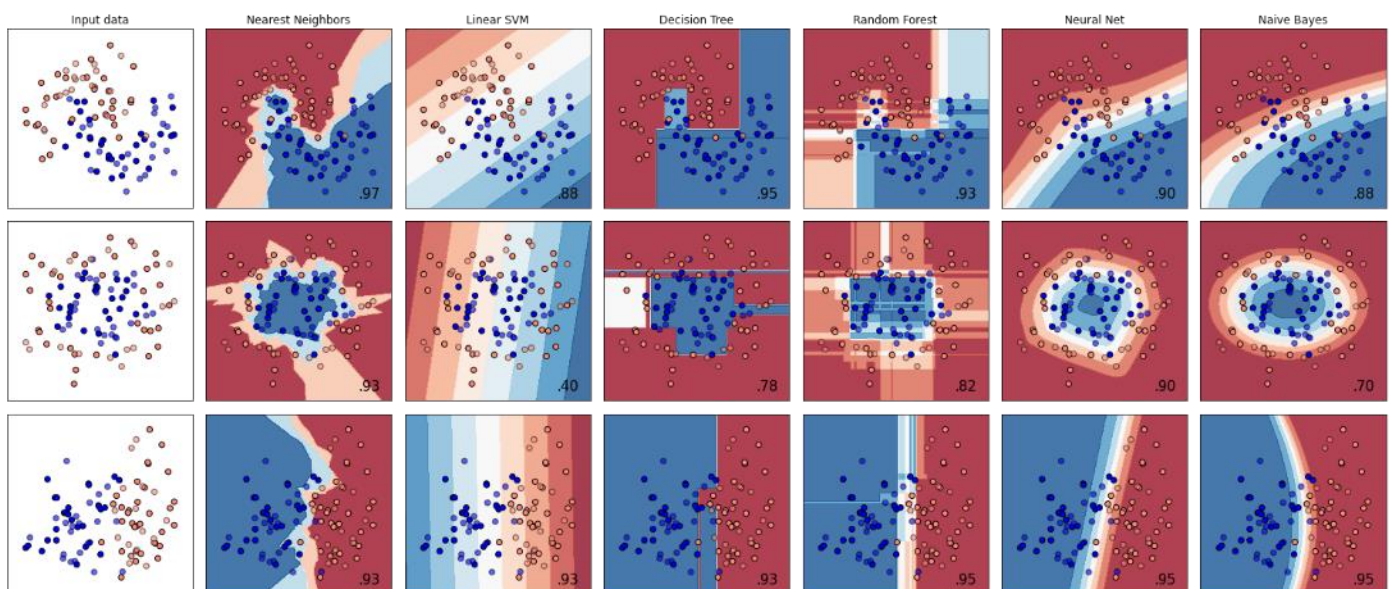


Einführung in Maschinelles Lernen und künstliche Intelligenz

Von Thomas Jörg, thomas@iludis.de



1: In dieser Tabelle sehen Sie die Klassifizierungsergebnisse der verschiedenen ML-Klassifikatoren, die in diesem Dokument verwendet werden. Auf der linken Seite ist der Originaldatensatz dargestellt. Abgeleitet von https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Eine Sammlung von Lektionen zur Einführung in die grundlegenden Konzepte der
Künstliche Intelligenz/Maschinelles Lernen im Schulunterricht.

Zielgruppe:

- Schüler der allgemeinbildenden Gymnasien ab 14 Jahren
- Schüler und Azubis der Berufsschulen,
- Teilnehmer der ZQ KI der IHK.

Inhaltsübersicht

Inhaltsübersicht	2
Warum sollten sich Schüler mit den Konzepten der künstlichen Intelligenz beschäftigen?	3
Ein Interview mit GPT-4, ob sich eine Beschäftigung mit klassischem ML noch lohnt:	4
Lektion 1: ML/AI/DL Was ist das?	5
Lektion 2: Klassifizierung und Vorhersage, wie steht es um den Datenschutz?	7
Lektion 3: ML/AI/DL: Was sind die Gemeinsamkeiten und Unterschiede?	8
Lektion 4: Was ist der Unterschied zwischen ‚klassischer Programmierung‘ und ‚KI‘?	9
Lektion 5: NLP und das GPT-3 zum Thema "Was ist maschinelles Lernen"?	15
Lektion 6: Teachable Machine und Scratch	18
Lektion 7: ML-Prinzipien unplugged	23
Einige nützliche Ressourcen für den KI-Unterricht	27
Lektion 8: Entscheidungsbäume	29
Lektion 9: Die logistische Regression für die Klassifizierung	37
Lektion 10: Naive Bayes Klassifikator	43
Lektion 11: Lineare Regression	55
Lektion 12: Unterscheiden Sie zwischen Korrelation und Kausalität!	63
Lektion 13: k-Means-Clustering.	64
Lektion 14: k-Nächster Nachbar	73
Lektion 15: Support-Vektor-Maschinen	80
Lektion 16: Orange Data Mining Projekt: Preisvorhersage für einen Gebrauchtwagen	84
Lektion 17: Orange Data Mining Project, wie überlebt man die Titanic?	92
Lektion 18: Zusammenfassung der gesamten Lektion	98
Didaktische Überlegungen zum Schluss	99



Abb. 1 Bild [[Gemeinfrei](#)] erzeugt mit [DALL-E](#); Prompt „a japanese girl and a robot, high quality photo“ von Jörg [[CC BY-SA 4.0 International](#)]

Warum sollten sich Schüler mit den Konzepten der künstlichen Intelligenz beschäftigen?

Einige einleitende Gedanken für Lehrkräfte

Künstliche Intelligenz entwickelt sich zu einer Technologie, die unser aller Leben derzeit grundlegend verändert, da sie den Kern des menschlichen Selbstverständnisses besetzt: Die Fähigkeit zu denken und sich dadurch weiterzuentwickeln. Die aktuelle Entwicklung der Sprachmodelle lässt erahnen, dass wir bald nicht mehr die intelligenteste Spezies auf diesem Planeten sein werden. Da sich KI ins Zentrum unseres Lebens rückt, wird das Wissen über KI damit zu einer existenziellen gesellschaftlichen Frage, die das Leben jedes Bürgers betrifft.

Elektronische Sensoren, die überall zu finden sind - von Temperatursensoren in Wohnräumen über RFID-Sensoren in der Fertigungsindustrie bis hin zu Blutdrucksensoren in Smartwatches - liefern kontinuierliche Datenströme über uns und unsere Umgebung. Diese Daten sind das "Rohöl" der KI, denn KI-Systeme sind datengesteuert: Eine KI lernt aus versteckten Mustern, die sie in den Daten entdeckt, erfasst deren Bedeutung und plant selbstständig daraus resultierende Aktionen. Manchmal mit, manchmal ohne menschliches Zutun. Auf diese Weise wird das, was früher "Big Data" war, zu "Smart Data".

Die junge Generation wird mit autonomen, intelligenten Systemen aufwachsen. Um diesen "Black Box"-Systemen nicht als einfacher Nutzer ausgeliefert zu sein, sollte jeder Schüler so früh wie möglich lernen, auf welchen Prinzipien künstliche Intelligenz beruht und wie man sie eigenständig gestaltet, nutzt und bewertet. Das Wissen um den Umgang mit KI wird damit zu einer der Kernkompetenzen des 21. Jahrhunderts. Relevante Fragen, die als Unterrichtsziele formuliert werden können, sind z.B.:


- Wie lernt eine Maschine?
- Was ist die Bedeutung von Daten? Und woher kommen die Daten?
- Wo liegen die Grenzen des maschinellen Lernens?
- Wie kann ich sie selbst beeinflussen?
- Wie kann ich mich unter bestimmten Umständen vor ihr schützen? Wo ist das notwendig?
- Welche Bedeutung werde ich als Mensch in einer Welt haben, in der Maschinen denken?

Das Erste, was ein junger Mensch verstehen sollte: Es ist ein Prozess, der mit der Aufnahme von Daten beginnt. Die gesammelten Daten werden in einem iterativen Lernzyklus verarbeitet, gefolgt von einer Bewertung des Lernerfolgs. Ist dieser Zyklus abgeschlossen, ist der nächste Schritt die Anwendung der gelernten Inhalte. Das "Herzstück" der künstlichen Intelligenz ist der jeweils verwendete Datenverarbeitungsalgorithmus, der je nach Problemstellung ausgewählt und angepasst wird.

Diese Sammlung von KI-Algorithmen verhält sich ähnlich wie ein großer Werkzeugkasten: Hier gibt es viele verschiedene Werkzeuge, die für eine bestimmte Aufgabe "passen". Sowohl der Prozess als auch die Werkzeuge sollten im Klassenzimmer gelehrt werden. Ziel ist es, dass ein junger Mensch das Wesen der KI versteht, indem er es selbst tut und erlebt.

Ein Interview mit GPT-4, ob sich eine Beschäftigung mit klassischem ML noch lohnt:

TH Liebe GPT-4, moderne KI-Systeme wie zum Beispiel Transformermodelle sind von ihrer Komplexität sehr weit weg von den Ursprüngen des maschinellen Lernens: Algorithmen wie kNN oder Decision Trees haben kaum mehr etwas gemeinsam mit hochkomplexen neuronalen Netzen. Ist es überhaupt noch sinnvoll, sich als Schüler mit maschinellern Lernen zu beschäftigen? Formuliere präzise, kurze und unterhaltsame Statements für unsere Leser:

 Natürlich! Hier sind einige präzise, kurze und unterhaltsame Statements zum Thema:

1. **Historische Wurzeln:** Das Studium der Grundlagen ist wie das Erlernen der Grundakkorde in der Musik. Bevor man ein Rockstar wird, muss man erst einmal "Twinkle Twinkle Little Star" spielen können!
2. **Verstehen statt Auswendiglernen:** Klassische Algorithmen wie kNN oder Decision Trees helfen dabei, die Intuition hinter maschinellern Lernen zu entwickeln. Es geht nicht nur darum, einen Button zu drücken und die Magie geschehen zu lassen!
3. **Nicht jeder braucht einen Jet:** Während moderne KI-Modelle wie Jets sind, können einfache Algorithmen wie Fahrräder oft genug sein, um von A nach B zu kommen.
4. **Kleine Daten, großer Impact:** Nicht alle Datenmengen sind riesig. Für kleinere Datenmengen können einfachere Modelle oft effektiver und schneller sein.
5. **Der Spaßfaktor:** Es gibt etwas Befriedigendes daran, einen Algorithmus von Grund auf zu erstellen und zu sehen, wie er funktioniert. Es ist wie das Bauen eines Lego-Schlusses statt nur eines vorgefertigten Spielzeugs.
6. **Wachsende Komplexität:** Wer mit den Grundlagen beginnt, hat es später leichter, komplexere Konzepte zu verstehen. Es ist wie das Lernen von Mathematik: Man fängt mit Addition an und arbeitet sich bis zur Algebra hoch.

Zusammengefasst: Maschinellern Lernen ist eine Reise, und jede Reise beginnt mit einem ersten Schritt. Auch wenn moderne Modelle beeindruckend sind, gibt es immer noch Wert im Verstehen der Grundlagen. Es ist wie beim Kochen: Man muss die Grundzutaten kennen, bevor man ein Meisterkoch wird!

Lektion 1: ML/AI/DL Was ist das?

Was die Schüler in dieser Lektion lernen sollten:

Die KI ist überall um uns herum, wir finden sie - manchmal unsichtbar, versteckt - an vielen Orten und in vielen Anwendungen. Es geht um Vorhersagen, Prognosen, Mustererkennung und regelbasierte Synthese neuer unbekannter Dinge (Style-GANs oder Transformer Models). Die Schüler sollten die Bedeutung der KI für ihr Leben verstehen.



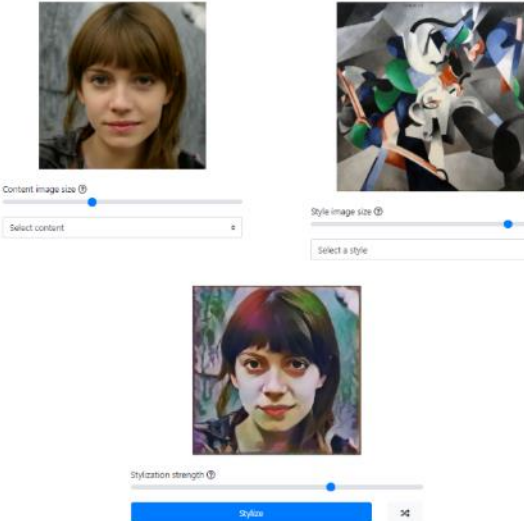
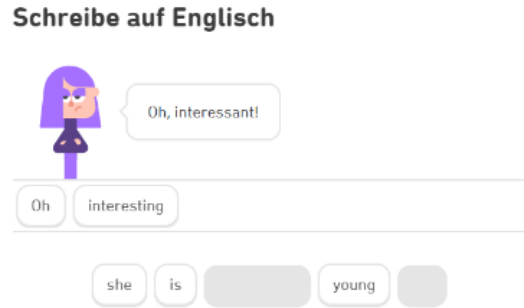
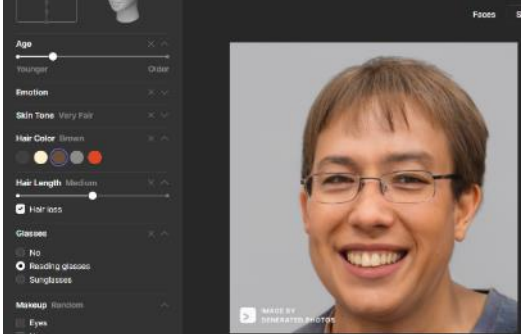
What if I told you, it's machine learning?
 What if I told you these people never existed?
 Visualisierung StyleGAN2 / Karras et al. and Nvidia / CC BY-SA 4.0
 Visualisierung Owlsmcgee / CC BY-SA 4.0

Mögliche Aktivitäten und Aufgaben der Schüler:

Die folgenden AI-Beispiele auf diesen Webseiten können für kurze Einführungspräsentationen verwendet werden. Eine Schülergruppe kann zu diesem Zweck in kleinere Gruppen von etwa zwei bis drei Schülern aufgeteilt werden. Jede Gruppe erhält eine dieser Webseiten zum Thema. Mögliche Fragen sind:

1. kurz darlegen, was der Zweck dieser KI ist.
2. Wie funktioniert diese KI bzw. welche Technologien werden hier eingesetzt?
3. Welche Probleme können damit gelöst werden?
4. Welche menschlichen Tätigkeiten könnten durch sie ersetzt werden?
5. Was sind die Möglichkeiten und Gefahren?

Thema	URL	Bildschirmfoto
Demo von handgeschriebenen Ziffern Erkennung	https://www.snaplogic.com/machine-learning-showcase/handwritten-digit-recognition	<p>The screenshot shows a simple interface for digit recognition. At the top, a blue handwritten digit '4' is displayed inside a black square box. Below the box are two green buttons: 'CLEAR' and 'READ'. Underneath the buttons, the text 'Your number is 4.' is displayed.</p>
Vorhersage von Zeichnungen	https://quickdraw.withgoogle.com/#	<p>The screenshot shows the Google Quick, Draw! interface. At the top, it says 'Gut gezeichnet!' (Well drawn!). Below that, it says 'Unser neuronales Netzwerk hat 6 deiner Zeichnungen erkannt. Wähl eine aus, um zu erfahren, was das genau war.' (Our neural network has recognized 6 of your drawings. Choose one to find out what it was exactly). There are six small boxes, each containing a drawing and a label: a hammer (✓ Hammer), a knife (✓ Knife), a fan (✓ Fan), a hockey stick (✓ Eishockeyschläger), a beach ball (✓ Beachball), and a mushroom (✓ Pilz). At the bottom, there is a button that says 'Zeichnungen teilen' (Share drawings).</p>

<p>Stil-Transfer</p>	<p>https://reiinakano.com/arbitrary-image-stylization-tfjs/</p>	
<p>Online-Sprachtraining (GPT-3)</p>	<p>https://www.duolingo.com/learn</p>	<p>Schreibe auf Englisch</p> 
<p>Text-Synthese (GPT-2)</p>	<p>https://textsynth.com/playground.html</p>	<p>TextSynth Home Playground Documentation Login Sign Up</p> <p>Text completion using large language models. GPT-2 is a 6 billion parameter English model. It has some abilities in other languages too, including computer languages. Boris is a fine-tuned version of GPT-2 for the French language. More information is available in the documentation.</p> <p>Type a text and let the neural network complete it. Each try returns a different randomly chosen completion.</p> <p>Model: GPT-2 (6B) top-k: 40 top-p: 0.9 temperature: 1.0 max tokens: 200 stop at: _____</p> <p>Cooking instructions</p> <p>Making an chocolate cookie is simple!</p> <p>1.</p> <p>Generate another More</p> <p>Completed Text:</p> <p>Making an chocolate cookie is simple!</p> <p>1. Make your dough. (Or if you'd rather, go here for a basic chocolate cookie dough recipe) Mix up the dough. Take a look at this picture for an example of a basic chocolate cookie dough.</p>
<p>GAN: Synthese von menschlichen Gesichtern</p>	<p>https://www.whichfaceisreal.com/ https://generated.photos/</p>	
<p>Texterkennung und Übersetzung.</p>	<p>https://deepl.com/translator</p>	

Lektion 2: Klassifizierung und Vorhersage, wie steht es um den Datenschutz?



Abb. 2 Bild [Gemeinfrei] erzeugt mit DALL-E; Prompt „several surveillance cameras on the red background, poster style“ von Jörg [CC BY-SA 4.0 International]

Was Schüler lernen sollten

„Daten werden von mir als Individuum über viele verschiedene Kanäle gesammelt: Über soziale Netzwerke, durch Bewegungen mit meinem Smartphone und sogar durch Bewegungsprofile in Supermärkten. All diese Daten werden von trainierten künstlichen Intelligenzen genutzt, um Persönlichkeitsprofile von mir zu erstellen. Diese machen es möglich, mein Denken und zukünftiges Handeln für externe Institutionen transparent zu machen. Und oft wissen andere über meine Gewohnheiten Bescheid, die mir selbst nicht bewusst sind. Hier muss man lernen, sich zu schützen.“

Mögliche Aktivitäten und Aufgaben der Schüler:

Die KI kann Ihr Verhalten aufgrund Ihrer (manchmal unbewussten) Gewohnheiten vorhersagen. Und KI ist oft unsichtbar, agiert im Hintergrund - daher ist es möglich, dass Sie eingestuft werden und nichts davon wissen.

Diskutieren Sie die folgenden drei Beispiele. Lesen Sie zunächst die Zeitungs-/Blogartikel und fassen Sie sie zusammen:

1. Welche Rolle spielt dabei die künstliche Intelligenz?
2. Sind sich die betroffenen Personen dessen bewusst?
3. Ist die Vorhersage, die die KI trifft, zum Nutzen der betroffenen Person?
4. Wie würden Sie sich fühlen, wenn eine KI-Vorhersagen über Ihren Charakter oder Ihr zukünftiges Verhalten machen würde?
5. Welche Möglichkeiten habe ich, darauf Einfluss zu nehmen?

"Wie Target herausfand, dass ein Teenager-Mädchen schwanger war, bevor es ihr Vater tat"

<https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=226677fa6668>

Gaggle: Echtzeit-Überwachung von Highschool-Schülern durch eine künstliche Intelligenz in den USA

<https://www.gaggle.net/>

<https://www.the74million.org/article/gaggle-surveillance-minneapolis-families-not-smart-ai-monitoring/>

<https://www.buzzfeednews.com/article/carolinehaskins1/gaggle-school-surveillance-technology-education>

Facebook: Die Entstehung von Liebe - Vorhersage von persönlichen Beziehungen zwischen Facebook-Nutzern.

https://m.facebook.com/nt/screen/?params=%7B%22note_id%22%3A10158928005273415%7D&path=%2Fnotes%2Fnote%2F&refsrc=deprecated&_rdr

Lektion 3: ML/AI/DL: Was sind die Gemeinsamkeiten und Unterschiede?

Was Schüler lernen sollten

"Grundsätzlich ist maschinelles Lernen eine Technik, mit der Computer auf der Grundlage früherer Beobachtungen Vorhersagen treffen können. Die KI-Technologie (Algorithmen) gibt Computern die Fähigkeit zu lernen, ohne explizit programmiert zu werden. Es ist ein anderer und unabhängiger Ansatz zur Nutzung von Computern im Vergleich zur traditionellen Programmierung. "

Mögliche Schüleraktivitäten und Aufgaben

Glauben Sie, dass es sich um künstliche Intelligenz handelt?	Warum glauben Sie das?	Welche Daten werden benötigt, damit es funktioniert?
Bild- und Gesichtserkennung		
E-Mail-Spam-Filterung		
Analyse des Risikos von Herzkrankheiten		
Klassifizierung der Gebrauchtwagenpreise		
Vorhersage des Energiebedarfs im nächsten Winter		
Berechnung der Ausfallwahrscheinlichkeit meines Autos		

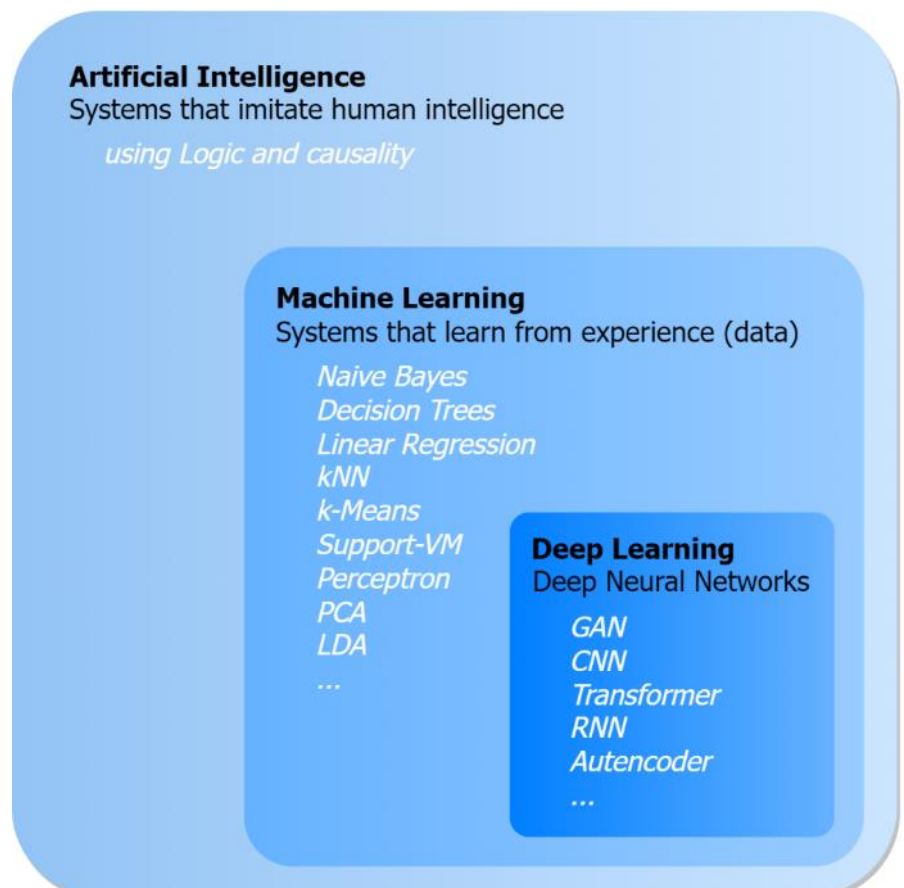
Was ist der Unterschied zwischen KI, maschinellem Lernen und Deep Learning?

Künstliche Intelligenz ist der Oberbegriff für alle Arten von maschinellem Verhalten, das die menschliche Intelligenz imitiert.

Es kann so interpretiert werden, dass alle Methoden des maschinellen Lernens eine Ebene logischer Schlussfolgerungen hinzufügen.

Maschinelles Lernen besteht aus Methoden (Algorithmen genannt), die aus Daten lernen. Daher müssen Daten gesammelt und aufbereitet werden. ML analysiert die Daten, um Muster und Regeln zu finden.

Deep Learning ist in der Regel der Bereich der neuronalen Netze.



Lektion 4: Was ist der Unterschied zwischen ‚klassischer Programmierung‘ und ‚KI‘?

Was Schüler lernen sollten

Wie unterscheidet sich die Computerprogrammierung mit Programmiersprachen (wie Java, Python oder Scratch) von Anwendungen des maschinellen Lernens? Was sind die Grenzen der traditionellen imperativen Programmierung?

Mögliche Aktivitäten und Aufgaben der Schüler:

Die Schüler sollen einem Computer beschreiben, was eine Katze ist.

Es hat spitze Ohren, es läuft auf vier Pfoten, es miaut, wenn es Hunger hat, und es fängt Mäuse. In klassischen Programmiersprachen werden Attribute und Methoden verwendet, um Objekte zu beschreiben:

- Die Katze hat vier Pfoten, zwei Augen und spitze Ohren.
- Katze kann klopfen, laufen, schnüffeln.

Um Katzen in der realen Welt zu erkennen, reicht das nicht aus, denn auch Hunde können die meisten dieser Eigenschaften haben. Anhand der bekannten Kaggle-Datensammlung "Cats and Dogs" sollen die Schüler herausfinden, wie diese beiden Tierarten unterschieden werden können. Die Frage wird wahrscheinlich schwer zu beantworten sein.

