Orange Data Mining für ABU-KIT

Fortbildung im Rahmen der Lehrplan-Multiplikation ABU-KIT

H. G. Stockmeier

Berufliche Schulen Landshut-Schönbrunn

24.3.2026

https://abukit.de

Übersicht

Übersicht

Blick in den Lehrplan

KIT12 Lernbereich 1: Künstliche Intelligenz und digitale Anwendungen (ca. 10 Std. h)

Kompetenzerwartungen

Die Schülerinnen und Schüler ...

- diskutieren Ansätze zur Definition des Begriffs Künstliche Intelligenz (KI), beschreiben verschiedene Verfahren der KI (u.a. maschinelles Lernen) sowie ihre Anwendungsgebiete.
- erheben mit Hilfe von digitalen Werkzeugen Daten, verarbeiten diese, werten sie aus und stellen sie sachgerecht dar.
- analysieren den Einfluss von Trainingsdaten und Parametern auf die Zuverlässigkeit der Ergebnisse des gewählten Verfahrens maschinellen Lernens, ggf. unter Verwendung eines geeigneten Werkzeugs.
- nehmen zu ausgewählten aktuellen Einsatzmöglichkeiten der Künstlichen Intelligenz Stellung, beurteilen (mithilfe fachlicher Kriterien) und bewerten (unter Berücksichtigung gesellschaftlicher Werte und Normen) Chancen und Risiken für Individuum und Gesellschaft.

Blick in den Lehrplan

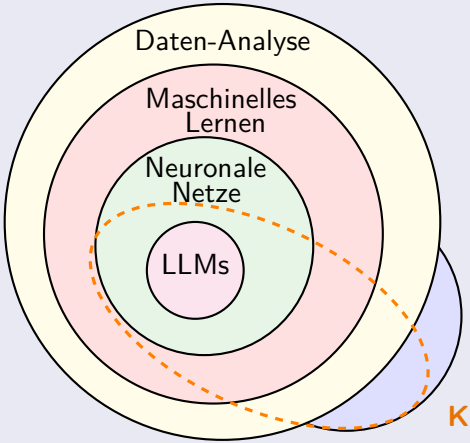
Berücksichtigung gesellschaftlicher Werte und Normen) Chancen und Risiken für Individuum und Gesellschaft.

Inhalte

- starke und schwache KI, Teilbereiche maschinellen Lernens (überwacht, unüberwacht, bestärkend)
- Datenverarbeitung, Datenauswertung, Datendarstellung, Tabellenkalkulation
- Menge, Auswahl und Qualität der Testdaten, Aufteilung in Trainings- und Testdaten, Über- bzw. Unteranpassung, Darstellungsmöglichkeiten
- Chancen, Risiken und Herausforderungen in Bezug auf technische (z. B. Zuverlässigkeit, Prognosen, Entscheidungsfindung) und gesellschaftliche (z. B. Personalisierung, Transparenz, Fairness, Gesetzgebung, Urheberrecht) Aspekte

Was ist KI?

Versuch einer Begriffsklärung ...



Daten-basiertes Analysieren

Iterative (konvergierende) Algorithmen

Biologie-inspirierte kleine Recheneinheiten

Große Sprachmodelle = Chat-bots

Regel-basiertes Analysieren

Fazit: Daten!

Was sind Daten?

- Antwort: Strukturierte Ansammlung von Informationen

Beispiel: Steine und Kartoffeln

Nr.	Volumen (cm^3)	Masse (g)	Kategorie	Nr.	Volumen (cm^3)	Masse (g)	Kategorie
1	710	570	Kartoffeln	13	711	740	Steine
2	576	350	Kartoffeln	14	575	850	Steine
3	539	350	Kartoffeln	15	536	700	Steine
4	392	310	Kartoffeln	16	390	370	Steine
5	359	250	Kartoffeln	17	357	600	Steine
6	258	180	Kartoffeln	18	263	300	Steine
7	226	150	Kartoffeln	19	227	448	Steine
8	181	120	Kartoffeln	20	180	270	Steine
9	175	145	Kartoffeln	21	177	290	Steine
10	111	80	Kartoffeln	22	112	160	Steine
11	95	57	Kartoffeln	23	95	140	Steine
12	68	45	Kartoffeln	24	67	90	Steine

Grundbegriffe I

Instanz

- Individuum, für das Messwerte (Merkmale) vorhanden sind
- Beispiel: Kartoffel Nr. 5
- Synonyme: Objekt, Zeile (engl.: row)

Merkmal

- Interessierende Eigenschaft der Objekte
- Beispiel: Volumen, Masse, Kategorie
- Synonyme: Attribut (engl.: feature), Variable, Spalte (engl.: column)
- **Zwei Sorten**:
  - **numerisch**: im Beispiel: Volumen, Masse
  - **nicht-numerisch**: im Beispiel: Kategorie
    (synonym: kategorial, nominal)

Grundbegriffe II

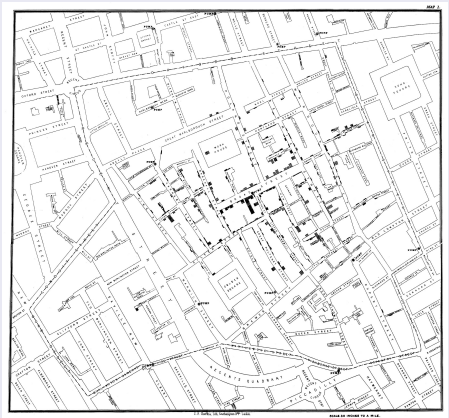
Ausprägung

- Wert eines Merkmals eines Objekts
- Synonym: Merkmalswert (engl.: value)
- Beispiel: Die Masse von Objekt Nr. 5 ist 250 g
- Beispiel: Die Kategorie von Objekt Nr. 5 ist “Kartoffel”

Übersicht

Daten-Visualisierung

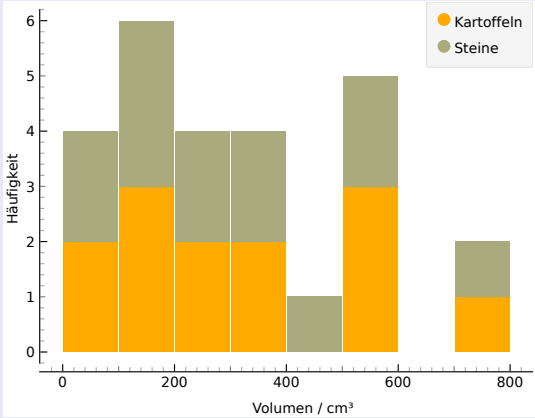
Cholera-Ausbruch London 1853

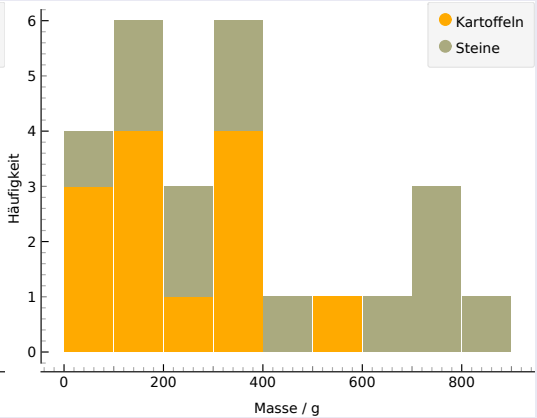


[1] John Snow (1855). *On the mode of communication of cholera*. second edition, much enlarged. London: John Churchill. 216 S.

Daten-Visualisierung

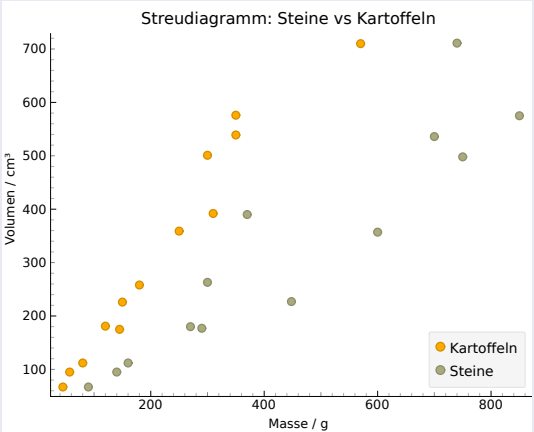
Häufigkeitsdiagramm





Daten-Visualisierung

Streudiagramm (Scatterplot)

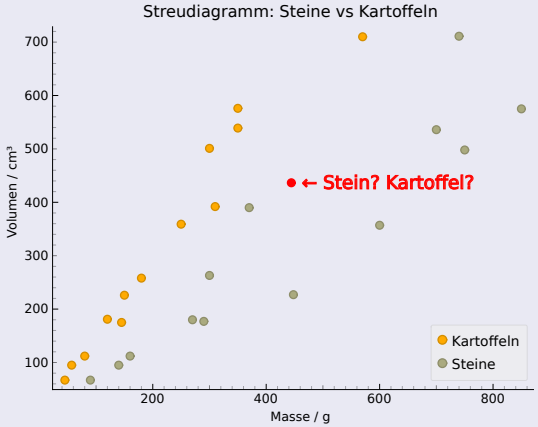


Übersicht

Vier Hauptaufgaben der Daten-Analyse

Klassifikation

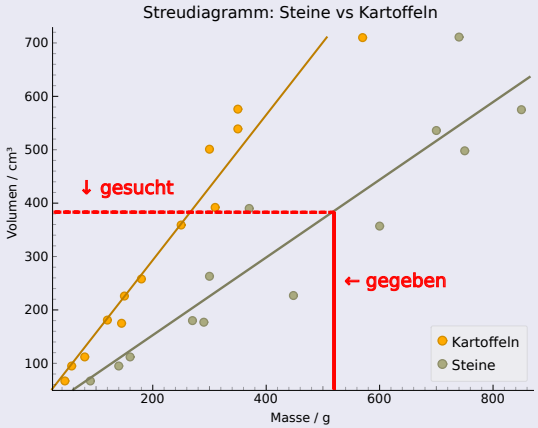
- Zuordnung eines Objekts zu einer **Klasse** (synonym: **Target**, **Zielmerkmal**)



Vier Hauptaufgaben der Daten-Analyse

Regression

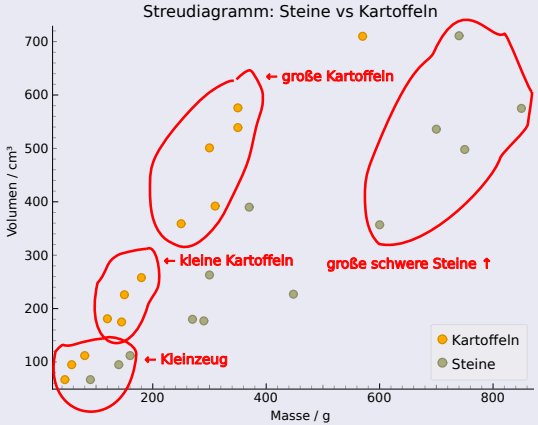
- Voraussage von Zahlenwerten ($\rightarrow$ "Regressions-Gerade")



Vier Hauptaufgaben der Daten-Analyse

Cluster-Analyse

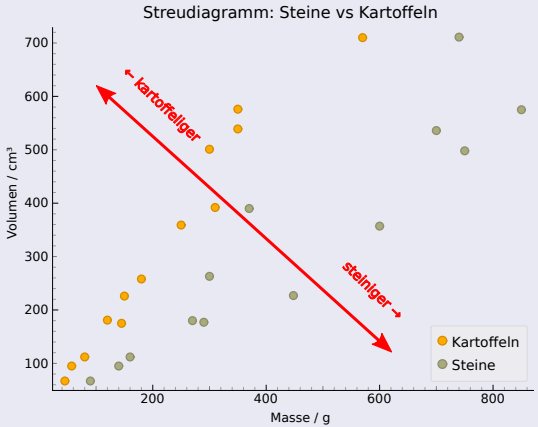
- Entdecken von Zusammengehörigkeiten unter den Objekten



Vier Hauptaufgaben der Daten-Analyse

Merkmals-Extraktion / Dimensions-Reduktion

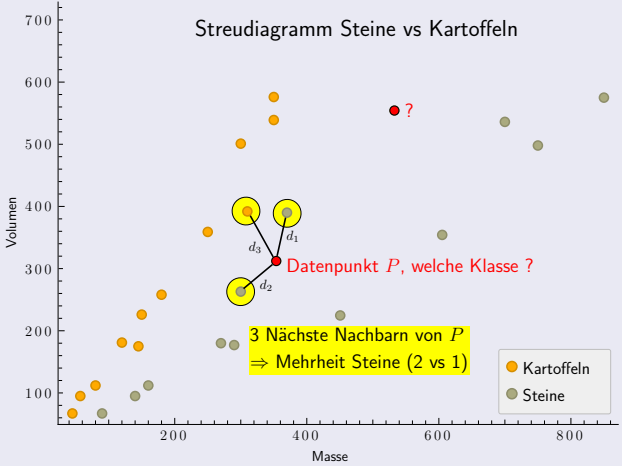
- Entdecken von neuen Merkmalen / Weglassen von unwichtigen Merkmalen



Übersicht

Der k-Nächste-Nachbarn-Algorithmus

Klassifikations-Problem

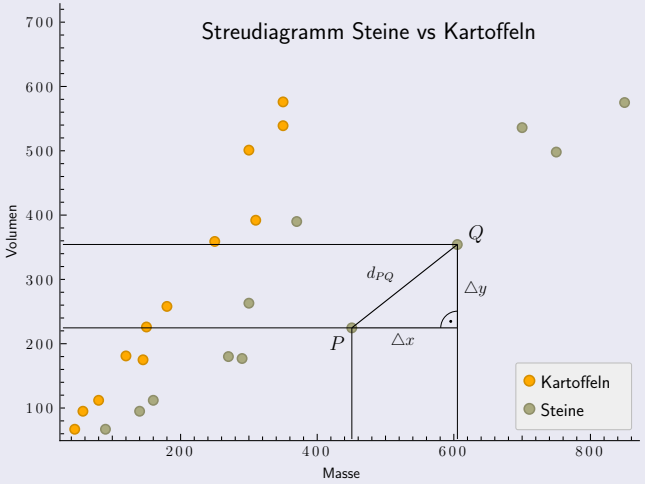


Algorithmus

1. Wähle Zahl k der nächsten Nachbarn (z.B. $k = 3$)
2. Miss die Abstände von P zu allen anderen Punkten
3. k Punkte mit kleinstem Abstand = nächste Nachbarn von P
4. **Mehrheit** der nächsten Nachbarn ⇒ Klasse von P

Der k-Nächste-Nachbarn-Algorithmus

Abstandsberechnung im Scatterplot („Merkmalsraum“)



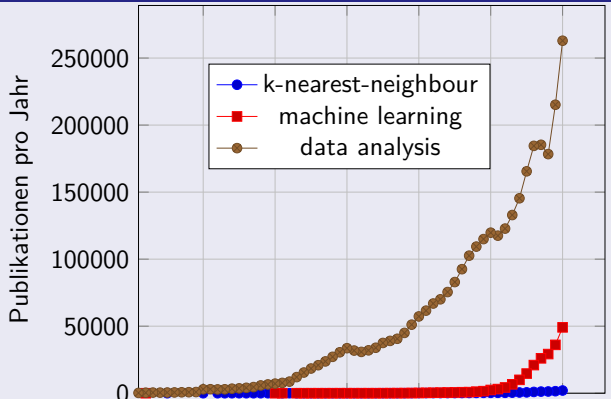
Abstand d_{PQ} zwischen 2 Punkten P und Q:

$$d_{PQ} = \sqrt{(\triangle x)^2 + (\triangle y)^2}$$

(Pythagoras)

Historische / Bedeutungs-Einordnung

Publikationen auf pubmed.gov mit dem Schlüsselwort „k-nearest-neighbour“u.ä.



