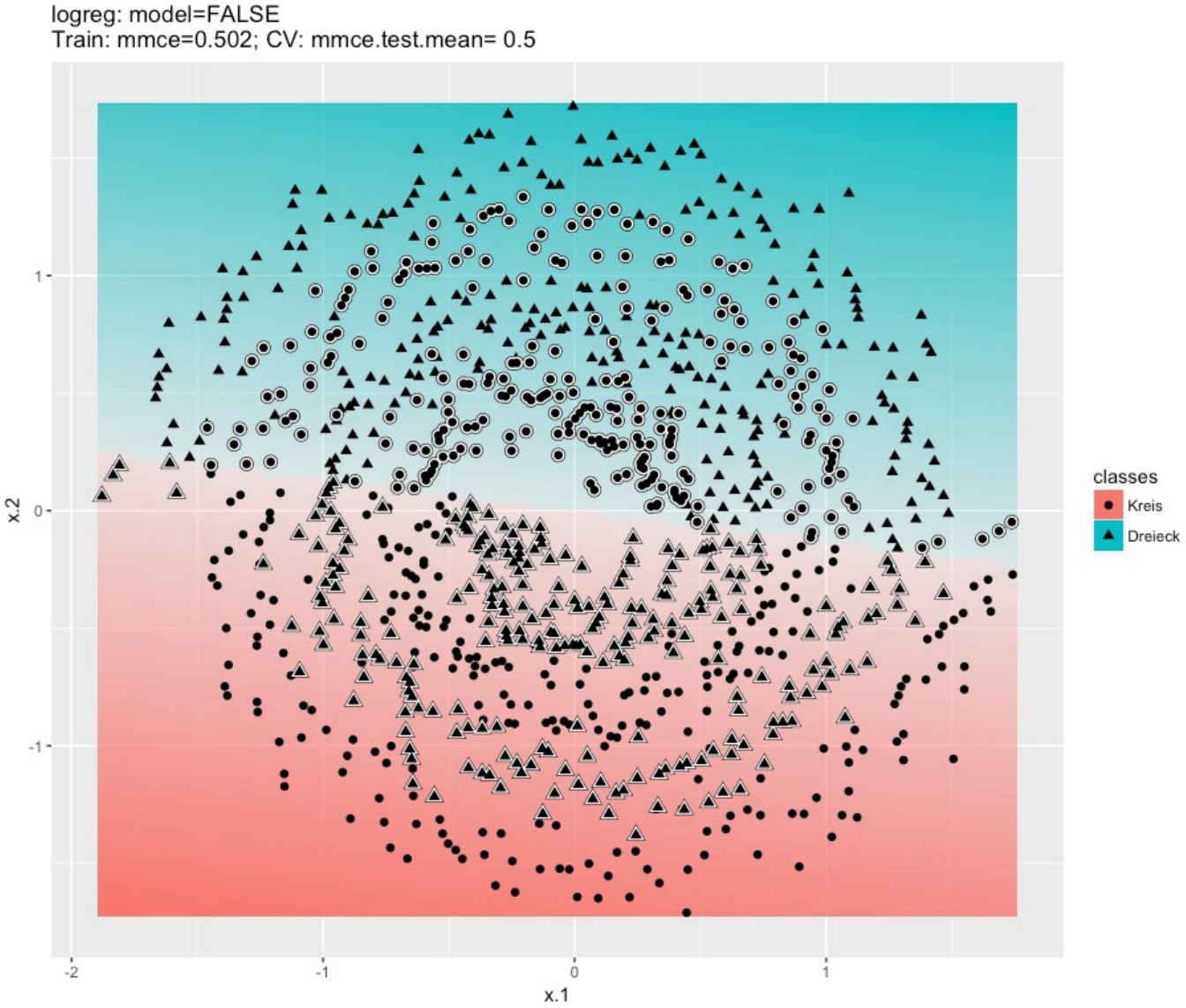
Boston Housing: *ntree* = 50 Bäume