Bei der künstlichen Intelligenz lernt ein Computer aus Erfahrung: Die (leicht unterschiedlichen) Merkmale vieler, vieler Katzen werden gesammelt, und der Computer wertet sie aus.

Teil I: Brainstorming: Beschreibe einem Computer, was eine Katze ist

Datensatz für Katzen und Hunde:

<https://www.kaggle.com/datasets/tongpython/cat-and-dog>



Du sollst einem Computer erklären, wie er eine Katze und einen Hund erkennt.

Sammeln Sie einige Stichworte, die beide Tierarten beschreiben.

Abb. 3 Bilder aus dem Datensatz
<https://www.kaggle.com/datasets/biaiscience/dogs-vs-cats> [Database
Contents License (DbCL) v1.0]

Die Aufgabe des Lehrers besteht darin, keine Verallgemeinerungen zuzulassen, sondern auf die Besonderheiten und Grenzen der Computerprogrammiersprachen zu achten. Eine Sache muss durch Eigenschaften beschrieben werden, nicht durch Näherungen.

Zulässig ist eine Aussage wie die folgende:

"Eine Katze hat ein Fell."

Unzulässig sind Aussagen wie:

"Ein Hund muss manchmal abends raus" oder "Eine Katze kann Hunde oft nicht leiden".

Wie können Sie diese unterscheiden?

Nach Ihren Beschreibungen: Kann der Computer nun Hunde und Katzen sicher erkennen?
Und darüber hinaus auch zwischen ihnen unterscheiden?

Sammeln und formulieren Sie Probleme, die ein Computer beim Erkennen und Unterscheiden haben könnte.



Abb. 4 Bilder aus dem Datensatz
<https://www.kaggle.com/datasets/biaiscience/dog-s-vs-cats> [Database Contents License (DbCL) v1.0]

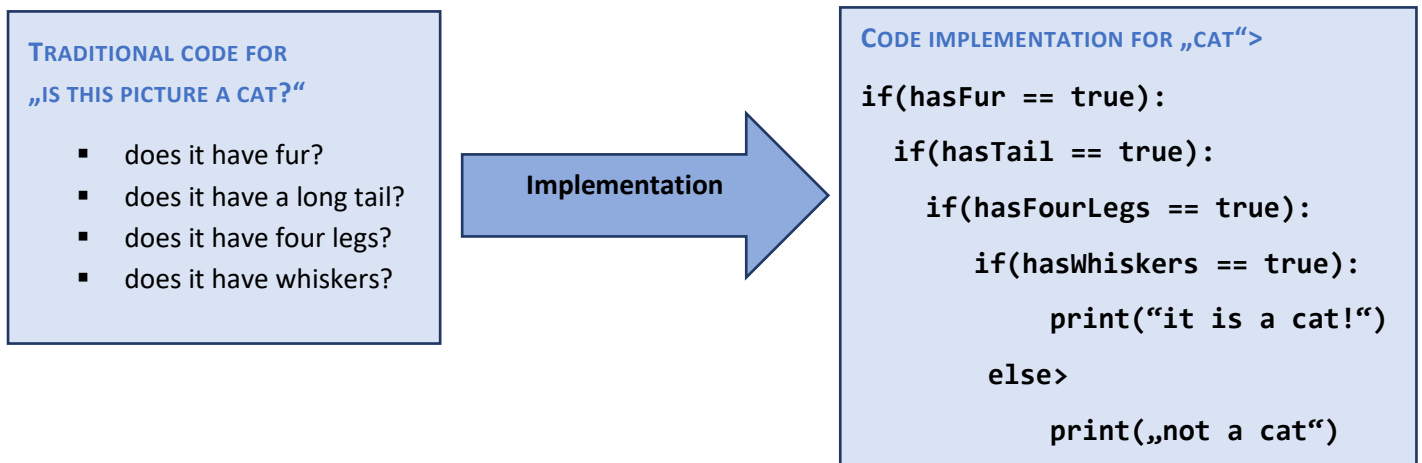
Teil II: Analyse mit Erläuterungen

[inspiriert von "Techgirlz: Wie Computer lernen", <https://www.techgirlz.org/>]



DALL-E2

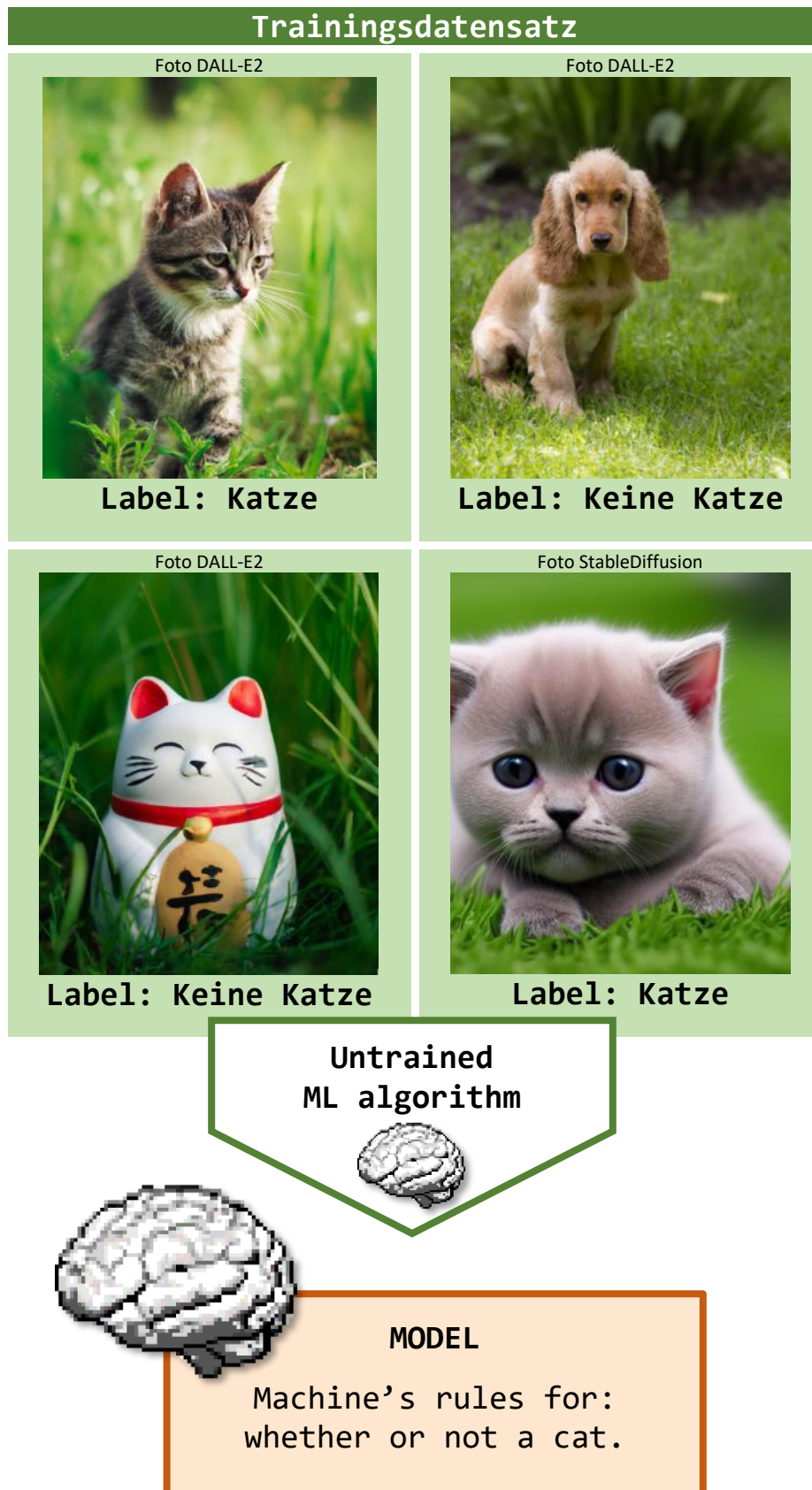
Nehmen wir eine Katze. Wenn wir wollen, dass ein Computer eine Katze identifiziert und sie z. B. von einem Hund oder einem Bären unterscheidet, müssen wir die Regeln implementieren - wenn wir die traditionelle Computerprogrammierung verwenden



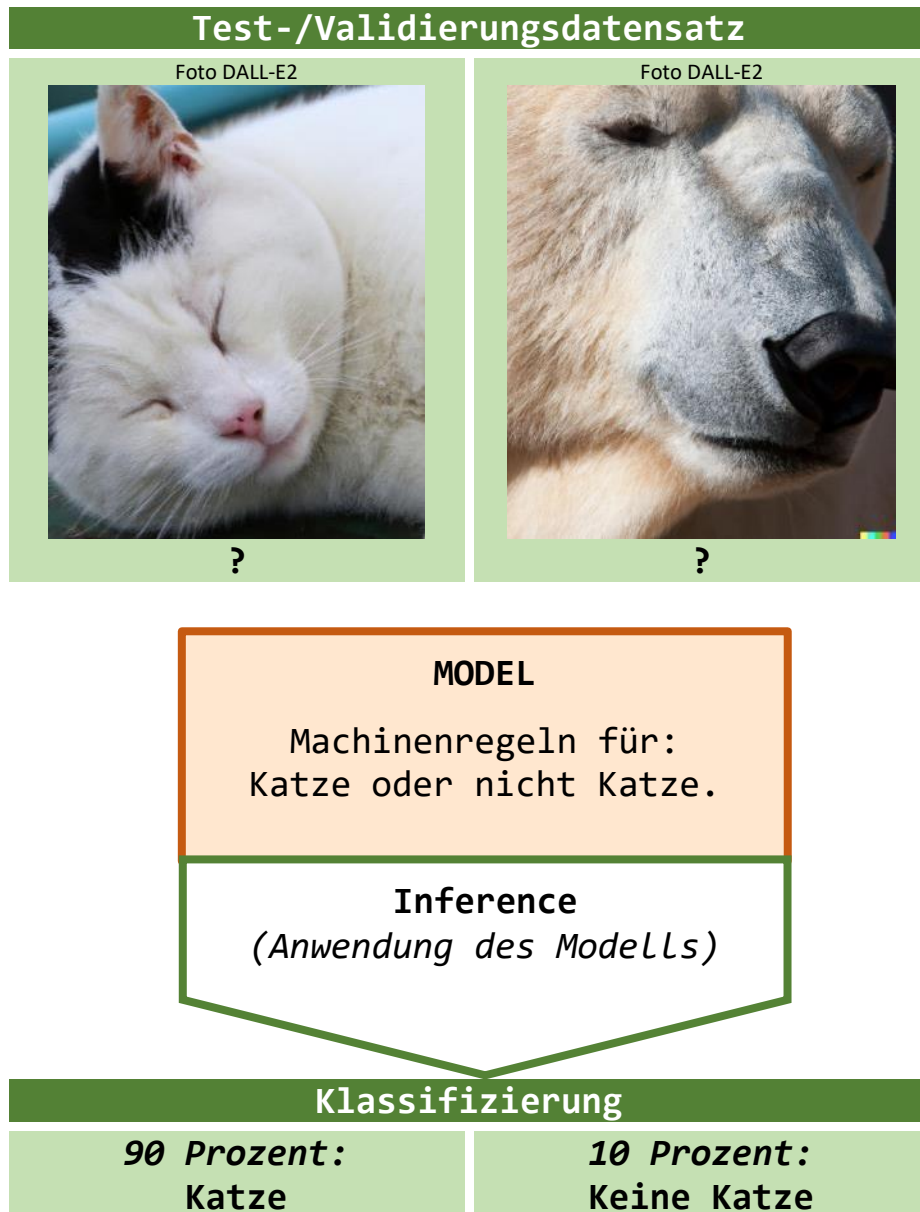
Bei der klassischen Programmierung würde die Klassifizierung durch exakten Vergleich mit genau festgelegten Parametern erfolgen. Im obigen Beispiel werden boolesche Operationen - d.h. Entscheidungen mit den beiden Möglichkeiten ja oder nein - verwendet. Besitzt das zu klassifizierende Objekt bestimmte, scharf definierte Eigenschaften oder nicht? Die Ergebnisse sind also ebenso scharf, exakt und meist viel zu eng gefasst. Das Erkennen einer beliebigen Katze in einer beliebigen Perspektive kann so nicht funktionieren.

Teil III: Prinzip des Trainings eines ML-Algorithmus mit Daten.

Wir nehmen einen Trainingsdatensatz, bei dem alle Daten vorklassifiziert sind. Dieser Datensatz wird als "gelabelter Datensatz" bezeichnet. Wir wenden darauf einen Algorithmus für maschinelles Lernen an (*in diesem speziellen Fall zum Beispiel ein Convolutional neural Network*):




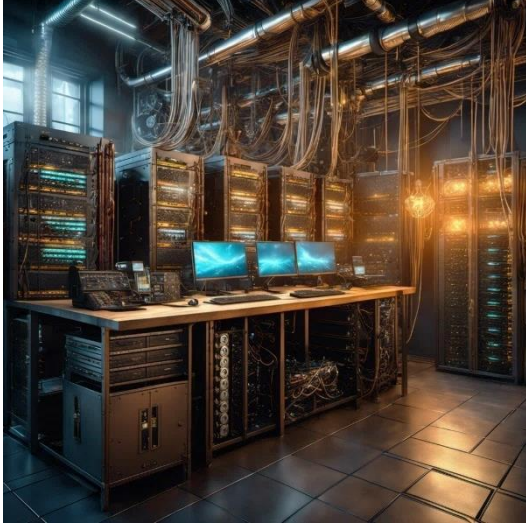
Teil IV: Testen und Validieren des vortrainierten Modells



Ein gut trainiertes maschinelles Lernmodell verallgemeinert: Es erkennt eine bestimmte Klasse von Dingen auf der Grundlage ähnlicher Merkmale und lässt daher Ungenauigkeiten zu. ML-Modelle arbeiten meist mit Wahrscheinlichkeiten: Es ist nicht möglich, mit 100%iger Sicherheit zu entscheiden, ob ein Bild eine Katze oder einen Hund zeigt. Bei gut trainierten Modellen sind Wahrscheinlichkeitswerte von 80-95% üblich und für den jeweiligen Zweck meist völlig ausreichend.

Teil V: Welche Zutaten benötigt man, damit das funktioniert?

Wir brauchen 3 Zutaten:

<p>1. Daten. Haufenweise Daten! Und noch besser: Jede Menge gelabelte Daten!</p>  <p>...</p> <p>...</p>	<p>2. Viele Computer mit vielen Ressourcen</p>  <p>Abb. 5 Bild [Gemeinfrei] erzeugt mit DALL-E; Prompt „An incredible complicated setup with servers and switches and routers in a server room, steampunk version, lots of wires, glowing electricity, high end studio photo, ((masterpiece)), (highest quality), cinematic lighting, octane render, hdr “ von Jörg [CC BY-SA 4.0 International]</p>	<p>3. Algorithmen zur Verarbeitung der Daten "Es gibt kein kostenloses Mittagessen"-Theorem" *</p> $Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$ $P(A B) = \frac{P(B A)P(A)}{P(B)}$ $d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$ $J = \sum_{i=1}^m \sum_{k=1}^K w_{ik} \ x^i - \mu_k\ $ $y = w_0 + \sum_{i=1}^n x_i w_i$ <p>...</p> <p>...</p>
---	--	--

*NFL-Theorem: <https://machinelearningmastery.com/no-free-lunch-theorem-for-machine-learning/>

Lektion 5: NLP und das GPT-3 zum Thema "Was ist maschinelles Lernen"?

Was die Schüler lernen sollten:

GPT-3 ist ein von der amerikanischen Non-Profit-Organisation OpenAI entwickeltes Sprachverarbeitungsmodell (die Abkürzung steht für "generative pretrained transformer"). Es nutzt Deep Learning, um Texte zu erstellen, zusammenzufassen, zu vereinfachen oder zu übersetzen. Es ist eines der ersten Modelle, das den Turing-Test (!) bestanden hat. Es ist (nach dem aktuellen Stand von 2021) eines der leistungsfähigsten KI-Modelle weltweit.

Mögliche Aktivitäten und Aufgaben der Schüler:

Die Schüler sollten den Text sorgfältig lesen. Der Inhalt ist weniger wichtig (der übrigens sehr informativ ist). Es geht um die Bewertung der sprachlichen Gestaltung. Diskutieren Sie die folgenden Fragen:

- Kann man hier noch erkennen, ob eine künstliche Intelligenz den Text geschrieben hat oder nicht?
- Und ob man es noch erkennen kann: Woran erkennt man das?
- Wo liegen mögliche Chancen, aber auch Gefahren?
- Was wäre, wenn dieses KI-Modell zum Beispiel dazu verwendet würde, Fake News zu produzieren?
- Wie kann ein solcher Missbrauch verhindert werden?

Gefunden auf R-Bloggers:

<https://www.r-bloggers.com/2021/05/the-most-advanced-ai-in-the-world-explains-what-ai-machine-learning-and-deep-learning-are/>

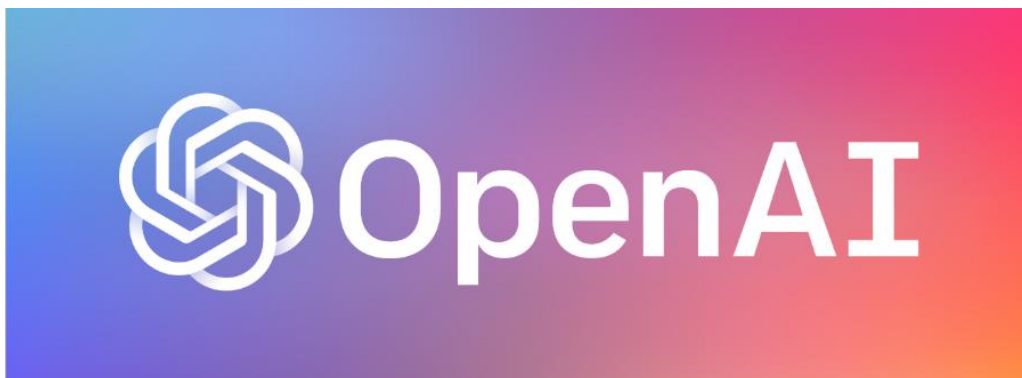


Abb. 6 Open AI Logo [[Apache License 2.0](#)]

Der Text auf der folgenden Seite wurde vollständig von einer künstlichen Intelligenz geschrieben, am 12.Mai 2021.

GPT-3 antwortet auf: "Was ist Künstliche Intelligenz"?

Künstliche Intelligenz (KI), maschinelles Lernen (ML) und Deep Learning (DL) sind verwandte Begriffe, die oft synonym verwendet werden. Es handelt sich aber auch um drei verschiedene und unterschiedliche Konzepte. In diesem Blogbeitrag werden wir künstliche Intelligenz, maschinelles Lernen und Deep Learning definieren und erklären, warum sie alle unterschiedlich sind und wie sie zusammenhängen.

Künstliche Intelligenz

KI ist ein weit gefasster und komplexer Begriff, den es schon seit Jahrzehnten gibt. KI wird verwendet, um ein Konzept oder ein System zu beschreiben, das die kognitiven Funktionen des menschlichen Gehirns nachahmt. Er kann verwendet werden, um eine Situation zu beschreiben, in der Maschinen in einer Weise handeln oder sich verhalten können, die das menschliche Verhalten nachahmt. KI wird häufig verwendet, um ein System zu beschreiben, das aus Erfahrungen lernen und Wissen nutzen kann, um Aufgaben auszuführen, zu denken und Entscheidungen zu treffen.

Es gibt viele verschiedene Arten von KI. Zum Beispiel gibt es Expertensysteme, neuronale Netze und Fuzzy-Logik. In diesem Blogbeitrag werden wir uns auf die verschiedenen Arten des maschinellen Lernens konzentrieren. Ein Modell für maschinelles Lernen ist ein KI-System, das aus einem Datensatz lernen kann und auf der Grundlage der Daten Vorhersagen oder Entscheidungen treffen kann.

Maschinelles Lernen

Maschinelles Lernen ist ein Teilbereich der KI und eine Methode, mit der Algorithmen aus Daten lernen können. Es kann verwendet werden, um Modelle zu erstellen, die zukünftiges Verhalten auf der Grundlage früherer Erfahrungen vorhersagen können. Maschinelles Lernen wird verwendet, um große Datensätze zu analysieren und Muster in den Daten zu finden. Ein Beispiel für ein maschinelles Lernmodell ist ein Spam-Filter, der lernt, zwischen Spam und Nicht-Spam-Nachrichten zu unterscheiden. Es gibt drei verschiedene Arten des maschinellen Lernens. Jede von ihnen wird für eine andere Art von Problem verwendet.

Überwachtes Lernen

Überwachtes Lernen ist die häufigste Form des maschinellen Lernens. Es wird verwendet, um Muster in Daten zu finden und zukünftiges Verhalten auf der Grundlage früherer Erfahrungen vorherzusagen. Beim überwachten Lernen werden die Daten in zwei Teile aufgeteilt, die als Trainingsmenge und Testmenge bezeichnet werden. Die Trainingsmenge wird zum Trainieren des Modells verwendet und die Testmenge zur Bewertung der Genauigkeit des Modells.

Ziel des überwachten Lernens ist es, eine Beziehung zwischen unabhängigen Variablen und abhängigen Variablen zu finden. Die unabhängigen Variablen sind die Dinge, die wir über die Daten wissen. Die unabhängigen Variablen sind zum Beispiel die Merkmale, die zur Beschreibung eines Kunden verwendet werden. Die abhängigen Variablen sind die Dinge, die wir über die Daten wissen wollen. Die abhängige Variable ist zum Beispiel der Gewinn, der mit einem bestimmten Kunden erzielt wird. Die Beziehung zwischen den unabhängigen Variablen und den abhängigen Variablen wird als Modell bezeichnet. Überwachtes Lernen kann verwendet werden, um den Gewinn eines Kunden auf der Grundlage der Merkmale des Kunden vorherzusagen.

Unüberwachtes Lernen

Unüberwachtes Lernen wird verwendet, um Strukturen in den Daten zu finden. Beim unüberwachten Lernen gibt es keinen Trainingssatz. Das Modell wird aus dem Testsatz gelernt. Es kann auch verwendet werden, um Gruppen oder Cluster in den Daten zu finden oder um Anomalien in den Daten zu identifizieren. Unüberwachtes Lernen kann verwendet werden, um Gruppen ähnlicher Kunden zu finden.

Verstärkendes Lernen

Verstärkungslernen ist eine Art des maschinellen Lernens, das dazu dient, gute Aktionen oder Entscheidungen auf der Grundlage von Daten zu finden. Verstärkungslernen wird verwendet, um eine optimale Aktion oder Entscheidung zu finden, die die Belohnung maximiert. Es wird eingesetzt, um die optimale Lösung für ein Problem zu finden. Die optimale Lösung hängt von der Belohnungsfunktion ab. Reinforcement Learning kann zur Optimierung verschiedener Problemtypen eingesetzt werden. So kann es beispielsweise zur Optimierung einer nichtlinearen Funktion oder zur Suche nach der kürzesten Route in einem Netzwerk eingesetzt werden (siehe auch Reinforcement Learning: Life is a Maze).

Tiefes Lernen

Deep Learning ist ein Teilbereich des maschinellen Lernens, der künstliche neuronale Netze verwendet. Künstliche neuronale Netze sind Berechnungsmodelle, die sich an der Architektur des menschlichen Gehirns orientieren. Sie werden verwendet, um Algorithmen zu entwickeln, die aus Daten lernen können (siehe auch Die Magie neuronaler Netze verstehen).

Deep Learning wird verwendet, um Modelle zu erstellen, die Daten klassifizieren oder Muster in den Daten finden können. Deep Learning wird verwendet, um komplexe Aufgaben wie Objekterkennung, Spracherkennung und Übersetzung durchzuführen. Deep Learning ist die beliebteste Art des maschinellen Lernens.

In diesem Blogbeitrag haben wir den Unterschied zwischen künstlicher Intelligenz, maschinellem Lernen und Deep Learning erklärt

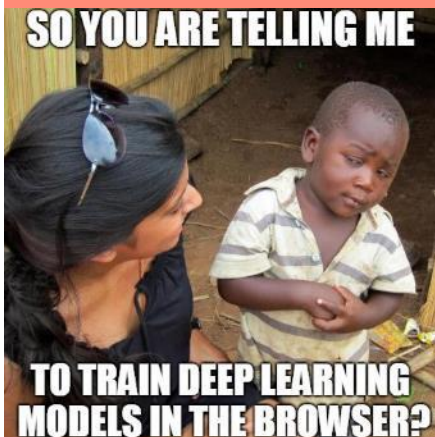


Abb. 7 [Mukul Devichand and Anisa Subedar, 'Sceptical Third World Child': What a viral picture tells us about child poverty in Africa, bbc.com](#)

Lektion 6: Teachable Machine und Scratch

Was die Schüler in dieser Lektion lernen sollten:

Trainieren Sie ein Modell mit einem selbst aufgenommenen Trainingsdatensatz: Um eine Klasse wie eine Tasse oder einen Schwamm zu verallgemeinern, ist es notwendig, viele verschiedene Objekte zu sammeln. Und diese Objekte müssen etwas ungenau sein. Wenn Sie beispielsweise einen Datensatz mit einer Tasse erstellen, müssen Sie die Tasse langsam vor der Kamera hin- und herbewegen, um verschiedene Perspektiven eines Objekts zu erzeugen. Der Algorithmus versucht, allgemeine Merkmale zu finden, die allen Objekten gemeinsam sind. Inferenz - die Vorhersagen/Klassifizierungen, die ein Modell nach dem Training durchführt - arbeitet mit Wahrscheinlichkeiten. In den meisten Fällen wird das Modell nicht mit 100%iger Sicherheit klassifizieren.