(abgerufen: 22.2.2026)

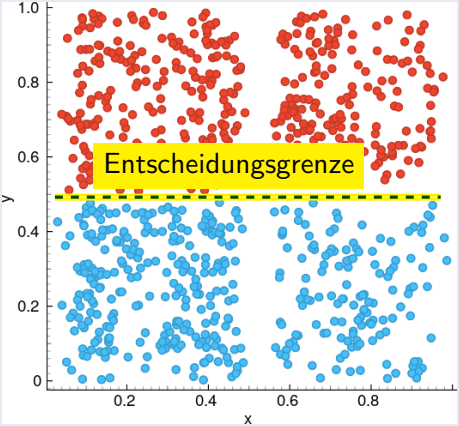
Übersicht

Entscheidungsbaum: Grund-Idee

1. Man trennt die Datensätze parallel zu den Merkmals-“Achsen“.
2. Die Trennung dokumentiert man in einem Baumdiagramm.

Entscheidungsbaum: Grund-Idee

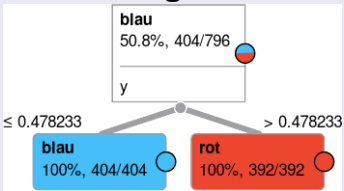
Leicht trennbares Beispiel



Erfundene Objekte:

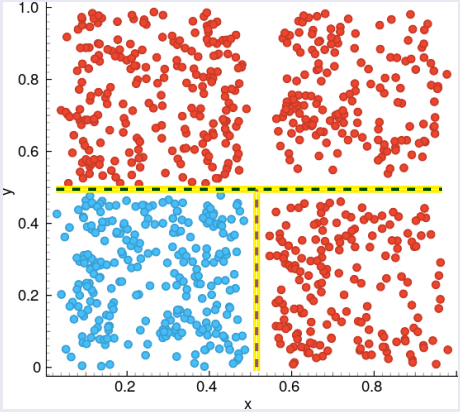
- 2 Klassen: „rot“ und „blau“
- 2 Merkmale: x und y

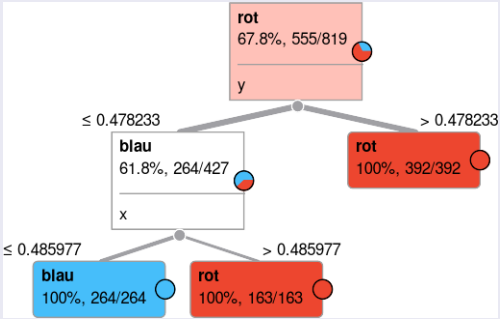
Entscheidungsbaum:



Entscheidungsbaum: Grund-Idee

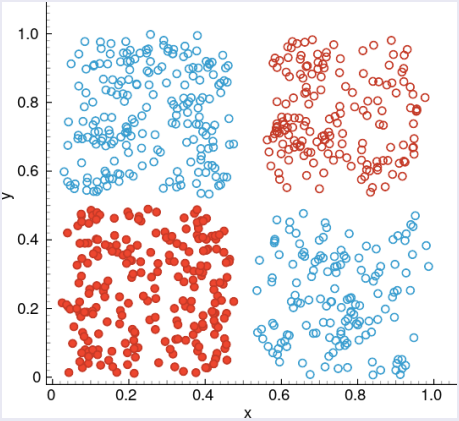
Nicht ganz so leicht trennbares Beispiel

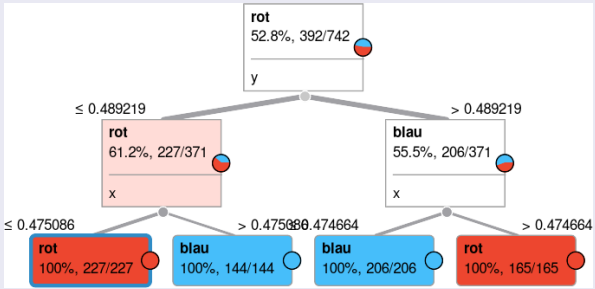




Entscheidungsbaum: Grund-Idee

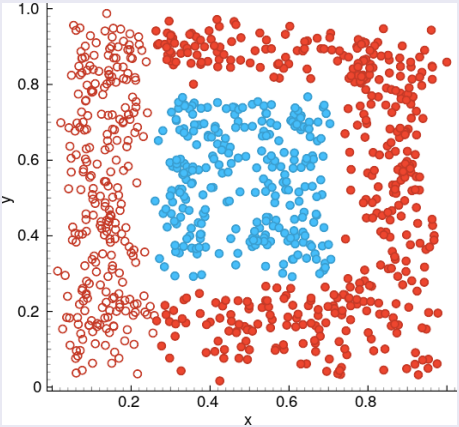
Weiteres Beispiel

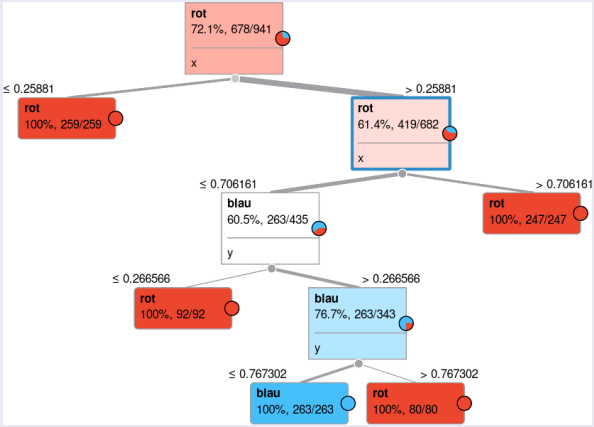




Entscheidungsbaum: Grund-Idee

Komplizierteres Beispiel



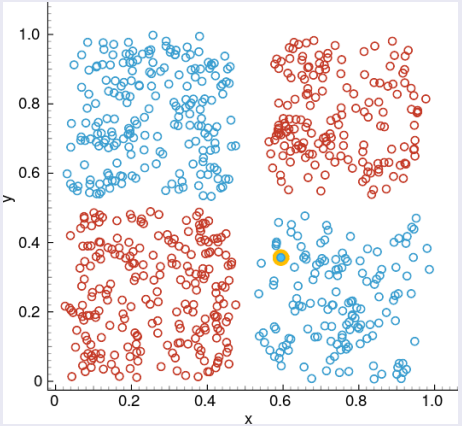


Entscheidungsbaum: Grund-Idee

- Das war das **Erstellen** des Entscheidungsbaums
- Wie sieht das **Klassifizieren** aus?

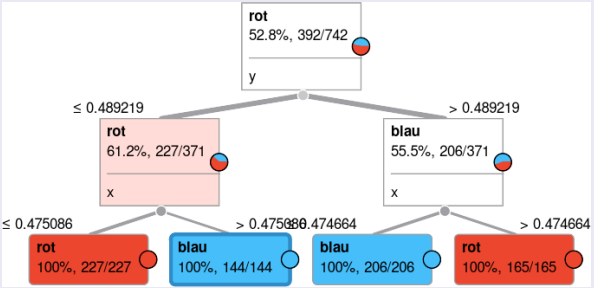
Entscheidungsbaum: Grund-Idee

Wie wird klassifiziert?



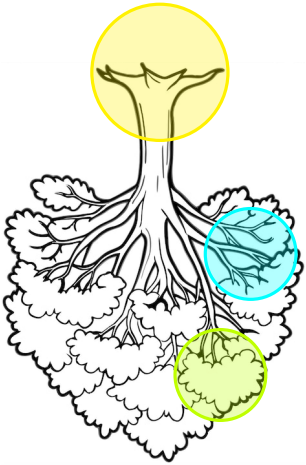
Gegeben:

Objekt mit $x = 0{,}594152$ und $y = 0{,}356703$



⇒ Objekt ist Klasse „**blau**“ !

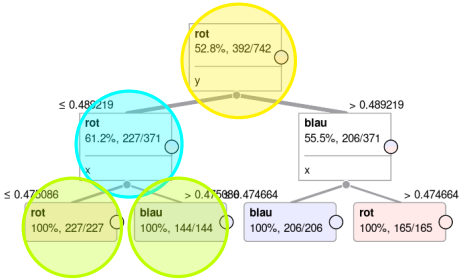
Einschub: Nomenklatur beim Entscheidungsbaum-Diagramm



Wurzel

Knoten

Blätter



Entscheidungsbaum: Grund-Idee

- Offene Fragen
  - Wie findet man die Entscheidungsgrenzen?
  - Wie wählt man die Merkmale aus?

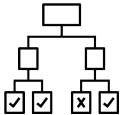
Übersicht

Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

UNIVERSITÄT PADERBORN ProDaBi

Datenbasierte Entscheidungsregeln

Wie lernt eine künstliche Intelligenz?



https://www.prodabi.de/materialien/entscheidungsbaeume

Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

- **So sehen die Spielkarten aus:**

Apfel

Nährwerte pro 100g	
Energie	52 kcal
Fett	0,2 g
davon gesättigte Fettsäuren	0,0 g
Kohlenhydrate	13,8 g
davon Zucker	11,0 g
Eiweiß	0,3 g
Salz	0,0 g

ProDaBi

Avocado

Nährwerte pro 100g	
Energie	160 kcal
Fett	13,0 g
davon gesättigte Fettsäuren	2,8 g
Kohlenhydrate	2,0 g
davon Zucker	0,7 g
Eiweiß	1,5 g
Salz	0,1 g

ProDaBi

Frikadellen

Nährwerte pro 100g	
Energie	294 kcal
Fett	22,0 g
davon gesättigte Fettsäuren	2,8 g
Kohlenhydrate	10,5 g
davon Zucker	2,0 g
Eiweiß	13,5 g
Salz	1,5 g

ProDaBi

Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

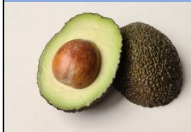
- Es gibt zwei Sorten von Karten:
  - Blauer Hintergrund: Erstellen des Baums
  - Gelber Hintergrund: Mit dem erstellten Baum klassifizieren

Apfel

Nährwerte pro 100g	
Energie	52 kcal
Fett	0,2 g
davon gesättigte Fettsäuren	0,0 g
Kohlenhydrate	13,8 g
davon Zucker	11,0 g
Eiweiß	0,3 g
Salz	0,0 g

ProDaBi

Avocado



Nährwerte pro 100g	
Energie	160 kcal
Fett	13,0 g
davon gesättigte Fettsäuren	2,8 g
Kohlenhydrate	2,0 g
davon Zucker	0,7 g
Eiweiß	1,5 g
Salz	0,1 g

ProDaBi

Frikadellen

Nährwerte pro 100g	
Energie	294 kcal
Fett	22,0 g
davon gesättigte Fettsäuren	2,8 g
Kohlenhydrate	10,5 g
davon Zucker	2,0 g
Eiweiß	13,5 g
Salz	1,5 g

ProDaBi

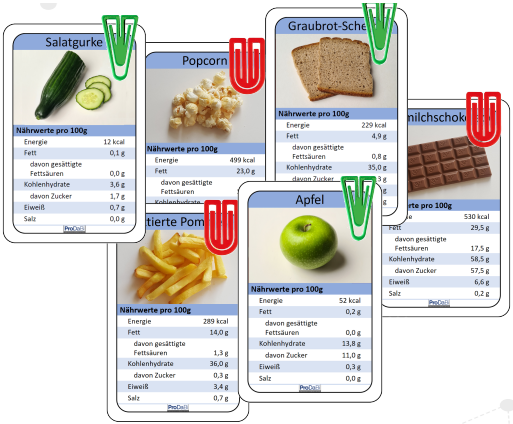
Zucchini

Nährwerte pro 100g	
Energie	19 kcal
Fett	0,4 g
davon gesättigte Fettsäuren	0,1 g
Kohlenhydrate	2,2 g
davon Zucker	1,6 g
Eiweiß	1,6 g
Salz	0,0 g

ProDaBi

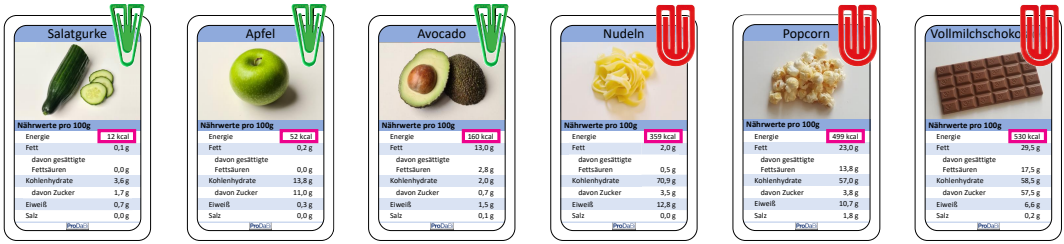
Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

- **Schritt 1:** SuS vergeben Klassen-„Klammern“ für die blauen Karten – grün = „gesund“, rot = „ungesund“



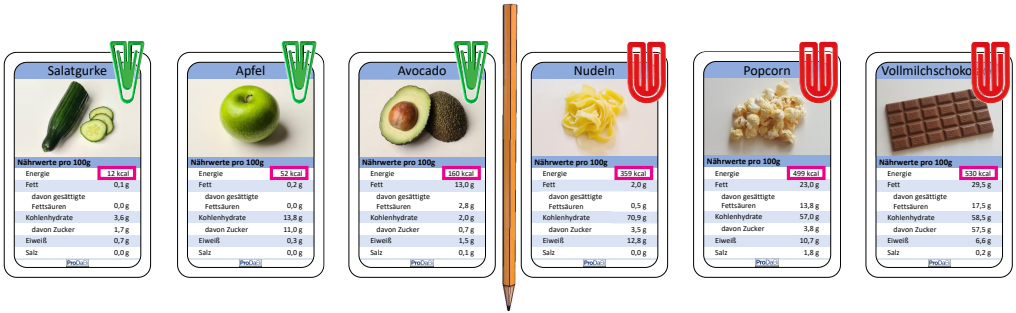
Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

- **Schritt 2:** Sortieren der Karten aufsteigend nach einem Merkmal, z. B. „Energie“



Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

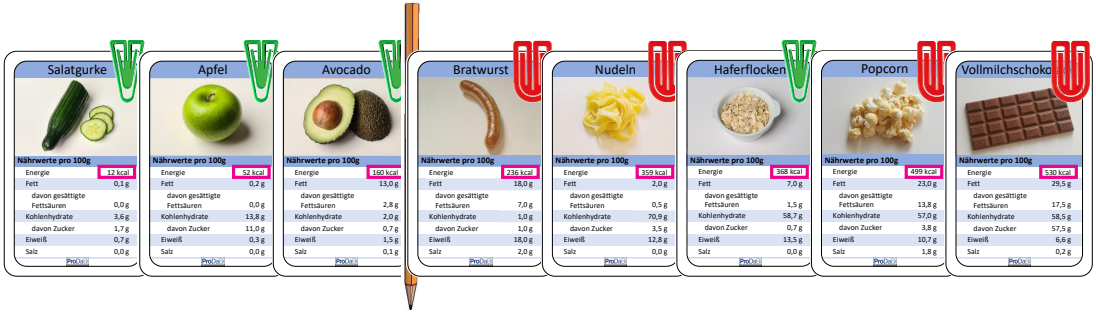
- **Schritt 3:** Suchen nach einer optimalen Entscheidungsgrenze mit der „Bleistift-Methode“(grün = „gesund“ rot = „ungesund“)



Fehler: 0 = Optimal! $\Rightarrow$ Entscheidungsgrenze = 260 kcal

Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

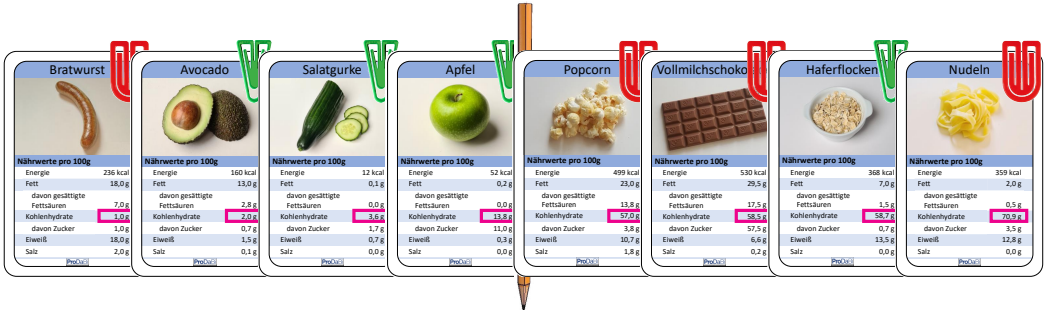
- Wo ist die optimale Entscheidungsgrenze in dieser Situation ?



