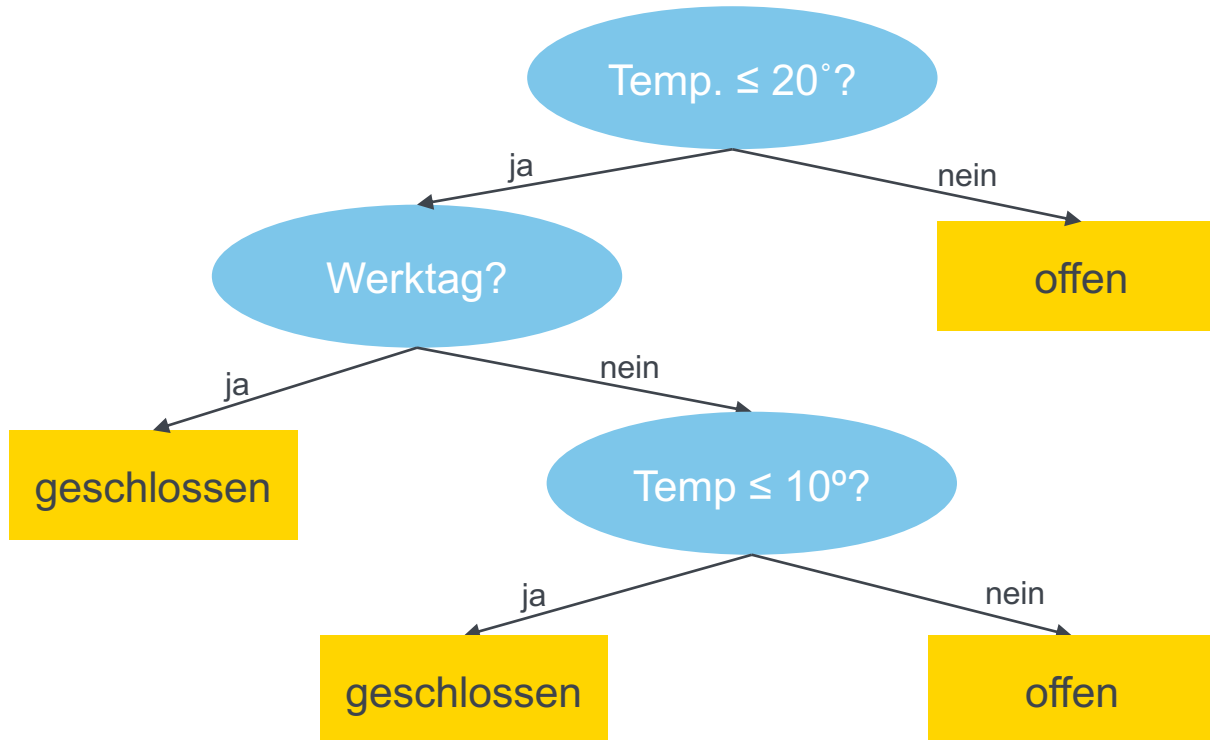


Klassifikationsbäume

Entscheidungsbäume für die Klassifikation

Der Entscheidungsbaum aus der Einführung zu KI-Modellen



Klassifikation

2 Klassen

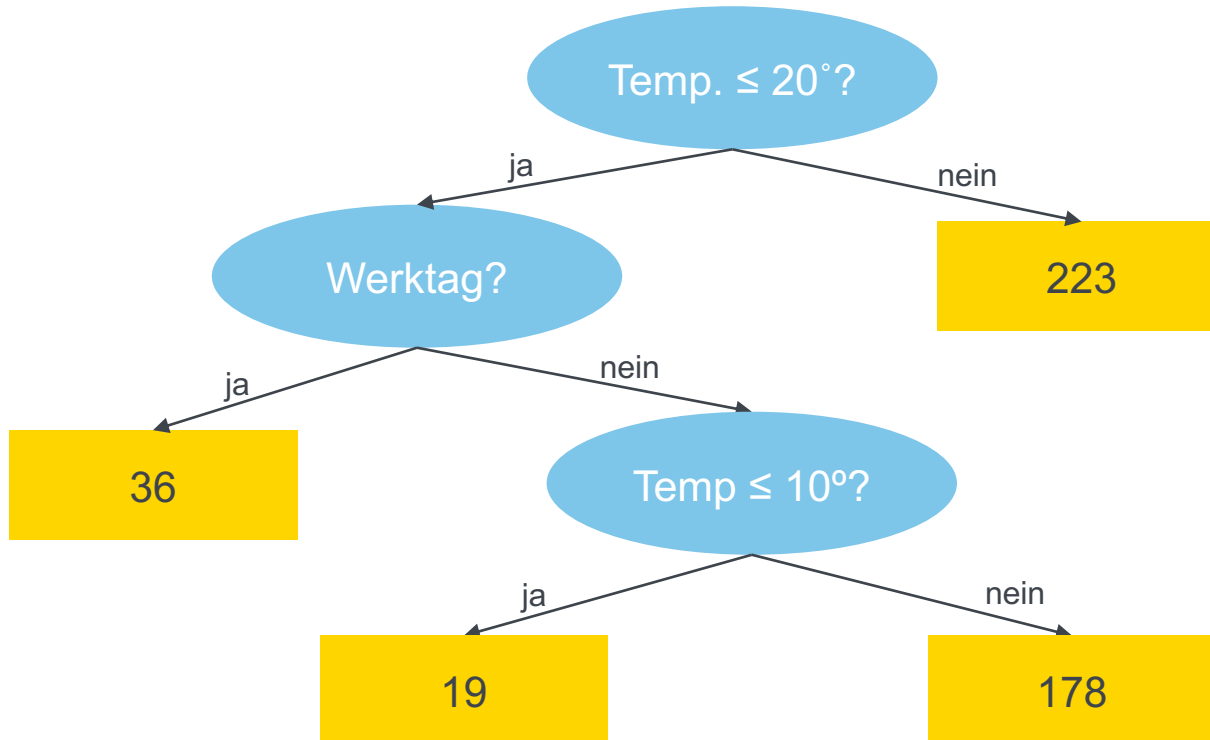
offen
geschlossen

**Klassifikationsbäume sind
Entscheidungsbäume, bei denen
das Ergebnis eine Klasse angibt.**

Rückblick: Wie kamen diese Klassen zustande?