Mögliche Aktivitäten und Aufgaben der Schüler:

Die Schüler führen ein paar einfache Schritte durch: Die Schüler wählen mehrere gleichartige Gegenstände aus ihrer unmittelbaren Umgebung aus, z. B. mehrere verschiedene Stifte, Scheren, Tassen, Bücher, Federmäppchen, Brillen usw.

Sie gruppieren diese Gegenstände zunächst in Klassen, programmieren diese Klasse in der Teachable Machine und erstellen Fotos von allen Gegenständen in dieser Gruppe.

Dabei wird bereits intuitiv auf die Datenanreicherung geachtet: Die Objekte werden vor der Kamera leicht verschoben, um die Perspektive zu ändern - dies verhindert eine Überanpassung und ermöglicht eine Verallgemeinerung.

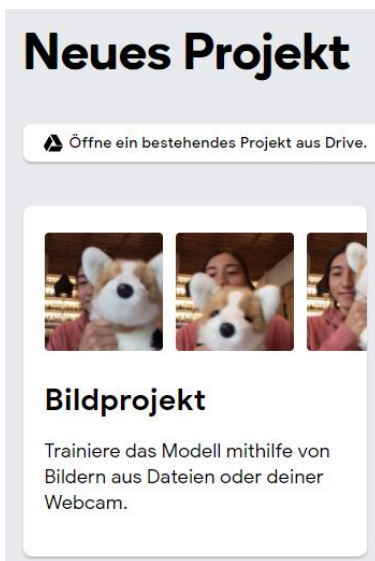
Nach der Datenerfassung trainieren die Schülerinnen und Schüler ihre Modelle und können dann mit ihnen herumspielen. Anschließend kann das Modell exportiert und z.B. in Scratch weiterverarbeitet werden.

Tutorial: TensorFlow-Modelle in Scratch ausführen: <https://dalelane.co.uk/blog/?p=4201>



1: Teachable Machine im Browser aufrufen
<https://teachablemachine.withgoogle.com/>

Abb. 8 Eigenes Screenshot vom Web Browser



2. ein neues Bildprojekt erstellen
<https://teachablemachine.withgoogle.com/train/image>

Es stehen verschiedene Möglichkeiten zur Auswahl - wir wollen ein Modell zur Objektklassifizierung trainieren

Abb. 9 Eigenes Screenshot vom Web Browser

3. Erfassen Sie Ihre Trainingsdaten

Klicken Sie dazu auf das Webcam-Symbol, starten Sie die Kamera und halten Sie ein beliebiges Objekt vor die Kamera. Während Sie die Maustaste gedrückt halten, können mehrere Bilder aufgenommen werden. Sie können eine beliebige Anzahl von Objektklassen festhalten.

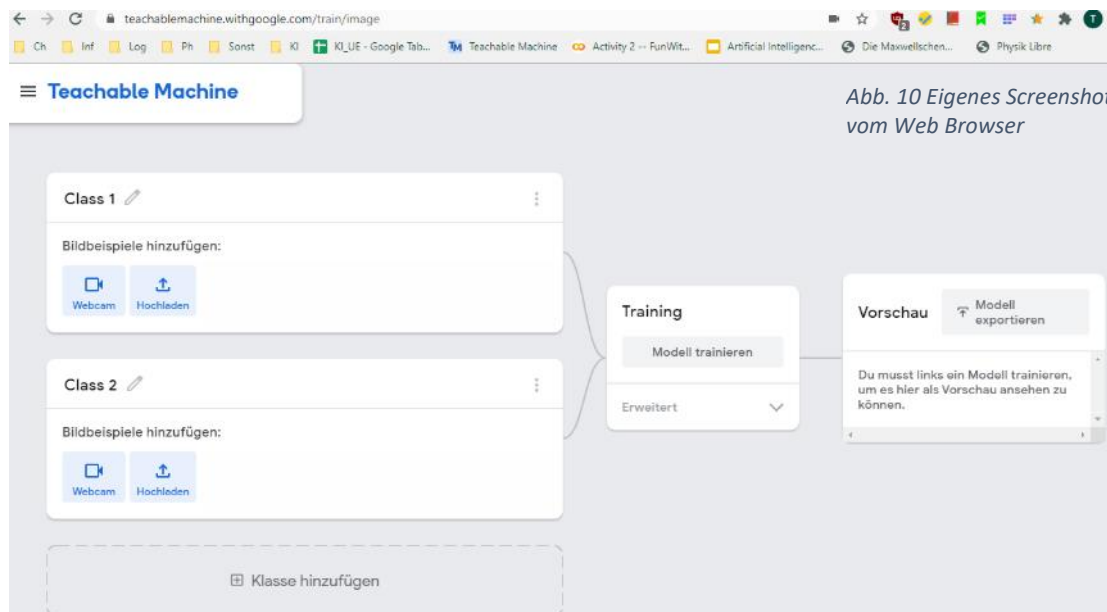
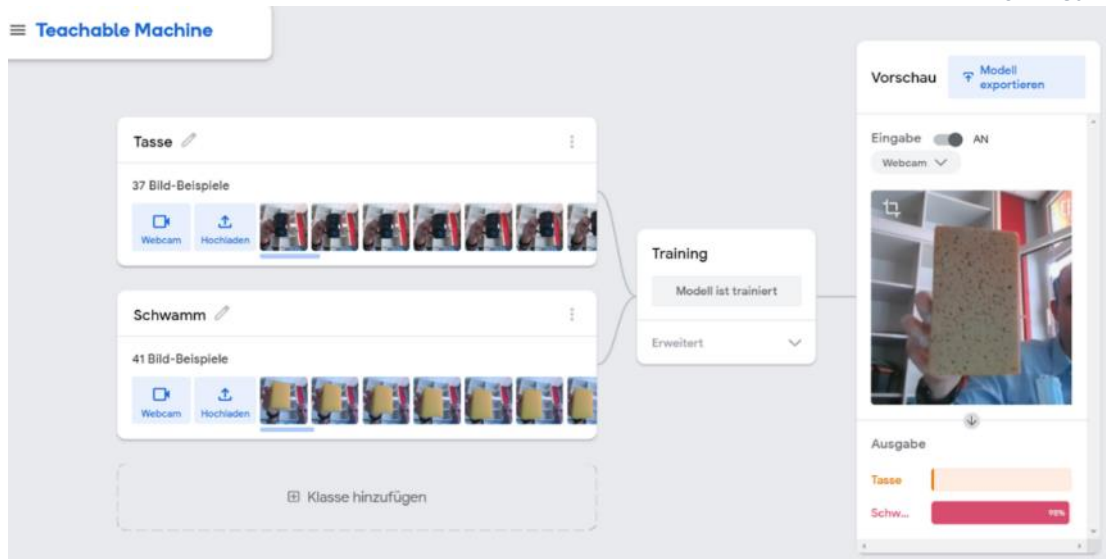


Abb. 10 Eigenes Screenshot vom Web Browser

Dann trainieren Sie das neuronale Netz und testen es:

Abb. 11 Eigenes Screenshot vom Web Browser



Google wird Bilder wie dieses aufnehmen: 224 x224 Pixel.



Abb. 12 Eigenes Screenshot vom Web Browser

Wenn Sie das Modell exportieren, sollten Sie es unter dem Standardlink von Google speichern:

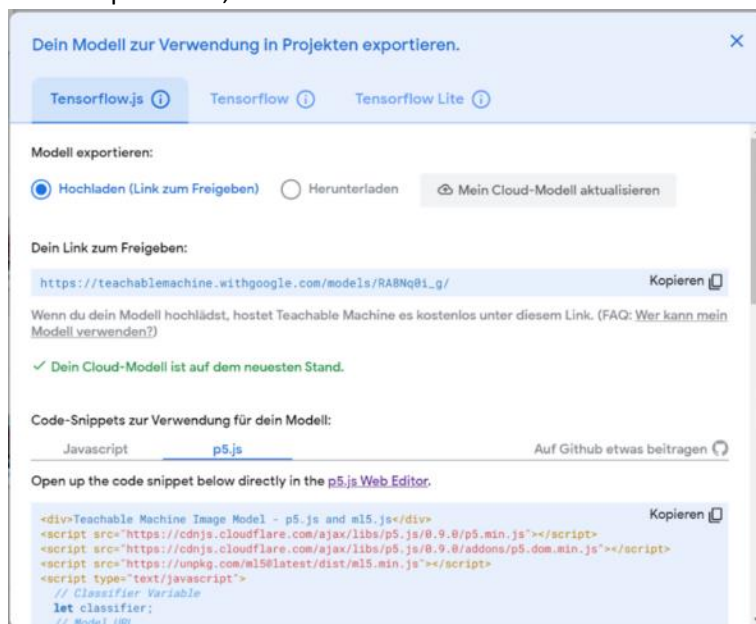


Abb. 13 Eigenes Screenshot vom Web Browser

4. Sie können das Modell in Scratch importieren

<https://machinelearningforkids.co.uk/#!/pretrained>

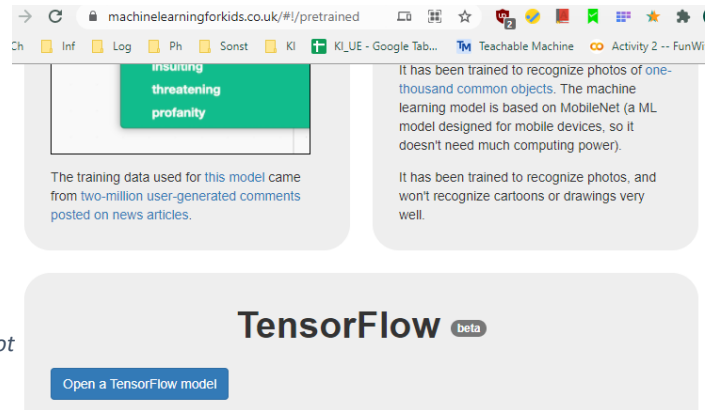


Abb. 14 Eigenes Screenshot vom Web Browser

Dann gibt man die URL deines TeachableMachine-Modells ein. Dieses wird dann in Scratch geladen

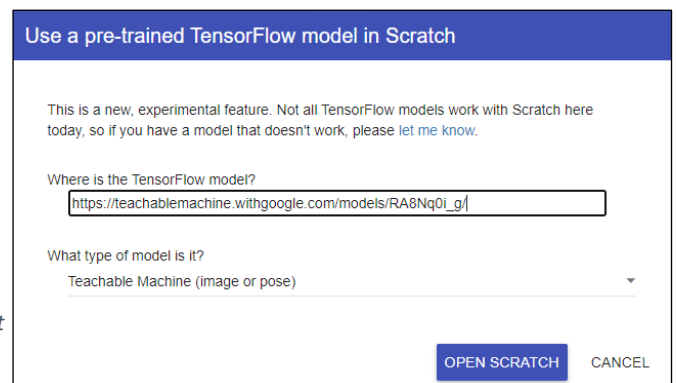


Abb. 15 Eigenes Screenshot vom Web Browser

Jetzt braucht man nur noch die Bildaufnahme als Plugin, dann kann man loslegen:

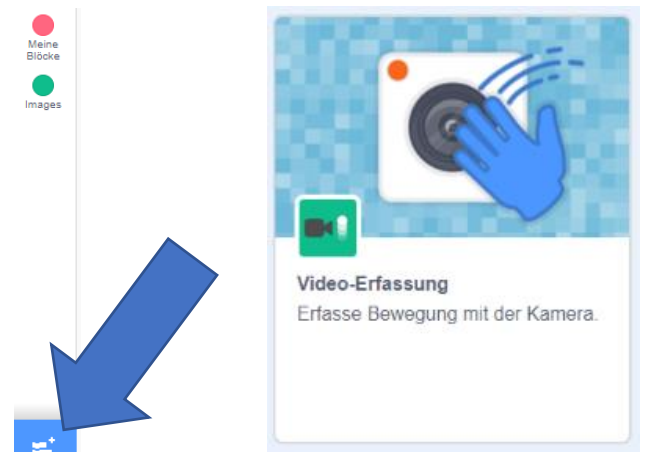


Abb. 16 Eigenes Screenshot vom Web Browser

Hier ist ein einfaches, aber funktionierendes Beispielprogramm.

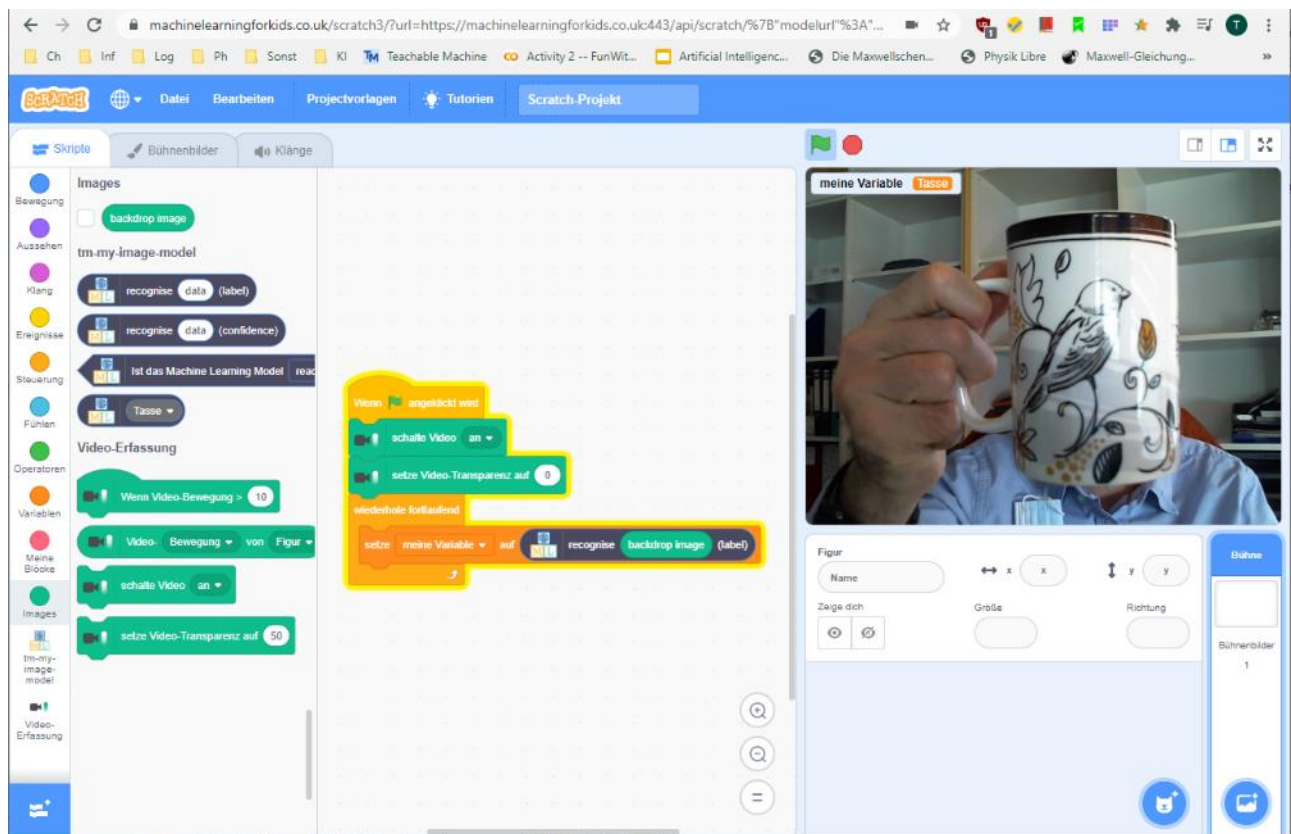


Abb. 17 Eigenes Screenshot vom Web Browser

Lektion 7: ML-Prinzipien unplugged



Abb. 28 Bild [\[Gemeinfrei\]](#) erzeugt mit [DALL-E](#);
 Prompt „sad boxer dog lying on wooden table,
 photorealistic“ von Jörg [\[CC BY-SA 4.0 International\]](#)

Was die Schüler lernen sollten:

KI bedeutet Vorhersage oder Klassifizierung durch ein Modell, das zuvor mit Trainingsdaten trainiert wurde. Diese Trainingsdaten müssen gesammelt werden und basieren auf Merkmalen: Zahlen oder Kategorien, die mit Hilfe von Sensoren gemessen werden müssen, z. B. Masse, Breite/Höhe-Verhältnis, Verformbarkeit, Farbe, Oberflächeneigenschaften und vieles mehr. Das Modell verallgemeinert diese vielen verschiedenen Merkmalswerte und kann nach diesem Trainingsprozess unbekannte Daten vorhersagen oder klassifizieren, was als Inferenz bezeichnet wird. Es reicht nicht aus, nur eine einzige Messung vorzunehmen, da mit nur einer Probe keine Verallgemeinerung oder Regelextraktion möglich ist. Um einen aussagekräftigen Datensatz zu bilden, werden viele Stichproben benötigt.

Mögliche Aktivitäten und Aufgaben der Schüler:

die Schülerinnen und Schüler sollen den Prozess der Modell Bildung nachspielen. dazu lösen sie ein einfaches Problem: wie lassen sich verschiedene alltagsgegenstände voneinander unterscheiden? dazu erhalten sie jeweils ein Ensemble aus einigen Tassen oder schwämmen oder stiften - je nachdem, was im Klassenzimmer bereits vorhanden ist. die Schülerinnen und Schüler sollen eigene Ideen entwickeln, mit welchen messbaren Eigenschaften der Objekte sich die verschiedenen Klassen voneinander trennen lassen. ein Becher ist z.B. in der Regel schwerer als ein Stift. ein schwamm hat in der Regel ein anderes Verhältnis von höhe zu breite als ein Stift. die Schülerinnen und Schüler werden daher aufgefordert, geeignete Eigenschaften zu konstruieren und zu nutzen. nachdem alle Objekte vermessen wurden, werden die Messwerte in Diagramme eingetragen. wenn die Eigenschaften gut gewählt wurden, lassen sich die Punktwolken z.B. durch Trennlinien leicht aufteilen - eine Klassifizierung ist erfolgt.

Teil I: Wie bringt man einer Maschine bei, zwischen einem Schwamm, einer Tasse oder einem Bleistift zu unterscheiden?



Abb. 39 Bild [\[Gemeinfrei\]](#) erzeugt mit [DALL-E](#); Prompt
 „sponge, coffee cup, pencil, grey background,
 photorealistic“ von Jörg [\[CC BY-SA 4.0 International\]](#)

Mögliche Merkmale:

1. Farbe
2. Masse
3. Oberflächentextur
4. Plastizität/Verformbarkeit
5. Verhältnis von Höhe zu Breite

Auswahl der besten Eigenschaften:

Wie einzigartig oder typisch sind die Merkmale? Wie leicht sind sie zu messen?

Verformbarkeit und Masse sind am besten geeignet.

→ Bestimmung dieser beiden Merkmale; Modellierung abgeschlossen.

Teil II: Generierung von Trainingsdaten:

Die Schüler können ihre eigenen Messungen anhand von 3-4 verschiedenen Objekten jeder Klasse durchführen.



Abb. 20 Bild [Gemeinfrei] erzeugt mit DALL-E; Prompt „coffe cup, sponge, digital scale, professional studio photography“ von Jörg [CC BY-SA 4.0 International]

Ergebnis: gelabelte Trainingsdaten

	Masse (g)	Verformbarkeit (%)
Becher 1	240	0
Becher 2	400	5
Becher 3	290	2
Becher 4	120	3
Schwamm 1	12	35
Schwamm 2	23	20
Schwamm 3	7	15
Schwamm 4	30	50

Sind diese Daten ausreichend (ist die Aussagekraft robust genug)?



Was ist, wenn ein Schwamm nass ist und eine viel höhere Masse hat?

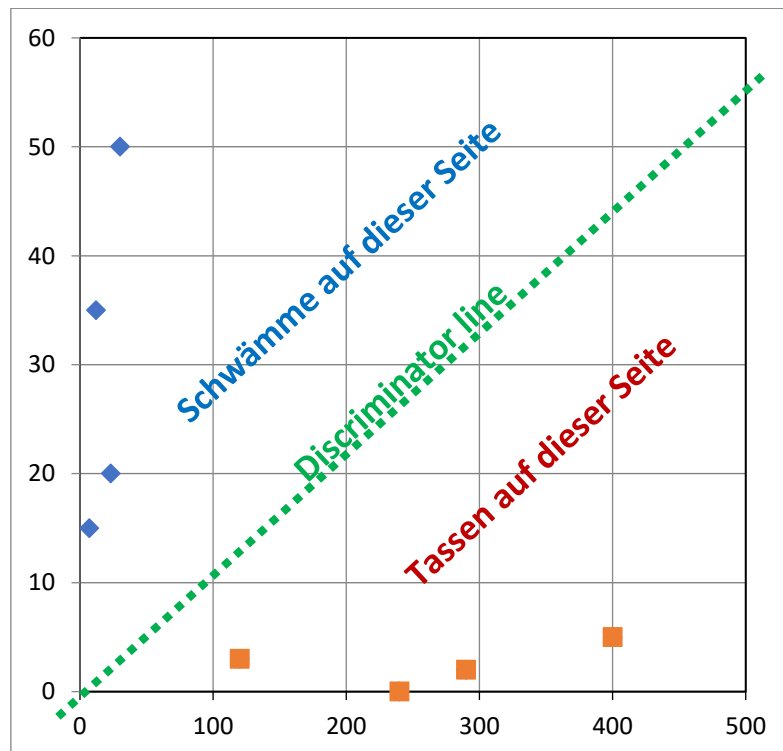
Was ist, wenn ein Becher aus einem verformbaren Material besteht?

Und: Hängt die Verformbarkeit von der Masse ab?



Abb. 21 Bild [\[Gemeinfrei\]](#) erzeugt mit [DALL-E](#); Prompt „colorful coffee cups, professional studio photography“ von Jörg [\[CC BY-SA 4.0 International\]](#)

Teil III: Dargestellt in einem Diagramm wie diesem:

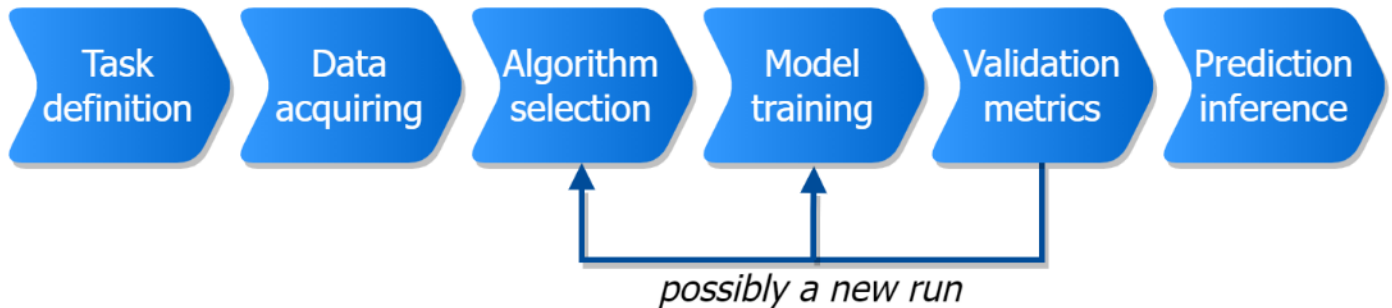


Die Maschine kann dieses Diagramm verwenden, um die beiden Gegenstandsklassen voneinander zu unterscheiden, und sie hat dies aus beschrifteten Trainingsdaten gelernt!

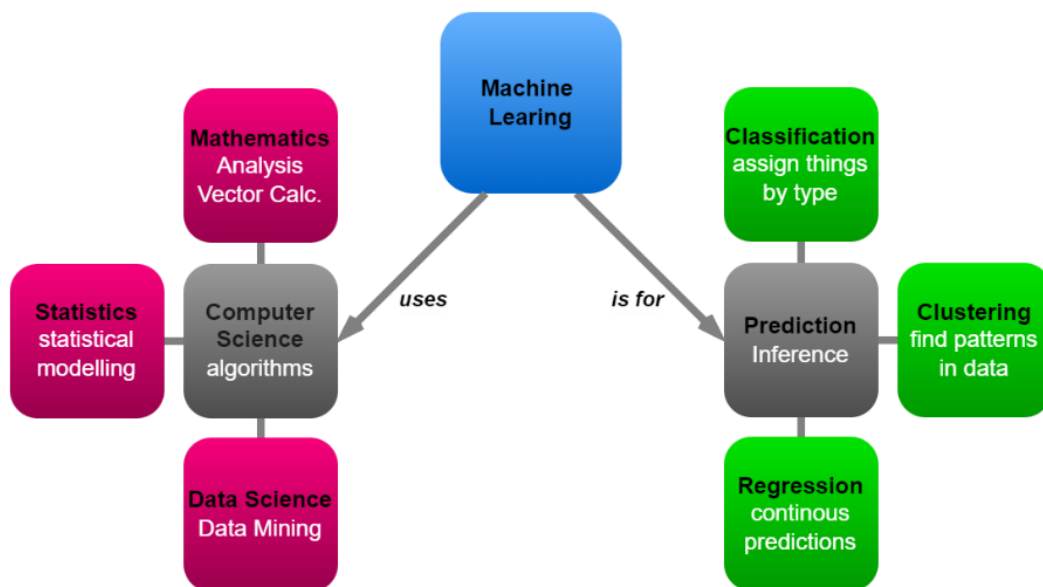
Teil IV: Zusammenfassung und Analyse der Lektion

Prozess des Trainings eines Machine-Learning-Modells: Wir müssen ...