Fehler: $1 \Rightarrow$ Entscheidungsgrenze $= 200$ kcal

Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

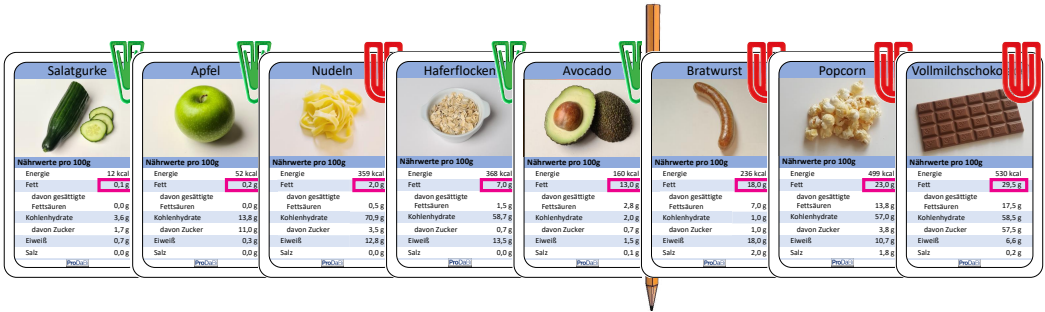
- **Schritt 4:** Weiterer Versuch mit einem anderen Merkmal
  - Sortieren nach z. B. „Kohlenhydrate“
  - Optimale Entscheidungsgrenze mit der „Bleistift-Methode“



Fehler: 2 $\Rightarrow$ Entscheidungsgrenze = 35 g Kohlenhydrate

Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

- **Schritt 4-2:** Weiterer Versuch mit einem anderen Merkmal
  - Sortieren nach z. B. „Fett“
  - Optimale Entscheidungsgrenze mit der „Bleistift-Methode“



Fehler: $1 \Rightarrow$ **Entscheidungsgrenze = 15 g Fett**

Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

- Suche dokumentieren auf Dokumentations-Bogen
  - Festhalten von 3 überprüften Merkmalen und ihren Entscheidungsergebnissen
  - Auswahl der besten Entscheidungsgrenze (wenigste Fehler)

Bisher von uns gefundene Entscheidungsgrenzen:

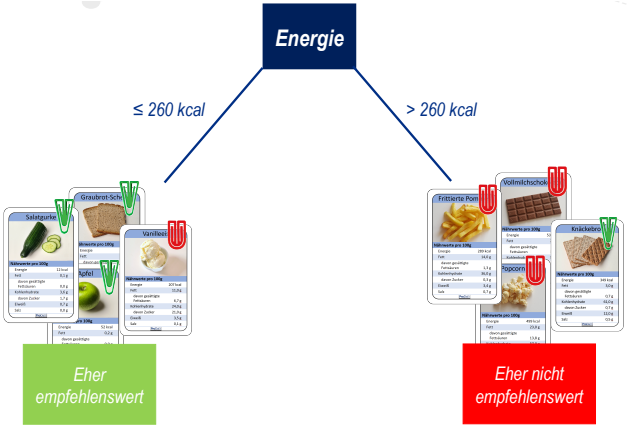
1. Energie: 200 kcal $\rightarrow$ 1 Fehler
2. Kohlenhydrate: 35 g $\rightarrow$ 2 Fehler
3. Fett: 15 g $\rightarrow$ 1 Fehler

$\Rightarrow$ Energie: 200 kcal $\rightarrow$ 1 Fehler

(Fett: 15 g $\rightarrow$ 1 Fehler wäre auch ok)

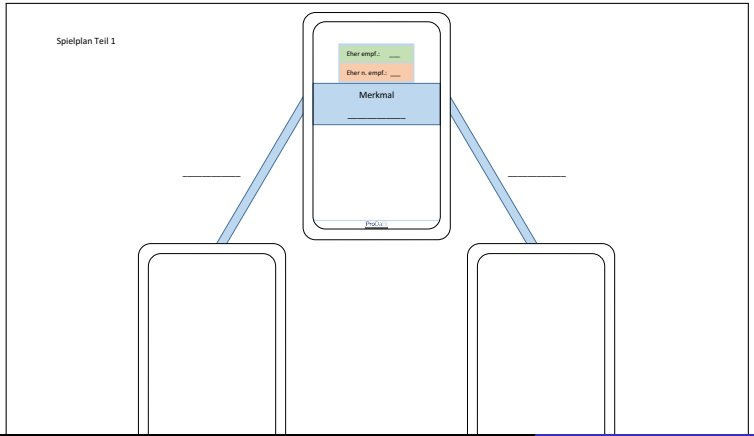
Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

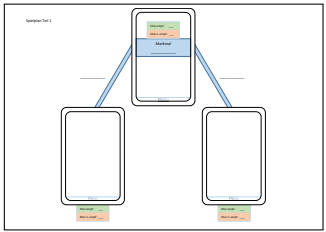
- Aufteilen der Karten in zwei Stapel entlang der besten Entscheidungsgrenze

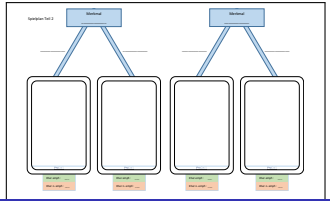


Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

- Für jeden der neuen Stapel neue Entscheidungs-Kriterien finden
- Entstehende Karten-Stapel nochmal aufteilen

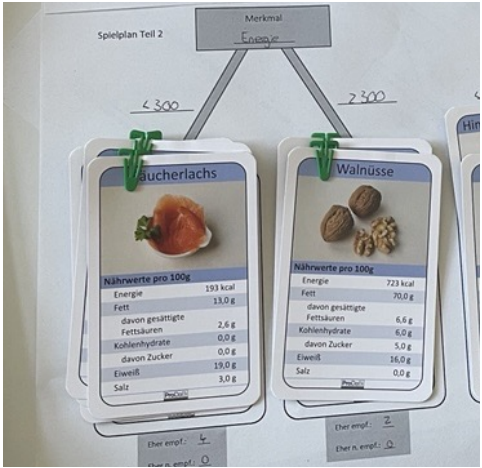






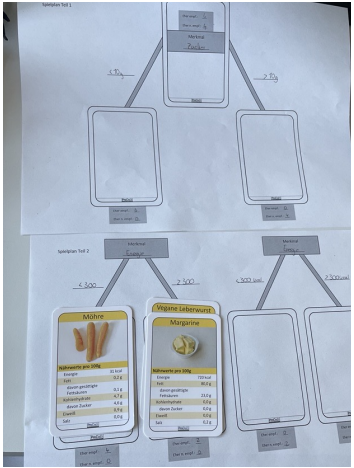
Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

- Komplett ausgefüllter Spielplan = fertiger Entscheidungsbaum



Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

- **Klassifizieren**: von der Lehrkraft werden „gelbe“ Karten ausgeteilt ...



Erstellen eines Entscheidungsbaums mit dem „ProDaBi“-Kartenspiel

Aufgabe Erstellen Sie einen Entscheidungsbaum mit „ProDaBi“-Karten

1. Sortieren der Karten aufsteigend nach einem Merkmal
2. Suche nach einer Entscheidungsgrenze mit möglichst geringem Klassifikationsfehler
3. Wiederholung der Schritte 1 und 2 für drei Merkmale
4. Auswahl des insgesamt günstigsten „Split-Kriteriums“ (Merkmal und Entscheidungsgrenze)
5. Aufteilen der Objekte nach dem „Split-Kriterium“ in 2 Stapel
6. Wiederholen des Ganzen für jeden der beiden Teil-Stapel

„Anleitung“zur Erstellung eines Entscheidungsbaums

Algorithmus (Entscheidungsbaum erstellen = „Training“)

1. Sortiere die Objekte nach einem Merkmal.
2. Suche nach einem Schwellwert, bei dem der Fehler (Falschklassifizierung) möglichst klein ist.
3. Wiederhole 1 und 2 für alle Merkmale.
4. Teile die Daten so auf (Merkmal und Schwellwert), dass der Fehler möglichst klein ist.
5. Wiederhole alles mit den geteilten Daten
   1. bis entweder kein Fehler mehr auftritt.
   2. bis eine maximale Tiefe des Baumes erreicht ist.

Übersicht

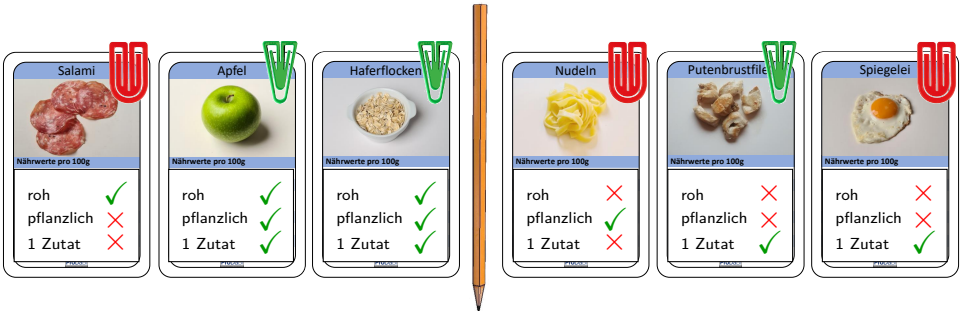
Entscheidungsbaum mit kategorialen Daten

- **Erinnerung**: Kategoriale Daten = nicht-numerisch
- $\Rightarrow$ Hier ProDabi-Karten auf nicht-numerische Merkmale „umgestellt“:

- Sortieren entfällt, da nicht-numerische Merkmale !

Entscheidungsbaum mit kategorialen Daten

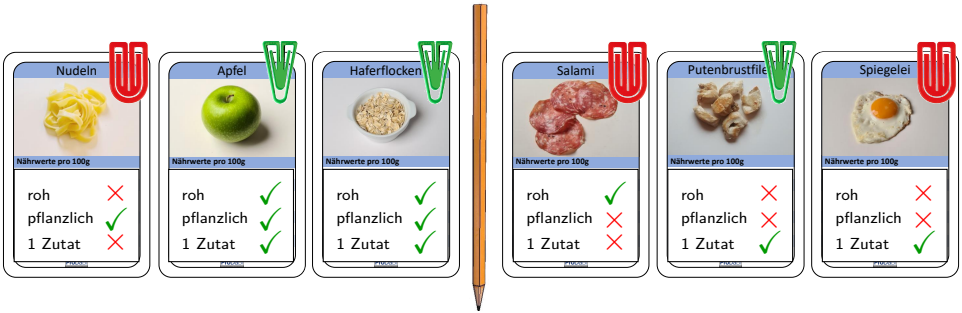
- Überprüfen Merkmal „roh“



Fehler: 2

Entscheidungsbaum mit kategorialen Daten

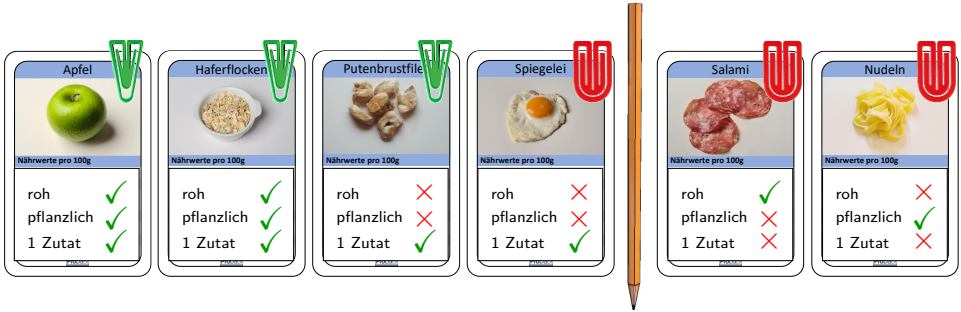
- Überprüfen Merkmal „pflanzlich“



Fehler: 2

Entscheidungsbaum mit kategorialen Daten

- Überprüfen Merkmal „1 Zutat“



Fehler: 1

$\Rightarrow$ das Merkmal „1 Zutat“„gewinnt“

Rechte Seite: kein weiteres Aufteilen nötig !

Entscheidungsbaum mit kategorialen Daten

- Aufteilen auf der linken Seite nach Merkmal „roh“

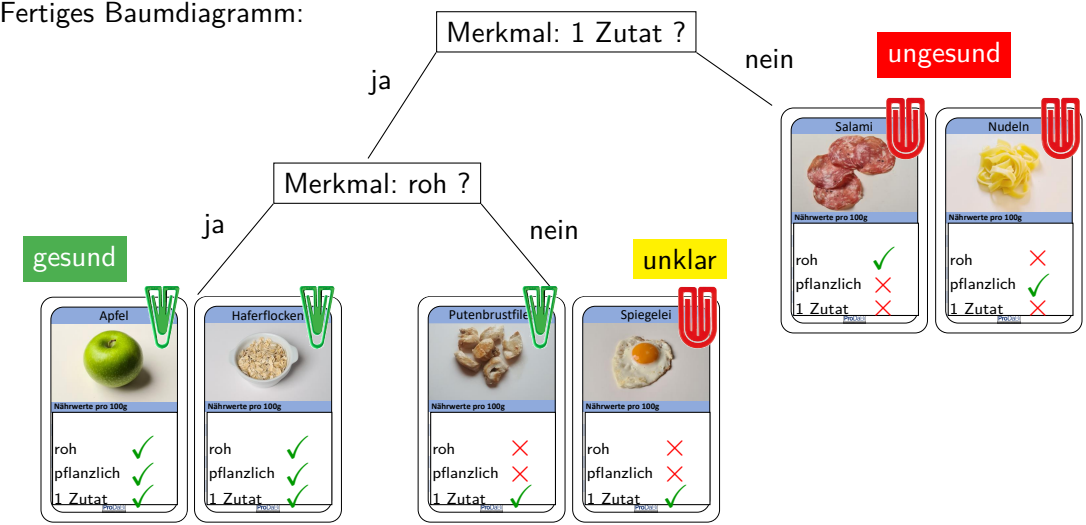
Fehler: 1

$\rightarrow$ links fertig !

$\rightarrow$ rechts kein weiteres Aufteilen möglich, da alle Merkmale gleich !

Entscheidungsbaum mit kategorialen Daten

Fertiges Baumdiagramm:



Entscheidungsbaum mit kategorialen Daten

Algorithmus Erstellen Entscheidungsbaum mit kategorialen Merkmalen

1. Merke die Fehlerrate für das erste Merkmal.
2. Wiederhole 1 für alle weiteren Merkmale.
3. Teile die Daten nach dem Merkmal auf, bei dem der Fehler am kleinsten ist.
4. Wiederhole alles mit den geteilten Daten
   1. bis entweder kein Fehler mehr auftritt.
   2. bis maximale Tiefe des Baumes erreicht ist.

Vorteile / Nachteile Entscheidungsbaum als Klassifikator

Pro

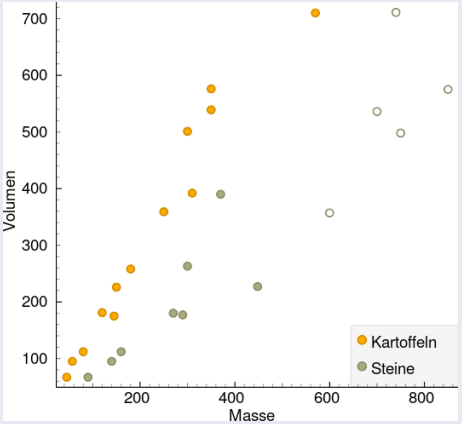
- Erkenntnis-Gewinn
  - Transparente Entscheidungsregeln
  - Wichtigkeit der Merkmale
- Erstellen des Baums und Klassifizieren getrennt
  - Vorteil bei großer Zahl an Objekten
- Weniger empfindlich bei ungleich verteilten Klassen in den Daten

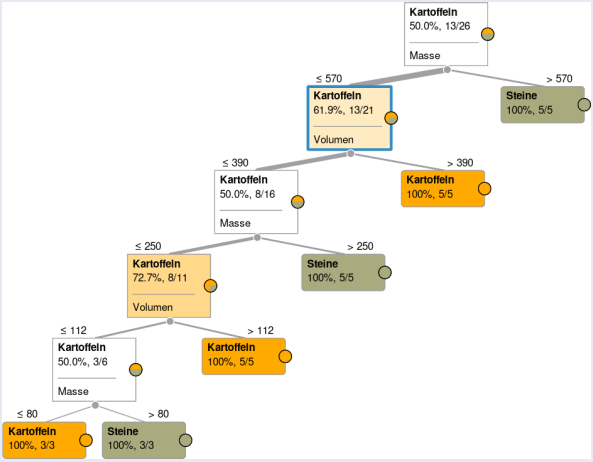
Contra

- Aufwand bei Erstellung des Baums
- Mehr Parameter
- Nicht für alle Daten geeignet

Vorteile / Nachteile Entscheidungsbaum als Klassifikator

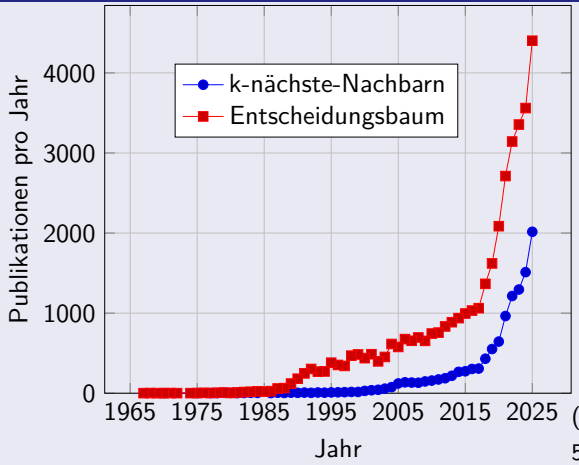
Problematischer Datensatz „Steine-Kartoffeln“





Historische / Bedeutungs-Einordnung

Publikationen auf pubmed.gov mit dem Schlüsselwort „decision tree“



(abgerufen: 5.3.2026)

Übersicht

Trainings- und Test-Daten

Problemstellung

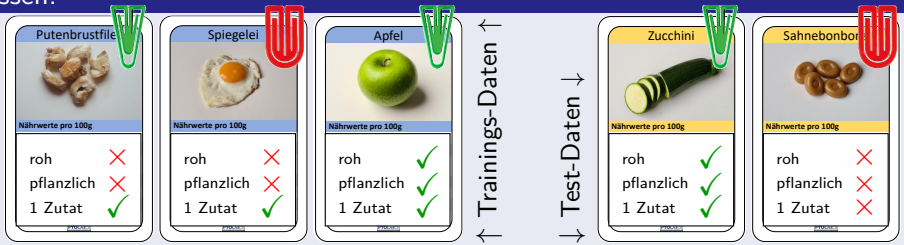
- **Gegeben**: Menge an Objekten mit bekannter Klasse
- **Gesucht**: Klasse eines Objekts mit unbekannter Klasse
- **Problem**: Woher wissen wir, wie oft der Klassifikator richtig liegt?