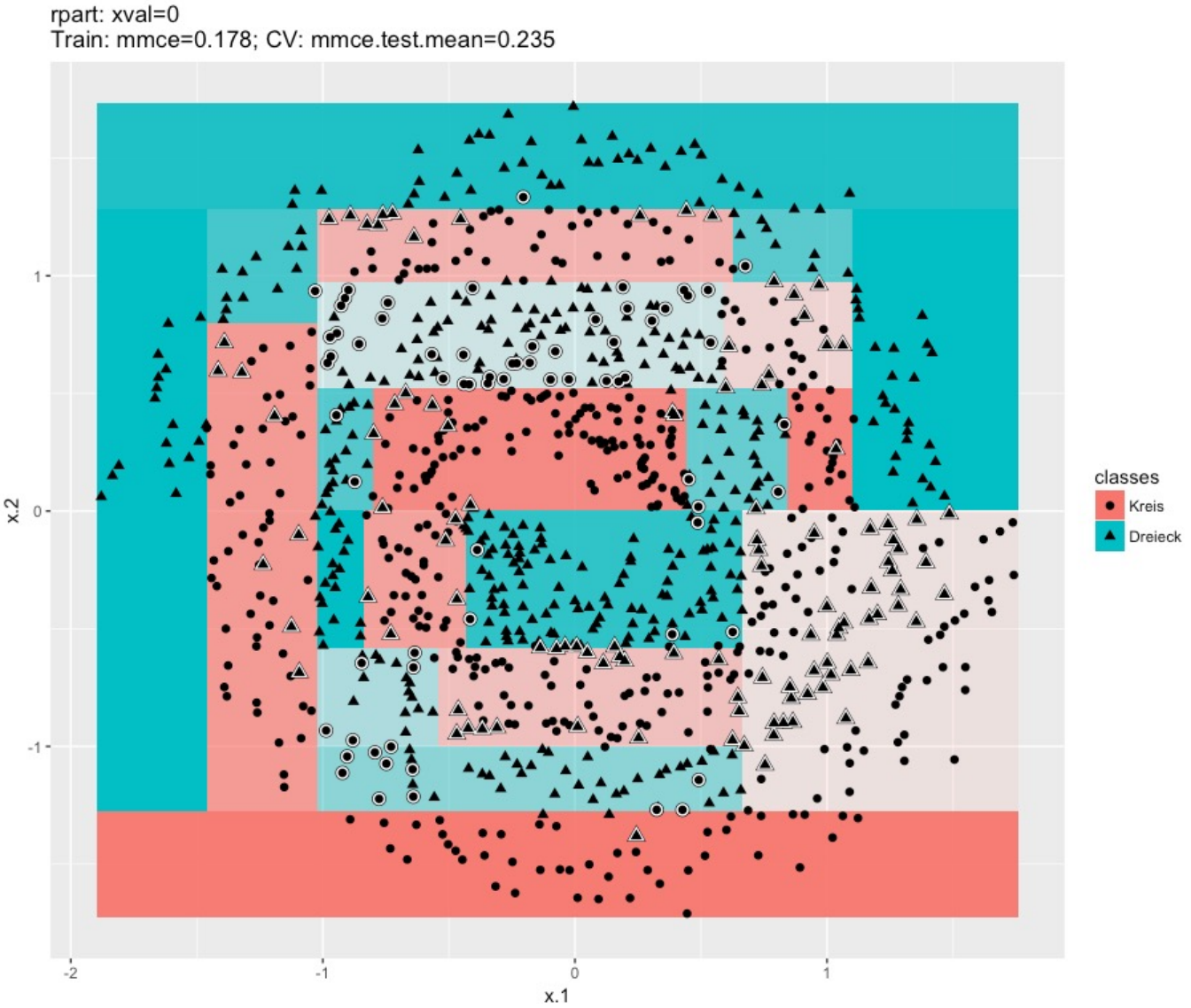
Boston Housing: *ntree* = 500 Bäume