- ... die Aufgabe genau definieren, damit wir die erforderlichen Daten sammeln können.
- ... Daten sammeln, um einen ML-Algorithmus trainieren zu können.
- ... einen passenden ML-Algorithmus finden, der am besten mit den Daten umgehen kann.
- ... das Modell mit den Daten und dem Algorithmus trainieren.
- ... messen, ob das Modell gute Vorhersagen macht, und es möglicherweise optimieren.
- ... das fertige Modell richtig verwenden.



Teil V: Worum geht es beim maschinellen Lernen?



Bei der KI geht es um Vorhersage, Klassifizierung und Clustering.

"Grundsätzlich ist maschinelles Lernen eine Technik, bei der Computer dazu verwendet werden, Dinge auf der Grundlage früherer Beobachtungen zu klassifizieren.

Die KI-Technologie (Algorithmen) verleiht Computern die Fähigkeit zu lernen, ohne ausdrücklich programmiert zu werden.

Es ist ein völlig anderer und unabhängiger Ansatz für die Nutzung von Computern im Vergleich zur traditionellen Programmierung. "

Einige nützliche Ressourcen für den KI-Unterricht

A: Datensätze

Kein maschinelles Lernen ohne Daten! Um als Lehrer mit Schülern die Grundlagen erarbeiten zu können, braucht man pädagogisch gut aufbereitete Datensätze. Hierfür ist es von Vorteil, möglichst viele verschiedene Datenquellen zu kennen, um die für den Unterricht am besten geeigneten herauszufiltern.

Gemeinsame Quellen für Datensätze:	
http://archive.ics.uci.edu/ml/index.php	Universität von Irvine
https://www.kaggle.com/	Größte Open Source Datenbank!
https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research	Liste der Datensatz-Quellen
https://appen.com/datasets-resource-center/	Kommerziell
https://data.world/	Gemeinschaft frei
https://datasetsearch.research.google.com/	Google-Repositories
http://yann.lecun.com/exdb/mnist/	Berühmte" MNIST DB
 Datensätze in Orange Data Mining	Integriert in die Anwendung

B: Simulationen

Menschen verarbeiten visuelle Informationen tausendmal schneller als schriftliche Informationen. Darauf beziehen sich Simulationen und Visualisierungen. Sie sind für Schüler genauso wichtig wie Unterricht oder Projektarbeit.

Visualisierungen, Demonstrationen, Simulationen	Thema
https://www.cs.ryerson.ca/~aharley/vis/conv/flat.html	CNN-Demo
https://poloclub.github.io/cnn-explainer/	CNN-Explainer
https://lecture-demo.ira.uka.de/convolution-demo/	Faltungs-Demo
https://cs.stanford.edu/~karpathy/svmjs/demo/demoforest.html	Entscheidungsbaum & RandomForest
https://playground.tensorflow.org/	Berühmte NeuralNetwork-Demo
https://iludis.de/kMeansClustering/index.html	k-Means-Clustering
https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/	k-Means-Clustering
https://lecture-demo.ira.uka.de/kmeans-demo/preset=1	k-Means-Clustering
https://iludis.de/kNNDemo/index.html	kNN
https://lecture-demo.ira.uka.de/knn-demo/#	kNN
https://iludis.de/LinearRegression/index.html	Lineare Regression
https://phet.colorado.edu/en/simulation/least-squares-regression	Lineare Regression
https://lecture-demo.ira.uka.de/neural-network-demo/	Perceptron-Demo
https://iludis.de/Perceptron/index.html	Perceptron mit Gradient dezent
https://ml4a.github.io/demos/	Repository mit vielen Demos
https://iludis.de/svmDemo/index.html	Support-Vektor-Maschine
https://jgreitemann.github.io/svm-demo	Support-Vektor-Maschine

C) Die Software: Orange Data Mining

In diesem Unterrichtskonzept wird sehr häufig das Softwarepaket "Orange Data Mining" verwendet. Es ermöglicht Studierenden, die noch nie programmiert haben, einen einfachen und intuitiven Zugang zum ML-Workflow. Es ist knotenbasiert und arbeitet daher visuell: Die Studierenden erkennen die bereits durchgeführten Arbeitsschritte, können darauf aufbauend weiter planen und experimentieren und haben darüber hinaus immer das große Ganze im Blick.

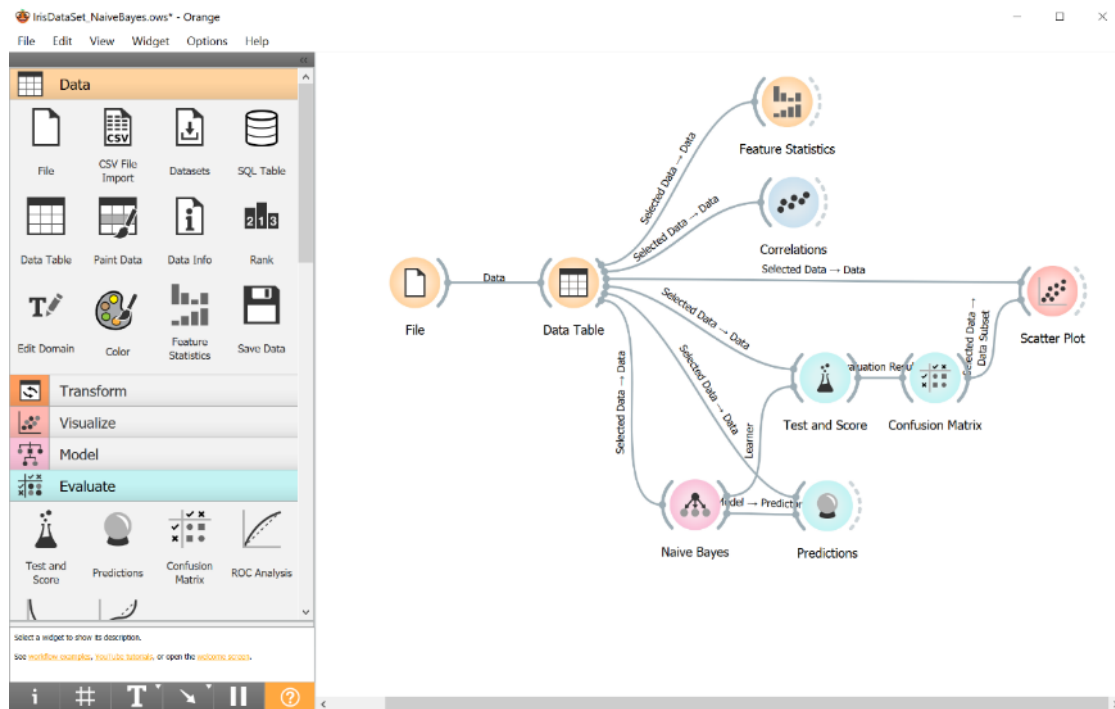


Abb. 22 Eigenes Screenshot Orange Datamining

Die Software ist kostenlos, benötigt keine Installationsrechte und läuft auch auf älteren Rechnern. (Nur bei neuronalen Netzen werden manchmal größere Rechenzeiten benötigt).



Abb. 23 Eigenes Screenshot Orange Datamining

Das Softwarepaket kann hier heruntergeladen werden:
<https://orangedatamining.com/>

Lektion 8: Entscheidungsbäume

Oder: Wie man Daten in Bäume umwandelt

Was Studenten lernen sollten:

Entscheidungsbäume sind der perfekte Einstieg in die Welt der maschinellen Lernalgorithmen. Sie veranschaulichen, wie eine Maschine eine Aufgabe zerlegt, strukturiert und Optimierungsberechnungen einsetzt, um das bestmögliche Modell für Vorhersagen zu erstellen. Darüber hinaus lassen sich Entscheidungsbäume sehr gut interpretieren: Das fertige Modell lässt sich leicht visualisieren und nachvollziehen.

Mögliche Schüleraktivitäten und Aufgaben:

Mögliche Aktivitäten und Aufgaben der Schüler:

Die Schüler lernen die Baumstruktur eines Entscheidungsbaums anhand eines sehr einfachen Einführungsquiz kennen. Anschließend erhalten sie eine Aufgabe zur Gruppenarbeit: Ein Alltagsszenario wird in Tabellenform erfasst und soll in einen Entscheidungsbaum übersetzt werden. Nachdem dieser Baum intuitiv erstellt wurde, werden die Studierenden angewiesen, diesen Baum zu optimieren. Dabei werden neben der Optimierung auch die Grundprinzipien des Gini-Index erklärt.



Abb. 24 Bild [Gemeinfrei] erzeugt mit DALL-E; Prompt „a cat smelling the gift, photorealistic“ von Jörg [CC BY-SA 4.0 International]

Teil I: Einleitendes Quiz

Lehrkraft:

"Sie kennen das Konzept der Entscheidungsbäume aus dem täglichen Leben. Lassen Sie uns also mit einem Beispiel beginnen:

[Die Lehrkraft überreicht einem Schüler einen geheimen Zettel mit 8 verschiedenen Tieren:]

Lehrkraft:

"Er/sie hat nun eine Liste mit 8 verschiedenen Tieren und wählt zufällig eines aus, das ich nicht kenne. Ich versuche zu erraten, welches es ist, indem ich nur drei Fragen stelle, eine nach der anderen - und die einzigen erlaubten Antworten sind "ja" oder "nein":

Liste mit Tieren:

- wolf
- Eichhörnchen,
- Huhn,
- Kuh,
- Marienkäfer,
- Biene,
- Silberfischchen,
- Bettwanze

Liste mit Fragen:

- 1** Hat er mehr als 5 Beine?
- 2a** lebt er in Ihrer Wohnung?
- 2b** lebt es auf einem Bauernhof?
- 3a** sammelt sie Pollen?
- 3b** ernährt er sich von Blut?
- 3c** ist es niedlich?
- 3d** Ist sie groß?

(Die Lehrkraft zeichnet den Entscheidungsbaum an die Tafel, beginnend mit den soeben gestellten Fragen.]

Lehrkraft: "Du hast immer noch die geheime Liste. Lass uns noch zwei weitere Tiere davon ausprobieren."

(Die Schüler vervollständigen den Baum gemeinsam mit dem Lehrer mit allen 8 Tieren.)

Lehrkraft: "Jetzt haben wir jedes Tier an seinem eigenen Ende des Baumes. Was ist mit den folgenden Beispielen?"

(Die Schüler ordnen alle fehlenden Tiere zu, die der Lehrer auffordert).

Spatz, Gans, Pferd, Humvee, Tausendfüßler, Kakerlake, Kopflaus usw.

Teil II: Visualisierung des ersten Baumes

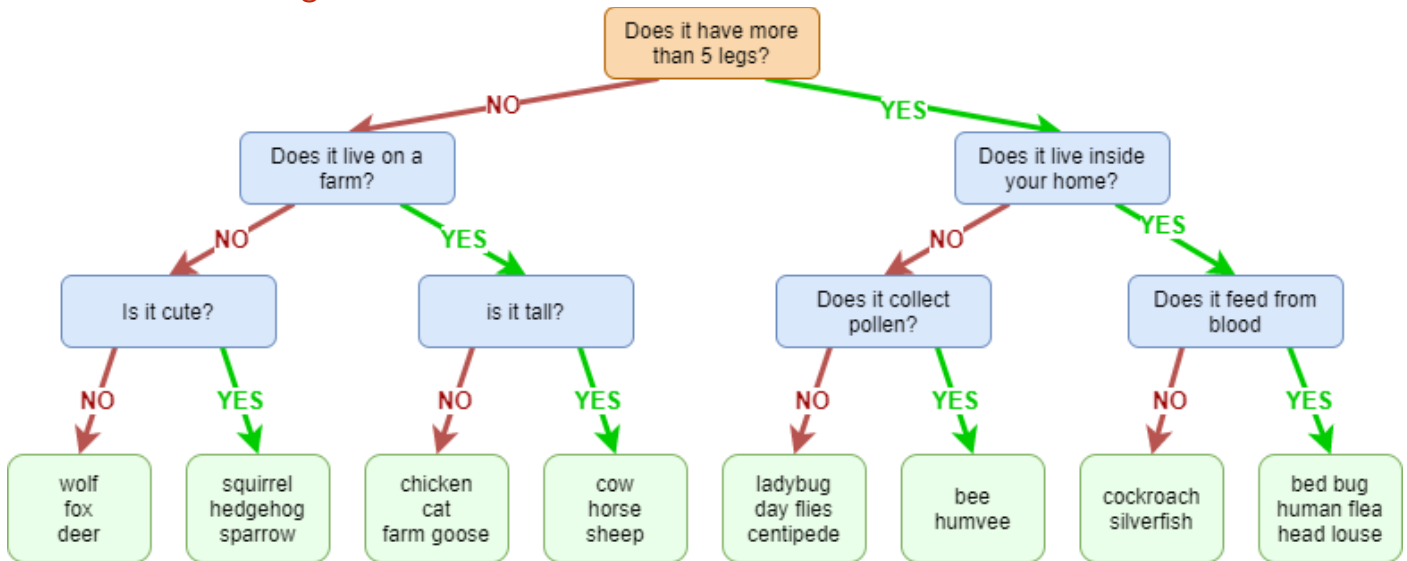


Abbildung 2: Tafelskizze: fertiger Entscheidungsbaum

Teil III: Grundlagen und Terminologie von Bäumen) Lehrkraft:

"Innerhalb dieses so genannten Entscheidungsbaums bilden einige Arten ihre eigenen Klassen. Wir können diesen Baum also zur Klassifizierung von Tieren verwenden - solange er für uns nützlich ist: Er ist nur ein Modell für eine bestimmte Aufgabe."

"Schauen wir uns einmal an, was ein Baum in der Informatik ist. Was ist sein Zweck und wie werden die verschiedenen Teile benannt? - ACHTUNG: TERMINOLOGIE!"

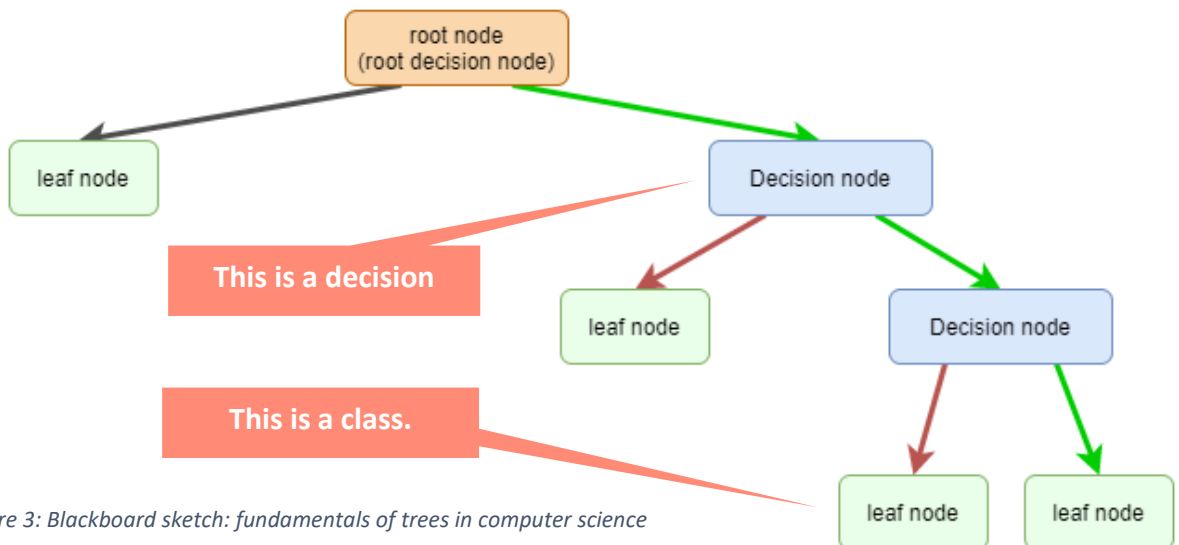


Figure 3: Blackboard sketch: fundamentals of trees in computer science

Teil IV: Problembasierte Teamarbeit: Erstellung eines Baums aus einem Datenrahmen

Lehrkraft:

"Nun zu etwas Teamarbeit. Stellen Sie sich folgendes Problem vor: Sie fragen sich, was Ihr Freund heute wohl tun wird, und Sie wollen ihn anrufen. Aber Sie zögern - wie wird er sich verhalten? Heute scheint die Sonne bis zum Nachmittag und es gibt eine Menge Hausaufgaben zu erledigen. Sie versuchen, seine Entscheidung aufgrund Ihrer Erfahrungen in der Vergangenheit vorherzusagen. Deshalb schreibst du deine Gedanken auf:"

Schon mit Freunden zusammen?	Scheint die Sonne?	Müssen Sie noch Hausaufgaben machen?	Beobachtetes Verhalten
Nein	Nein	Nein	Videospiele
Nein	Nein	Ja	Hausaufgaben machen
Nein	Ja	Nein	Draußen spielen
Nein	Ja	Ja	Draußen spielen
Ja	Nein	Nein	Videospiele
Ja	Nein	Ja	Videospiele
Ja	Ja	Nein	Draußen spielen
Ja	Ja	Ja	Draußen spielen

Abbildung 4: Arbeitsblatt mit Datenrahmen

[Die Schüler konvertieren den Trainingsdatensatz in einen Entscheidungsbaum. Die ersten drei Spalten sind die Daten/Merkmale, die letzte Spalte ist der Zieldatensatz. Verschiedene Ansätze der Schüler sind möglich]

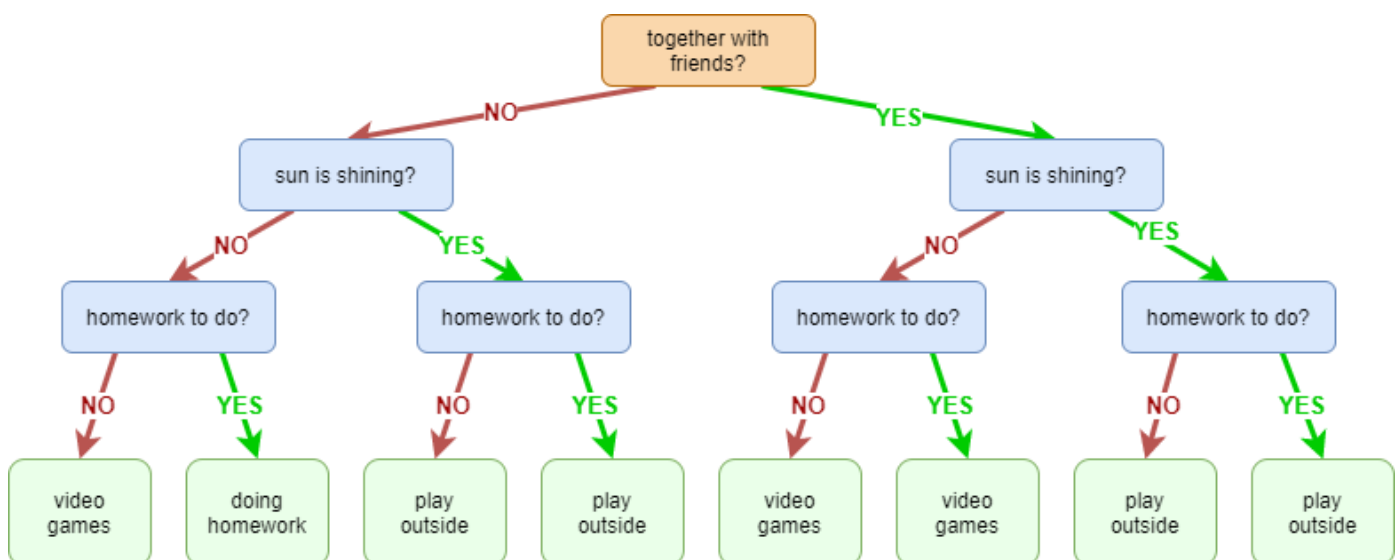


Abbildung 5: Tafelskizze für den fertigen Entscheidungsbaum

Teil V: Optimierung eines Entscheidungsbaums durch Vereinfachung.

Lehrkraft:

"Es gibt einige Entscheidungen, die vereinfacht werden können. Kannst du sie finden? Vereinfachen Sie also bitte den Baum und erklären Sie, was Sie tun."

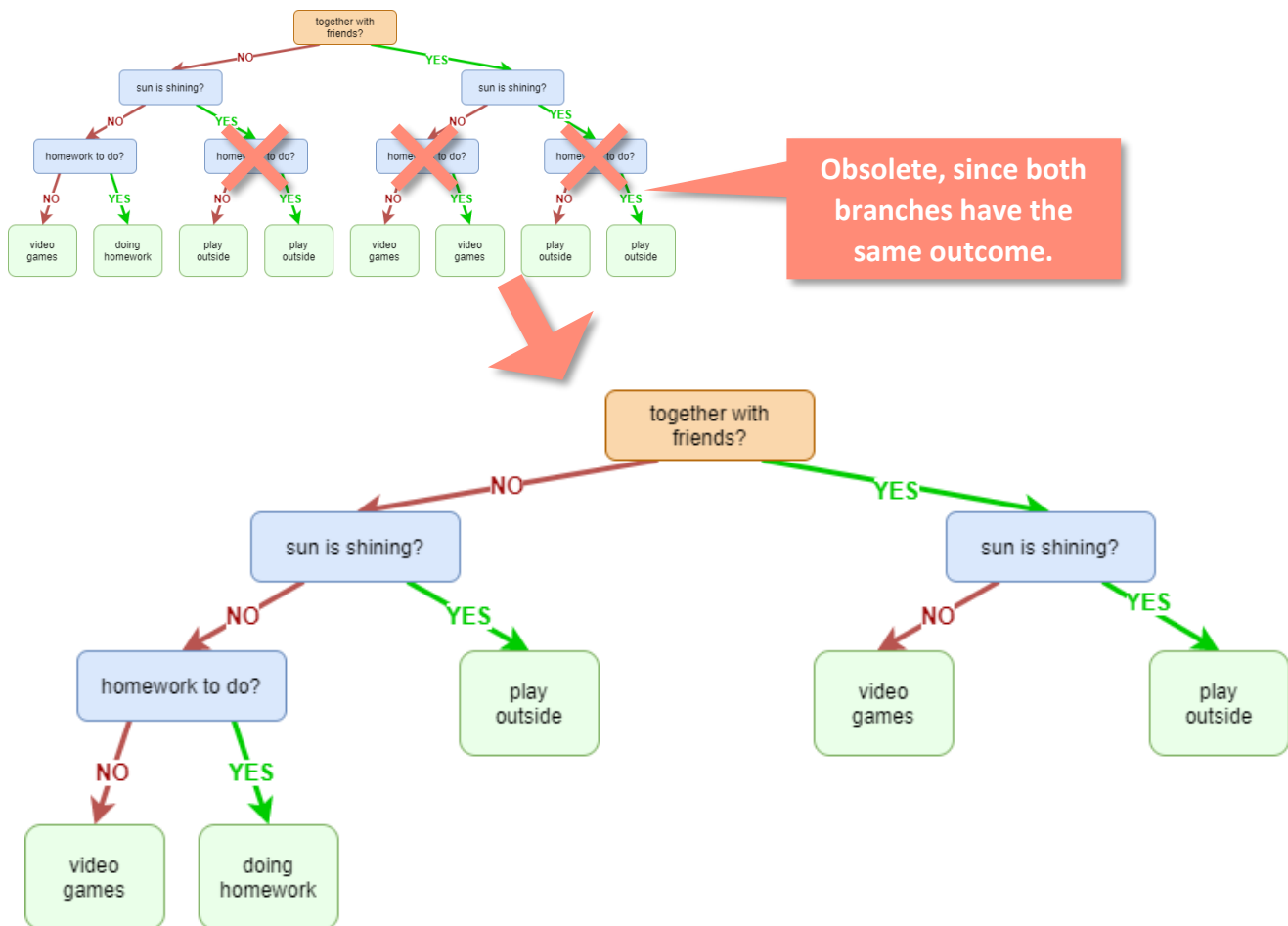


Abbildung 6: verbesserte, aber nicht optimierte Entscheidungsäume

Lehrkraft:

"Warum ist es besser?"

Studenten:

"Es sind weniger Entscheidungen notwendig, um dieselbe Aufgabe zu bewältigen. Es gibt im Durchschnitt 12 Entscheidungen geteilt durch 5 Ergebnisse = 2,4 Entscheidungen pro Ergebnis"

Einige vereinfachte Bäume sind möglich, und einige von ihnen sollten in die Übungsbücher der Schüler aufgenommen werden, da sie wichtig für das Verständnis sind: Es gibt keine einzige Methode, um das Problem zu lösen.

Teil VI: Welches ist die beste Entscheidung für den Anfang? Entropie-gesteuerte Optimierung.