Aufteilung der Daten in Trainings-Daten und Test-Daten

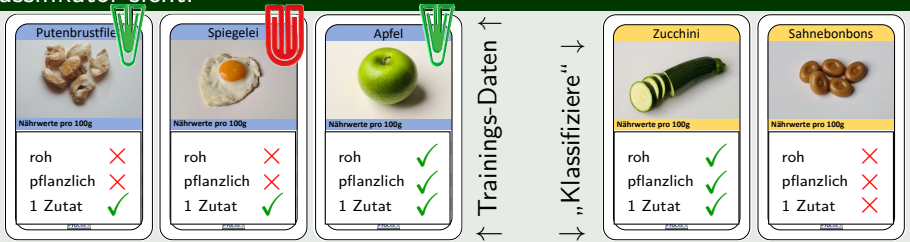
- **Trainings-Daten-Satz**
  - Objekte werden Klassifikator mit Klassen-Label übergeben.
- **Test-Daten-Satz**
  - Objekte werden dem Klassifikator ohne Klassen-Label übergeben.
  - **Ziel**: Überprüfen der Voraussage-Richtigkeit

Trainings- und Test-Daten

Wir wissen:



Der Klassifikator sieht:



Korrekt-/Falsch-Klassifikations-Rate

Korrekt-Klassifikations-Rate

$$\mathrm{KKR} = \frac{\text{Zahl richtig Klassifizierte im Test-Datensatz}}{\text{Zahl aller Objekte im Test-Datensatz}}$$

Synonyme: Klassifikationsrate, classification accuracy („CA“)

Beispiel

Unser Entscheidungsbaum soll als Test 5 Lebensmittel in gesund / ungesund einteilen. Bei 4 Lebensmitteln macht er das richtig. Wie groß ist die KKR ?

$\Rightarrow \mathrm{KKR} = \frac{4}{5} = 80\%$

Korrekt-/Falsch-Klassifikations-Rate

Falsch-Klassifikations-Rate

$$\text{FKR} = \frac{\text{Zahl falsch Klassifizierte}}{\text{Zahl Test-Objekte}}$$

Synonyme: Klassifikationsfehler, Fehlerrate, error, error rate

Hinweis: $\text{KKR} + \text{FKR} = 100\%$

Beispiel

Unser Entscheidungsbaum hat eine KKR von 80% erzielt. Wie groß ist die FKR ?

$\Rightarrow$ 20%

Korrekt-/Falsch-Klassifikations-Rate

Aufgabe

Wir haben 10 Äpfel (A) und 14 Birnen (B) und einen Klassifikator, der beides unterscheiden lernen soll. Wir geben dem Klassifikator als Trainingsdaten 8 Äpfel und 8 Birnen – mit dem Rest testen wir, wie gut er klassifiziert. Vom Obst im Testdatensatz werden 2 Äpfel und 4 Birnen richtig erkannt. Berechne die KKR und die FKR.

Lösung

Größe Datensatz insgesamt: $10 + 14 = 24$
Größe Trainingsdatensatz: $8 + 8 = 16$
Größe Testdatensatz: $24 - 16 = 8$
Davon richtig klassifiziert: $2 + 4 = 6$
$\mathrm{KKR} = \frac{6}{8} = \frac{3}{4} \qquad \mathrm{FKR} = \frac{8-6}{8} = \frac{2}{8} = \frac{1}{4}$

Konfusionsmatrix

- **Ziele**:
  - Übersichtlicher
  - Mehr Information darüber, wo Klassifikationsfehler gemacht werden.

Beispiel von oben (Äpfel und Birnen)

vorausgesagte Klasse (predicted)				
tatsächliche Klasse (actual)		A	B	$\sum$
	A	2	0	2
	B	2	4	6
	$\sum$	4	4	8

richtig klassifiziert

falsch klassifiziert

Einsicht: Die Äpfel werden schon gut erkannt. Bei den Birnen gibt es noch Probleme.

Konfusionsmatrix

Beispiel

In einer Ziegelei werden 1000 Dachziegel von einem automatischen System in fehlerfrei (A) und fehlerbehaftet (B) eingeteilt. Es werden 12 Dachziegel als B eingestuft – 2 davon allerdings fälschlicherweise. Sonst werden keine Fehler bei der Klassifikation gemacht. Erstellen Sie eine vollständig ausgefüllte Konfusionsmatrix.

Lösung

Testdaten	vorausgesagte Klasse			
tatsächliche Klasse		A	B	$\sum$
	A	988	2	990
	B	0	10	10
	$\sum$	988	12	1000

Übersicht

Trainings-Test-Schema „Leave-one-out“

Problem

- Bei manueller Auswahl von Test- und Trainings-Set kann man “Pech” haben und eine nicht repräsentative Auswahl erwischen. D.h. die KKR wird keine gute Schätzung dafür sein, wie gut unbekannte Daten klassifiziert werden.

Lösung

- Man macht mehrere Durchläufe mit verschiedenen Test- und Trainings-Sets.
- Die Auswahl der Test- und Trainings-Sets erfolgt nicht zufällig, sondern systematisch.
- Ein Datenpunkt ist das Test-Set, der Rest das Trainingsset $\rightarrow$ “Leave-one-out”-Schema

Trainings-Test-Schema „Leave-one-out“

Beispiel

Gegeben: Datensatz mit Objekten 1, 2, 3, 4.

Leave-One-Out-Schema:

			Klasse Test-Objekt		
Durchgang Nr.	Test-Set	Trainings-Set	tasächlich	Voraussage	KKR
I	1	2, 3, 4	A	A	100%
II	2	1, 3, 4	B	B	100%
III	3	1, 2, 4	A	B	0%
IV	4	1, 2, 3	B	B	100%
				Gesamt-KKR:	75%

Trainings-Test-Schema „Leave-one-out“

Algorithmus

1. Wähle einen Datenpunkt als Test-Set für den ersten Durchgang aus.
   Der Rest der Datenpunkte ist das Trainings-Set für diesen Durchgang.
2. Ermittle Korrekt-Klassifikations-Rate für diesen Durchgang (0% oder 100%).
3. Gehe zurück zu Schritt 1., wähle aber den 2., 3. usw. Datenpunkt der Reihe nach aus, bis alle Datenpunkte dran waren.
4. Berechne die Gesamt-Korrekt-Klassifikations-Rate:
   Mittelwert der einzelnen Durchgänge

Übersicht

Übersicht

Was ist Orange Data Mining?

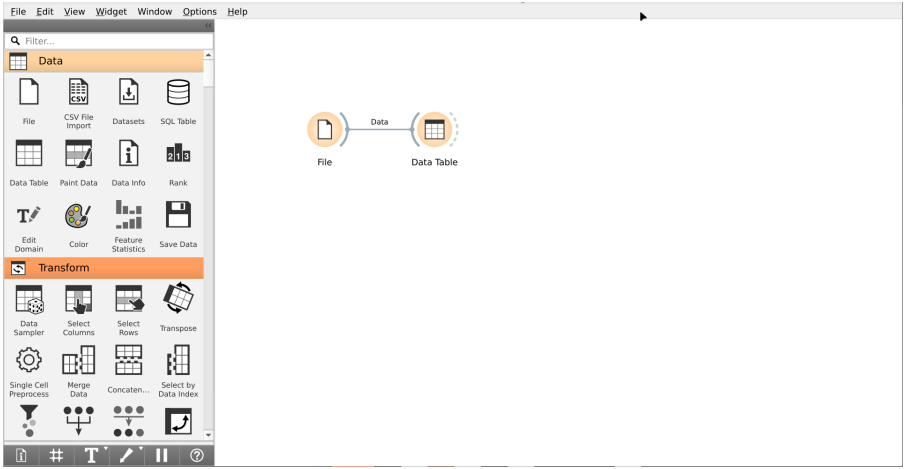
- Werkzeug für
  - Daten-Visualisierung
  - Maschinelles Lernen
  - „Visuelles Programmieren“
- Entwickler
  - Bioinformatics Lab der Universität Ljubljana, Slowenien
  - Open Source Community
- Technischer Hintergrund
  - Python
    - NumPy, scikit-learn, Qt
    - Eigenentwicklung

Unterstützungs-Material

- Prof. Engel Video-Reihe (deutsch): `https://youtu.be/HNOiE8pD6gs`
- Uni Ljubliana Video-Tutorials (engl.): `https://youtu.be/HXjnDIgGDuI`
- `https://orangedatamining.com/widget-catalog/`

Benutzer-Oberfläche

- links: Widget-Catalog, rechts: „Workflow“



Widgets und Datenfluss

- Widgets machen „irgendwas“ mit Daten
- Datenfluss wird durch Linie markiert
- Daten-Fließ-Richtung: von links nach rechts

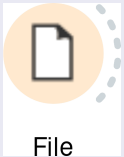
Daten laden und anschauen

„File“-Widget 

- Lädt Daten
- Versteht verschiedene Datei-Formate
  - .CSV: „comma separated values“
    - einfach aufgebaut
    - mögliche Probleme mit Kommas in Namen oder Zahlen
  - .XLSX: Excel
    - verbreitet, meist übersichtlich
    - mögliche Probleme durch Excel-“Eigenwilligkeiten“ (z. B. „Um-Interpretation“ von Zahlen- und Datums-Formaten)
- Lädt auch direkt aus dem Internet (URL)
  - z. B.: `https://abukit.de/daten/steinekartoffeln.csv`
- Daten-Übersicht (Zahl Instanzen, Zahl Merkmale)
- Daten-Vorschau

Daten laden und anschauen

„File“-Widget



File View Window Help

Source

File: datasets/iris.tab ... Reload

URL: https://abukit.de/daten/steinekartoffeln.csv

File Type

Automatically detect type

Info

26 instances
3 features (no missing values)
Data has no target variable.
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	**Klasse**	C categori...	**target**	**Kartoffeln, Steine**
2	Masse	N numeric	feature	
3	Volumen	N numeric	feature	

Daten laden und anschauen

„Data Table“-Widget 

- Daten in Tabellen-Form ansehen
- Tipp: „Visualize numeric values“ und „Color by instance class“ anhaken

Daten laden und anschauen

Aufgabe: Workflow nachbauen, Daten anschauen in „Data Table“-Widget

- Lade: `https://abukit.de/daten/steinekartoffeln.csv`

File Edit View Window Help

Info

26 instances (no missing data)
2 features
Target with 2 values
No meta attributes.

Variables

- ☑ Show variable labels (if present)
- ☑ Visualize numeric values
- ☑ Color by instance classes

Selection

- ☑ Select full rows

Restore Original Order

☑ Send Automatically

26 | 26 | 26

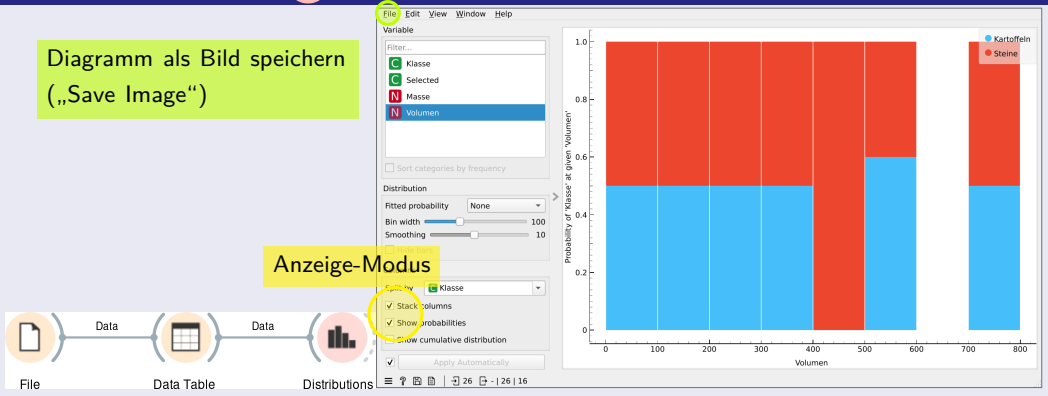
	Klasse	Masse	Volumen
1	Kartoffeln	80	112
2	Steine	160	112
3	Kartoffeln	45	67
4	Steine	90	67
5	Kartoffeln	180	258
6	Steine	300	263
7	Kartoffeln	150	226
8	Steine	448	227
9	Kartoffeln	250	359
10	Steine	600	357
11	Kartoffeln	350	539
12	Steine	700	536
13	Kartoffeln	120	181
14	Steine	270	180
15	Kartoffeln	57	95
16	Steine	140	95

Übersicht

Häufigkeits-Diagramme

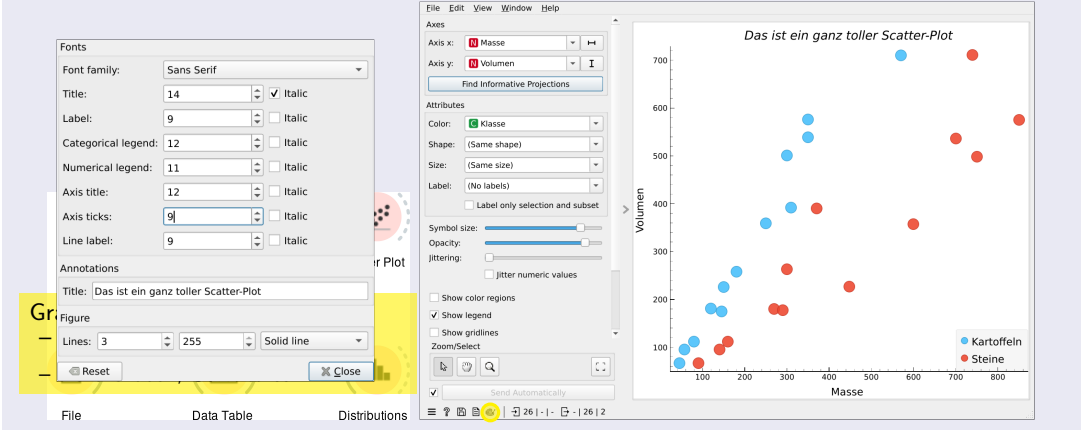
„Distributions“-Widget

Diagramm als Bild speichern („Save Image“)



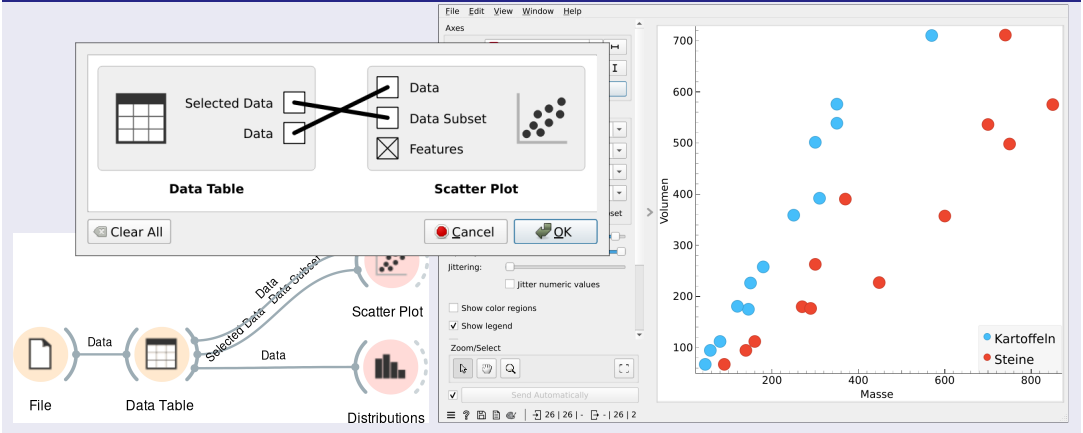
Streu-Diagramme

„Scatter-Plot“-Widget



Streu-Diagramme

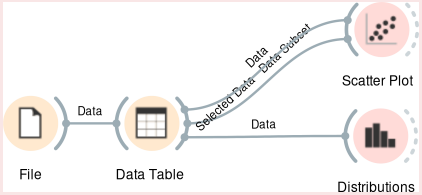
Auswählen („Selektieren“) von Instanzen: 1. „Selected Data“mit „Subset“verbinden

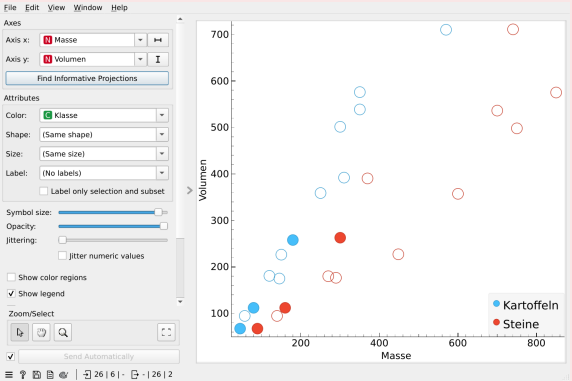


Streu-Diagramme

Aufgabe: Workflow nachbauen, Visualisierungs-Einstellungen verändern

- Lade: `https://abukit.de/daten/steinekartoffeln.csv`
- Experimentiere mit Häufigkeits-Diagrammen
- Visualisiere ausgewählte Instanzen in einem Streudiagramm





Übersicht

Mehr als 2 Merkmale

Beispiel: Iris-Datensatz[2]

> "Der Datensatz besteht aus je 50 Proben von jeder der **drei Schwertlilienarten** (Iris setosa, Iris virginica und Iris versicolor). Bei jeder der Proben wurden **vier Merkmale** gemessen: Die Länge und Breite von Kelchblatt und Kronblatt in Zentimetern."

`https://de.wikipedia.org/wiki/Schwertlilien-Datensatz` (21.10.2025)

Iris versicolor (Wikipedia)

[2] R. A. Fisher (Sep. 1936). "The use of multiple measurements in taxonomic problems". en. In: *Annals of Eugenics* 7.2, S. 179–188.