- Idee war: Eisstand öffnen, wenn mindestens 100 verkaufte Portionen zu erwarten sind
- Eigentlich schon fast ein Regressionsproblem 😊
- Könnte ersetzt werden durch 2 Schritt-Lösung:
 - Regressionsmodell
 - Plus eine Regel: offen, wenn Regressionsmodell ≥ 100 Portionen vorhersagt
- Und auch umgekehrt: Klassifikationsbaum umwandelbar in Regressionsbaum
 - Ersetze Klassen durch die jeweiligen mittleren Verkäufe

Ein Regressionsbaum



Regression
Zahlen
(messbare Größe)

Achtung!
Entscheidungsbäume
sind oft keine guten
Regressionsmodelle

**Regressionsbäume sind
Entscheidungsbäume, bei denen
das Ergebnis eine messbare
Größe angibt.**

Ein aus Daten gelernter Klassifikationsbaum

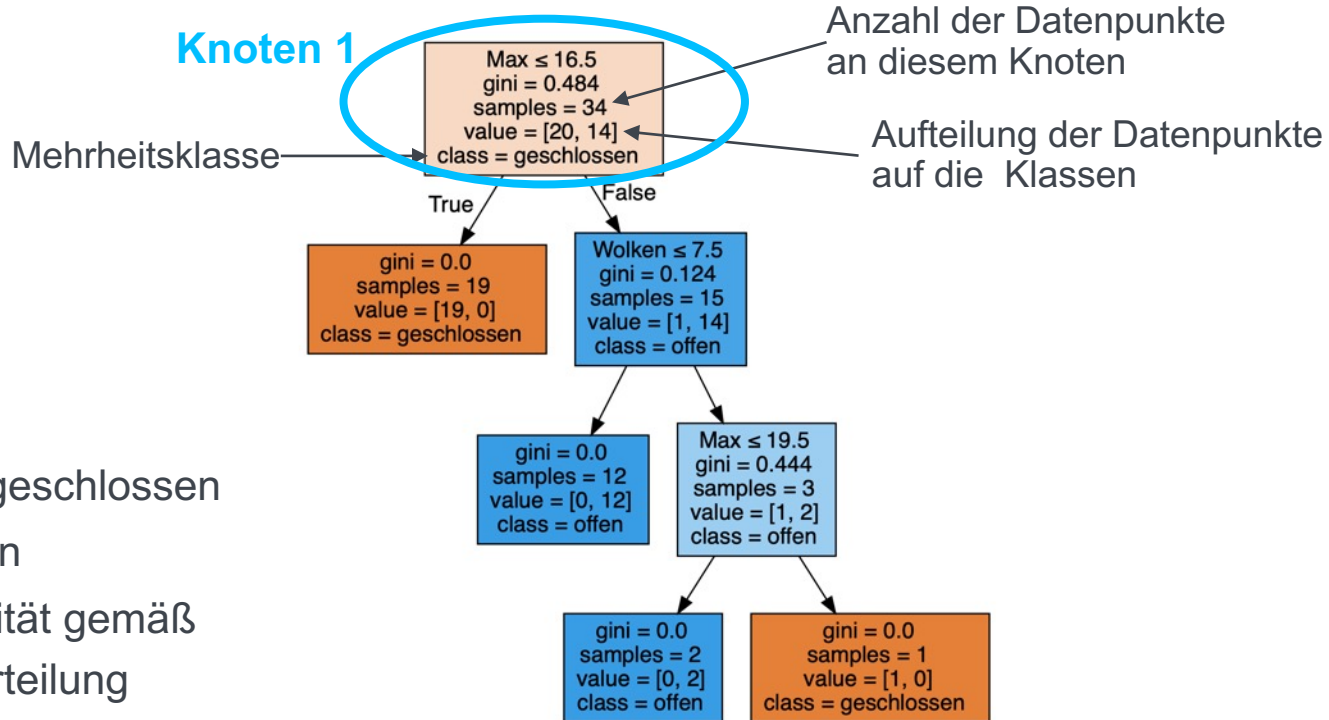
Die Beispieldaten

- Daten aus praktischem Beispiel vom Anfang
- Höchsttemperatur (Max), Bewölkung (Wolken), Wochenende, Verkaufte Portionen
- hier: Wolken messbare Größe (Bedeckungsgrad) zwischen 0 und 8