Lehrkraft: "Wir haben nicht bedacht, dass wir auch die Reihenfolge der verschiedenen Fragen variieren können. Ist es wichtig, die drei Fragen so zu stellen, wie wir es bisher getan haben?"

Studenten: "Nein, das ist nicht nötig."

Lehrkraft: "Also, variieren wir die Ausgangsfrage und suchen wir die beste Lösung. Wie viele Varianten sind möglich?"

Studenten: "3 verschiedene Fragen zu Beginn, dann 2 weitere verschiedene Fragen und eine feste Frage zum Schluss:
 Schluss: $3 \cdot 2 \cdot 1 = 6$ verschiedene Entscheidungsbäume"

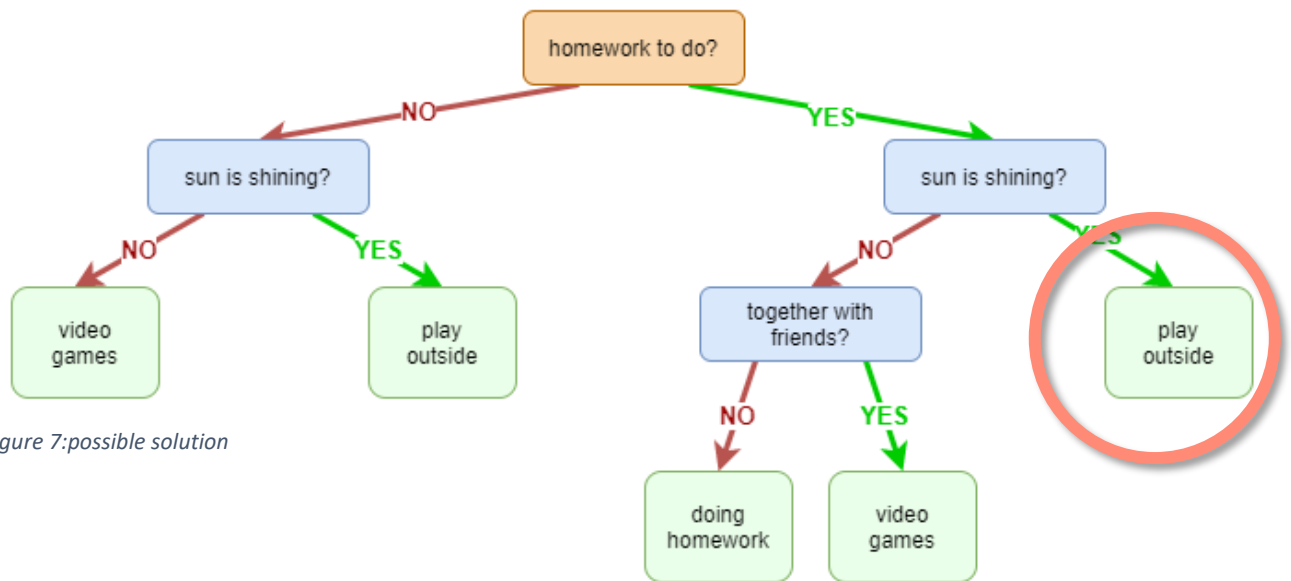
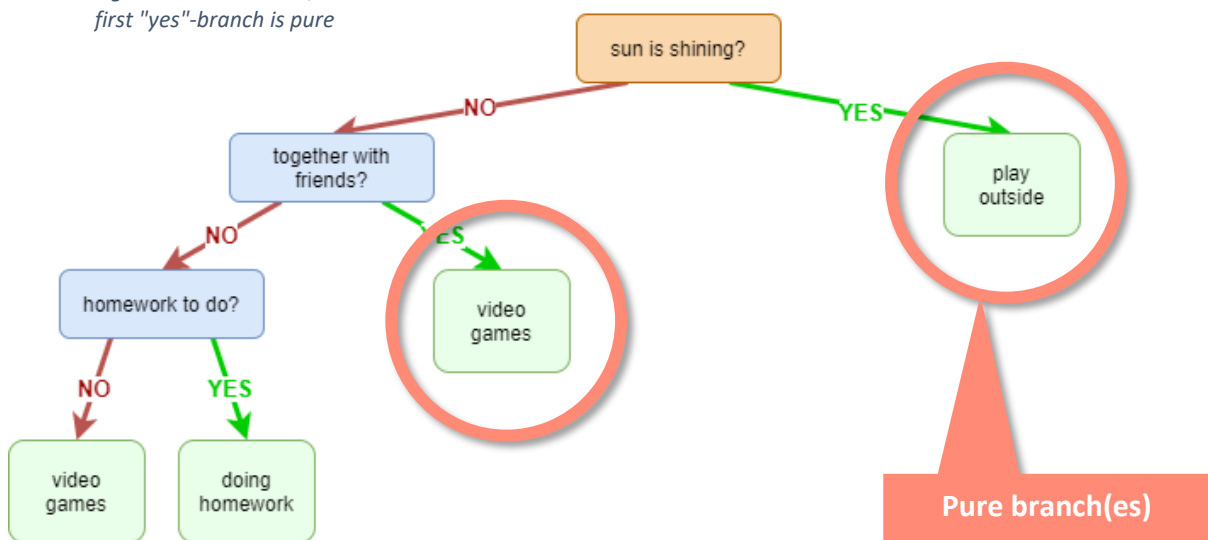


Figure 7: possible solution

Lehrkraft:

"Wenn wir so optimieren können, dass eine Entscheidung ohne weitere Aufteilung eine fertige Klassifizierung ergibt, wird das Ergebnis "rein" genannt (TERMINOLOGIE). Je reiner die Entscheidung nach einer Aufteilung, desto besser. Aber: Was ist Reinheit? Kann sie gemessen werden?"

Figure 8: best solution, since the first "yes"-branch is pure





Teil VII: Wie kann man sich vorstellen, was Reinheit ist, und wie kann sie gemessen werden .

Beispiel: Kennen Sie den Kirsch-Bananen-Shake?

Es ist eine Mischung aus zwei verschiedenen Saftbestandteilen.

Beziehung zu unserem Problem:

Es gibt viele Möglichkeiten, die Reinheit zu messen, die gebräuchlichsten sind der so genannte GINI-Koeffizient und der so genannte ENTROPY. Beide sind unabhängig voneinander und beschreiben den Grad der Reinheit, den eine Entscheidung bewirkt. Sie variieren zwischen 0 und 1.

Im Allgemeinen gilt: je niedriger der GINI-Koeffizient, desto reiner die Entscheidung - und desto besser!

Abb. 25 Bild [\[Gemeinfrei\]](#) erzeugt mit [DALL-E](#); Prompt „bananas, cherries and a red smoothie in a glass, professional studio photography“ von Jörg [\[CC BY-SA 4.0 International\]](#)

keine Vermischung,	Leichte Vermischung	starke Vermischung	vollständige Vermischung
Gini = 0	Gini ≈ 0,15	Gini ≈ 0,3	Gini = 0,5
Entropie=0	Entropie=0,3	Entropie=0,7	Entropie=1
Bezieht sich auf eine Entscheidung, die Splitting ist:			
Perfekt	Nahezu perfekt	Größtenteils akzeptabel	Sinnlos

```
import numpy as np
import matplotlib.pyplot as plt
import math as math

def entropy(p):
    return - p * math.log2(p) - (1-p)
    * math.log2(1 - p)

x = np.arange(0.001, 1.0, 0.001)
print(x)
ent = [entropy(p) for p in x]
print(ent)

fig = plt.figure(figsize=(8,4))
ax = plt.subplot(1,1,1)
ax.plot(x, ent, color = 'red', lw = 2)
plt.ylim([0, 1.01])
plt.xlim([-0.01, 1.01])
plt.xlabel('p(x)')
plt.ylabel('Entropy')
plt.show()
```

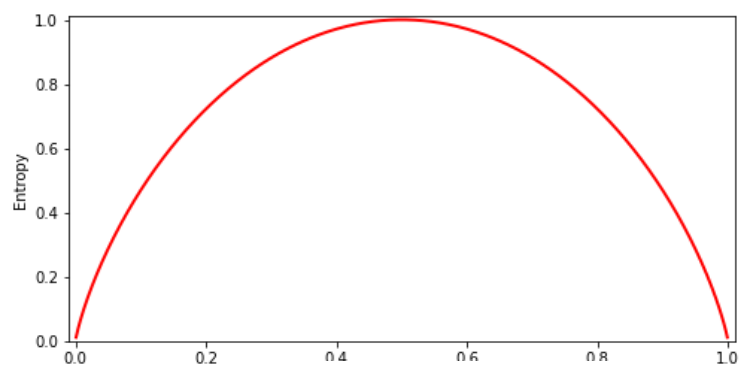


Figure 9: Entropy plot

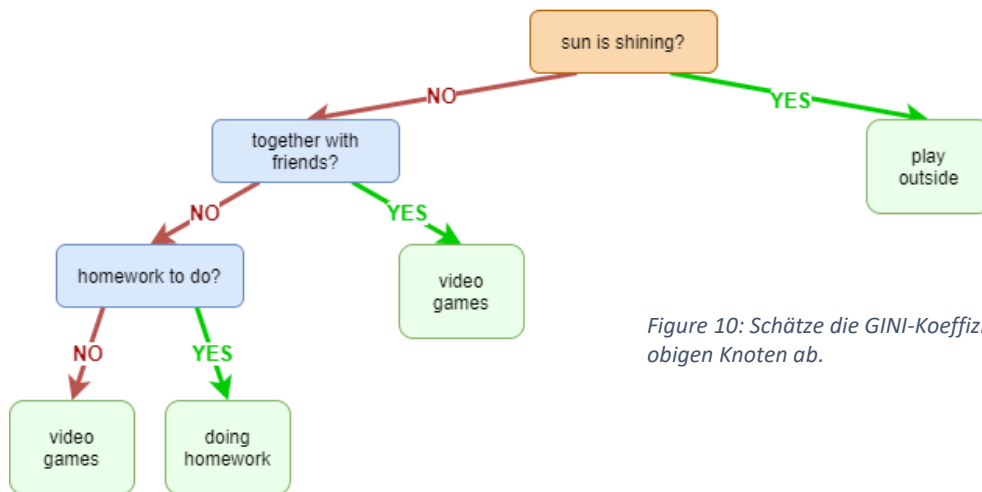


Figure 10: Schätze die GINI-Koeffizienten für die obigen Knoten ab.

Teil VIII: Zusammenfassung

Entscheidungsbäume sind ein algorithmisches Werkzeug zur Vorhersage oder Klassifizierung neu erscheinender Daten auf der Grundlage von in der Vergangenheit gesammelten Trainingsdaten. Er strukturiert den Weg zur richtigen Klassifizierung durch eine hierarchische Kette von Entscheidungsaufteilungen. Diese Teilungen können mit Hilfe von Gini- oder Entropie-gesteuerten Messungen optimiert werden. Der Entscheidungsbaum-Algorithmus nimmt Eingabedaten auf, berechnet die besten Aufteilungen und gibt einen Baum aus, der leicht visualisiert werden kann.

Teil IX: Überanpassung



Abb. 26 Jason Mayes [@jason_mayes]. [The best way to explain #overfitting in #MachineLearning](#), 2020,21.8.

From <https://en.wikipedia.org/wiki/Overfitting>

The green line represents an overfitted model and the black line represents a regularized model. While the green line best follows the training data, it is too dependent on that data and it is likely to have a higher error rate on new unseen data, compared to the black line.

Was würde passieren, wenn wir jeden einzelnen Datenpunkt berücksichtigen würden, ohne zu verallgemeinern?

In diesem Fall hätte keine Verallgemeinerung stattgefunden: Es wäre ein Fall der sogenannten "Überanpassung".

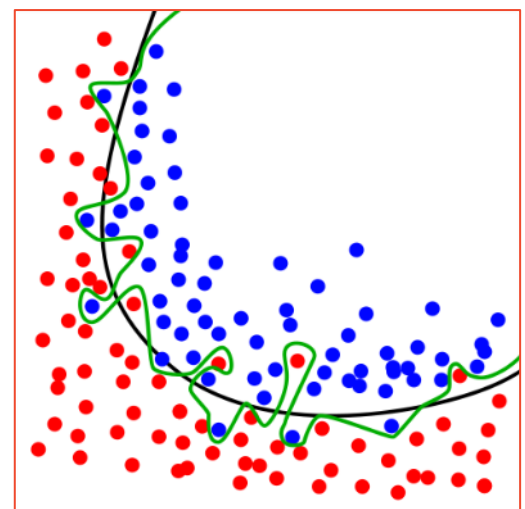


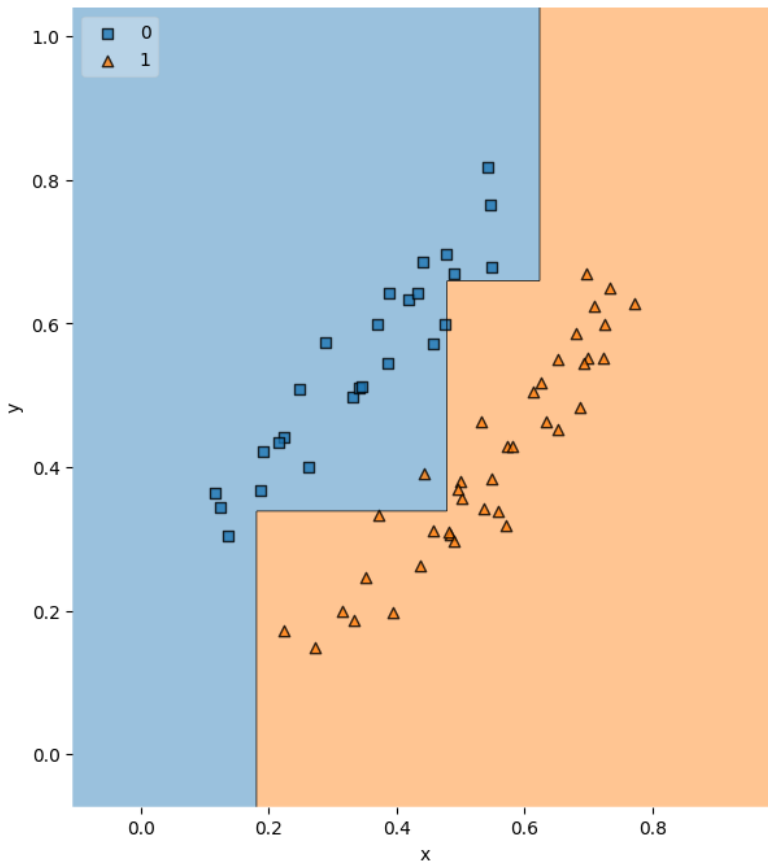
Abb. 27 <https://en.wikipedia.org/wiki/Overfitting#/media/File:Overfitting.svg> [CC BY-SA 4.0]

Teil X: Decision Boundaries: Wie zerlegt ein Baum den Bereich?

Gegeben ist der folgende Baum, der aus einem eigentlich leicht separierbaren Datensatz erstellt wurde. Die einzelnen Datenpunkte sind im Scatterplot auf der rechten Seite zu sehen:



Klassen A und B mit DecTree



Möglicher Lösungsansatz:

Im Wurzelknoten wird separiert für x -Werte, die kleiner gleich 0,479 sind. Das lässt sich so interpretieren, indem man einen Bereich links von $x=0,479$ und einen Bereich rechts davon abtrennt.

Geht man in die nächste Node nach links (rot), dann wird die nächste Unterteilung oberhalb von $y=0,339$ und unterhalb davon gezogen.

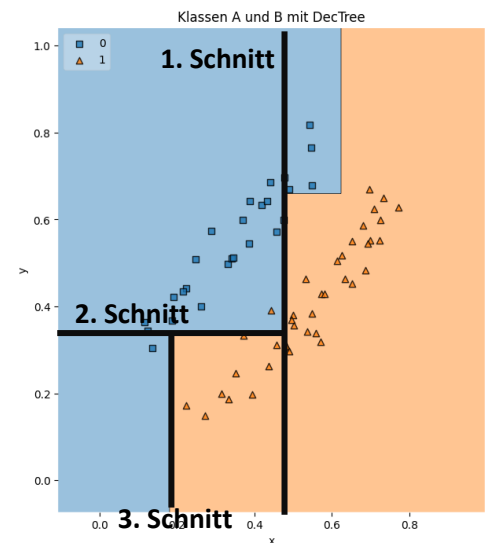


Abb. 28 Eigene Screenshots aus Jupyter Notebook

Lektion 9: Die logistische Regression für die Klassifizierung

Was die Schüler lernen sollten:



 DALL-E 3
Created 2 images

Benutze die Darstellung einer Waage, um eine symbolische Repräsentation der logistischen Regression in Form eines Bildes zu erzeugen.

Die logistische Regression separiert Klassen mittels einer Trennlinie oder eines Trennbereichs. Hierbei repräsentiert die "1" die Zugehörigkeit zu einer Klasse und die "0" das Gegenteil.

In der Praxis sind jedoch reine "Entweder-Oder"-Entscheidungen selten. Daher werden Wahrscheinlichkeiten benötigt, um Unsicherheiten abzubilden. Die logistische Regression gibt deshalb nicht nur eine Klassentrennung vor, sondern auch die Wahrscheinlichkeit einer Klassenzugehörigkeit.

Die Sigmoid-Funktion modelliert diesen kontinuierlichen Wahrscheinlichkeitsübergang zwischen 0 und 1. Ein wichtiger Aspekt ist die Entscheidungsgrenze, oft bei 0.5, die bestimmt, welcher Klasse eine Beobachtung zugeordnet wird, basierend auf der vorhergesagten Wahrscheinlichkeit.

Nicht mehr Bestandteil der Einführung: Die Methode der Maximum-Likelihood-Schätzung (MLE) optimiert die Gewichtungen entsprechend den beobachteten Daten.

Mögliche Aktivitäten und Aufgaben der Schüler:

Anhand eines Alltagsproblems – wie unterscheidet man die sehr ähnlichen Blumenarten ‚Margeriten‘ und ‚Gänseblümchen‘ – wird das Konzept der Trennlinie eingeführt. Ein ähnliches Problem findet sich auch in Lektion 8 (ML-Prinzipien unplugged), in denen Tassen von Schwämmen anhand zweier Features getrennt wurden. Hier wird angeknüpft und anhand eines synthetischen Datensatzes zu den Blumen eine logistische Regression erstellt. Im Anschluss werden mit den gleichen Features Gänseblümchen und Asters getrennt. Hier separieren sich die Klassen nicht sauber, weshalb man mit Wahrscheinlichkeiten arbeiten muss. Die Schüler leiten hieraus die Idee ab, einen Wahrscheinlichkeitsübergang zu definieren.

Teil I: "Wie soll man diese Blumen unterscheiden? – einfacher Fall"



https://praxistipps.focus.de/margeriten-so-unterscheiden-sie-sich-von-gaensebluemchen_159181

Margeriten und Gänseblümchen sind sich von ihrem Aussehen her sehr ähnlich: Mit optischen Erkennungs-Systemen lassen sie sich nur schwer auseinanderhalten. Ein Fachmann oder eine Fachfrau weiß aber, dass sich beide anhand ihrer Größe voneinander trennen lassen. Dazu gehören sowohl die Stängellänge als auch die Blütengröße.

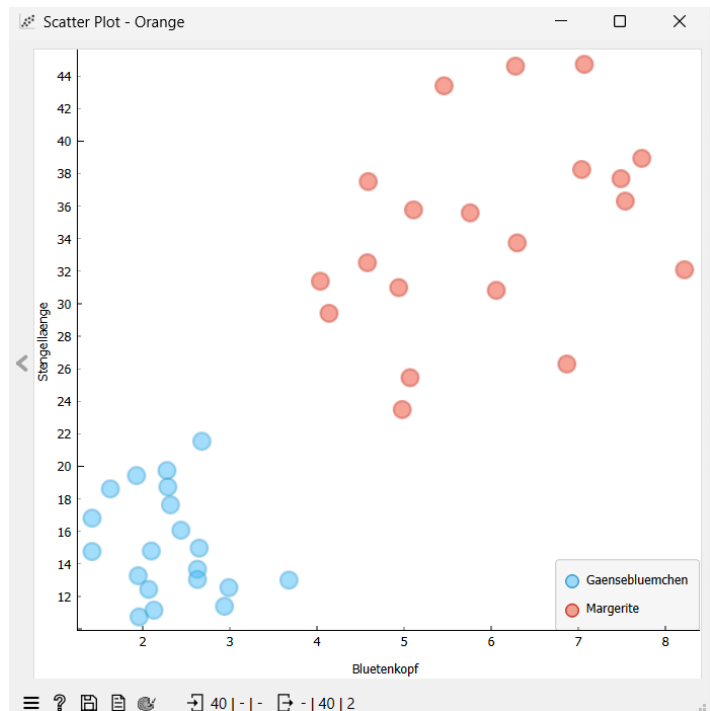
hyperrealistic art close-up photo of a flower vase on a marble table, in a clean kitchen, Leucanthemum, Bellis perennis, Aster, daisies. extremely high-resolution details, photographic, realism pushed to extreme, fine texture, incredibly lifelike

Klasse, Blüte, Stengel

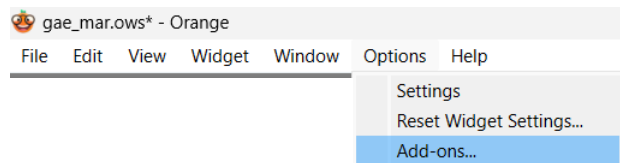
Margerite	7.54	36.31
Margerite	4.04	31.38
Margerite	5.76	35.59
Margerite	4.59	37.51
Margerite	7.07	44.72
Margerite	6.28	44.61
Margerite	6.3	33.74
Margerite	4.98	23.49
Margerite	4.58	32.52
Margerite	7.49	37.69
Margerite	7.04	38.25
Margerite	4.14	29.41
Margerite	5.07	25.45
Margerite	4.94	30.99
Margerite	6.87	26.29
Margerite	7.73	38.94
Margerite	8.22	32.09
Margerite	5.11	35.77
Margerite	5.46	43.4
Margerite	6.06	30.82
GaenseB	2.29	18.73
GaenseB	1.95	13.27
GaenseB	2.07	12.43
GaenseB	2.99	12.54
GaenseB	3.68	13
GaenseB	2.94	11.39
GaenseB	1.63	18.61
GaenseB	2.32	17.63
GaenseB	1.42	14.76
GaenseB	2.28	19.74
GaenseB	2.63	13.67
GaenseB	2.44	16.07
GaenseB	1.96	10.74
GaenseB	2.68	21.54
GaenseB	1.42	16.81
GaenseB	2.65	14.97
GaenseB	1.93	19.43
GaenseB	2.1	14.79
GaenseB	2.63	13.04
GaenseB	2.13	11.15

Tabelle mit jeweils 20 Margeriten und Gänseblümchen. Kann so direkt in Excel oder Orange Data Mining übernommen werden.

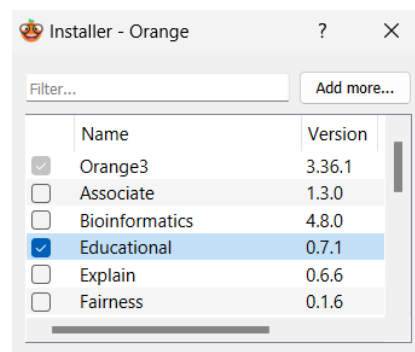
Alle Angaben sind in cm.



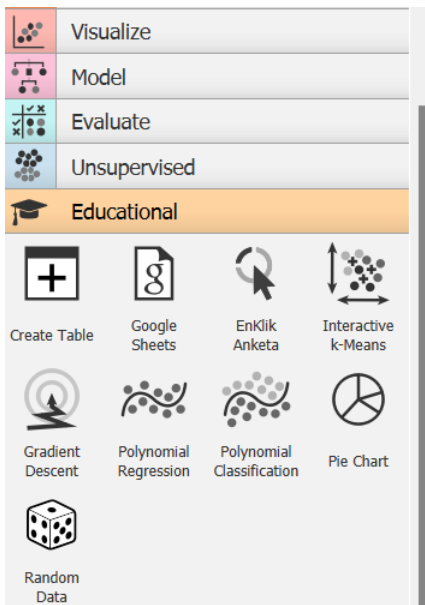
2: Scatterplot des Datensatzes aus Orange Data Mining. Man sieht deutlich die Möglichkeit einer Trennlinie. Diese sollten von den Schülern eingetragen werden.



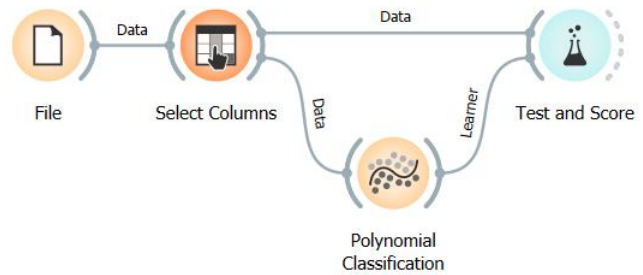
1: Um den Datensatz mittels logistischer Regression bearbeiten zu können, kann zwar das entsprechende Widget verwendet werden, allerdings ist es für die didaktische Heranführung besser, das "Educational"-Widget zu installieren.