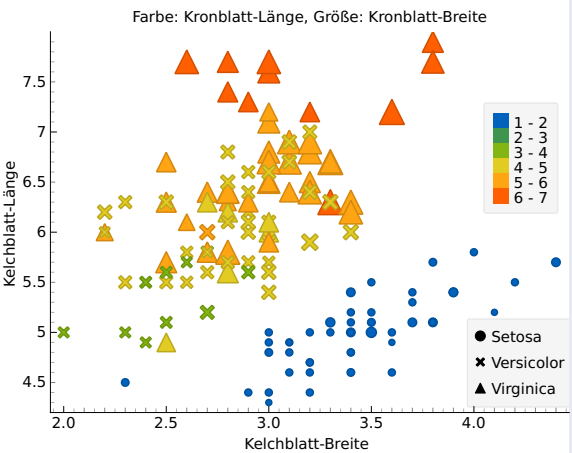
Mehr als 2 Merkmale

Problem

- Der Datensatz hat **vier** Merkmale.
- Ein Streudiagramm hat nur **zwei** Achsen.
- Wie visualisieren?

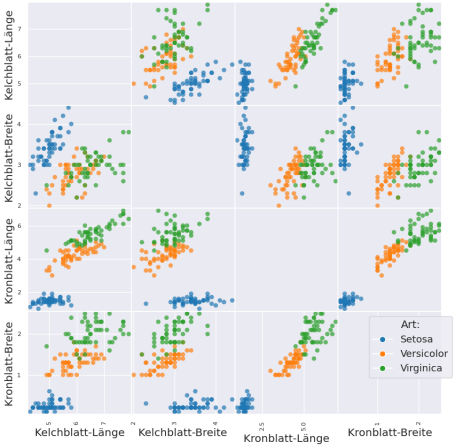
Mehr als 2 Merkmale visualisieren am Beispiel Iris-Datensatz

Iris-Datensatz: Versuch der Darstellung aller 4 Merkmale (***nicht*** nachmachen)



Mehr als 2 Merkmale visualisieren am Beispiel Iris-Datensatz

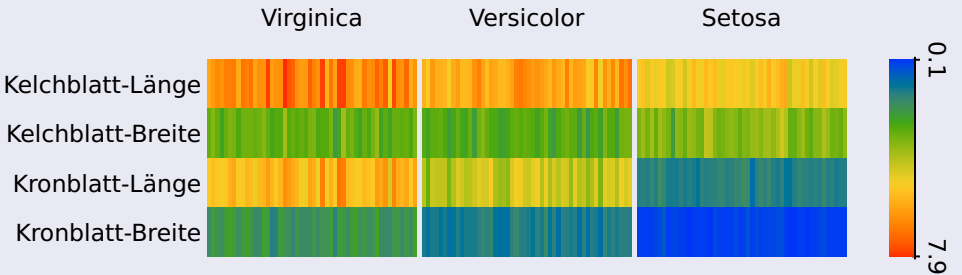
Streudiagramm-Matrix: in Orange nicht möglich, redundant, unübersichtlich



Mehr als 2 Merkmale visualisieren am Beispiel Iris-Datensatz

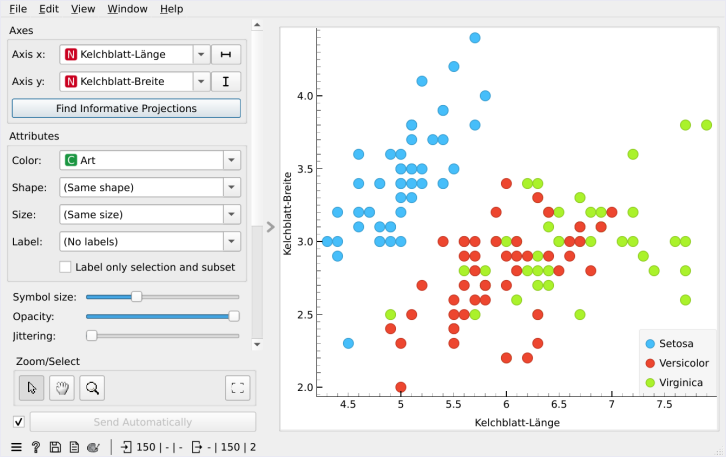
Heat-Map mit dem „Heat Map“-Widget

- Zahlenwerte werden als Farben dargestellt
- Jede Instanz entspricht einem senkrechten, nach Merkmalen unterteilten Strich
- Nachteil: **Nur bei wenigen Datensätzen brauchbar** (= u. U. Versuch wert)



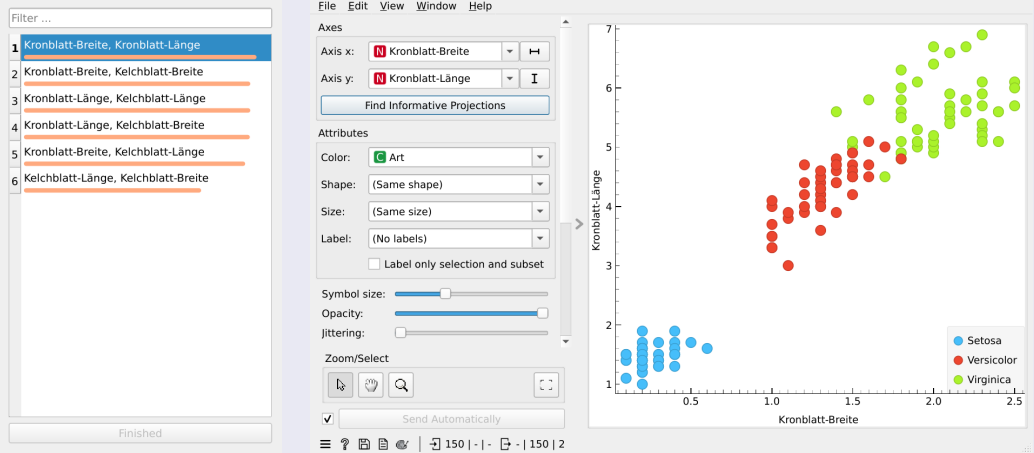
Mehr als 2 Merkmale visualisieren am Beispiel Iris-Datensatz

Not / Tugend: **Gezielte Merkmals-Auswahl** im Streudiagramm-Widget



Mehr als 2 Merkmale visualisieren am Beispiel Iris-Datensatz

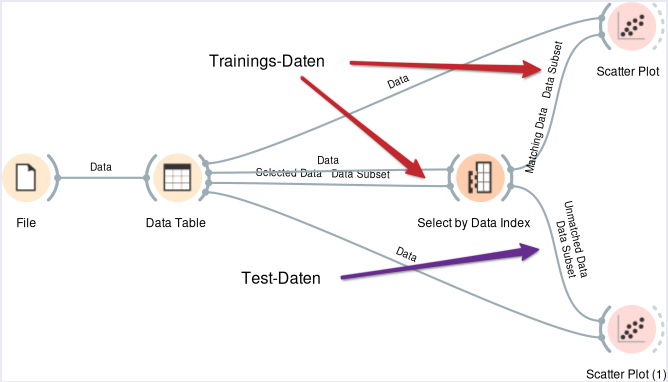
Dazu: „**Find Informative Projections**“-Vorschläge im Streudiagramm-Widget



Übersicht

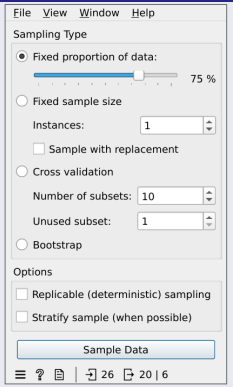
Aufteilen von Datensätzen in Trainings- und Test-Datensätze

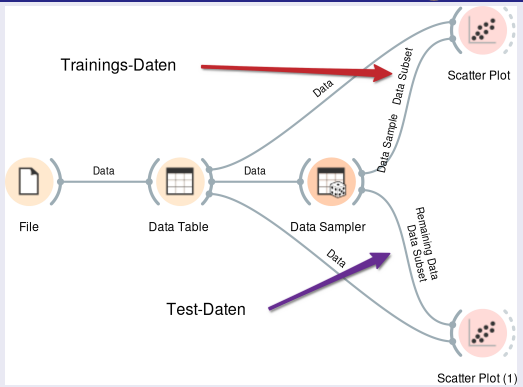
Manuelles Aufteilen mit dem „Select by Data Index“-Widget

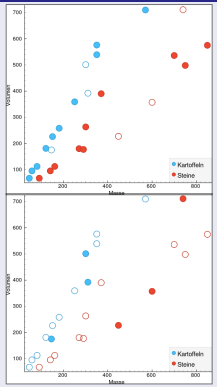


Aufteilen von Datensätzen in Trainings- und Test-Datensätze

Automatisches Aufteilen mit dem „Data Sampler“-Widget



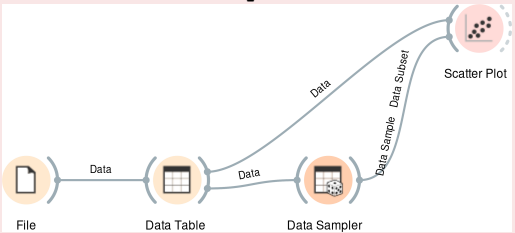


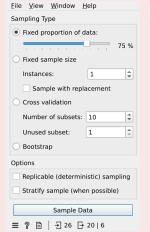


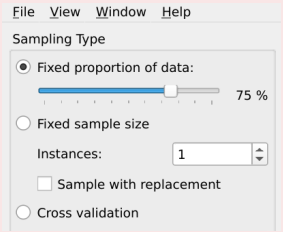
Aufteilen von Datensätzen in Trainings- und Test-Datensätze

Aufgabe: „Data Sampler“ in Aktion

- Bauen Sie den Workflow nach, um automatisiert eine Auswahl von Instanzen aus einem Datensatz zu gewinnen.
- Beobachten Sie die Zufälligkeit der Auswahl in einem Scatterplot durch wiederholtes Neu-Auswürfeln der Auswahl („Sample Data“-Button).
- Experimentieren Sie mit verschiedenen Auswahl-Größen.
- Verwenden Sie: `https://abukit.de/daten/steinekartoffeln.csv`



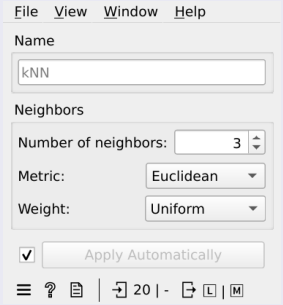


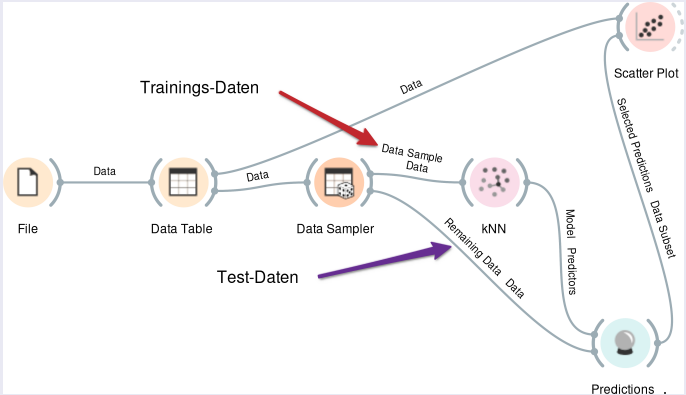


Übersicht

Einfache Klassifikations-„Pipeline“

Klassifizieren mit dem „kNN“-Widget und dem „Predictions“-Widget

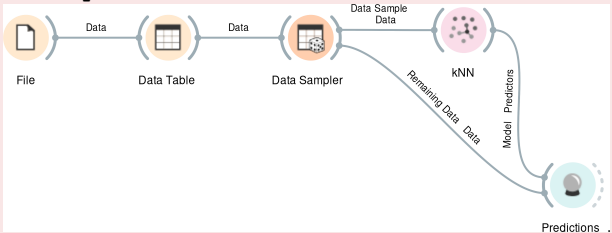




Einfache Klassifikations-„Pipeline“

Aufgabe: Einfachen Klassifikations-Workflow erstellen

- Bauen Sie einen einfachen Klassifikations-Workflow.
- Beobachten Sie in einem Predictions-Widget die Klassifikations-Ergebnisse, wenn Sie per „Data Sample“-Widget Trainings- und Test-Datensatz neu zuteilen.
- Benutzen Sie: `https://abukit.de/daten/steinekartoffeln.csv`

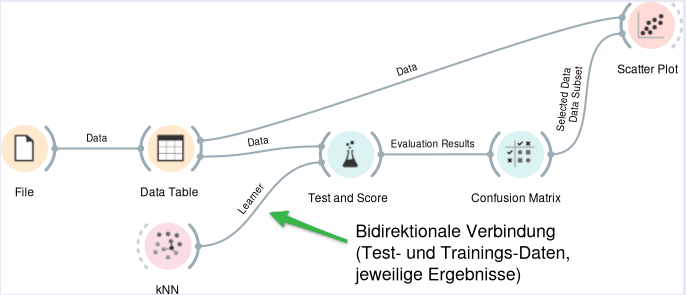


⚠ Im File-Widget muss evtl. „Klasse” von „feature” auf “target” gesetzt werden! ⚠

Übersicht

Klassifikation mit Qualitäts-Schätzung

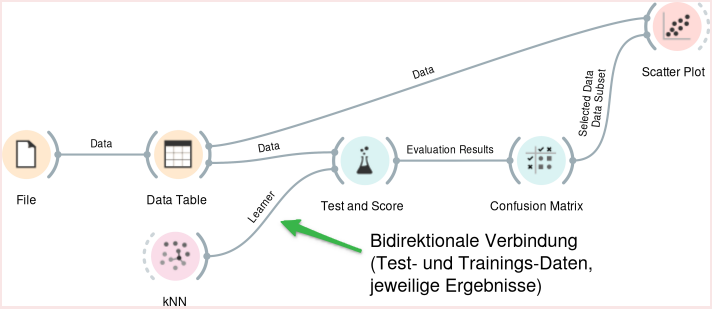
Klassifikation mit „Test and Score“-Widget und „Confusion Matrix“-Widget



Klassifikation mit Qualitäts-Schätzung

Aufgabe: kNN-Klassifikation mit Leave-one-out

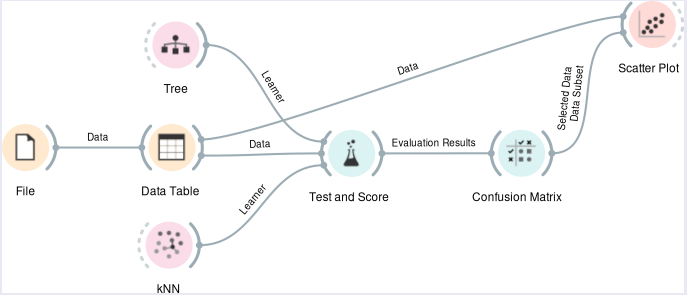
- Bauen Sie einen Klassifikations-Workflow mit Leave-one-out-Validierung („Test & Score“-Widget)
- Benutzen Sie: `https://abukit.de/daten/steinekartoffeln.csv`
- Variieren Sie den Parameter k des kNN-Klassifikators. Beobachten Sie dabei in einem Streudiagramm, welche Instanzen jeweils nicht richtig erkannt werden.



Übersicht

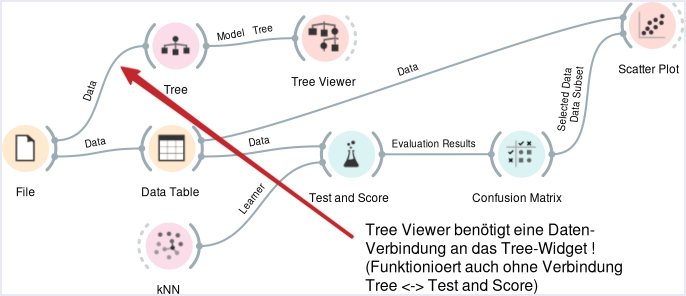
Visualisieren und Klassifizieren mit Entscheidungsbaum

Klassifizieren mit „Tree“-Widget



Visualisieren und Klassifizieren mit Entscheidungsbaum

Klassifizieren mit „Tree“-Widget – Visualisieren mit „Tree Viewer“-Widget



Visualisieren und Klassifizieren mit Entscheidungsbaum

Aufgabe: Erstellen eines Entscheidungsbaum-Workflows

- Erstellen Sie einen Workflow zur Erzeugung und Visualisierung eines Entscheidungsbaums.
- Experimentieren Sie mit den Einstellungen im „Tree Viewer“-Widget, insbesondere „Zoom“ und „Width“.
- Verändern Sie im „Tree“- und im „Tree Viewer“-Widget die „Depth“ des Baums und beobachten Sie das Ergebnis.
- Benutzen Sie `https://abukit.de/daten/steinekartoffeln.csv` und `https://abukit.de/daten/iris.csv`

Übersicht

Aufgaben basierend auf: „Fische-Datensatz“

Übersicht Daten

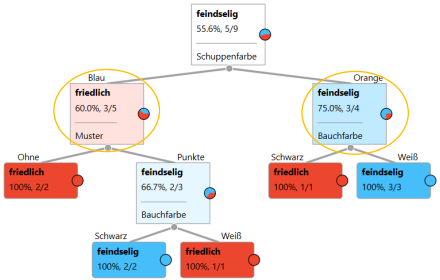
- Ursprgl. Quelle: Dr. Wolfgang Pfeffer, Tobias Fuchs, Uni Passau[3]
- Synthetische Daten
- 5 kategoriale Merkmale
- 16 Instanzen

[3] Wolfgang Pfeffer und Tobias Fuchs (2026a). *Handbuch ENTER - ENTscheidungsbaum-ERsteller*. URL: https://kex.fim.uni-passau.de/ENTER_Entscheidungsbaum_Erstellter/files/Handbuch_ENTER_Entscheidungsbaum-Ersteller.pdf (besucht am 20. 03. 2026).