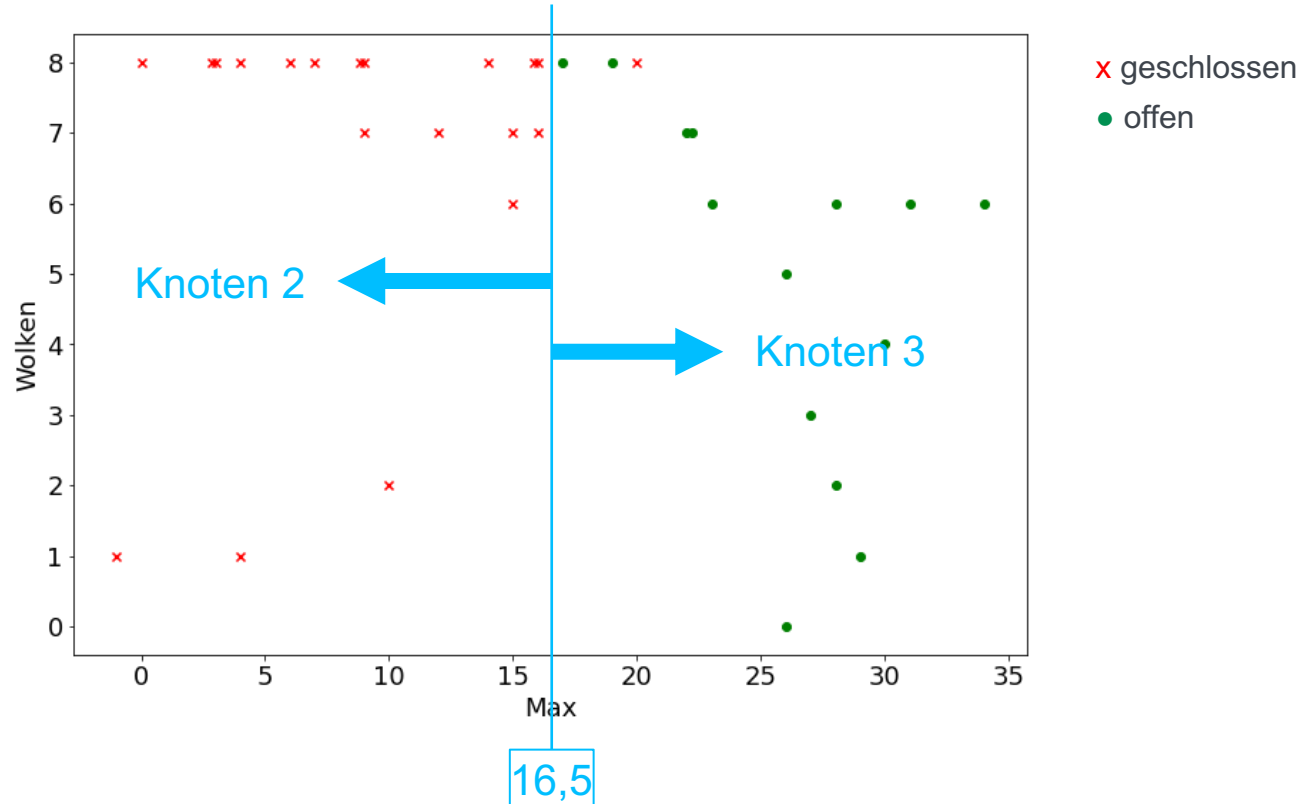
	Datum	Max	Wolken	Wochenende	Portionen	Zustand
9	2018-04-10	20.0	8.0	0	84.0	geschlossen
10	2018-04-21	27.0	3.0	1	270.0	offen
11	2018-05-02	16.0	8.0	0	45.0	geschlossen
12	2018-05-13	17.0	8.0	1	138.0	offen
13	2018-05-24	22.0	7.0	0	100.0	offen
14	2018-06-04	30.0	4.0	0	247.0	offen

Klassen nachträglich zum Lernen ergänzt: das wäre der optimale Zustand gewesen

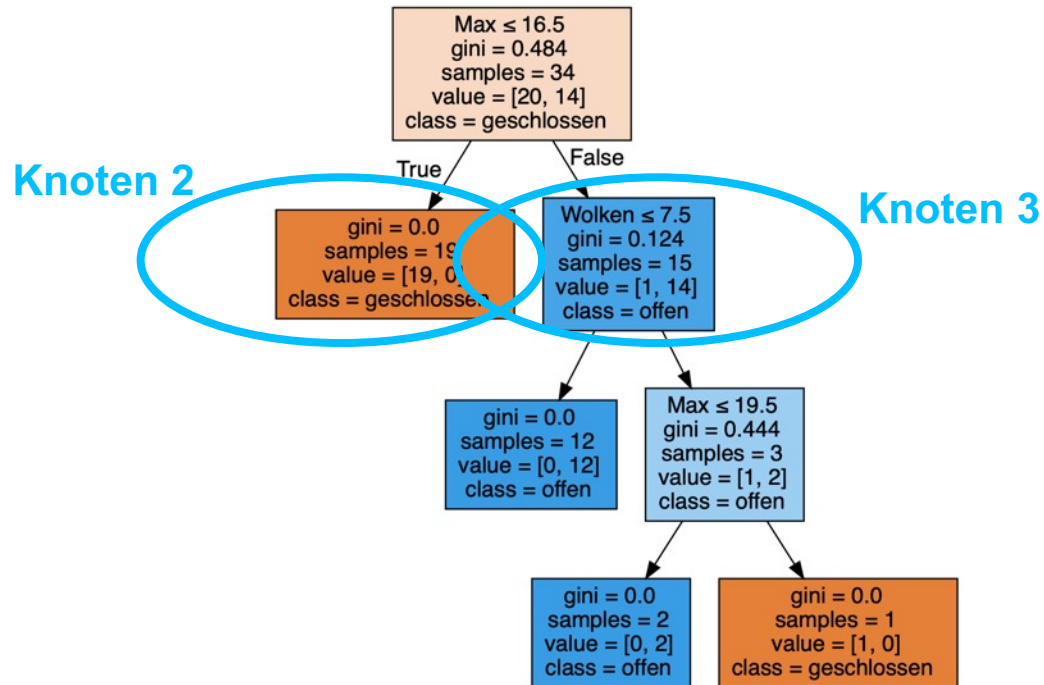
Klassifikation: Vorhersage des optimalen Zustands