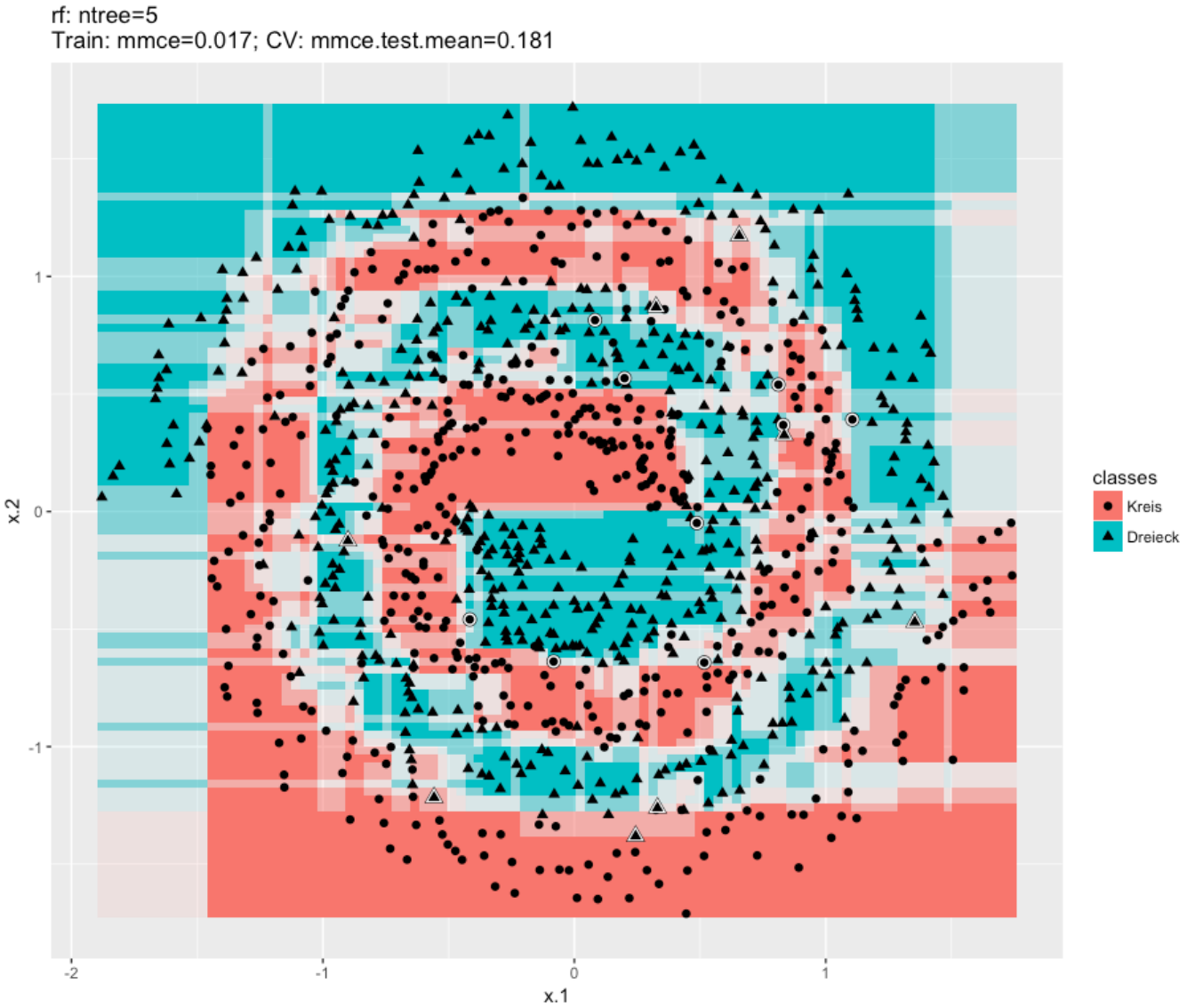
Nonlineare Zusammenhänge - Zum Vergleich: logistische Regression