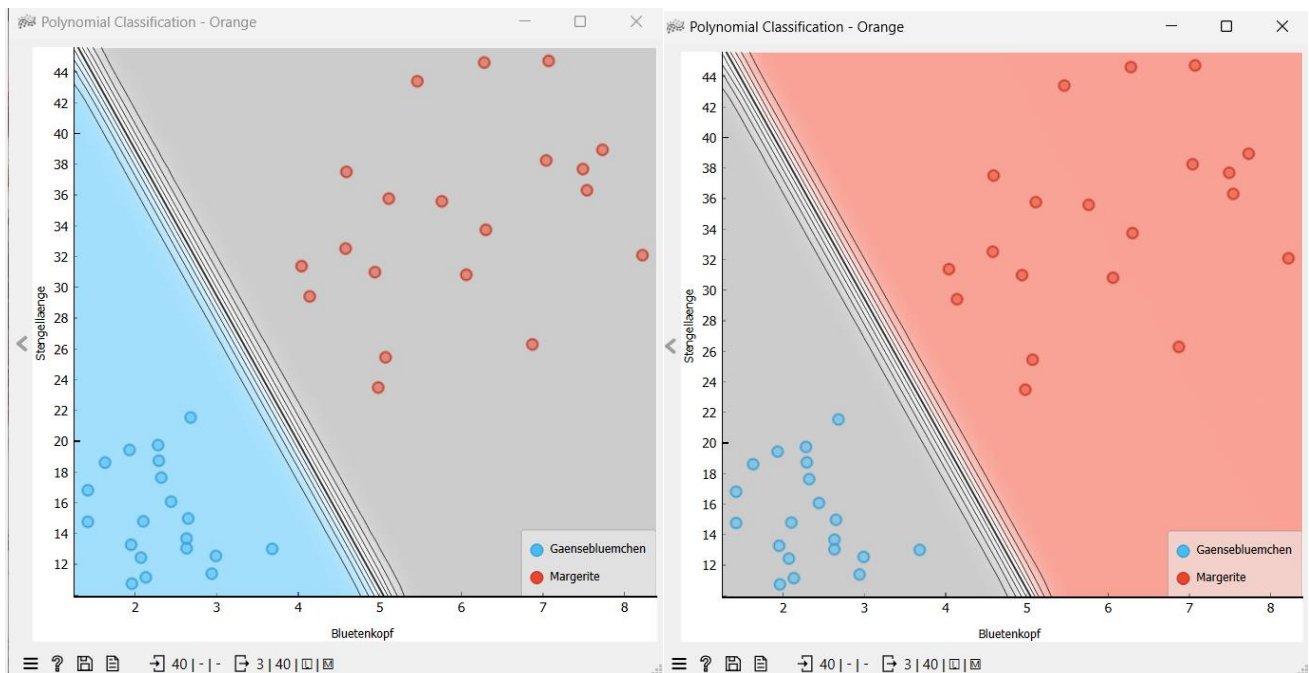
3: Dieses Findet sich unter den Add-ons als "Educational". Wählt man es an und bestätigt mit 'ok', dann wird es installiert.



5 Im Educational-Widget wird nun die 'Polynomial Classification' ausgewählt ...



4... und damit das folgende Setup erstellt. Die "Select Columns"-Node ist für den Fall, dass Orange die Target-Spalte nicht erkennt. Diese kann hier manuell gesetzt werden.



6. Beim Doppelklick auf die 'Polynomial Classification' Node ergibt sich der folgende Conturplot. Ändert man das „Target“, so wechseln die farbig gekennzeichneten Flächen.

Fährt man mit der Maus über die Fläche, so zeigen sich unterschiedliche Zahlenwerte. Die Schüler sollen an dieser Stelle diese Zahlenwerte interpretieren und in Worte fassen. Mögliche Antwort:

Für das jeweils gesetzte Target werden Werte zwischen 0 und 1 angezeigt. Dabei liegt der Wert bei 1, wenn man sich sicher im Bereich des Targets befindet. Umgekehrt wird die 0 angezeigt, wenn man sich im jeweils komplementären Bereich befindet.

Teil II: "Wie soll man diese Blumen unterscheiden? – schwieriger Fall"

Klasse, Blüte, Stängel
Margerite, 7.54, 36.31
Margerite, 4.04, 31.38
Margerite, 5.76, 35.59
Margerite, 4.59, 37.51
Margerite, 7.07, 44.72
Margerite, 6.28, 44.61
Margerite, 6.3, 33.74
Margerite, 4.98, 23.49
Margerite, 4.58, 32.52
Margerite, 7.49, 37.69
Margerite, 7.04, 38.25
Margerite, 4.14, 29.41
Margerite, 5.07, 25.45
Margerite, 4.94, 30.99
Margerite, 6.87, 26.29
Margerite, 7.73, 38.94
Margerite, 8.22, 32.09
Margerite, 5.11, 35.77
Margerite, 5.46, 43.4
Margerite, 6.06, 30.82
Aster, 3.11, 58.76
Aster, 4.4, 46.67
Aster, 5.78, 45.44
Aster, 2.84, 49.73
Aster, 3.02, 31.87
Aster, 3.57, 45.06
Aster, 4.21, 37.41
Aster, 4.46, 45.79
Aster, 4.07, 32.59
Aster, 3.56, 40.04
Aster, 4.37, 58.81
Aster, 3.57, 55.47
Aster, 2.67, 38.25
Aster, 5.82, 40.22
Aster, 4.68, 51.88
Aster, 3.8, 60.15
Aster, 4.68, 59.32
Aster, 3.65, 33.53
Aster, 1.56, 41.93
Aster, 3.46, 51.86

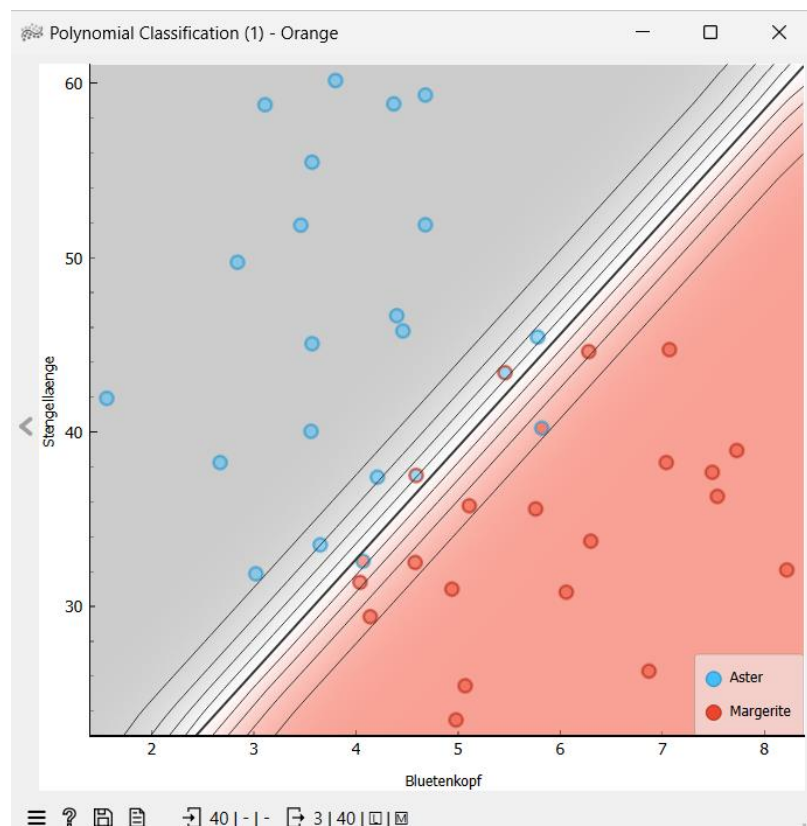
Tabelle mit jeweils 20 Asten und Margeriten. Kann so direkt in Excel oder Orange Data Mining übernommen werden.

Alle Angaben sind in cm.

Nun ist ein weniger eindeutiger Datensatz gegeben, nämlich einer von Margeriten und Asten. Diese besitzen leider recht ähnliche Blütengrößen und Stängel-Längen. Die Unterscheidung ist hier weniger gut möglich, weil sich die Bereiche der beiden Blumen überlappen. Man muss mit einer ‚Unschärfe‘ in der Aussagekraft rechnen:



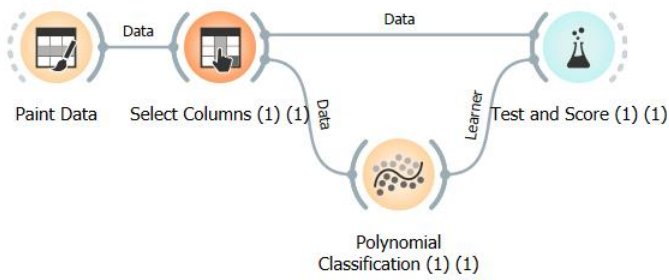
7. hyperrealistic art close-up photo of a flower vase on a marble table, in a clean kitchen, Aster, Bellis perennis. extremely high-resolution details, photographic, realism pushed to extreme, fine texture, incredibly lifelike



7: Logistische Regression für Asten und Margeriten. Die Schüler sollen auch hier mit der Maus über die unterschiedlichen Bereiche des Diagramms fahren und mit Worten beschreiben, was die unterschiedlichen Zahlenwerte bedeuten. Ausserdem ist der Übergangsbereich deutlich breiter, weil sich die beiden Einzelbereiche überlappen.

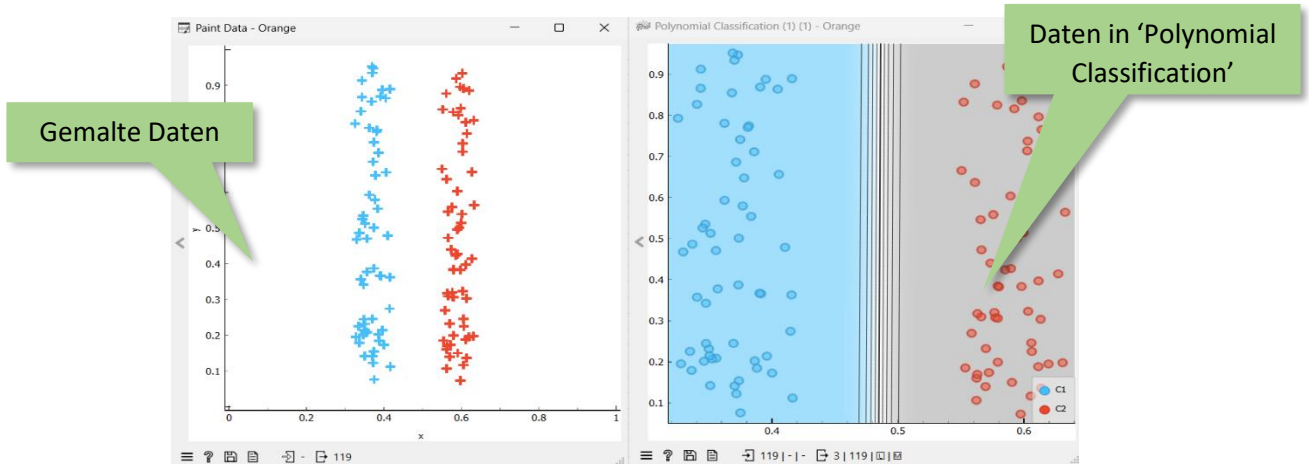
Teil III: Die logistische Funktion

Wovon hängt die breite des Bereichs ab? Dazu machen wir Orange Data Mining ein Experiment:

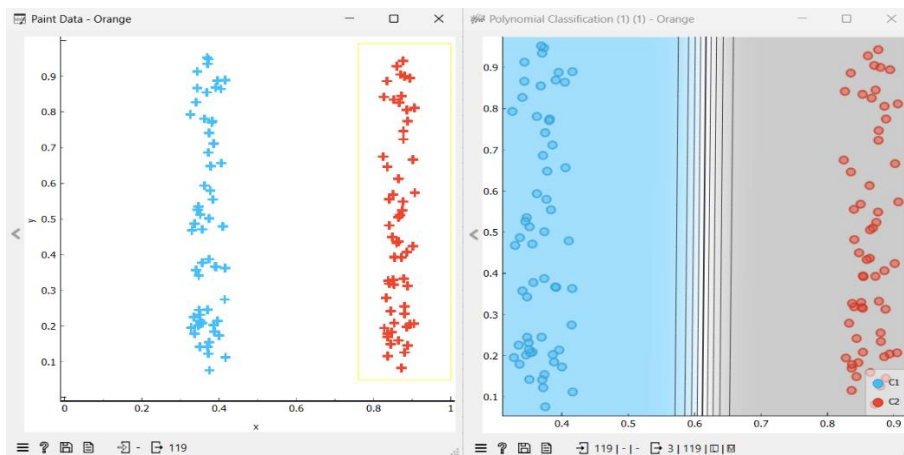


Wir erzeugen uns mittels der ‚Paint Data‘-Node einen Datensatz aus zwei Klassen, und zwar so:

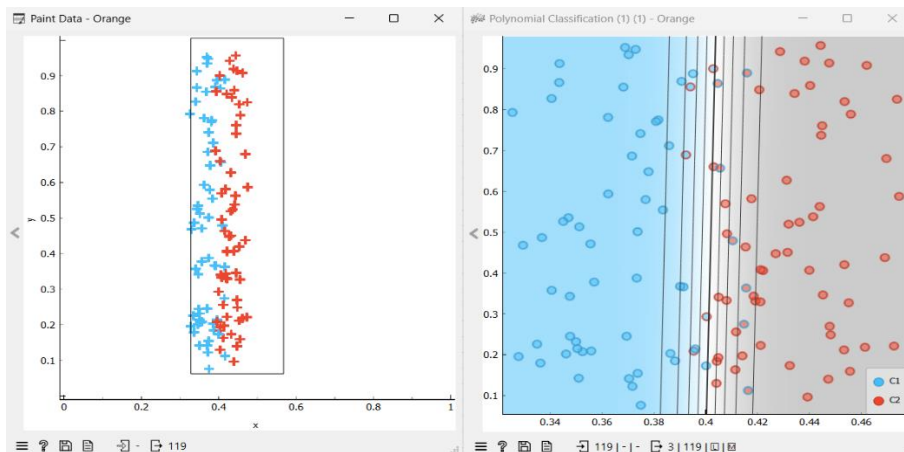
Beide Klassen sollten möglichst senkrecht verteilt sein, sodass sich in der ‚Polynomial Classification‘-Node ein möglichst senkrecht verlaufender Trennbereich ergibt. Um das zu erreichen, legen wir beide Plotfenster nebeneinander:



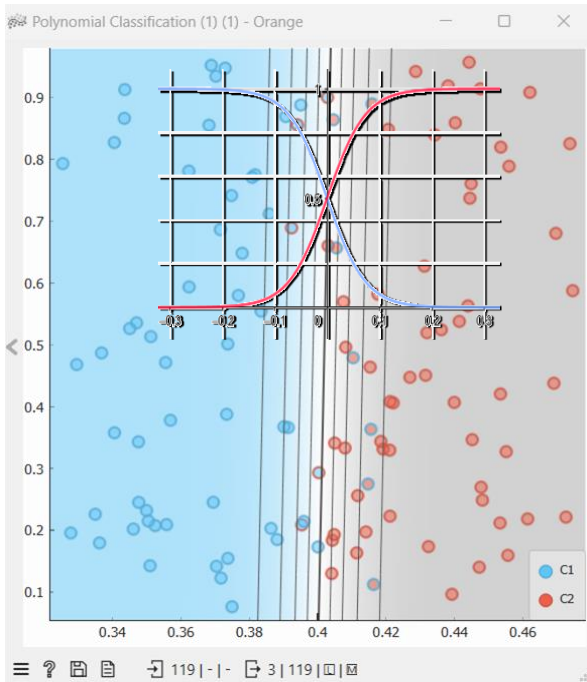
Schritt 1: Rote Daten selektieren und verschieben: Wie ändert sich der Übergangsbereich und warum?



Schritt 2: Rote Daten selektieren und überlappen lassen: Wie ändert sich der Übergangsbereich und warum?



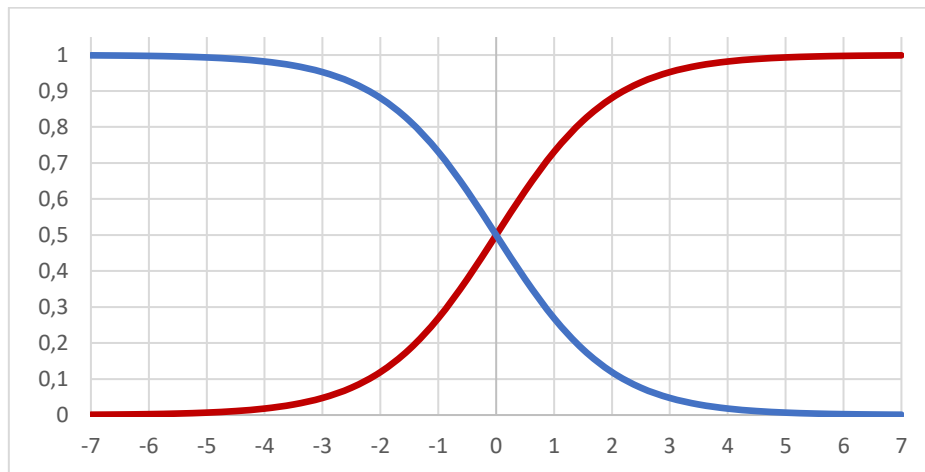
Formuliere einige ‚Je-Desto‘-Sätze, um die Beobachtungen zu erklären.



erstelle ein kleine Tabelle, in der du die Werte der Konturen abliest und auf der X-Achse aufträgst:

X-Achsenwert	Übergangswert
0,38	0,9
0,39	0,8
0,395	0,6
0,4	0,5
0,405	0,7
0,41	0,3
0,52	0,1

Wiederhole für das andere Target. Du erhältst, wenn du daraus ein Diagramm baust, in etwa die folgende Funktion:



8: Die logistische Funktion, einmal für das rote, einmal für das blaue Target.

Welche Eigenschaften hat die Funktion? Beschreibe:

- Den Wertebereich der roten Kurve: bei welchem Wert beginnt sie, wo endet sie?
- Interpretiere die unterschiedlichen Werte: Was bedeutet die ‚0‘, was bedeutet die ‚0,5‘ und was die ‚1‘?
- Wie stehen beide Kurven zueinander in Bezug? Wenn die eine Kurve steigt, dann ...
- Welchen Zahlenwert erhält man, wenn man zu gegebenen X-Achsenwert die Werte von blau und rot addiert?
- Wie ändert sich die Kurve, wenn die Trennung der beiden Klassen schärfer wird? Oder uneindeutiger?
- Den Wert von ‚0,5‘ bezeichnet man als „Entscheidungsgrenze“. Kannst du dir erklären, warum das so ist?

Lektion 10: Naive Bayes Klassifikator

"Sag mir nie, wie hoch meine Chancen sind!"

Was die Schüler lernen sollten:

Die Schüler lernen das Konzept der bedingten Wahrscheinlichkeit anhand der berühmten psychologischen Testfrage "Bibliothekar oder Bauer" kennen. Das überraschende Ergebnis eröffnet den Schülerinnen und Schülern neue Einsichten über die Grenzen ihres eigenen Denkens. In dem darauf aufbauenden Beispiel wird die Aussagekraft eines typischen Corona-Schnelltests konkret analysiert, wobei die Schülerinnen und Schüler die Fachbegriffe wahr positiv/ falsch negativ/ wahr negativ/ falsch positiv erarbeiten. Die bekannte Formel des Bayes-Theorems ergibt sich in direkter Folge. In der anschließenden Vorstellung des Klassifikators werden die Schüler an die Integration verschiedener Merkmale herangeführt. Dies kann vertieft werden durch Textmining oder ein kleines oranges Data-Mining-Projekt vertieft werden.

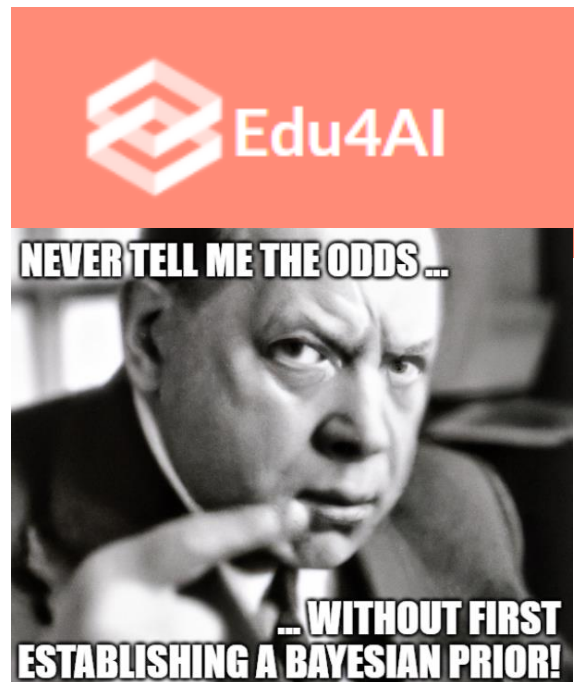


Abb. 29 Bild [Gemeinfrei] erzeugt mit DALL-E; Prompt „angry elderly businessman raises his index finger, 1950s archival photo, closeup“ von Jörg [CC BY-SA 4.0 International]

Mögliche Aktivitäten und Aufgaben der Schüler:

Hier geht es darum, neben dem grundlegend wichtigen Naive-Bayes-Klassifikator auch die wichtigsten statistischen Grundkonzepte zu erlernen und anzuwenden: Mit Hilfe der ihnen bekannten Coronatest-Statistik erarbeiten die Schüler den Verzweigungsbaum aller vier möglichen Testergebnisse. Indem sie diese als Wahr Positiv, Falsch Negativ, Wahr Negativ und Falsch Positiv klassifizieren, leiten sie die bedingte Wahrscheinlichkeitsgleichung, das Bayes Theorem, ab. Die Schüler wenden die Konzepte der Empfindlichkeit (= Recall) und der Spezifität in ihren eigenen Übungen an. Später wird dieser noch eindimensionale Ansatz für mehrere Merkmale erweitert, bis man schließlich zur allgemeinen Version des Klassifikators gelangt. Darüber hinaus werden die Konfusionsmatrix und die wichtigen Metriken wie Precision, Recall und Accuracy vorgestellt.

Teil I: Einleitende Frage: Bibliothekar oder Landwirt?

Bitte gehen Sie bei der nächsten Frage davon aus, dass Steve nach dem Zufallsprinzip aus einer repräsentativen Stichprobe ausgewählt wurde:

Eine Person wurde von einem Nachbarn wie folgt beschrieben:

"Steve ist sehr schüchtern und zurückhaltend, immer hilfsbereit, aber wenig interessiert an Menschen oder an der Welt der Realität. Er ist ein sanftmütiger und ordentlicher Mensch mit einem Bedürfnis nach Ordnung und Struktur und einer Leidenschaft für Details."

Ist Steve eher ein Bibliothekar oder ein Landwirt?