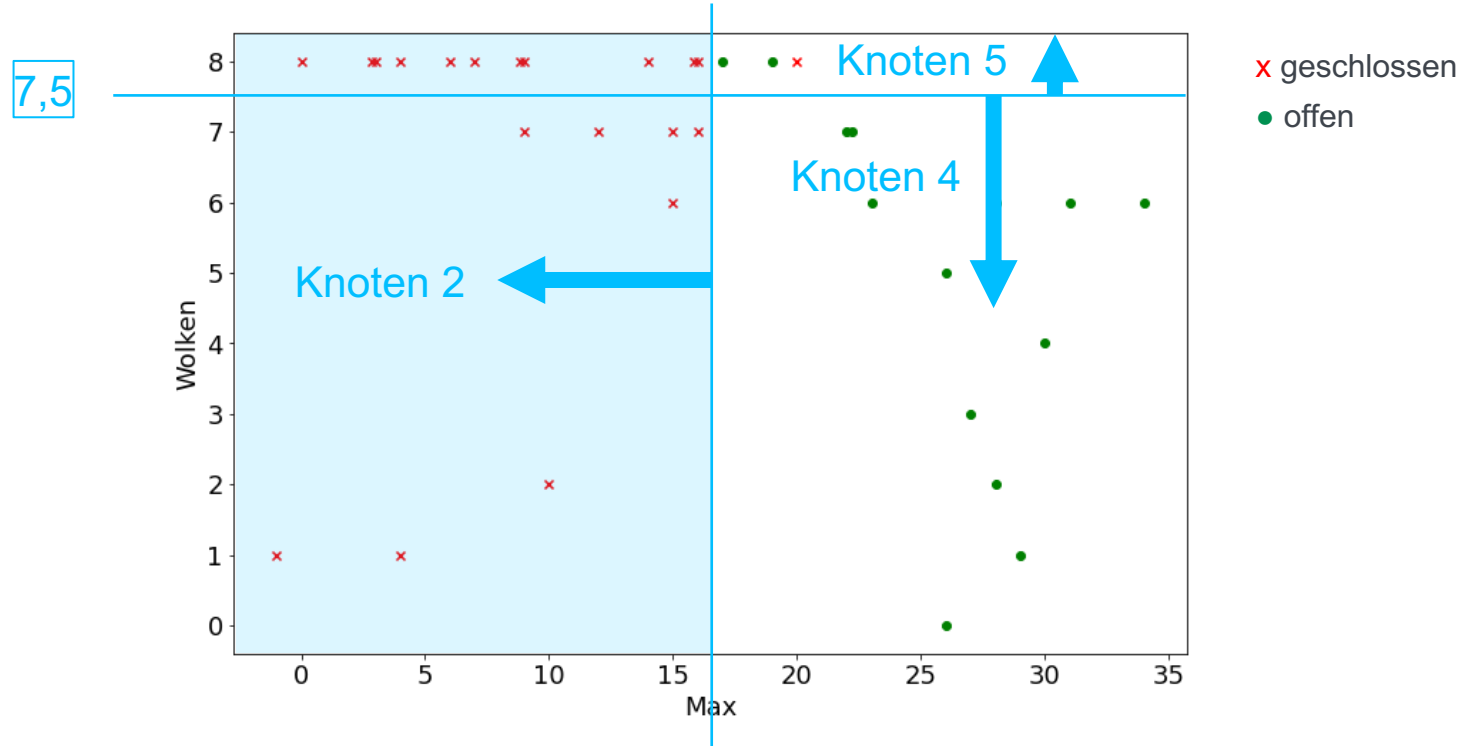
Datenpunkte an Knoten 1



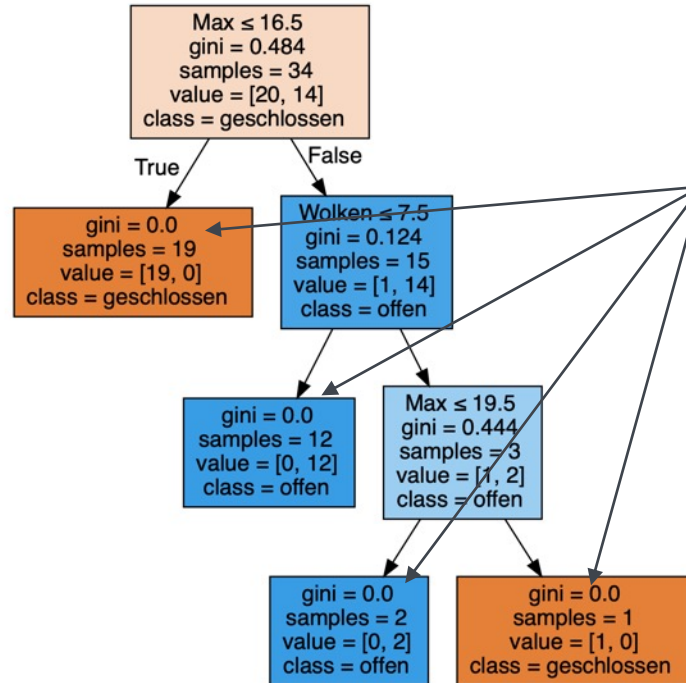
Klassifikation: Vorhersage des optimalen Zustands