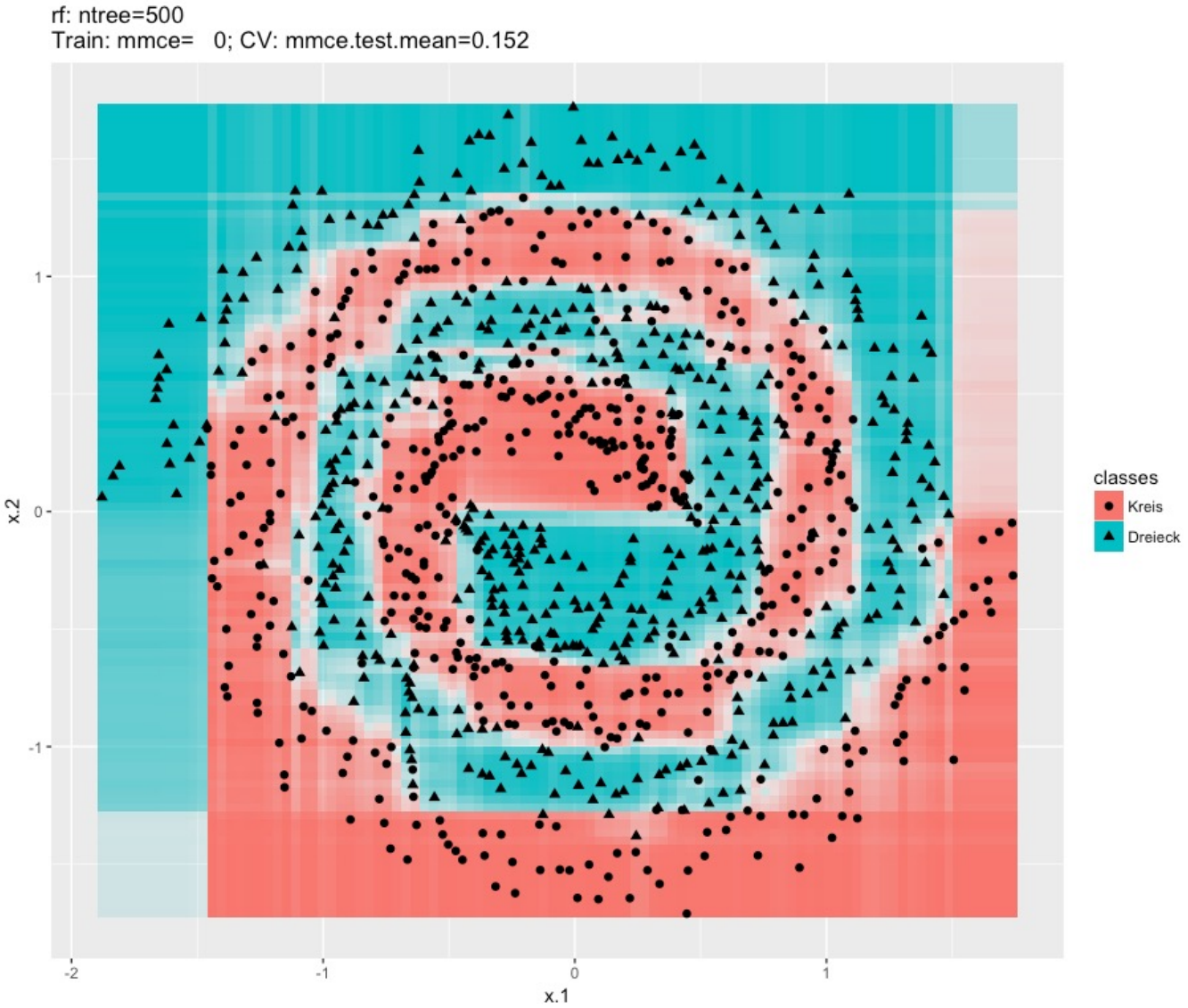
Nonlineare Zusammenhänge: Einzelner Entscheidungsbaum (CART)