Aufgaben basierend auf: „Fische-Datensatz“

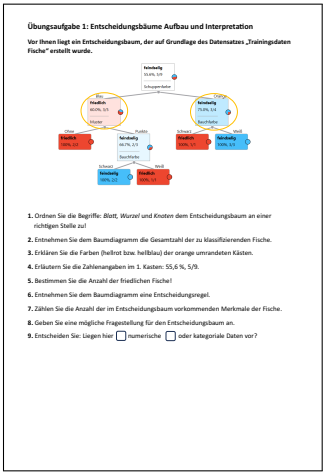
Übungsaufgabe 1: Entscheidungsbäume Aufbau und Interpretation

Vor Ihnen liegt ein Entscheidungsbaum, der auf Grundlage des Datensatzes „Trainingsdaten Fische“ erstellt wurde.



1. Ordnen Sie die Begriffe: *Blatt, Wurzel* und *Knoten* dem Entscheidungsbaum an einer richtigen Stelle zu!

2. Entnehmen Sie dem Baumdiagramm die Gesamtzahl der zu klassifizierenden Fische.



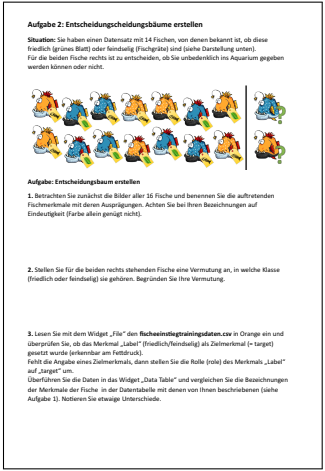
Aufgaben basierend auf: „Fische-Datensatz“

Aufgabe 2: Entscheidungscheidungsbäume erstellen

Situation: Sie haben einen Datensatz mit 14 Fischen, von denen bekannt ist, ob diese friedlich (grünes Blatt) oder feindselig (Fischgräte) sind (siehe Darstellung unten).
Für die beiden Fische rechts ist zu entscheiden, ob Sie unbedenklich ins Aquarium gegeben werden können oder nicht.

Aufgabe: Entscheidungsbaum erstellen

1. Betrachten Sie zunächst die Bilder aller 16 Fische und benennen Sie die auftretenden Fischmerkmale mit deren Ausprägungen. Achten Sie bei Ihren Bezeichnungen auf Eindeutigkeit (Farbe allein genügt nicht).



Aufgabe 2: Entscheidungscheidungsbäume erstellen

Situation: Sie haben einen Datensatz mit 14 Fischen, von denen bekannt ist, ob diese friedlich (grünes Blatt) oder feindselig (Fischgräte) sind (siehe Darstellung unten).
Für die beiden Fische rechts ist zu entscheiden, ob Sie unbedenklich ins Aquarium gegeben werden können oder nicht.

Aufgabe: Entscheidungsbaum erstellen

1. Betrachten Sie zunächst die Bilder aller 16 Fische und benennen Sie die auftretenden Fischmerkmale mit deren Ausprägungen. Achten Sie bei Ihren Bezeichnungen auf Eindeutigkeit (Farbe allein genügt nicht).

2. Stellen Sie für die beiden rechts stehenden Fische eine Vermutung an, in welche Klasse (friedlich oder feindselig) sie gehören. Begründen Sie Ihre Vermutung.

3. Lesen Sie mit dem Widget „File“ den **fischeeinstiegtrainingsdaten.csv** in Orange ein und überprüfen Sie, ob das Merkmal „Label“ (friedlich/feindselig) als Zielmerkmal (= target) gesetzt wurde (erkennbar am Fettdruck).
Fehlt die Angabe eines Zielmerkmals, dann stellen Sie die Rolle (role) des Merkmals „Label“ auf „target“ um.
Überführen Sie die Daten in das Widget „Data Table“ und vergleichen Sie die Bezeichnungen der Merkmale der Fische in der Datentabelle mit denen von Ihnen beschriebenen (siehe Aufgabe 1). Notieren Sie etwaige Unterschiede.

Aufgaben basierend auf: „Praktikums-Datensatz“

Übersicht Daten

- Ursprgl. Quelle: Dr. Wolfgang Pfeffer, Tobias Fuchs, Uni Passau[4]
- Fiktive SuS-Daten von Bewerbungen auf eine Praktikumsstelle
- 13 numerische und kategoriale Merkmale, u.a. Schulnoten in M und D
- Zielmerkmal: „Praktikumsangebot“ (= Einstellungsangebot wurde unterbreitet)
- 2000 Instanzen

[4] Wolfgang Pfeffer und Tobias Fuchs (2026b). *Mebis-Kurs mit Materialien für Lehrkräfte.* Zusatzmaterial für Lehrkräfte. URL: https://kex.fim.uni-passau.de/Zusatzmaterial/html/Zusatzmaterial.html (besucht am 21. 03. 2026).

Aufgaben basierend auf: „Praktikums-Datensatz“

Von einem Unternehmen wurden aus zahlreichen Bewerbungsschreiben von Praktikanten, die sich um einen Praktikumsplatz bewarben, ein Datensatz erstellt. An Informationen wurden beispielsweise Noten, Sprachkenntnisse und Angaben zur Bewerbung erfasst. Ziel ist es nun, mithilfe von Data Mining herauszufinden, welche Faktoren für den Erhalt eines Praktikumsplatzes eine Rolle spielen und ob sich vorhersagen lässt, ob eine Person ein Praktikumsangebot erhält.

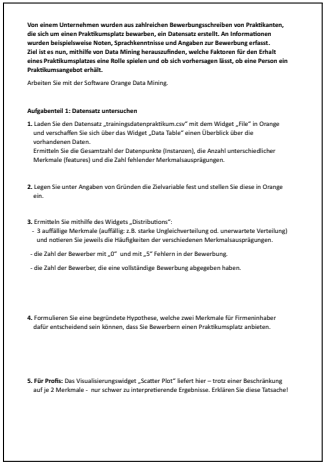
Arbeiten Sie mit der Software Orange Data Mining.

Aufgabenteil 1: Datensatz untersuchen

1. Laden Sie den Datensatz „trainingsdatenpraktikum.csv“ mit dem Widget „File“ in Orange und verschaffen Sie sich über das Widget „Data Table“ einen Überblick über die vorhandenen Daten.
Ermitteln Sie die Gesamtzahl der Datenpunkte (Instanzen), die Anzahl unterschiedlicher Merkmale (features) und die Zahl fehlender Merkmalsausprägungen.

2. Legen Sie unter Angaben von Gründen die Zielvariable fest und stellen Sie diese in Orange ein.

3. Ermitteln Sie mithilfe des Widgets „Distributions“:
- 3 auffällige Merkmale (auffällig: z.B. starke Ungleichverteilung od. unerwartete Verteilung) und notieren Sie jeweils die Häufigkeiten der verschiedenen Merkmalsausprägungen.


Von einem Unternehmen wurden aus zahlreichen Bewerbungsschreiben von Praktikanten, die sich um einen Praktikumsplatz bewarben, ein Datensatz erstellt. An Informationen wurden beispielsweise Noten, Sprachkenntnisse und Angaben zur Bewerbung erfasst. Ziel ist es nun, mithilfe von Data Mining herauszufinden, welche Faktoren für den Erhalt eines Praktikumsplatzes eine Rolle spielen und ob sich vorhersagen lässt, ob eine Person ein Praktikumsangebot erhält.

Arbeiten Sie mit der Software Orange Data Mining.

Aufgabenteil 1: Datensatz untersuchen

1. Laden Sie den Datensatz „trainingsdatenpraktikum.csv“ mit dem Widget „File“ in Orange und verschaffen Sie sich über das Widget „Data Table“ einen Überblick über die vorhandenen Daten.
Ermitteln Sie die Gesamtzahl der Datenpunkte (Instanzen), die Anzahl unterschiedlicher Merkmale (features) und die Zahl fehlender Merkmalsausprägungen.

2. Legen Sie unter Angaben von Gründen die Zielvariable fest und stellen Sie diese in Orange ein.

3. Ermitteln Sie mithilfe des Widgets „Distributions“:
- 3 auffällige Merkmale (auffällig: z.B. starke Ungleichverteilung od. unerwartete Verteilung) und notieren Sie jeweils die Häufigkeiten der verschiedenen Merkmalsausprägungen.
- die Zahl der Bewerber mit „0“ und mit „5“ Fehlern in der Bewerbung.
- die Zahl der Bewerber, die eine vollständige Bewerbung abgegeben haben.

4. Formulieren Sie eine begründete Hypothese, welche zwei Merkmale für Firmeninhaber dafür entscheidend sein können, dass Sie Bewerbern einen Praktikumsplatz anbieten.

5. Für Profis: Das Visualisierungswidget „Scatter Plot“ liefert hier – trotz einer Beschränkung auf je 2 Merkmale - nur schwer zu interpretierende Ergebnisse. Erklären Sie diese Tatsache!

Aufgaben basierend auf: „Bewerbungs-Datensatz“

Übersicht Daten

- Ursprgl. Quelle: Dr. Wolfgang Pfeffer, Tobias Fuchs, Uni Passau[5]
- Synthetische Bewerber-Daten einer fiktiven Firma
- 5 kategoriale Merkmale, 1 numerisches
- Zielmerkmal: Score (= Einstellungs-Eignung) „hoch“ oder „niedrig“
- 5000 Instanzen

[5] Wolfgang Pfeffer und Tobias Fuchs (2026b). *Mebis-Kurs mit Materialien für Lehrkräfte*. Zusatzmaterial für Lehrkräfte. URL: https://kex.fim.uni-passau.de/Zusatzmaterial/html/Zusatzmaterial.html (besucht am 21. 03. 2026).

Aufgaben basierend auf: „Bewerbungs-Datensatz“

Einladung zu einem Bewerbungsgespräch: Ja oder nein?

Situation: In einem Betrieb wurden von den vorhandenen Mitarbeiterinnen und Mitarbeitern jeweils folgende Daten erhoben:

Staatsangehörigkeit, Geschlecht, Englischkenntnisse, weitere Sprachen, akademischer Abschluss und Berufserfahrung (über 5 Jahre).

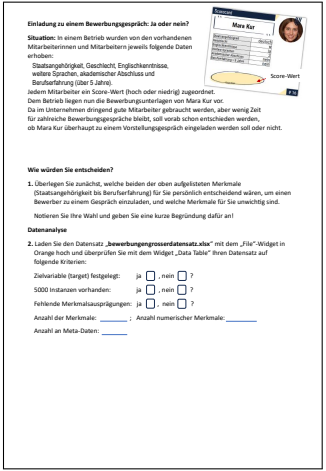
Jedem Mitarbeiter ein Score-Wert (hoch oder niedrig) zugeordnet.
Dem Betrieb liegen nun die Bewerbungsunterlagen von Mara Kur vor.
Da im Unternehmen dringend gute Mitarbeiter gebraucht werden, aber wenig Zeit für zahlreiche Bewerbungsgespräche bleibt, soll vorab schon entschieden werden, ob Mara Kur überhaupt zu einem Vorstellungsgespräch eingeladen werden soll oder nicht.

Wie würden Sie entscheiden?

1. Überlegen Sie zunächst, welche beiden der oben aufgelisteten Merkmale (Staatsangehörigkeit bis Berufserfahrung) für Sie persönlich entscheidend wären, um einen Bewerber zu einem Gespräch einzuladen, und welche Merkmale für Sie unwichtig sind.

Notieren Sie Ihre Wahl und geben Sie eine kurze Begründung dafür an!

Datenanalyse



Aufgaben basierend auf: „Philadelphia-Crime-Datensatz“

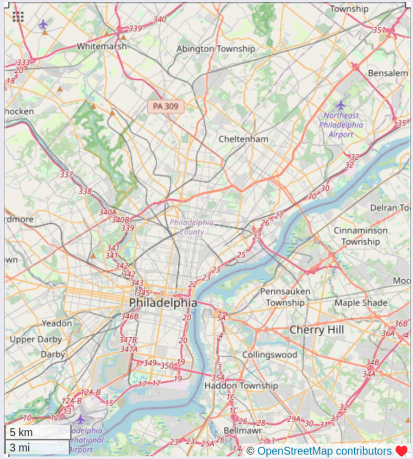
Übersicht Daten

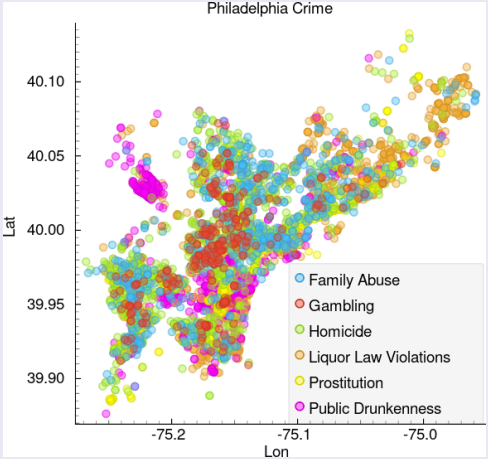
- Polizei-Einsatz-Daten der Stadt Philadelphia[6]
- Ursprgl. Quelle: https://opendataphilly.org/datasets/
- Jahre 2006 - 2016
- 6 verschiedene Vergehen (Zielmerkmal)
- Merkmale:
  - Orts-Daten (Breitengrad, Längengrad)
  - Uhrzeit und Datum
- 9666 Instanzen

[6] Thomas Rau (Mai 2022). *Geänderte Datei mit Daten zu Philadelphia Crimes, in der ich Datum und Uhrzeit, im Original eine gemeinsame Spalte, in zwei aufgeteilt habe.* URL: https://www.herr-rau.de/wordpress/2022/05/orange-data-mining.htm (besucht am 02. 03. 2026).

Aufgaben basierend auf: „Philadelphia-Crime-Datensatz“

Philadelphia (-75,30° bis -74,95° Länge und 39,85° bis 40,15° Breite)



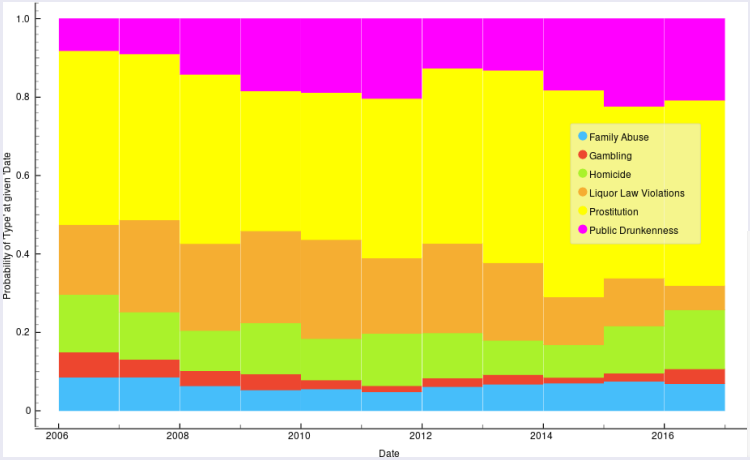


Aufgaben basierend auf: „Philadelphia-Crime-Datensatz“

Häufigkeit Vergehen *absolut* nach Jahren

Aufgaben basierend auf: „Philadelphia-Crime-Datensatz“

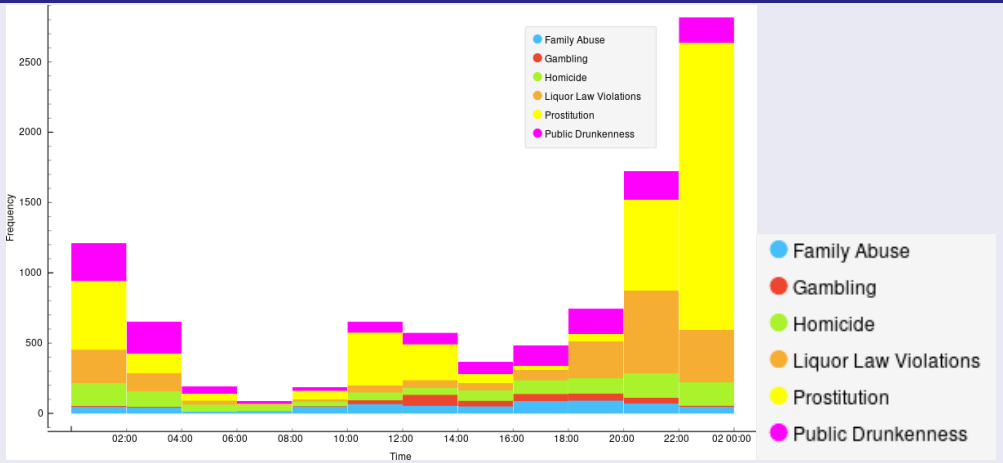
Häufigkeit Vergehen *relativ* nach Jahren



Family Abuse
Gambling
Homicide
Liquor Law Violations
Prostitution
Public Drunkenness

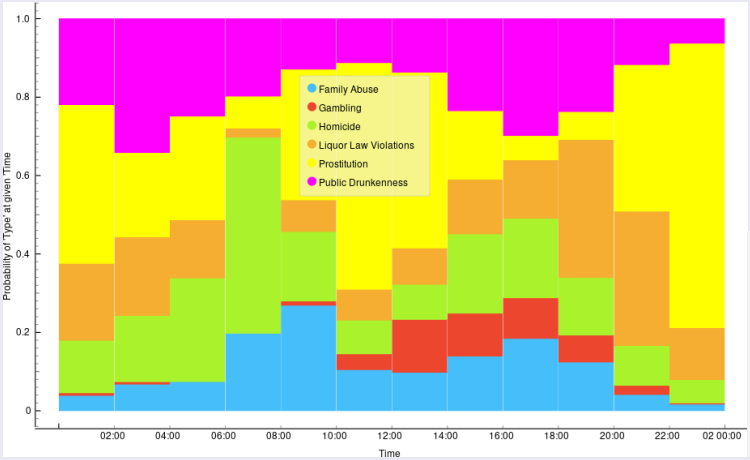
Aufgaben basierend auf: „Philadelphia-Crime-Datensatz“

Häufigkeit Vergehen *absolut* nach Uhrzeit



Aufgaben basierend auf: „Philadelphia-Crime-Datensatz“

Häufigkeit Vergehen *relativ* nach Uhrzeit



Aufgaben basierend auf: „Philadelphia-Crime-Datensatz“

Predictive Policing

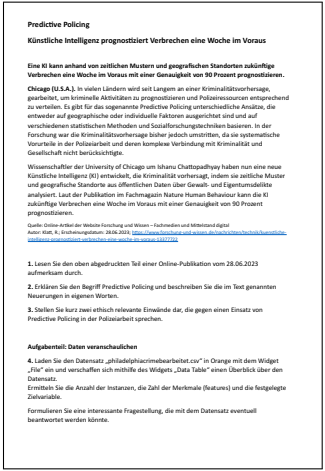
Künstliche Intelligenz prognostiziert Verbrechen eine Woche im Voraus

Eine KI kann anhand von zeitlichen Mustern und geografischen Standorten zukünftige Verbrechen eine Woche im Voraus mit einer Genauigkeit von 90 Prozent prognostizieren.