Datenpunkte an Knoten 3



Klassifikation: Vorhersage des optimalen Zustands



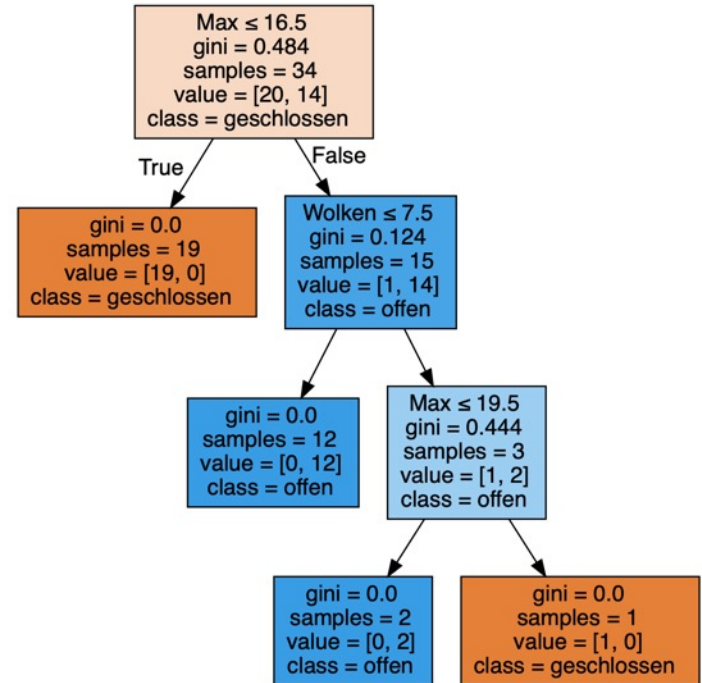
Gini-Index:

wie ungleich sind die Klassen verteilt?

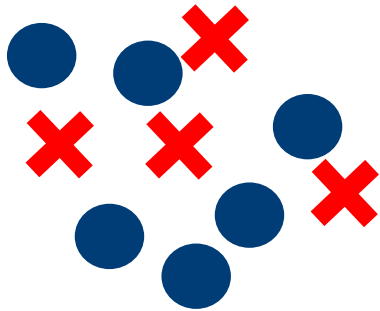
0 = alle Datenpunkte aus derselben Klasse

Der Gini-Index einer Menge

Wie wahrscheinlich ist es, dass ein zufällig gewählter Datenpunkt aus der Menge nicht korrekt klassifiziert wird, wenn er gemäß der Klassenverteilung der Menge zufällig klassifiziert wird?



Gini-Index



Verteilung der Klassen:

x 0.4 • 0.6

Gezogen: Klassif. als: Wahrscheinlichkeit:

x 0.4 x 0.4 $0.4 * 0.4 = 0.16$

x 0.4 • 0.6 $0.4 * 0.6 = 0.24$

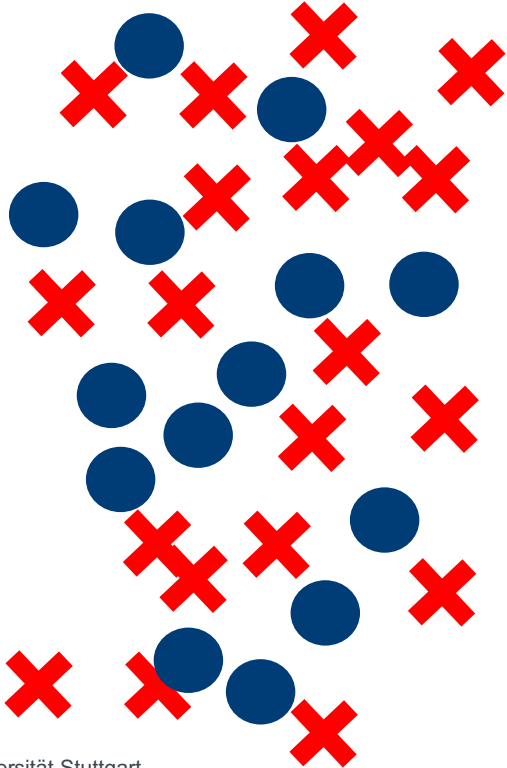
• 0.6 x 0.4 $0.6 * 0.4 = 0.24$

• 0.6 • 0.6 $0.6 * 0.6 = 0.36$

falsch

Gini-Index: $0.24 + 0.24 = 0.48$

Gini-Index



Verteilung der Klassen:

x $20/34 = 0.588$

• $14/34 = 0.412$

Gezogen: Klassif. als: Wahrscheinlichkeit:

x 0.588 **x** 0.588 $0.588 * 0.588 = 0.346$

x 0.588 **•** 0.412 $0.588 * 0.412 = 0.242$

• 0.412 **x** 0.588 $0.412 * 0.588 = 0.242$

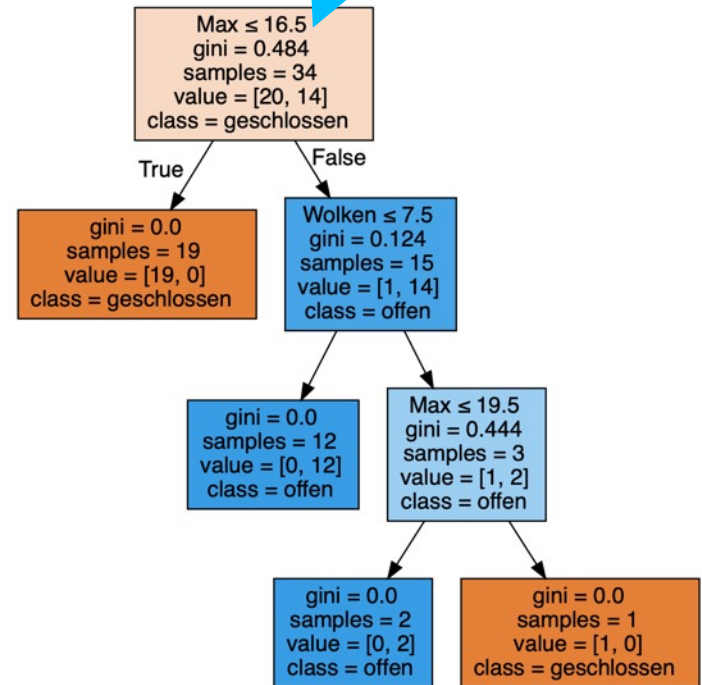
• 0.412 **•** 0.412 $0.412 * 0.412 = 0.170$

falsch

Gini-Index: $0.242 + 0.242 = 0.484$

Der Gini-Index einer Menge

Wie wahrscheinlich ist es, dass ein zufällig gewählter Datenpunkt aus der Menge nicht korrekt klassifiziert wird, wenn er gemäß der Klassenverteilung der Menge zufällig klassifiziert wird?