Nonlineare Zusammenhänge: Random Forest (5 Bäume)



Nonlineare Zusammenhänge: Random Forest (500 Bäume)



Vor– und Nachteile von Random Forests

Vorteile:

- Alle Vorteile einzelner Entscheidungsbäume (mit Ausnahme der guten Interpretierbarkeit und der grafischen Veranschaulichung des Modells)
- Vorhersagegüte oft vergleichbar mit deutlich aufwändigeren Machine Learning Verfahren (Einer der besten „off-the-shelf“ Algorithmen)
  - Meist kein kompliziertes Tuning der Hyperparameter notwendig
    → niedriger Bias und niedrige Varianz mit Standardeinstellungen
  - Meist keine zusätzliche Vorauswahl von Prädiktorvariablen notwendig
    → effektive Nutzung einer großen Zahl von Prädiktorvariablen

Nachteil:

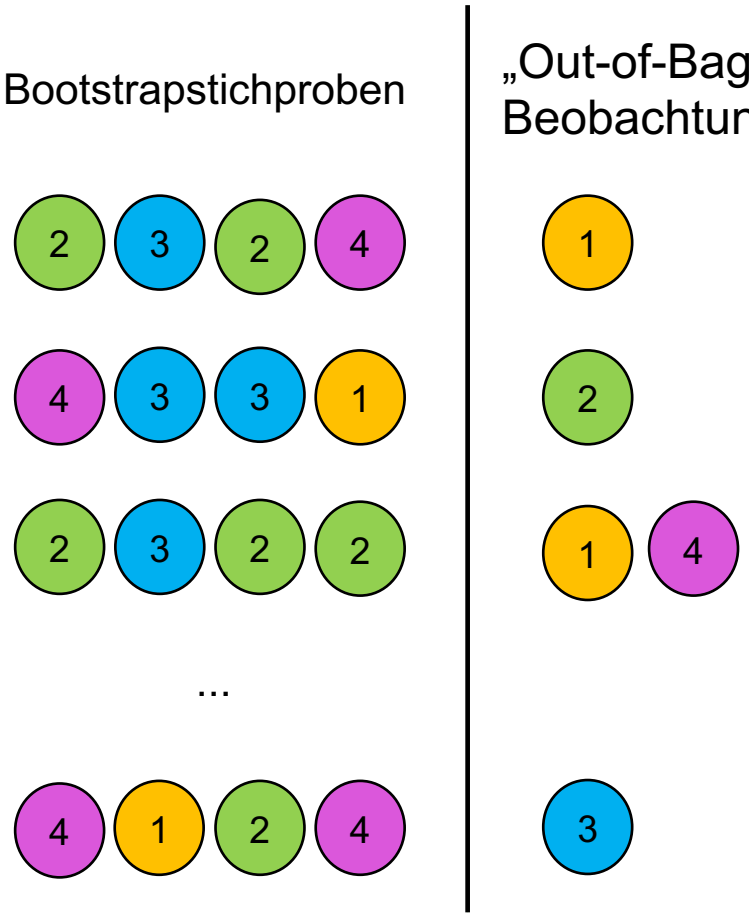
- Modell ist schwer zu interpretieren...
  (Eine grafische Interpretation hunderter tiefer Entscheidungsbäume mit teilweise stark unterschiedlicher Struktur ist in der Praxis unmöglich)
  - Aber: siehe aktuelle Entwicklungen zum „interpretable machine learning (IML)“, bzw. „explainable machine learning (xML)“ / „explainable AI (xAI)“

Interpretierbarkeit nonlinearer prädiktiver Modelle

- Zwei wichtige Aspekte für die Interpretation eines prädiktiven Modells:
  1. Welche Prädiktorvariablen haben den größten Einfluss?
  2. Auf welche Art beeinflusst eine Prädiktorvariable die Vorhersagen?
     - Richtung des Effekts: positiv vs. negativ
     - Form des Effekts: linear vs. nonlinear
     - Art des Effekts: Haupteffekte vs. Interaktionseffekte