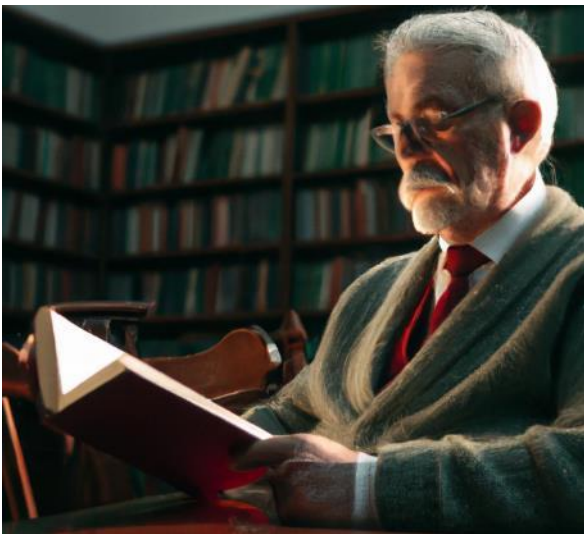
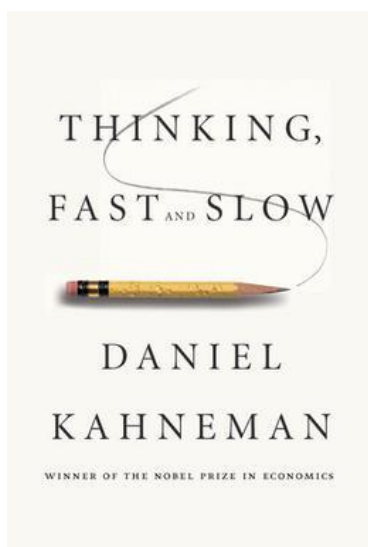


Abb. 30 Bild [\[Gemeinfrei\]](#) erzeugt mit [DALL-E](#); Prompt „old librarian reading inside his library, high quality photo, dramatic lightning“ von Jörg [\[CC BY-SA 4.0 International\]](#)



Abb. 31 Bild [\[Gemeinfrei\]](#) erzeugt mit [DALL-E](#); Prompt „old farmer and a goat, high quality photo, dramatic lightning“ von Jörg [\[CC BY-SA 4.0 International\]](#)

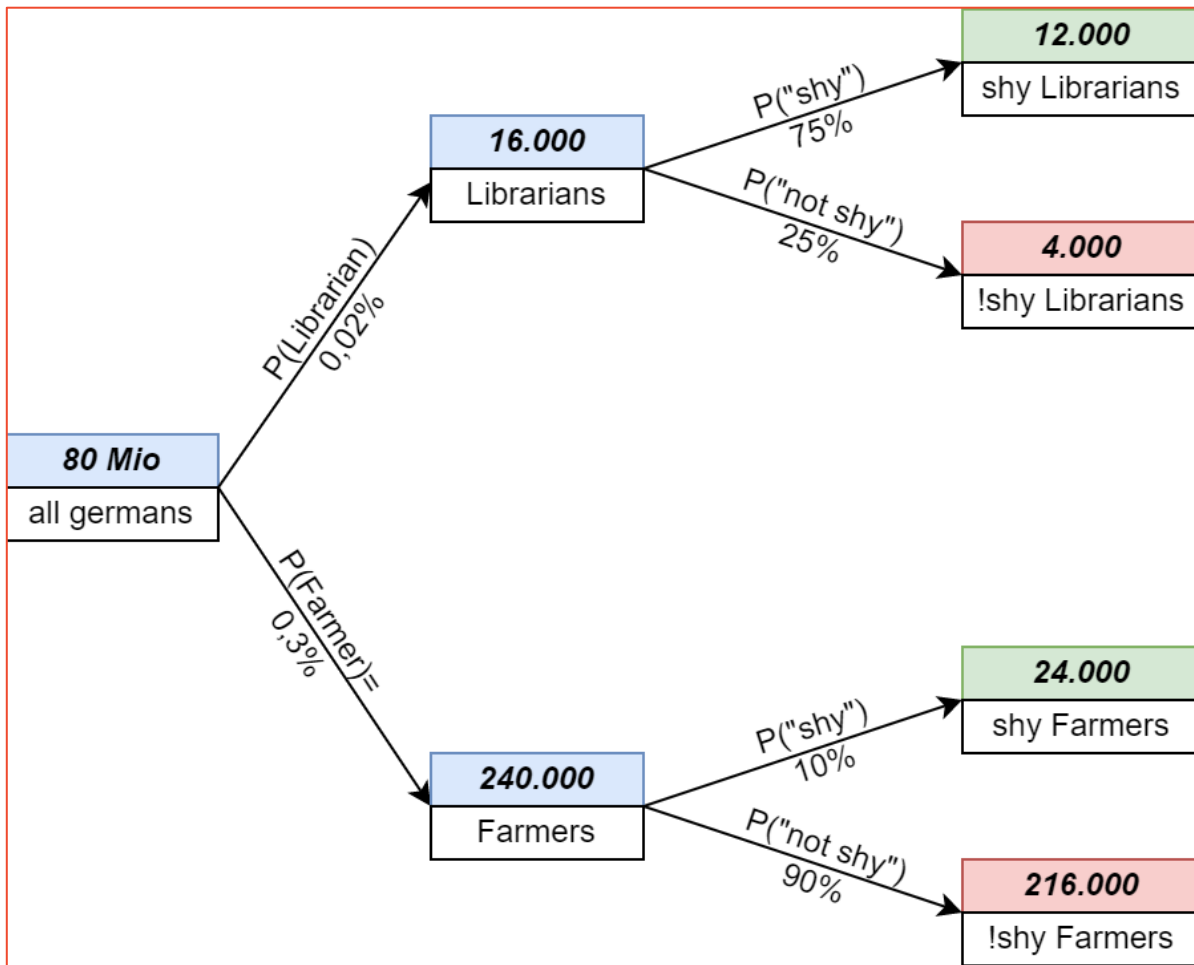


Dieses berühmte Beispiel stammt aus dem Buch

"Thinking, Fast and Slow"

von Daniel Kahnemann und wird verwendet, um die Grenzen des eigenen Denkens zu erkennen.

Menschen neigen nicht dazu, in bedingten Wahrscheinlichkeiten zu denken, es sei denn, sie wurden speziell darauf trainiert. Für den Naive Bayes Classifier müssen die Schüler jedoch genau diese Denkweise entwickeln.



An der Tafel wird der Entscheidungsbaum gemeinsam mit den Schülern entwickelt. An ihm können viele Berechnungen durchgeführt und die wichtigen Konzepte der bedingten Wahrscheinlichkeit abgeleitet werden:

- Wie groß ist die Wahrscheinlichkeit, einen schüchternen Menschen zu finden, wenn er Bibliothekar ist?
- Wie hoch ist die Wahrscheinlichkeit, in der Menge aller Bauern einen schüchternen Menschen zu finden?
- Wie hoch ist also die Wahrscheinlichkeit, einen Bibliothekar unter den schüchternen Menschen zu finden?

Antwort: Sehr gering, denn: Im Allgemeinen gibt es deutlich weniger Bibliothekare als Landwirte,

es gibt also in absoluten Zahlen deutlich weniger schüchterne Bibliothekare als schüchterne Bauern.

Teil II: Die Mathematik eines gemeinsamen "Coronatest"

Die meisten Schülerinnen und Schüler werden die typischen Corona-Schnelltests entweder aus ihrem häuslichen Umfeld oder aus der Schule kennen: Stäbchen, die man sich in die Nase steckt und nach der Behandlung mit Testflüssigkeit abwartet, ob ein damit benetzter Teststreifen die benötigte zweite Markierung anzeigt. Das führt dann zu einer Quarantänezeit. Die Schüler sollten eine sehr emotionale Beziehung zu diesem Test haben.

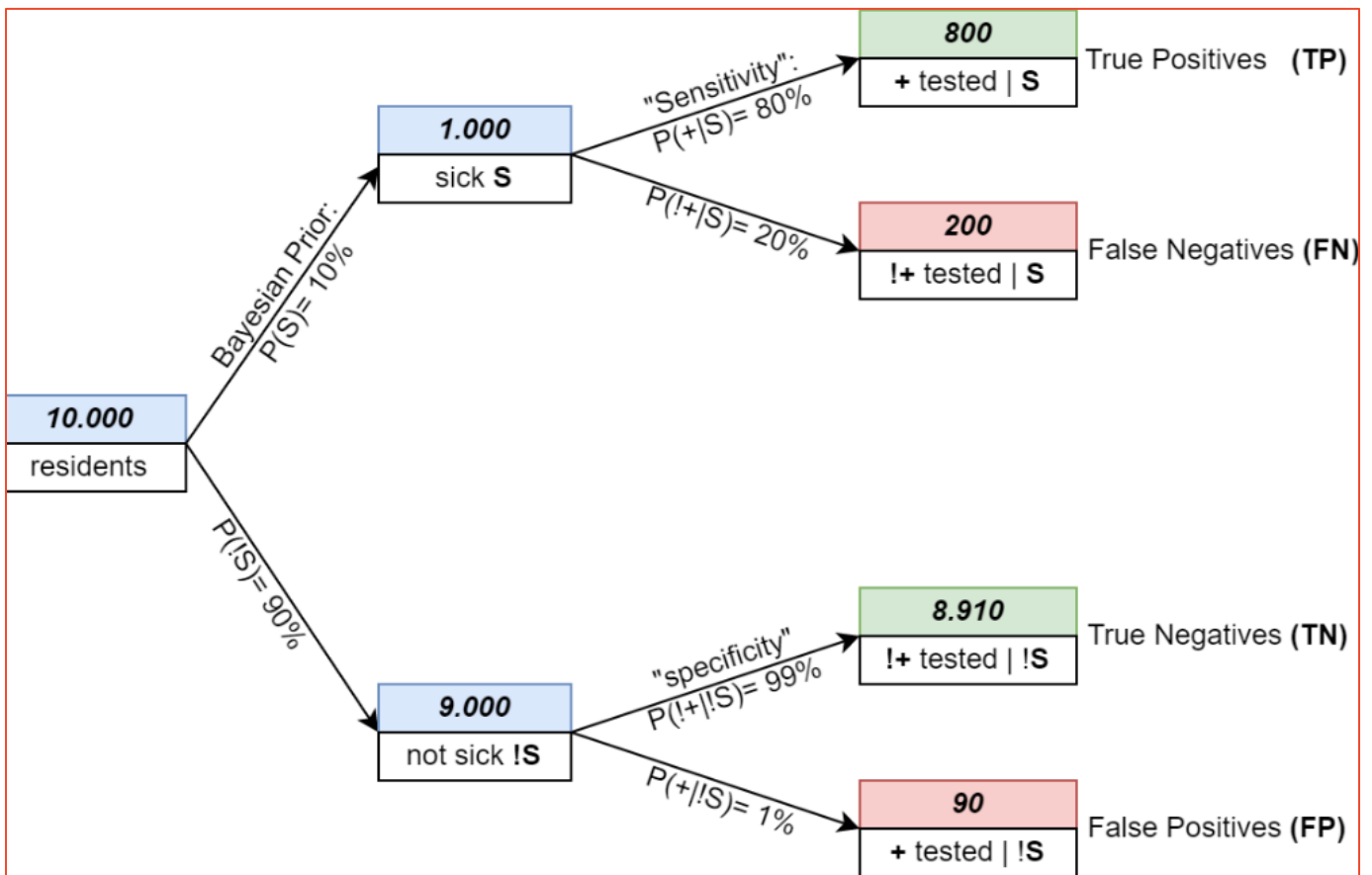


Abb. 32 Bild [\[Gemeinfrei\]](#) erzeugt mit [DALL-E](#); Prompt „doctor examines a male patient with a stethoscope, professional studio photography“ von Jörg [\[CC BY-SA 4.0 International\]](#)

Diese Tests sind jedoch relativ unsicher (weshalb in der Praxis häufig die PCR eingesetzt wurde, um zuverlässige Aussagen zu machen). Die Testeigenschaften sind:

Sensitivität (Recall): Dies ist die Wahrscheinlichkeit, mit der eine kranke Testperson positiv getestet wird.

Spezifität: Dies ist die Wahrscheinlichkeit, mit der eine gesunde Testperson negativ getestet wird.

In beiden Fällen, richtig positiv und richtig negativ getestet, gibt es auch die beiden komplementären Ereignisse, nämlich ein falsch positives oder falsch negatives Testergebnis.

Welches ist ein falsches Positiv und welches ist ein falsches Negativ?

Teil III: Einführung der Konfusionsmatrix

Die Ergebnisse des Entscheidungsbaums können in die so genannte Vierfeldertafel (mathematischer Fachbegriff) oder in die so genannte Konfusionsmatrix eingetragen werden.

Hier ist die gleiche Konfusionsmatrix in doppelter Ausführung:

In die obere werden die Zahlenwerte eingetragen, in die untere die entsprechenden Wahrscheinlichkeiten und Proportionen bzw. Raten.

Das Symbol für 'Not' wurde aus der Programmier-Syntax übernommen, wo es üblicherweise als Ausrufezeichen ("!") dargestellt wird.

Detaillierte Bezeichnung mit konkreten Zahlenwerten

	Tatsächlich krank (S)	Eigentlich nicht Sick (!S)	Summe
positiv getestet	800 " Echt positiv" TP	90 "Falsch positiv" FP	890 = 10.000 Einwohner - P(+)
negativ getestet	200 "Falsch-negative" FN	8.910 "Wahres Negativ" TN	9.110 = 10.000 Einwohner - P(!+)
Summe	1.000 = 10.000 - P(S)	9.000 = 10.000 - P(!S)	10.000 Einwohner

Symbolanzeige mit Kursen oder Anteilen (nummernunabhängig):

	S	!S	Summe
+	$P(S) - P(+ S) = 0,1 - 0,8$ TP-Rate	$P(!S) - P(+ !S) = 0,9 - 0,01$ FP-Rate	$P(+)$ = 0,089
!+	$P(S) - P(!+ S) = 0,1 - 0,2$ FN-Rate	$P(!S) - P(!+ !S) = 0,9 - 0,99$ TN-Rate	$P(!+)$ = 0,911
Summe	$P(S) = 0,1$	$P(!S) = 0,9$	1

Verwendete Terminologie:

Begriff	In normaler Sprache	Terminologie
P(S)	Wahrscheinlichkeit, krank zu sein	<i>Prior</i>
P(+)	Wahrscheinlichkeit, positiv getestet zu werden	<i>Evidenz</i>
P(!S)	Wahrscheinlichkeit, nicht krank zu sein	
P(!+)	Wahrscheinlichkeit, nicht positiv getestet zu werden	
P(+ S)	Wahrscheinlichkeit, positiv getestet zu werden, wenn man krank ist	<i>Sensitivität</i>
P(!+ S)	Wahrscheinlichkeit, nicht positiv getestet zu werden, wenn man krank ist	
P(+ !S)	Wahrscheinlichkeit, positiv getestet zu werden, wenn man nicht krank ist	
P(!+ !S)	Wahrscheinlichkeit, nicht positiv getestet zu werden, wenn man nicht krank ist	
P(S +)	Wahrscheinlichkeit, bei einem positiven Test tatsächlich krank zu sein	<i>Posterior</i>

Woran sind wir interessiert? Die Ableitung des Bayes-Theorems:

Im Anteil der Personen, die unter der Bedingung, dass sie positiv getestet wurden, tatsächlich krank sind, P(S | +):

$$P(S | +) = \frac{TP}{TP + FP}$$

Wahre Positive geteilt durch alle positiv getesteten Personen (alle Positiven: Wahre Positive plus Falsch-Positive)

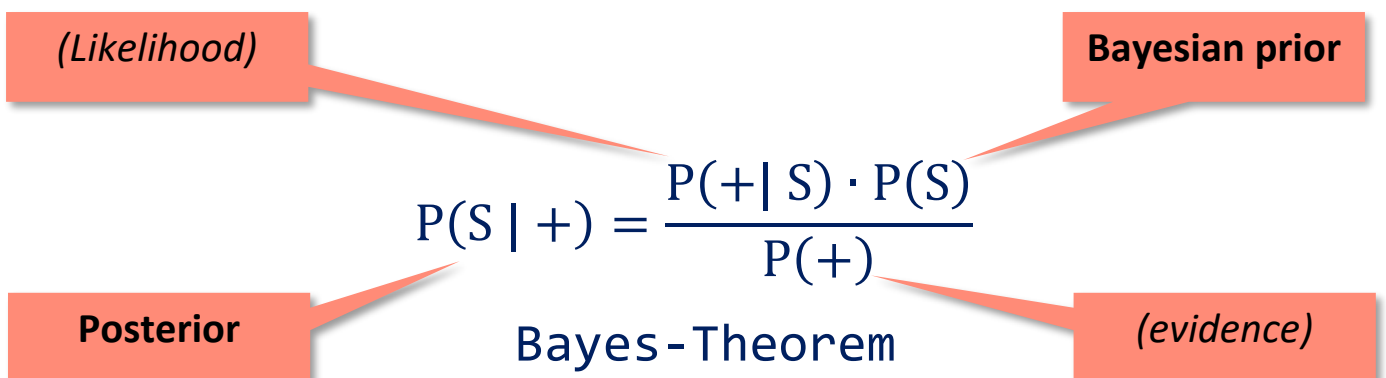
$$P(S | +) = \frac{TP \text{ rate}}{TP \text{ rate} + FP \text{ rate}}$$

Durch die Division durch alle Einwohner wird die absolute Zahl der Personen eliminiert

$$P(S | +) = \frac{P(+ | S) \cdot P(S)}{P(+)}$$

Sehen Sie sich den Baum oben an: TP-Rate = P(S) - P(+ | S). TP-Rate + FP-Rate = Wahrscheinlichkeit, positiv getestet zu werden, P(+)

$$P(S | +) = \frac{800}{800 + 90} = \frac{0.1 \cdot 0.8}{0.089} = 0.898$$



Teil VI: Warum ist der Bayes'sche Prior so wichtig?

Berechnen Sie neu für eine Wahrscheinlichkeit von 30% krank zu sein, $P(S) = 0,3$

	Tatsächlich krank (S)	Eigentlich nicht krank (IS)	Summe
positiv getestet	2400 TP	70 FP	2.470
negativ getestet	600 FN	6.930 TN	7.530
Summe	3.000	7.000	10.000

$$P(S | +) = \frac{2400}{2400 + 70} = \frac{0.3 \cdot 0.8}{0.2470} = 0.971$$

Berechnen Sie neu für eine Wahrscheinlichkeit von 3%, krank zu sein, $P(S) = 0,03$

	Tatsächlich krank (S)	Eigentlich nicht krank (IS)	Summe
positiv getestet	240 TP	97 FP	337
negativ getestet	60 FN	9.603 TN	9.663
Summe	300	9.700	10.000

$$P(S | +) = \frac{240}{240 + 97} = \frac{0.03 \cdot 0.8}{0.0337} = 0.712$$

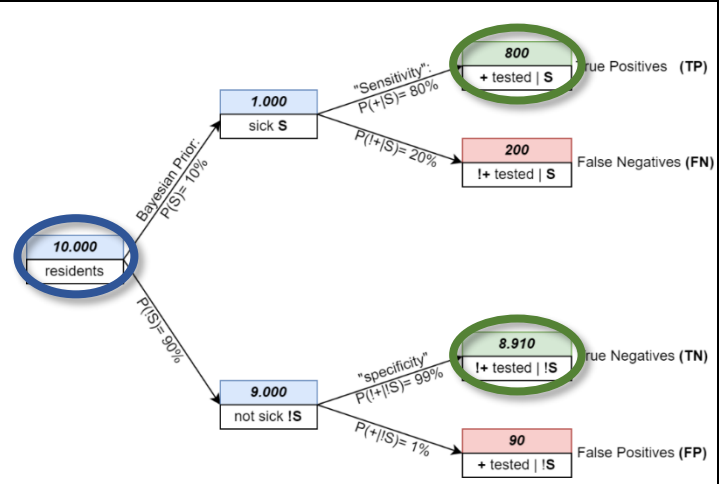
Vergleicht man die drei verschiedenen Berechnungen für das Posterior, so zeigt sich eine unterschiedliche Aussagekraft in Abhängigkeit von den Priorwerten. Je höher die Wahrscheinlichkeit des Priors ist, desto mehr Aussagekraft hat das Posterior.

Teil V: Wichtige Begriffe der Leistungsmetrik:

Accuracy

$$= \frac{\text{Alle korrekt klassifizierte Objekte}}{\text{Alle Objekte}}$$

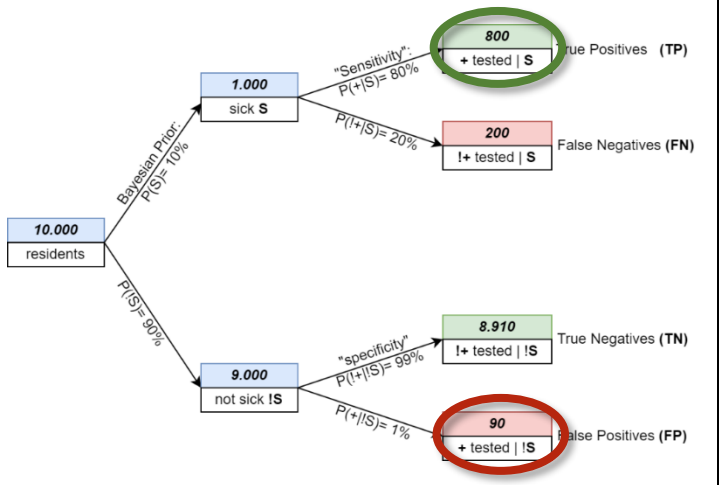
$$= \frac{TP + TN}{TP + TN + FP + FN}$$



Precision

$$= \frac{\text{Alle korrekt klassifizierte positiven Objekte}}{\text{Alle positiv klassifizierte Objekte}}$$

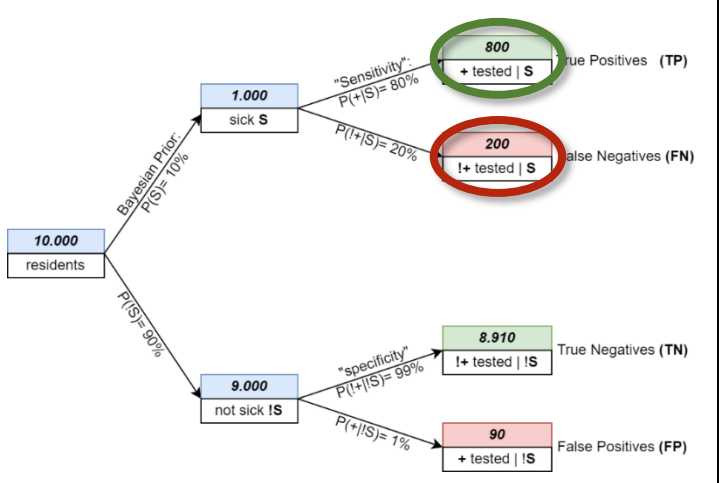
$$= \frac{TP}{TP + FP}$$



Recall (Sensitivity)

$$= \frac{\text{Alle korrekt klassifizierte positiven Objekte}}{\text{Alle tatsächlich positiven Objekte}}$$

$$= \frac{TP}{TP + FN}$$



Teil IV: Vom Bayes-Theorem zum maschinellen Lernen: Naive-Bayes-Klassifikator

Wenn wir unser einfaches Beispiel aus der letzten Lektion wiederholen, erkennen wir, dass das tatsächliche Ergebnis die Klassifizierung ist, ob jemand krank ist oder nicht. Die Vorhersage ist, ob jemand ein positives Corona-Testergebnis hat oder nicht.

Daher bezeichnen wir das Kranksein als unser gewünschtes Ziel und die Testergebnisse als unser zugehöriges Merkmal.

Wir bezeichnen unser Ziel mit dem kleinen Buchstaben y und unsere Merkmale mit dem großen Buchstaben X , was eine sehr gebräuchliche und weit verbreitete Terminologie im maschinellen Lernen ist. Wir können also unsere Bayes-Theorem-Formel umformulieren:

$$P(S | +) = \frac{P(+ | S) \cdot P(S)}{P(+)} \quad \longrightarrow \quad P(y | X) = \frac{P(X | y) \cdot P(y)}{P(X)}$$

Mit y als Zielbezeichner
und X als Merkmalsbezeichner

"Wahrscheinlichkeit, krank zu sein bei einem positiven Testergebnis."

"Wahrscheinlichkeit einer bestimmten Klassifizierung bei einem bestimmten Merkmalswert."

Jetzt können wir den naiven Bayes-Klassifikator einsetzen.

Stellen Sie sich eine Polizeistatistik über gestohlene Autos vor. Für jedes Auto wurden die Umstände erfasst und tabellarisch dargestellt:

Nr.	Geparkt	Typ	Farbe	Gestohlen?
1	auf der Straße	Sport	Rot	Ja
2	auf der Straße	Sport	Rot	Nein
3	auf der Straße	Sport	Rot	Ja
4	auf der Straße	Sport	Blau	Nein
5	in der Garage	Sport	Blau	Ja
6	in der Garage	Kompakt	Blau	Nein
7	in der Garage	Kompakt	Rot	Ja
8	auf der Straße	Kompakt	Rot	Nein
9	in der Garage	Kompakt	Rot	Nein
10	in der Garage	Sport	Rot	Ja

X
(Merkmale)
 y
(Ziel)

Hier haben wir drei Merkmale anstelle eines Merkmals und müssen daher unsere Formel für die Wahrscheinlichkeit anpassen

$$P(y | X) = \frac{P(X | y) \cdot P(y)}{P(X)}$$

$$P(\text{gestohlen} | \text{Geparkt, Typ, Farbe}) = \frac{P(\text{Geparkt, Typ, Farbe} | \text{gestohlen}) \cdot P(\text{gestohlen})}{P(\text{Geparkt, Typ, Farbe})}$$

Wir suchen wieder nach der bedingten Wahrscheinlichkeit auf der linken Seite der Gleichung, der Nachfolgewahrscheinlichkeit:

Die Wahrscheinlichkeit, dass ein Auto gestohlen wird, wenn man weiß,

- **wie es geparkt war,**
- **welchen Typ es hatte und**
- **welche Farbe es hatte.**

Und hier kommt der naive Teil des Naive Bayes Klassifikators:

Wir gehen davon aus, dass alle Merkmale unabhängig voneinander sind. Das bedeutet in diesem Beispiel:

Ob ein bestimmtes Auto vom Typ "Sport" ist, ist völlig unabhängig von seiner Farbe und unabhängig davon, wie es geparkt wurde. Diese Merkmale können als unabhängig interpretiert werden, wie drei gleichzeitig geworfene Spielwürfel. Die Wahrscheinlichkeiten können also mit Hilfe der Kettenregel umgeschrieben werden:

$$P(\text{Geparkt, Typ, Farbe}) = P(\text{Geparkt}) \cdot P(\text{Type}) \cdot P(\text{Farbe})$$

$$P(\text{Geparkt, Typ, Farbe} | \text{Stolen}) = P(\text{Geparkt} | \text{gestohlen}) \cdot P(\text{Typ} | \text{gestohlen}) \cdot P(\text{Farbe} | \text{gestohlen})$$

In tabellarischer Form sehen die Rechnungen so aus:

Farbe	Gestohlen?
Blau	Nein
Blau	Nein
Blau	Ja
Rot	Nein
Rot	Nein
Rot	Nein
Rot	Ja
Rot	Ja
Rot	Ja
Rot	Ja

	Gestohlen ("Ja")	Nicht gestohlen	Summe
Rot	4	3	7
Blau	1	2	3
Summe	5	5	10

	P(Ja)	P(Nein)
Rot	4/7	3/7
Blau	1/3	2/3

Typ	Gestohlen?
Kompakt	Nein
Kompakt	Nein
Kompakt	Nein
Kompakt	Ja
Sport	Nein
Sport	Nein
Sport	Ja
Sport	Ja
Sport	Ja
Sport	Ja

	Gestohlen ("Ja")	Nicht gestohlen	Summe
Kompakt	1	3	4
Sport	4	2	6
Summe	5	5	

	P(Ja)	P(Nein)
Kompakt	1/4	3/4
Sport	4/6	2/6

Geparkt	Gestohlen?
in der Garage	Nein
in der Garage	Nein
in der Garage	Ja
in der Garage	Ja
in der Garage	Ja
auf der Straße	Nein
auf der Straße	Nein
auf der Straße	Nein
auf der Straße	Ja
auf der Straße	Ja

	Gestohlen ("Ja")	Nicht gestohlen	Summe
Garage	3	2	5
Straße	2	3	5
Summe	5	5	

	P(Ja)	P(Nein)
Garage	3/5	2/5
Straße	2/5	3/5

Bevor wir uns in die Berechnungen stürzen, können wir die Formel noch einmal vereinfachen:

$P(\text{Geparkt})$, $P(\text{Typ})$ und $P(\text{Farbe})$ ändern sich nicht, sie sind Konstanten. Daher können wir sie vernachlässigen:

$$P(\text{Parked}) \cdot P(\text{Type}) \cdot P(\text{Color}) = \text{constant}$$

Unsere Formel ändert sich von...

$$P(\text{Stolen} | \text{Parked, Type, Color}) = \frac{P(\text{Parked, Type, Color} | \text{Stolen}) \cdot P(\text{Stolen})}{P(\text{Parked}) \cdot P(\text{Type}) \cdot P(\text{Color})}$$

...zu:

$$P(\text{Stolen} | \text{Parked, Type, Color}) \propto P(\text{Parked, Type, Color} | \text{Stolen}) \cdot P(\text{Stolen})$$

Berechnen Sie die Wahrscheinlichkeit, ob ein Auto gestohlen wird oder nicht, wenn es rot ist, einen kompakten Typ hat und auf der Straße parkt:

$$P(\text{Yes} | \text{Street, Compact, Red}) = P(\text{Street} | \text{Yes}) \cdot P(\text{Compact} | \text{Yes}) \cdot P(\text{Red} | \text{Yes}) \cdot P(\text{Yes})$$

$$P(\text{Yes} | \text{Street, Compact, Red}) = \frac{2}{5} \cdot \frac{1}{4} \cdot \frac{4}{7} \cdot \frac{1}{2} = \frac{8}{280}$$

Dies ist zu vergleichen mit:

$$P(\text{No} | \text{Street, Compact, Red}) = P(\text{Street} | \text{No}) \cdot P(\text{Compact} | \text{No}) \cdot P(\text{Red} | \text{No}) \cdot P(\text{No})$$

$$P(\text{No} | \text{Street, Compact, Red}) = \frac{3}{5} \cdot \frac{3}{4} \cdot \frac{3}{7} \cdot \frac{1}{2} = \frac{27}{280}$$

Da $27/280 > 8/280$ ist, stufen wir unser Auto mit diesem speziellen Merkmalsatz als NICHT gestohlen ein.