Der Gini-Index eines Knotens ist null, wenn alle Datenpunkte an diesem Knoten zur selben Menge gehören.

Er ist nur wenig größer als null, wenn fast alle Elemente einer Menge zur selben Klasse gehören.

Ein klassisches Klassifikationsproblem

Beispiel: die “Irisdaten”

- Bekannte gemeinfreie Datenbank mit Daten zur Klassifizierung von Schwertlilien: die “Irisdaten”
- Erhältlich z.B. hier <https://www.kaggle.com/datasets/arshid/iris-flower-dataset>
- Blattlängen und –breiten verschiedener Klassen (Arten) von Schwertlilien:



Iris setosa

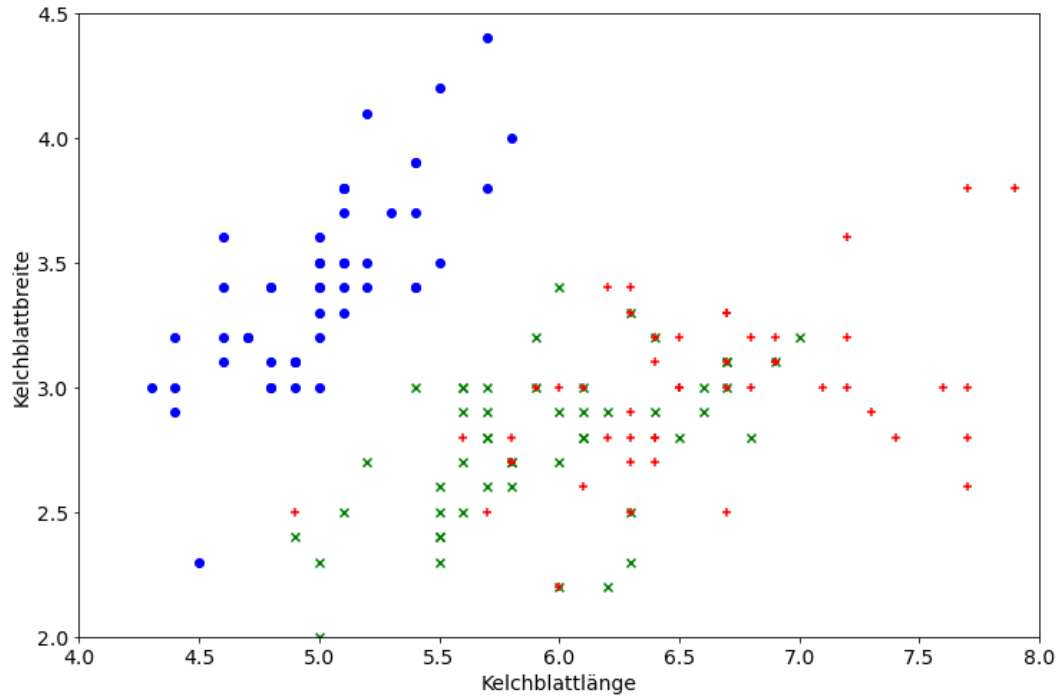


Iris versicolor



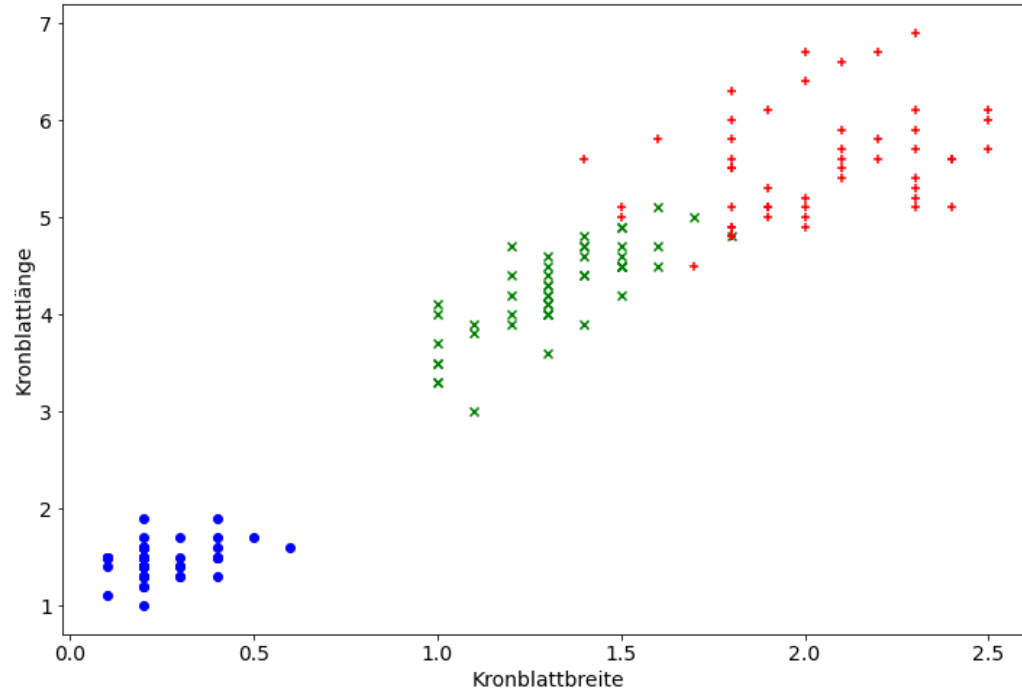
Iris virginica

Kelchblattmaße



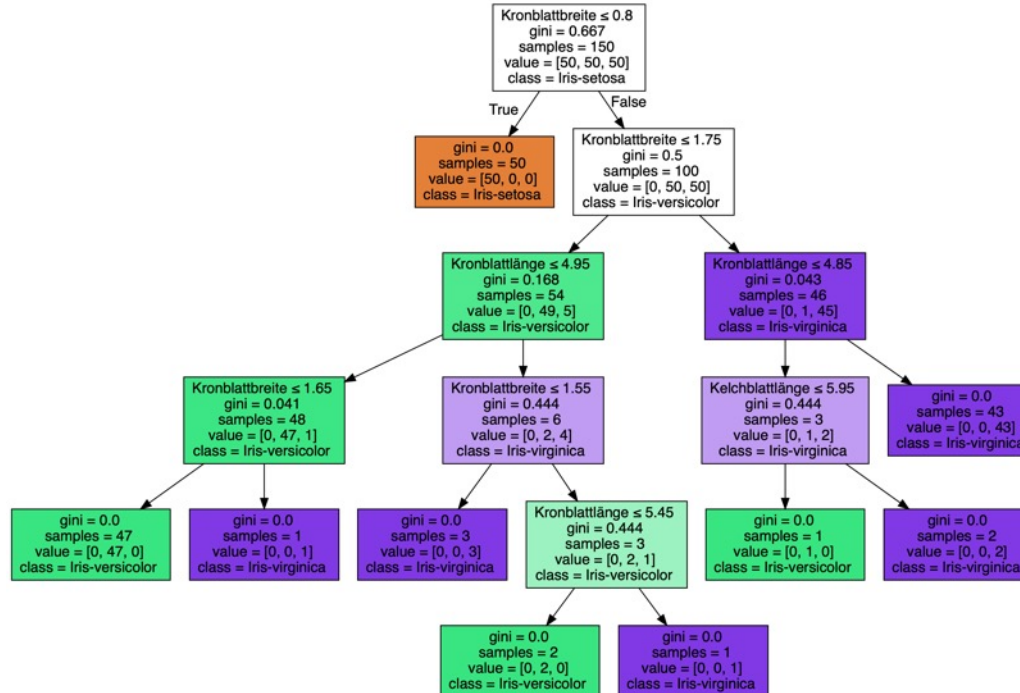
- Setosa
- x Versicolor
- + Virginica

Kronblattmaße

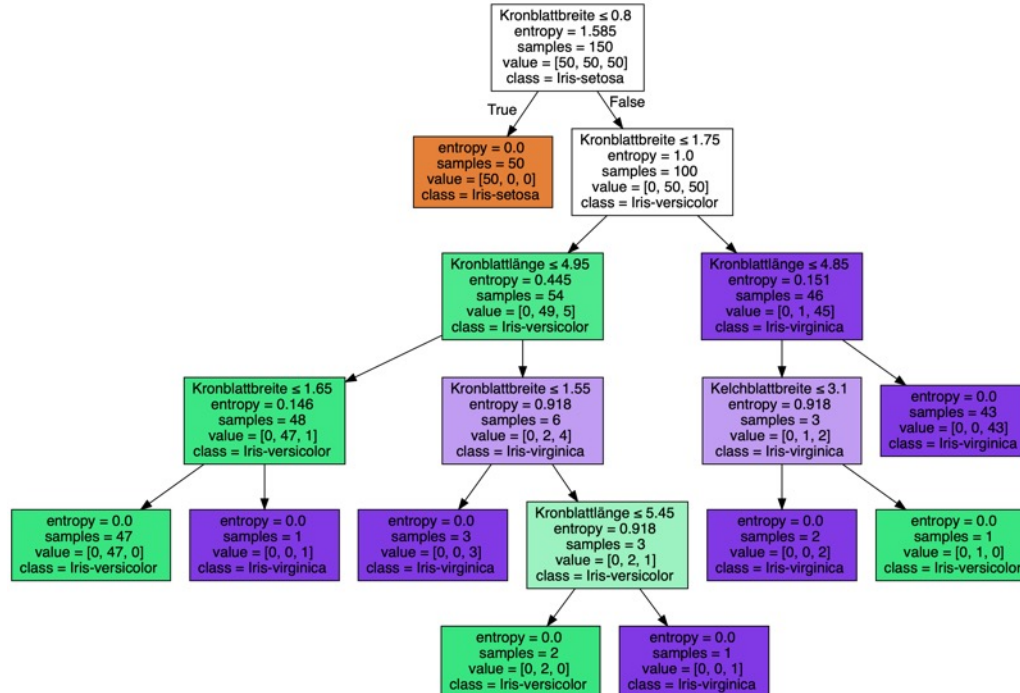


- Setosa
- x Versicolor
- + Virginica

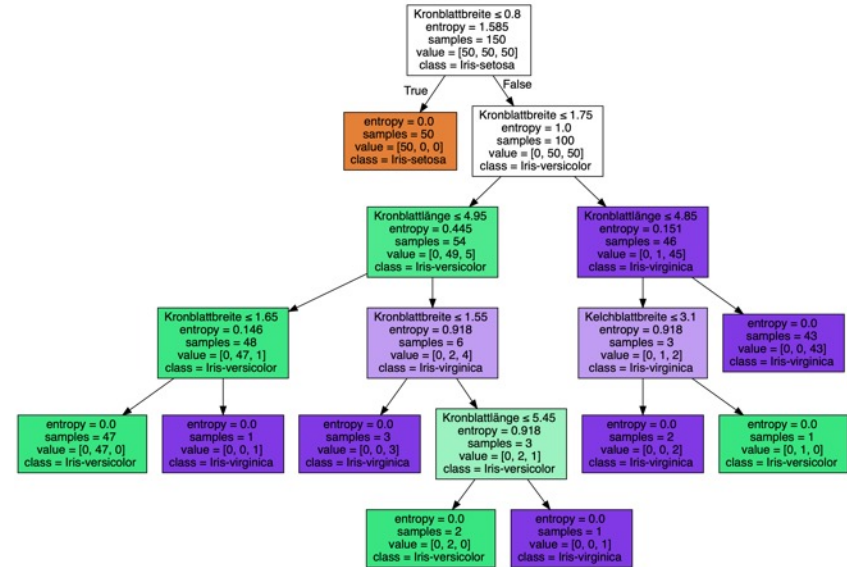
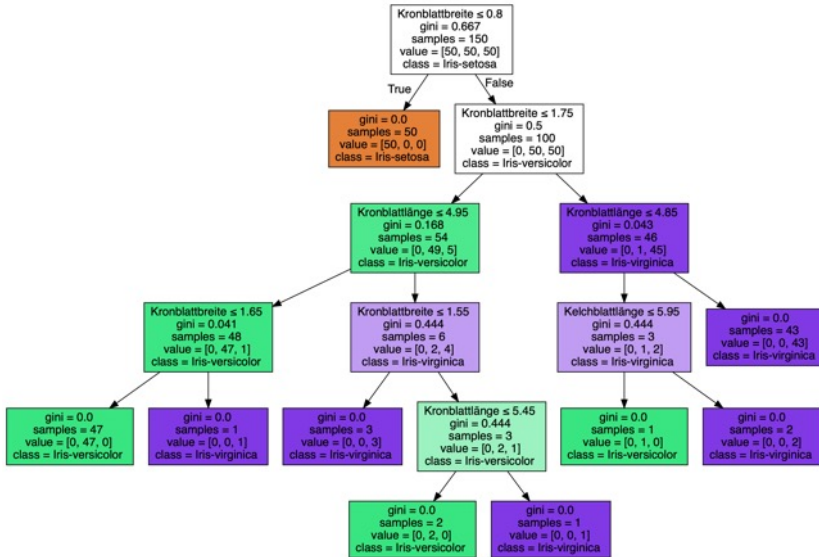