- Mögliche Lösungen:
  - zu 1.: Berechnung von „Variable Importance“ Maßen
    → für den Random Forest relativ einfach (siehe im Anschluss)
  - zu 2.: Deskriptive Analyse von „Individual Conditional Expectations“
    → für Interessierte, siehe z.B. Pargent, Schoedel, & Stachl (2022): https://doi.org/10.31234/osf.io/89snd

Variable Importance: Out-of-Bag Beobachtungen



- In jeder Bootstrapstichprobe fehlen im Durchschnitt ca. 1/3 der Beobachtungen der gesamten Stichprobe, welche bei jeder Bootstrapziehung ein Testset darstellen
- Diese Beobachtungen werden auch als „Out-of-Bag“ (OOB) Beobachtungen bezeichnet
- Für jeden einzelnen Baum lässt sich anhand der OOB-Beobachtungen eine realistische Abschätzung des erwarteten Vorhersagefehlers berechnen

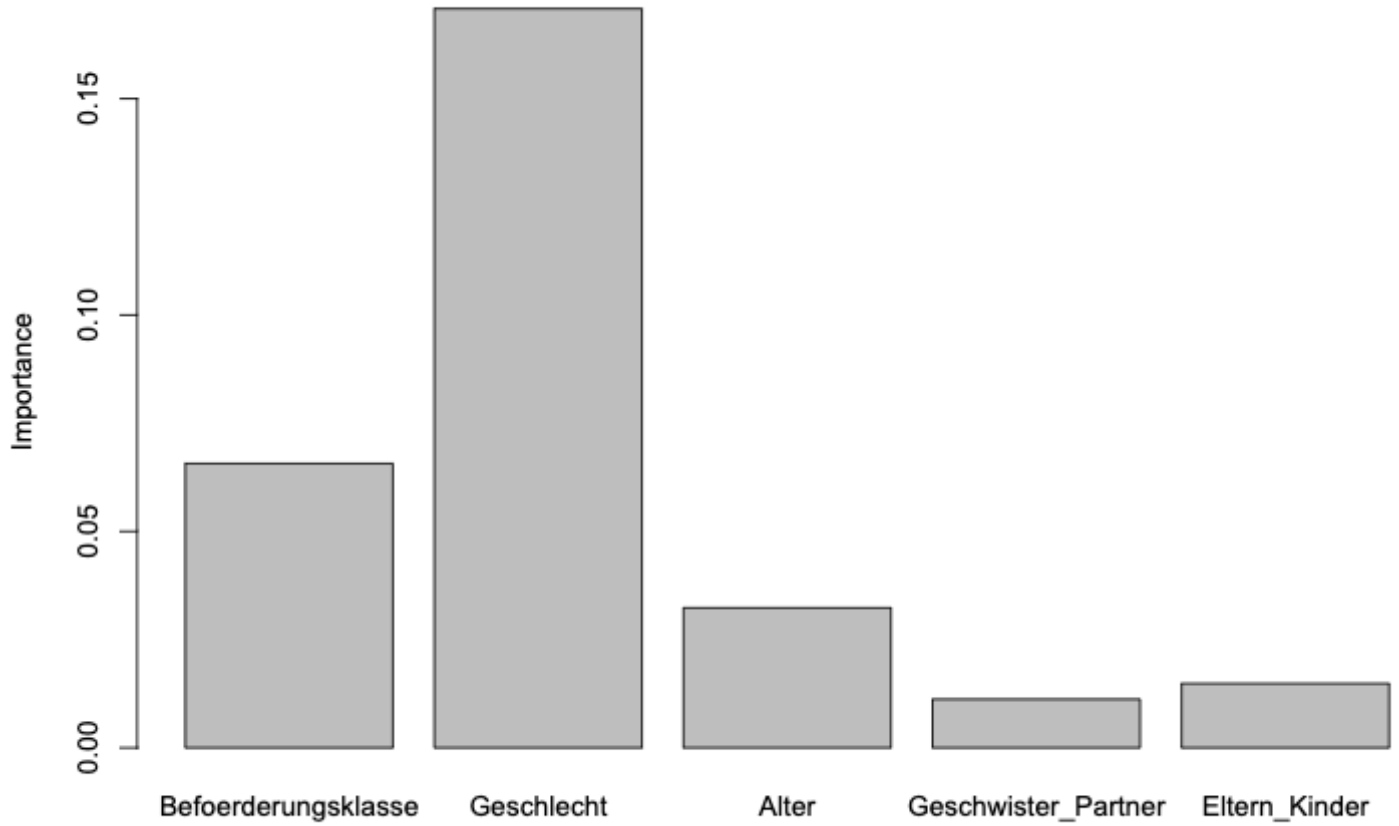
Permutation Variable Importance

Berechnung der „Permutation Variable Importance“ für eine Prädiktorvariable:

- Für jeden Baum im Random Forest ...
  - berechne MSE_{OOB} (bzw. $MMCE_{OOB}$) für die OOB-Beobachtungen
  - durchmische („permutiere“) zufällig die Werte der OOB-Beobachtungen auf der interessierenden Prädiktorvariable
    → Zerstörung der enthaltenen Information über die Kriteriumsvariable
  - berechne $MSE_{OOB}^{(perm)}$ für die OOB-Beobachtungen mit permutierter Prädiktorvariable (alle anderen Prädiktorvariablen bleiben unverändert)
  - berechne die Differenz $MSE_{OOB}^{(perm)} - MSE_{OOB}$: Um wie viel wird die OOB-Vorhersage schlechter, wenn wir eine Prädiktorvariable „kaputt“ machen?
- Variable Importance: Mittelwert der Differenzen über alle Bäume im Forest

Je höher die Variable Importance einer Prädiktorvariable, desto wichtiger ist diese Prädiktorvariable bei der Vorhersage der Kriteriumsvariable.
→ Berücksichtigung des Haupteffekts und aller Interaktionen mit dieser Variable.