Chicago (U.S.A.). In vielen Ländern wird seit Langem an einer Kriminalitätsvorhersage, gearbeitet, um kriminelle Aktivitäten zu prognostizieren und Polizeiressourcen entsprechend zu verteilen. Es gibt für das sogenannte Predictive Policing unterschiedliche Ansätze, die entweder auf geographische oder individuelle Faktoren ausgerichtet sind und auf verschiedenen statistischen Methoden und Sozialforschungstechniken basieren. In der Forschung war die Kriminalitätsvorhersage bisher jedoch umstritten, da sie systematische Vorurteile in der Polizeiarbeit und deren komplexe Verbindung mit Kriminalität und Gesellschaft nicht berücksichtigte.

Wissenschaftler der University of Chicago um Ishanu Chattopadhyay haben nun eine neue Künstliche Intelligenz (KI) entwickelt, die Kriminalität vorhersagt, indem sie zeitliche Muster und geografische Standorte aus öffentlichen Daten über Gewalt- und Eigentumsdelikte analysiert. Laut der Publikation im Fachmagazin Nature Human Behaviour kann die KI zukünftige Verbrechen eine Woche im Voraus mit einer Genauigkeit von 90 Prozent prognostizieren.

Quelle: Online-Artikel der Website Forschung und Wissen – Fachmedien und Mittelstand digital
Autor: Klatt, R.; Erscheinungsdatum: 28.06.2023; https://www.forschung-und-wissen.de/nachrichten/technik/kuenstliche-intelligenz-prognostiziert-verbrechen-eine-woche-im-voraus-13377722



Predictive Policing

Künstliche Intelligenz prognostiziert Verbrechen eine Woche im Voraus

Eine KI kann anhand von zeitlichen Mustern und geografischen Standorten zukünftige Verbrechen eine Woche im Voraus mit einer Genauigkeit von 90 Prozent prognostizieren.

Chicago (U.S.A.). In vielen Ländern wird seit Langem an einer Kriminalitätsvorhersage, gearbeitet, um kriminelle Aktivitäten zu prognostizieren und Polizeiressourcen entsprechend zu verteilen. Es gibt für das sogenannte Predictive Policing unterschiedliche Ansätze, die entweder auf geographische oder individuelle Faktoren ausgerichtet sind und auf verschiedenen statistischen Methoden und Sozialforschungstechniken basieren. In der Forschung war die Kriminalitätsvorhersage bisher jedoch umstritten, da sie systematische Vorurteile in der Polizeiarbeit und deren komplexe Verbindung mit Kriminalität und Gesellschaft nicht berücksichtigte.

Wissenschaftler der University of Chicago um Ishanu Chattopadhyay haben nun eine neue Künstliche Intelligenz (KI) entwickelt, die Kriminalität vorhersagt, indem sie zeitliche Muster und geografische Standorte aus öffentlichen Daten über Gewalt- und Eigentumsdelikte analysiert. Laut der Publikation im Fachmagazin Nature Human Behaviour kann die KI zukünftige Verbrechen eine Woche im Voraus mit einer Genauigkeit von 90 Prozent prognostizieren.

Quelle: Online-Artikel der Website Forschung und Wissen – Fachmedien und Mittelstand digital
Autor: Klatt, R.; Erscheinungsdatum: 28.06.2023; https://www.forschung-und-wissen.de/nachrichten/technik/kuenstliche-intelligenz-prognostiziert-verbrechen-eine-woche-im-voraus-13377722

1. Lesen Sie den oben abgedruckten Teil einer Online-Publikation vom 28.06.2023 aufmerksam durch.

2. Erklären Sie den Begriff Predictive Policing und beschreiben Sie die im Text genannten Neuerungen in eigenen Worten.

3. Stellen Sie kurz zwei ethisch relevante Einwände dar, die gegen einen Einsatz von Predictive Policing in der Polizeiarbeit sprechen.

Aufgabenteil: Daten veranschaulichen

4. Laden Sie den Datensatz „philadelphiacrimebearbeitet.csv“ in Orange mit dem Widget „File“ ein und verschaffen sich mithilfe des Widgets „Data Table“ einen Überblick über den Datensatz.
Ermitteln Sie die Anzahl der Instanzen, die Zahl der Merkmale (features) und die festgelegte Zielvariable.

Formulieren Sie eine interessante Fragestellung, die mit dem Datensatz eventuell beantwortet werden könnte.

Aufgaben basierend auf: „Milch-Fettsäuren-Datensatz“

Übersicht Daten

- Quelle: Dr. Joachim Molkantin, Max-Rubner-Institut Kiel[7][8]
- Massenanteile verschiedener Fettsäuren an allen Fettsäuren in Milch
- 157 Proben biologisch und konventionell erzeugter Milch aus 4 Bundesländern
- Merkmale:
  - 66 numerische Merkmale (Fettsäure-Anteile)
  - 2 kategoriale (Bundesland, Erzeugung), als Zielmerkmal verwendbar
  - 1 Meta-Merkmal (Erzeugungsdatum, 2006 -)

[7] Joachim Molkentin (Feb. 2026). *Milk fatty acids in Germany*. en. S2411. DOI: 10.25826/Data20260209-090013-0. URL: https://doi.org/10.25826/Data20260209-090013-0.

[8] Joachim Molkentin (11. Feb. 2009). “Authentication of Organic Milk Using δ^{13}C and the α-Linolenic Acid Content of Milk Fat”. In: *Journal of Agricultural and Food Chemistry* 57.3, S. 785–790. ISSN: 0021-8561, 1520-5118. DOI: 10.1021/jf8022029. URL: https://pubs.acs.org/doi/10.1021/jf8022029 (besucht am 10. 03. 2026).

Aufgaben basierend auf: „Milch-Fettsäuren-Datensatz“

Wie sich falsche Bio-Milch verrät

Martin Rötzschke: Wie sich falsche Bio-Milch verrät. 2. März 2009 Internetpublikation unter: www.wissenschaft.de/erde-umwelt/wie-sich-falsche-bio-milch-verraet

„Die Wissenschaft hat einen Weg gefunden, um zu testen, ob Milch wirklich "bio" ist. Ein deutscher Wissenschaftler hat ein zuverlässiges Verfahren gefunden, um Etikettenschwindel bei Bio-Milch zu entlarven: Ökologisch erzeugte Milch lässt sich anhand des Gewichtsverhältnisses der enthaltenen Kohlenstoffatome von konventionell produzierter unterscheiden. Auch der Anteil bestimmter Fettsäuren ist verschieden, zeigt die Arbeit von Joachim Molkentin vom Max-Rubner-Institut in Kiel.

[...] Über 18 Monate deckte sich der Forscher darum im Kieler Einzelhandel und bei einer Bioland-Molkerei mit Milch ein: Alle zwei Wochen kaufte er Proben von sechs vertrauenswürdigen Milch-Sorten, drei davon Bio-Milch. Beim Vergleichen der Proben fand er chemische Eigenschaften, mit denen sich Milch aus biologischer Herstellung von konventioneller Milch eindeutig unterscheiden ließ.

Während ökologisch gehaltene Kühe vorwiegend frisches Gras oder Heu zu fressen bekommen, enthält die Nahrung von anderen Kühen meist große Anteile an Mais aus dem Silo. Mais verwertet Kohlenstoffdioxid aus der Luft auf andere Weise als die meisten anderen Futterpflanzen. Dabei sammeln sich in der Pflanze besonders schwere Kohlenstoffatome an, die sogenannten C-13-Isotopen. Da Bio-Kühe weniger Mais fressen, besitzt ihre Milch auch einen geringeren C-13-Anteil, erklärt Molkentin.

Für sein Testverfahren nutzt der Forscher noch einen weiteren Unterschied: Bio-Milch besitzt wiederum einen höheren Anteil einer Fettsäure namens C18:3-omega-3. Der Grund dafür

Wie sich falsche Bio-Milch verrät

Martin Rötzschke: Wie sich falsche Bio-Milch verrät. 2. März 2009 Internetpublikation unter: www.wissenschaft.de/erde-umwelt/wie-sich-falsche-bio-milch-verraet

„Die Wissenschaft hat einen Weg gefunden, um zu testen, ob Milch wirklich "bio" ist. Ein deutscher Wissenschaftler hat ein zuverlässiges Verfahren gefunden, um Etikettenschwindel bei Bio-Milch zu entlarven: Ökologisch erzeugte Milch lässt sich anhand des Gewichtsverhältnisses der enthaltenen Kohlenstoffatome von konventionell produzierter unterscheiden. Auch der Anteil bestimmter Fettsäuren ist verschieden, zeigt die Arbeit von Joachim Molkentin vom Max-Rubner-Institut in Kiel.

[...] Über 18 Monate deckte sich der Forscher darum im Kieler Einzelhandel und bei einer Bioland-Molkerei mit Milch ein: Alle zwei Wochen kaufte er Proben von sechs vertrauenswürdigen Milch-Sorten, drei davon Bio-Milch. Beim Vergleichen der Proben fand er chemische Eigenschaften, mit denen sich Milch aus biologischer Herstellung von konventioneller Milch eindeutig unterscheiden ließ.

Während ökologisch gehaltene Kühe vorwiegend frisches Gras oder Heu zu fressen bekommen, enthält die Nahrung von anderen Kühen meist große Anteile an Mais aus dem Silo. Mais verwertet Kohlenstoffdioxid aus der Luft auf andere Weise als die meisten anderen Futterpflanzen. Dabei sammeln sich in der Pflanze besonders schwere Kohlenstoffatome an, die sogenannten C-13-Isotopen. Da Bio-Kühe weniger Mais fressen, besitzt ihre Milch auch einen geringeren C-13-Anteil, erklärt Molkentin.

Für sein Testverfahren nutzt der Forscher noch einen weiteren Unterschied: Bio-Milch besitzt wiederum einen höheren Anteil einer Fettsäure namens C18:3-omega-3. Der Grund dafür sind vermutlich Auswirkungen der Futterzusammensetzung auf die Verdauung. Obwohl der Gehalt an C18:3-omega-3 und an C-13-Isotopen über die Jahreszeiten variiert, fand der Forscher Schwellenwerte, mit denen sich fast alle Proben korrekt als Bio-Milch oder konventionelle Milch einordnen lassen. „[...]

Originalartikel: Joachim Molkentin (Max-Rubner-Institut, Kiel): Journal of Agricultural and Food Chemistry, Bd. 57, 785-790

Arbeitsaufträge:

1. Lesen Sie den oben abgedruckten Teil einer Online-Publikation vom 02.03.2009 aufmerksam durch.

1.1 Nennen Sie 2 Gründe, warum ein zuverlässiges Verfahren zur Unterscheidung von konventionell und ökologisch erzeugter Milch notwendig ist.

1.2 Zählen Sie die beiden im Text genannten Möglichkeiten auf, wie ökologisch und konventionell erzeugte Milch unterschieden werden kann. Nennen Sie kurz die Hintergründe für das Auftreten der jeweiligen Marker.

Übersicht

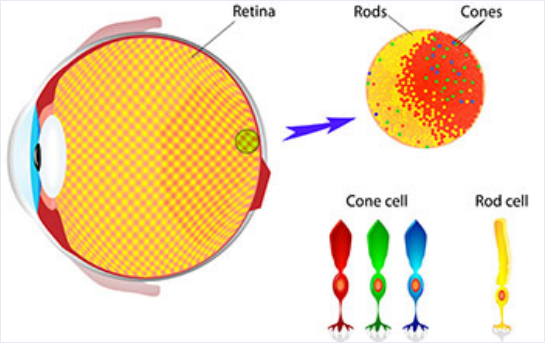
Bild-Segmentierung in Orange

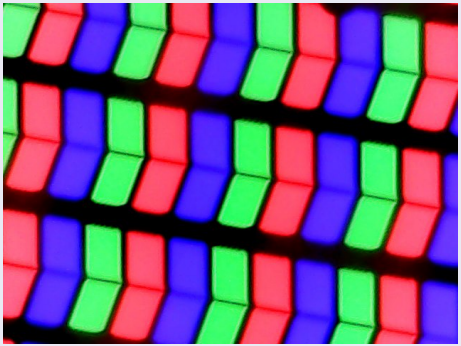
Projekt-Idee

- Segmentierung = Entdecken / Isolieren von Objekten in Bildern
- Bilder = reichlich vorhandenes Material bzw. leicht zu generieren
- 1 Pixel besteht aus 5 Merkmalen (siehe nächstes Dia)
  - Klassifikator auf zu entdeckende Objekte (Pixel) trainieren
  - Objekte (Pixel) in anderen Bildern entdecken

Bild-Segmentierung in Orange

Links: Photorezeptoren in der Netzhaut; Rechts: Pixel-Raster eines Laptops





$\Rightarrow$ Jeder Bildpunkt besteht aus 5 Merkmalen:
x-Koordinate, y-Koordinate, Rot-Wert, Grün-Wert, Blau-Wert