Ein Klassifikationsbaum für Irisarten



Ein Klassifikationsbaum für Irisarten – mit Entropie statt Gini-Index



Die beiden Bäume im Vergleich



Es gibt verschiedene Maße, mit denen man bewerten kann, wie divers die Klassen an den Knoten von Entscheidungsbäumen sind.

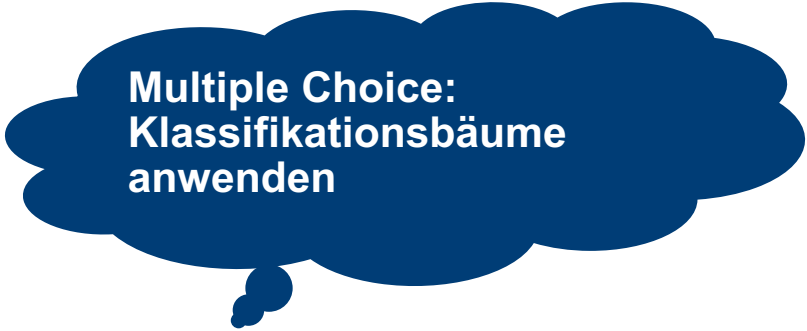
Der Gini-Index und die Entropie sind zwei Beispiele für solche Maße.



**Question Set:
Klassifikationsbäume**



**Multiple Choice:
Klassifikationsbäume
interpretieren**



**Multiple Choice:
Klassifikationsbäume
anwenden**

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung



Universität Stuttgart

Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering



IHK Industrie- und Handelskammer
Reutlingen

Reutlingen | Tübingen | Zollernalb



IHK Region Stuttgart



IHK Industrie- und Handelskammer
Karlsruhe



Quellenverzeichnis

Titelfoto: *Iris virginica*, von C T Johansson, lizenziert unter CC BY 3.0

<<https://creativecommons.org/licenses/by/3.0>>, via Wikimedia Commons, Bildausschnitt verändert

Seite 20, *Iris versicolor*, von terri bateman, gemeinfrei, via Wikimedia Commons, Bildausschnitt verändert

Seite 20, *Iris setosa*, von Денис Анисимов, gemeinfrei, via Wikimedia Commons, Bildausschnitt verändert

Seite 20, *Iris virginica*, von ksandsman, lizenziert unter CC BY 4.0

<<https://creativecommons.org/licenses/by/4.0>>, Bildausschnitt verändert

Lizenzhinweise

„Klassifikationsbäume“ von Antje Schweitzer, KI B³ / Uni Stuttgart

Das Werk - mit Ausnahme der folgenden Elemente:

- Logos der Verbundpartner und des Förderprogramms
- im Quellenverzeichnis aufgeführte Medien

ist lizenziert unter:

 [CC BY 4.0 \(https://creativecommons.org/licenses/by/4.0/deed.de\)](https://creativecommons.org/licenses/by/4.0/deed.de)

(Namensnennung 4.0 International)

Bitte beachten Sie auch die Lizenzangaben im Quellenverzeichnis.