Permutation Variable Importance am Beispiel Titanic



Importance = Reduktion in OOB-Vorhersagegüte
≙ Erhöhung in den Fehlklassifikationen (MMCE) bei Randomisierung eines Prädiktors

Ein kleiner Nachtrag: Tortured phrases

„We have been able to spot fraudulent research thanks in large part to one key tell that an article has been artificially manipulated: The nonsensical “**tortured phrases**” that fraudsters use in place of standard terms to avoid anti-plagiarism software.“

Tortured phrase	Actual phrase
arbitrary woods	
irregular backwoods	
random timberland	
random wooded area	
machines getting to know	
selection bushes	
education set	

https://thebulletin.org/2022/01/bosom-peril-is-not-breast-cancer-how-weird-computer-generated-phrases-help-researchers-find-scientific-publishing-fraud/
https://pubpeer.com/publications/71D829DEB0F2BEC5A7B27C90588BC0

Ein kleiner Nachtrag: Tortured phrases

„We have been able to spot fraudulent research thanks in large part to one key tell that an article has been artificially manipulated: The nonsensical "**tortured phrases**" that fraudsters use in place of standard terms to avoid anti-plagiarism software.

Tortured phrase	Actual phrase
arbitrary woods	random forest
irregular backwoods	random forest
random timberland	random forest
random wooded area	random forest
machines getting to know	machine learning
selection bushes	decision trees
education set	training set

https://thebulletin.org/2022/01/bosom-peril-is-not-breast-cancer-how-weird-computer-generated-phrases-help-researchers-find-scientific-publishing-fraud/
https://pubpeer.com/publications/71D829DEB0F2BEC5A7B27C90588BC0

Nächste Woche

- Welche Modellklasse liefert für eine konkrete Fragestellung die präzisesten Vorhersagen?

→ **Benchmarking**