Bild-Segmentierung in Orange

Orange Workflow, um in Bildern Objekte anhand ihrer typischen Farbe zu erkennen

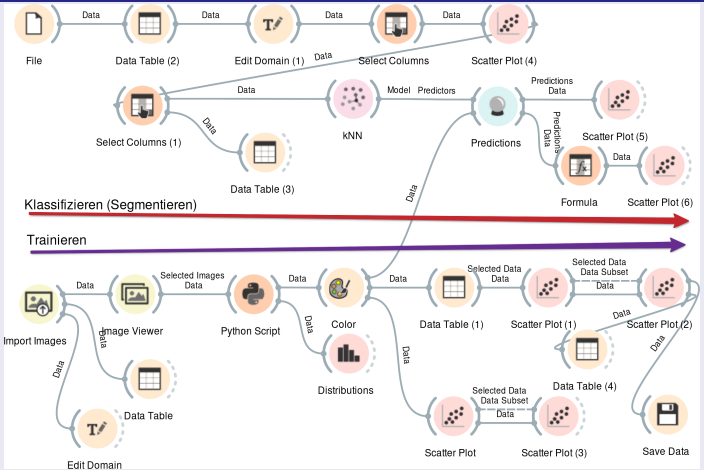


Bild-Segmentierung in Orange

Die Pixelfarben dieses Bildes werden als Trainingsdaten verwendet

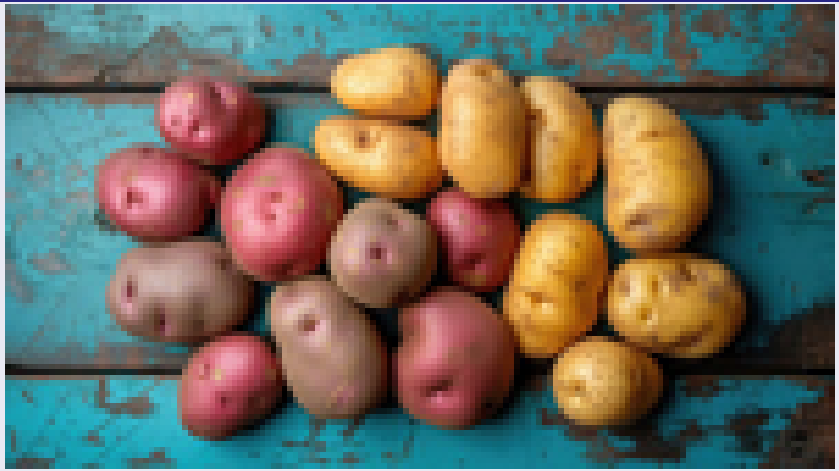


Bild-Segmentierung in Orange

Scatter-Plot-Widget zur Darstellung des Rot-Kanals

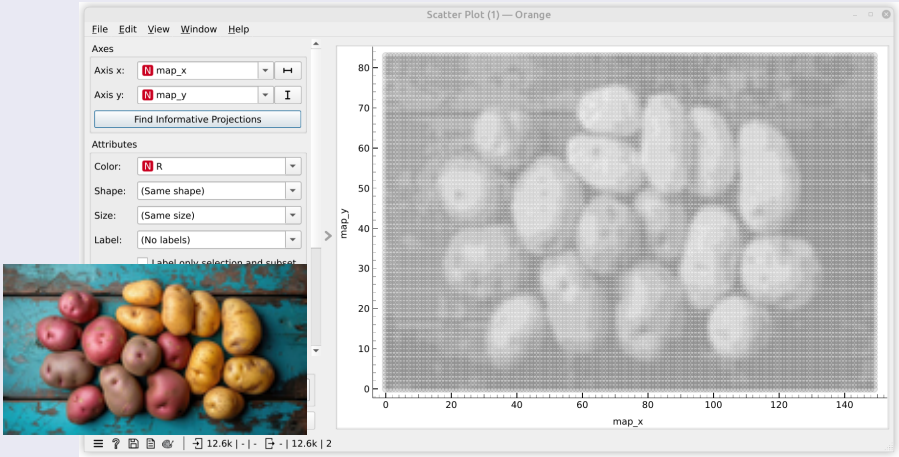


Bild-Segmentierung in Orange

Im Scatter-Plot-Widget werden die braunen Kartoffeln „selektiert“

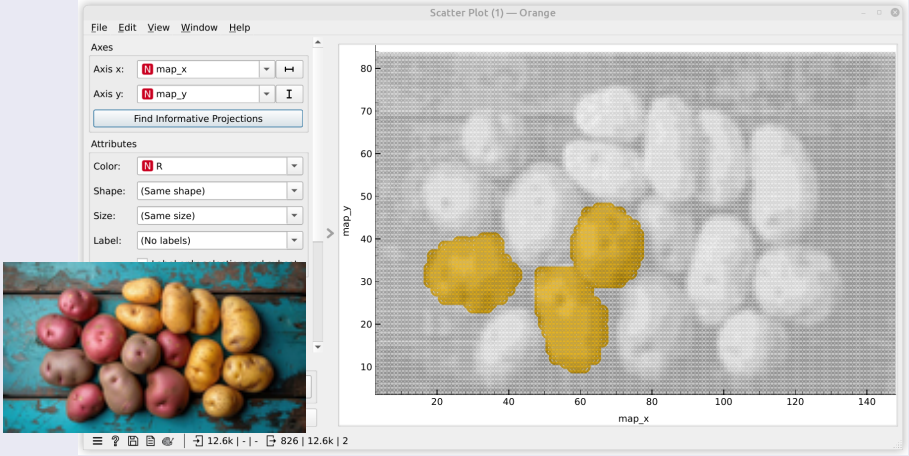


Bild-Segmentierung in Orange

Streudiagramm des Rot- und Grün-Kanals; rot: selektierte Pixel

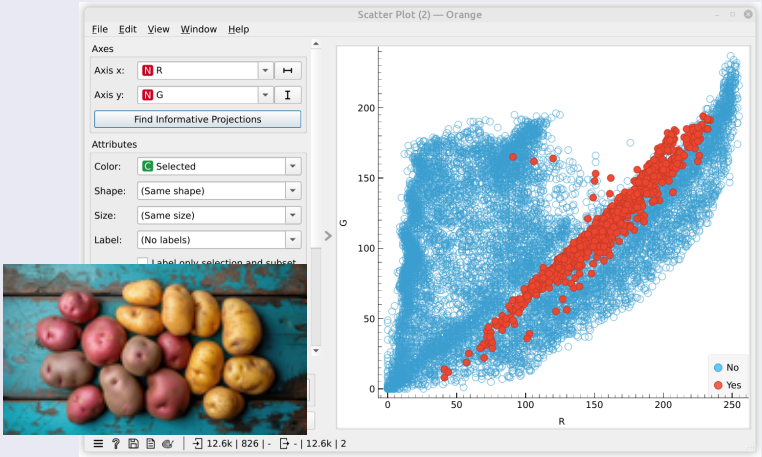


Bild-Segmentierung in Orange

Wie vor, nur: Rot = vom kNN (k=3) markierte Pixel !

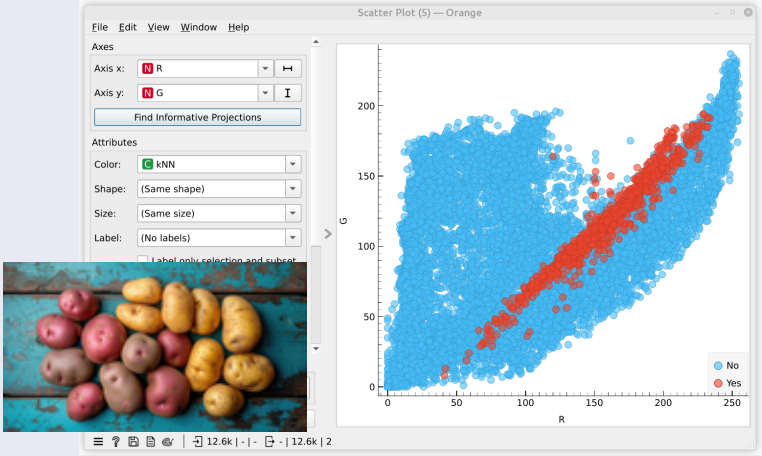


Bild-Segmentierung in Orange

Vom kNN markierte Pixel im Trainingsbild (Rot-Kanal)

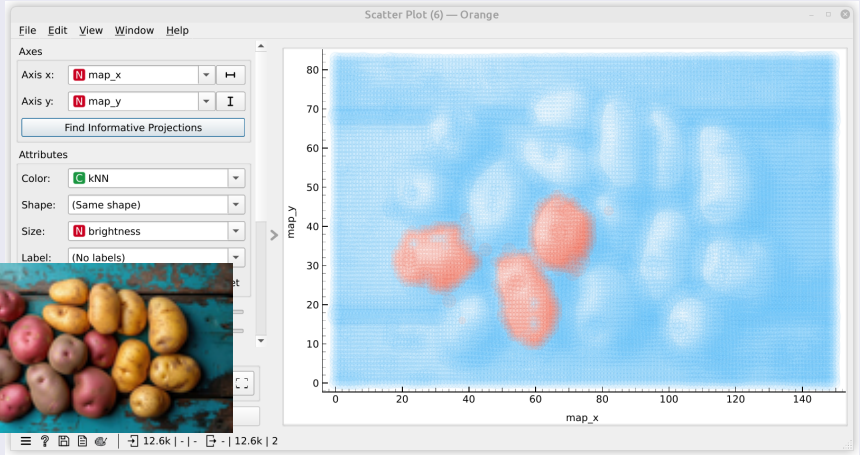
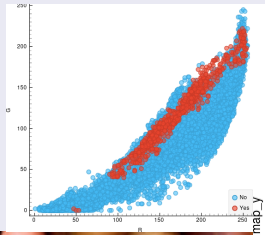


Bild-Segmentierung in Orange

Vom kNN markierte Pixel in Testbild Nr. 4 (Rot-Kanal)



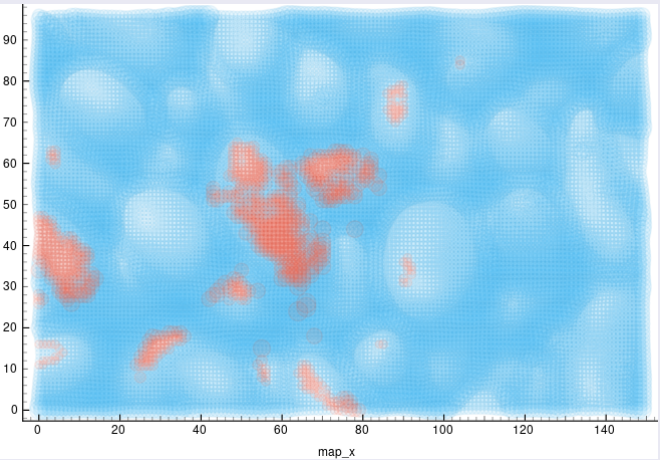


Bild-Segmentierung in Orange

Vom kNN markierte Pixel in Testbild Nr. 6 (Rot-Kanal)

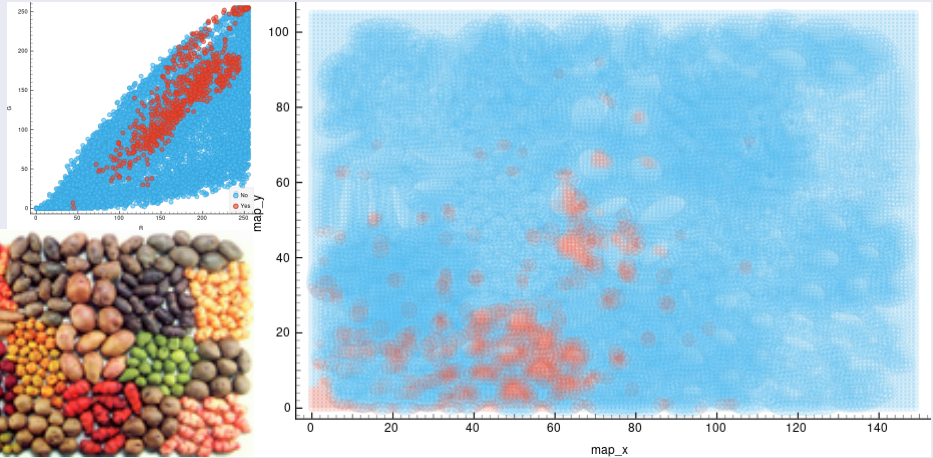
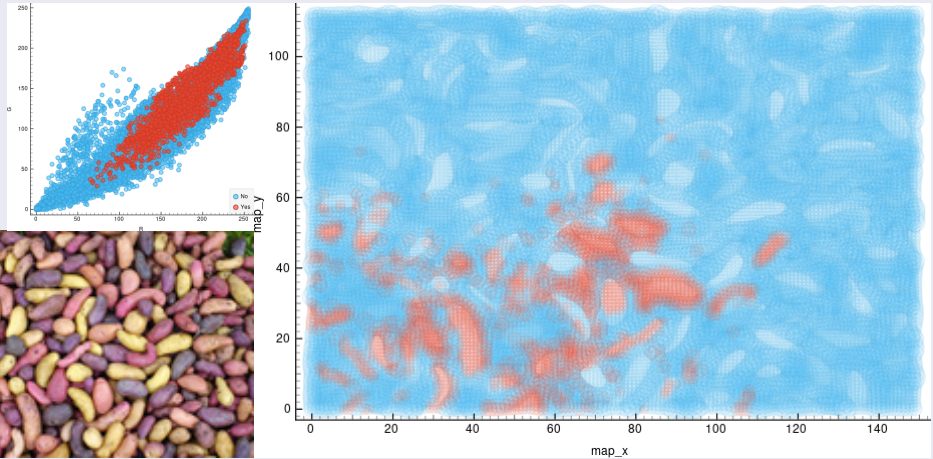


Bild-Segmentierung in Orange

Vom kNN markierte Pixel in Testbild Nr. 1 (Rot-Kanal)



1-Pixel-Farbsensor

Projekt-Idee

- 1-Pixel-Farbsensor zur Merkmals-Gewinnung
- Vorteile:
  - Komplexität der x-/y-Koordinaten fällt weg
  - Definierte Aufnahme-Bedingungen, da immer gleicher Sensor
  - Höhere Empfindlichkeit:
    - Sensor kann statt 255 Helligkeiten pro Kanal (übliche Kamera) 65536 unterscheiden
    - Es gibt (von anderen Firmen) auch 1-Pixel-Sensoren mit 10 Farb-Kanälen $\rightarrow$ können 10 statt 3 Farben aufnehmen $\rightarrow$ „hyperspektral“-Aufnahme
- Mögliche Ziele:
  - Obst-Arten, -Sorten, -Reife-Grade unterscheiden

„Color-Picker-Bricklet 2.0“ der Fa. Tinkerforge GmbH

https://www.tinkerforge.com/en/doc/Hardware/Bricklets/Color_V2.html (23.3.2926)

NutriScore / OpenFoodFacts

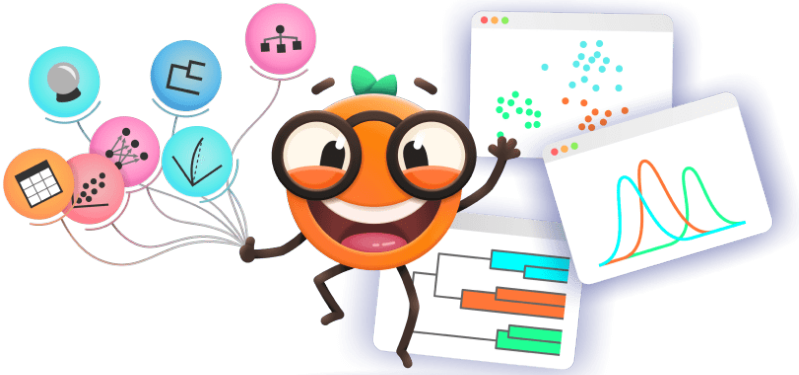
Projekt-Idee

- NutriScore:
  - omni-präsent, weithin bekannt, hochrelevant
  - veröffentlichte, verstehbare Berechnungs-Formel(n)
    - (als Entscheidungsbaum darstellbar)
- openfoodfacts.org:
  - franz. NGO, die Lebensmittel-Bestandteile aus Inhaltslisten (Verpackung) sammelt
  - stellt die erfassten Lebensmittel-Daten als Download zur Verfügung
    - 398 555 Instanzen = Produkte (22.3.2026)
    - üblicherweise > 5 Merkmale je Lebensmittel
    - Merkmal „NutriScore“ = ideales Zielmerkmal $\rightarrow$ NUTRI-SCORE A B C D E

NutriScore / OpenFoodFacts

Probleme

- Daten-Format
  - Daten nicht als zweidimensionale Tabelle gespeichert (sondern „noSQL“)
    - Vorhandensein von Merkmalen kann von anderen Merkmalen abhängen
  - Datei-Formate schwierig: mongodbdump, jsonl, parquet oder API; „defektes“ CSV
- Daten-Qualität
  - Daten von Nutzern erhoben (per Webseite / App)
    - keine / kaum Qualitätskontrolle
    - eigene / eigenwillige Auslegung der Merkmals-Bezeichnungen und Einheiten
    - Nutzung aller möglichen Schriften und Sonderzeichen (z. B. Emojis)
    - Lücken / Fehler / Doubletten / ...
- Berechnung des NutriScore
  - Hängt von der Lebensmittel-Kategorie ab → oft falsche / unklare Zuordnung
  - Änderung Berechnungs-Methode und Zuordnung in 2023 → 



Danke für die Aufmerksamkeit!

Literaturverzeichnis I

- Fisher, R. A. (Sep. 1936). "The use of multiple measurements in taxonomic problems". en. In: *Annals of Eugenics* 7.2, S. 179–188.
- Molkentin, Joachim (11. Feb. 2009). "Authentication of Organic Milk Using δ^{13}C and the α-Linolenic Acid Content of Milk Fat". In: *Journal of Agricultural and Food Chemistry* 57.3, S. 785–790. ISSN: 0021-8561, 1520-5118. DOI: 10.1021/jf8022029. URL: https://pubs.acs.org/doi/10.1021/jf8022029 (besucht am 10. 03. 2026).
- — (Feb. 2026). *Milk fatty acids in Germany*. en. S2411. DOI: 10.25826/Data20260209-090013-0. URL: https://doi.org/10.25826/Data20260209-090013-0.

Literaturverzeichnis II

Pfeffer, Wolfgang und Tobias Fuchs (2026a). *Handbuch ENTER - ENTscheidungsbaum-ERsteller*. URL: https://kex.fim.uni-passau.de/ENTER_Entscheidungsbaum_Erstellter/files/Handbuch_ENTER_Entscheidungsbaum-Ersteller.pdf (besucht am 20. 03. 2026).

— (2026b). *Mebis-Kurs mit Materialien für Lehrkräfte*. Zusatzmaterial für Lehrkräfte. URL: https://kex.fim.uni-passau.de/Zusatzmaterial/html/Zusatzmaterial.html (besucht am 21. 03. 2026).

Rau, Thomas (Mai 2022). *Geänderte Datei mit Daten zu Philadelphia Crimes, in der ich Datum und Uhrzeit, im Original eine gemeinsame Spalte, in zwei aufgeteilt habe*. URL: https://www.herr-rau.de/wordpress/2022/05/orange-data-mining.htm (besucht am 02. 03. 2026).

Literaturverzeichnis III


Snow, John (1855). *On the mode of communication of cholera*. second edition, much enlarged. London: John Churchill. 216 S.