Lektion 11: Lineare Regression

Was sollen die Schüler lernen?

Der zweite Teil der "großen Drei" Klassifizierung, Clustering und Regression: Datenanalyse von korrelierten Daten. Kontinuierliche Vorhersage mit Hilfe von Inter- und Extrapolationen. Daher müssen die Schüler eine "Trendlinie" in einer Tabellenkalkulationssoftware wie Microsoft Excel, LibreOffice Calc oder Google Tables erstellen. Die Qualität der Vorhersage kann mithilfe des sogenannten Korrelationskoeffizienten geschätzt werden.



Abb. 33 Bild [Gemeinfrei] erzeugt mit DALL-E; Prompt „man with beard and open mouth, very surprised, photorealistic“ von Jörg [CC BY-SA 4.0 International]

Mögliche Aktivitäten und Aufgaben der Schüler:

Wann wird mehr Speiseeis verkauft? Wenn es warm ist. Können Sie daraus ein Theorem formulieren? Je höher die Temperatur, desto mehr Eis wird verkauft? Die Daten geben Aufschluss über die Hypothese: Eine lineare Regression mit einem hohen Korrelationskoeffizienten bedeutet, dass das Modell gut zur Vorhersage verwendet werden kann. Was bedeuten in diesem Zusammenhang Korrelationskoeffizient, Residuen, Summe der kleinsten Quadrate, y-Achsenabschnitt und Steigung? Die Schüler erarbeiten die grundlegenden Konzepte anhand eines einfachen Eisverkäufer-Beispiels, lernen die mathematische Interpretation anhand einer Simulation und wenden sie später in einem kleinen Projekt an. Dabei wird die Bedeutung der Korrelationskoeffizienten deutlich. Für ältere Schülerinnen und Schüler kann die Methode des Gradientenabstiegs erklärt werden; die mathematischen Grundlagen sollten ab Klasse 10 vorhanden sein.

Teil I: Die Geschichte

Ein Eisverkäufer möchte die voraussichtlichen Verkaufszahlen für das nächste Wochenende vorhersagen. Das Einzige, was er weiß, ist die Wettervorhersage: Es wird ein sonniger Nachmittag und die Temperatur wird wahrscheinlich 13°C am Samstag und 21°C am Sonntag betragen. Bitte helfen Sie ihm, dieses Wochenende zu organisieren und zu planen. Er hat eine Tabelle mit den Werten der letzten Wochen:

Temperatur (°C)	Anzahl der verkauften Eiscreme
12	180
14	220
15	330
16	325
17	400
18	425
20	480
13	?
21	?



Abb. 34 Bild [Gemeinfrei] erzeugt mit DALL-E; Prompt „ice cream parlor with many types of ice cream, sunshine“ von Jörg [CC BY-SA 4.0 International]

Teil II: Modellierung in GoogleDocs / Microsoft Excel / LibreOffice Calc usw.

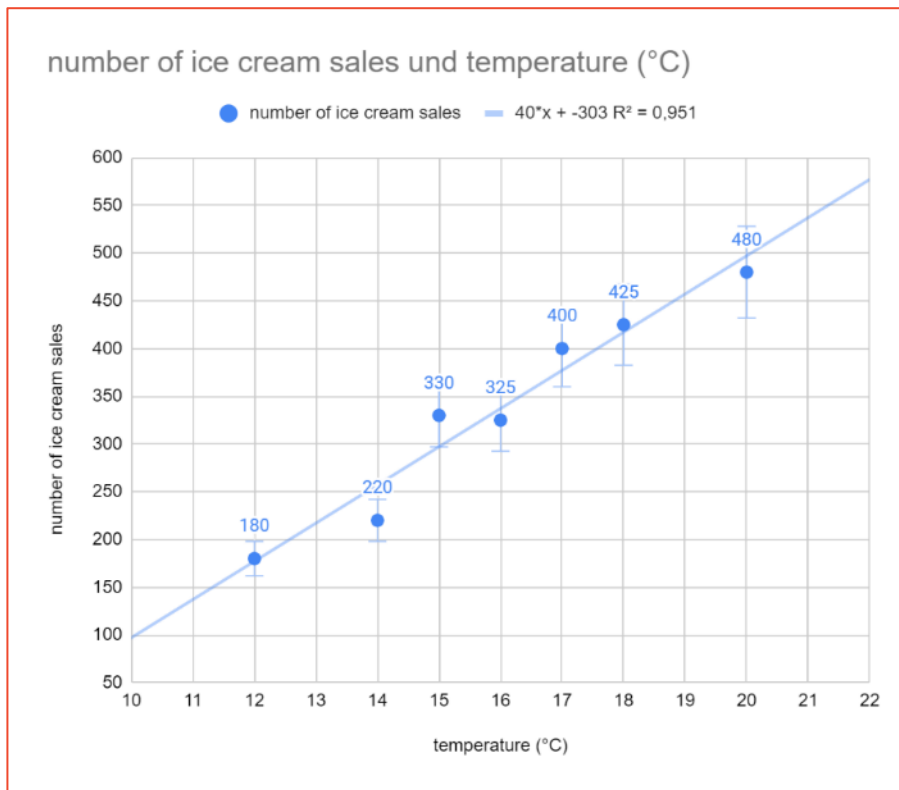


Abbildung 11: Das in Google Tables erstellte Streudiagramm. Die Regressionsberechnung ist bereits durchgeführt, und die Gleichung wird angezeigt. Dies ist die Ausgangsbasis für die Übungen.

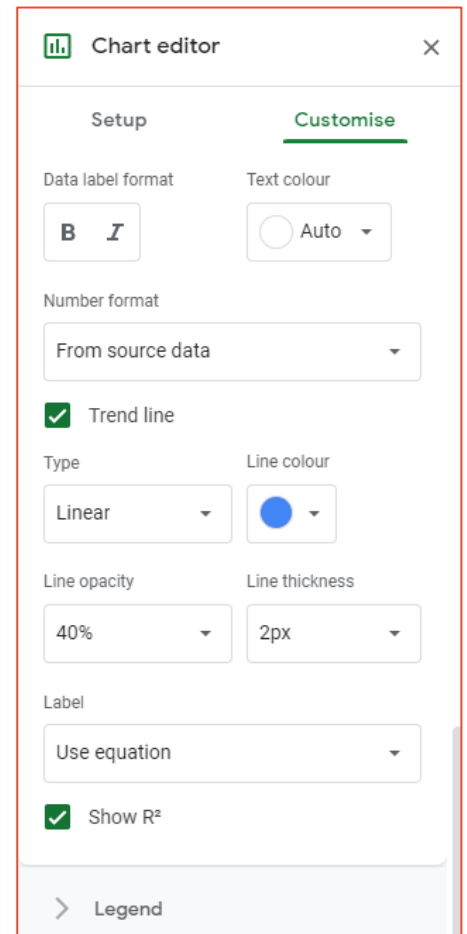


Figure 12: Necessary settings for calculating the regression equation. First you have to build a scatter plot from the data table. Then you have to add a trendline to the plot. At last, you need to format the trendline for displaying equation and squared correlation coefficient.

Regressionsfunktion:

$$y = 40 \cdot x - 303$$

Teil III: Übungen:

Berechnen Sie den wahrscheinlichen Umsatz von Speiseeis bei der folgenden Temperatur:

- a) 0°C Antwort: -303 Stück Eiscreme, macht keinen Sinn
- b) 10°C Antwort: $400 - 303 = 97$ Stück Speiseeis
- c) 13 und 21°C Antwort: 13°C: 217 Eiscreme | 21°C: 537 Eiscreme

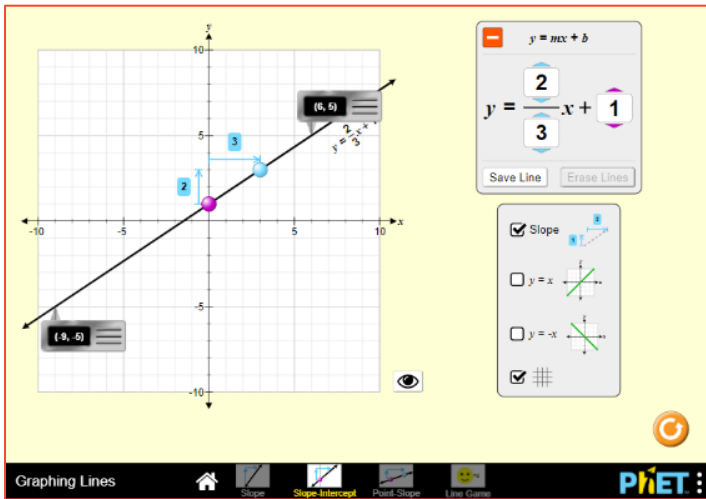
Berechnen Sie einen **Rückstand nach** Ihrer Wahl. Zum Beispiel, bei 15°C:

$$y_{\text{Predicted}} = 40 \cdot 15 - 303 = 600 - 303 = 297$$

$$y_{\text{measured}} = 330$$

$$\text{Residue}_{x=15} = 330 - 297 = 33$$

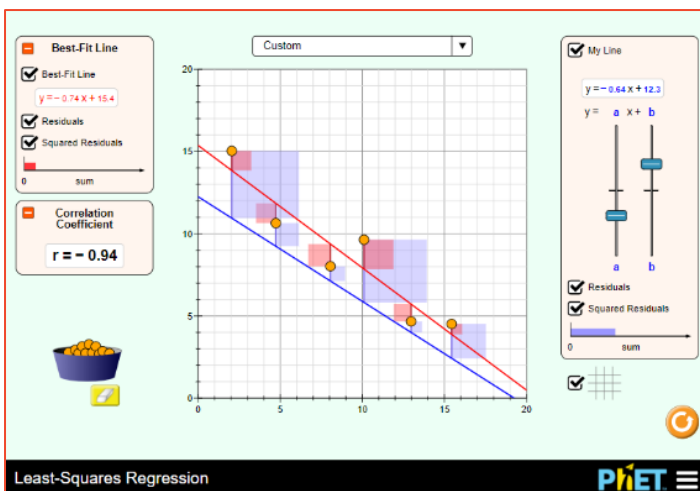
Teil IV: Wie man lineare Regression simuliert. Phet und Gradientenabstieg



Lineare Gleichungen

https://phet.colorado.edu/sims/html/graphing-lines/latest/graphing-lines_en.html

Wenn es notwendig erscheint, die Grundlagen der linearen Gleichungen zu wiederholen oder zu vertiefen, können Sie mit dieser Simulation beginnen. Die Grundlagen von Y-Achsenabschnitt und Steigung werden dabei interaktiv erarbeitet und geübt. Ein interessantes Rätsel ("Linienspiel") rundet die Simulation ab.



Regression der kleinsten Quadrate

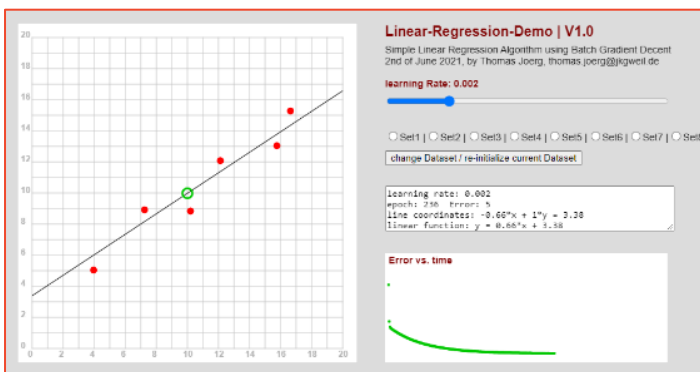
<https://phet.colorado.edu/en/simulation/least-squares-regression>

Anhand vieler integrierter Datensätze können die Grundlagen der Regression erarbeitet werden, z. B.:

kleinste Summe der Quadrate, Residuen und quadrierte Residuen (sie werden visualisiert!)
Korrelationskoeffizient (einschließlich Vorzeichen)

Die Möglichkeit, eigene Datensätze einzugeben und eigene Geraden zu schätzen, macht die Anwendung enorm vielseitig

Abb. 35 Eigene Screenshots aus dem Webbrowser



Lineare Regression mit Gradientenabstieg

<https://iludis.de/LinearRegression/index.html>

Der Unterschied zwischen der klassischen linearen Regression in Statistik und ML besteht darin:

1. In der Statistik sucht man nach Korrelationen, beim maschinellen Lernen sucht man nach Vorhersagen.
2. Die statistische Mathematik arbeitet analytisch, d.h. die Regressionsgleichungen werden mit Formeln berechnet.

Beim maschinellen Lernen verwendet man Optimierungsalgorithmen, wie z.B. die Gradientenabstiegsmethode. Dabei wird zunächst eine Gerade zufällig bestimmt und eine Fehlerfunktion für die Datenpunkte berechnet. Der Gradientenabstieg verwendet die Restquadrate: Die Formel für die Berechnung wird abgeleitet - daraus wird der Gradient bestimmt. Will man die Geradengleichung weiter optimieren, muss man die Parameter in Richtung des negativen Gradienten verändern. Mathematisch gesehen läuft man also genau in die Richtung, in der die Restquadrate abnehmen. Und genau so kommt man iterativ zu der bestmöglichen Geraden. Diese wird also nicht analytisch berechnet, sondern iterativ optimiert: Das ist das Prinzip des Gradientenabstiegs, das auch in vielen anderen Bereichen des maschinellen Lernens immer wieder angewendet wird.

Teil V (fakultativ) Ableitung des Gradientenabstiegs bei der linearen Regression

1. Kostenfunktion KF() in der linearen Regression: KF() ist der Mittelwert aller Restquadrate
2. Speicher Mittelwert: Wenn 'n' Werte vorhanden sind, muss durch 'n' geteilt werden
3. Residuum: Die Differenz zwischen Datenwert und Regressionswert ($y = mx + c$)
4. Die Kostenfunktion bildet eine Parabel, und diese Parabel kann abgeleitet werden.
5. Die Optimierung erfolgt durch Anpassung der Steigung (m) und der Y-Achsenabschnitte (c).
6. Sie muss daher nach diesen beiden Komponenten abgeleitet werden (höhere Mathematik mit sogenannten partiellen **Ableitungen muss nicht verstanden werden**)

$$KF() = \frac{1}{n} \sum_{i=1}^n (y_{i, \text{DataPoint}} - y_{i, \text{RegressionLine}})^2$$

$$KF() = \frac{1}{n} \sum_{i=1}^n (y_{i, \text{DataPoint}} - (mx_i + c))^2$$

Ausmultiplizieren, $y_{\text{DataPoint}}$ ist jetzt y , die Indizes 'i' werden vernachlässigt:

$$\begin{aligned} (y_{\text{DataPoint}} - y_{\text{RegressionLine}})^2 &= \\ (y - (mx + c))^2 &= \\ (y - mx - c)^2 &= \\ (y - mx - c) \cdot (y - mx - c) &= \\ y^2 - ymx - yc - ymx + m^2x^2 + mxc - yc + mxc + c^2 &= \\ y^2 - 2ymx - 2yc + m^2x^2 + 2mxc + c^2 & \end{aligned}$$

deriving according to 'm': $\frac{\text{Derivative KF()}}{\text{according to m}} = -2yx + 2mx^2 + 2xc =$

$$\begin{aligned} -2x \cdot (y - mx - c) &= \\ -2x \cdot (y - (mx + c)) &= \\ -2x \cdot (y_{\text{DataPoint}} - y_{\text{RegressionLine}}) &= \\ -2x \cdot \text{ResidualValue} & \end{aligned}$$

deriving according to 'c': $\frac{\text{Derivative KF()}}{\text{according to c}} = -2y + 2mx + 2c =$

$$\begin{aligned} -2 \cdot (y - mx - c) &= \\ -2 \cdot (y - (mx + c)) &= \\ -2 \cdot \text{ResidualValue} & \end{aligned}$$

Iterativ wird dann die LearningRate nach der folgenden Formel optimiert:

$$\begin{aligned} c_{\text{new}} &= c_{\text{old}} - 2 \cdot \text{ResidualValue} \cdot \text{LearningRate} \\ m_{\text{new}} &= m_{\text{old}} - 2 \cdot x_i \cdot \text{ResidualValue} \cdot \text{LearningRate} \end{aligned}$$

Einige Formeln für lineare Regressionsberechnungen

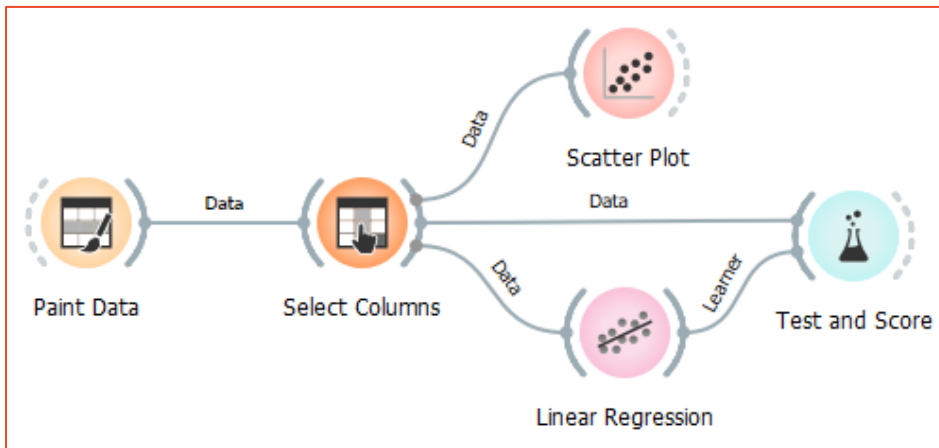
(geheimes Wissen der Lehrer, das die Schüler nicht zu kennen brauchen)

<p>Mittlerer Wert von x: \bar{x} [hier: die Temperatur]</p> $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 16$	<p>Mittlerer Wert von y: \bar{y} [hier: der Speiseeisverkauf]</p> $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = 337,14$
<p>Standardabweichung von x: s_x</p> $s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = 2,45$	<p>Standardabweichung von y: s_y</p> $s_y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} = 100,49$
<p>Kovarianz: Cov_{xy}</p> $Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) = 240$	<p>Korrelation R_{xy} = standardisierte Kovarianz Cov_{xy}</p> $R_{xy} = \frac{Cov_{xy}}{S_x \cdot S_y} = 0,97$
<p>Regressionskoeffizienten (a: Koeffizient, b: Konstante)</p> $y = a \cdot x + b$ $a = \frac{S_y}{S_x} \cdot R_{xy} = 40 \quad b = -\frac{S_y}{S_x} \cdot R_{xy} \cdot \bar{x} + \bar{y} = -303$	<p>Bestimmungskoeffizient (R-Quadrat):</p> $R^2 = (R_{xy})^2 = 0,951$

Teil VI: Interpretation des Korrelationskoeffizienten im Orange Data Mining

Je besser die Datenpunkte an die Regressionslinie angepasst sind:

- desto kleiner die Summe der quadrierten Residuen
- desto näher liegt der Korrelationskoeffizient an einem Wert von +1 für eine positive Steigung oder -1 für eine negative Steigung



Das komplette Setup

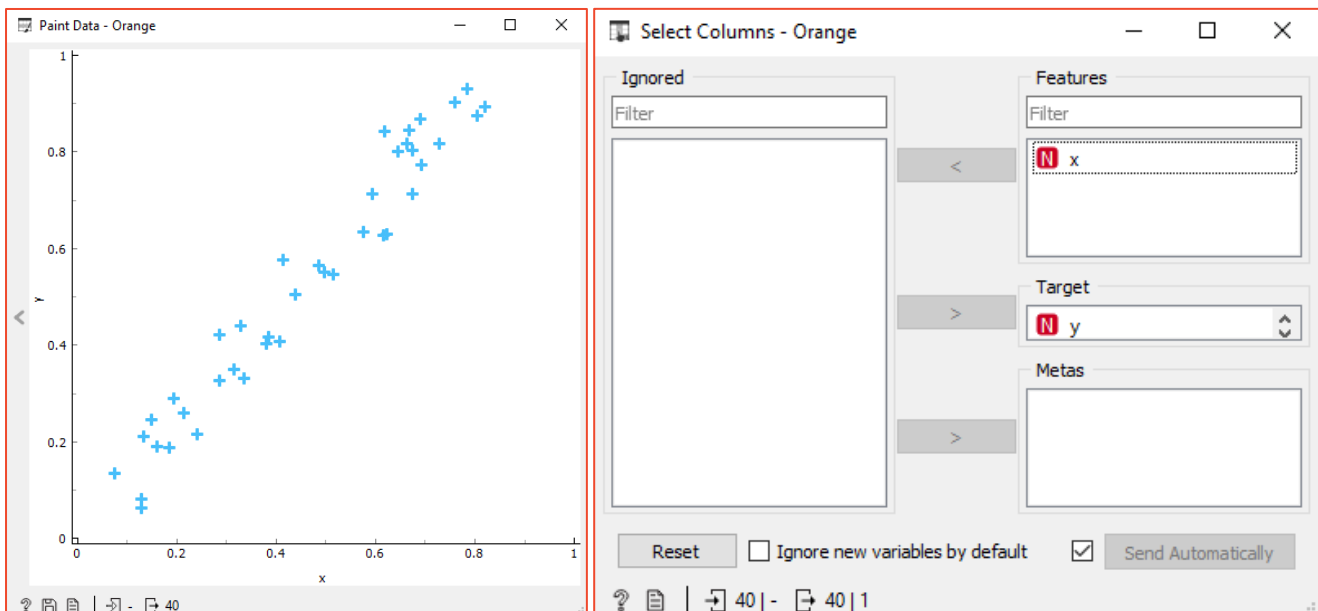
des Projekts: Ein Scatterplot zur Visualisierung der Regressionslinie, ein Linearer Regressionsknoten zusammen mit einem Test- und Score-Knoten für Leistungsmetriken.

Abb. 36 Eigene Screenshots aus Orange Datamining

Erläuterung der Einrichtung und der Knotenpunkte:

Daten malen und Spalten auswählen: Vorbereiten des Testdatensatzes

Abb. 37 Eigene Screenshots aus Orange Datamining



Einige synthetische Daten werden gezeichnet, da wir nur die Bedeutung des Wertes des Korrelationskoeffizienten verstehen wollen.

Variante 1: Starke positive Korrelation von ca. +1 (+0,98)

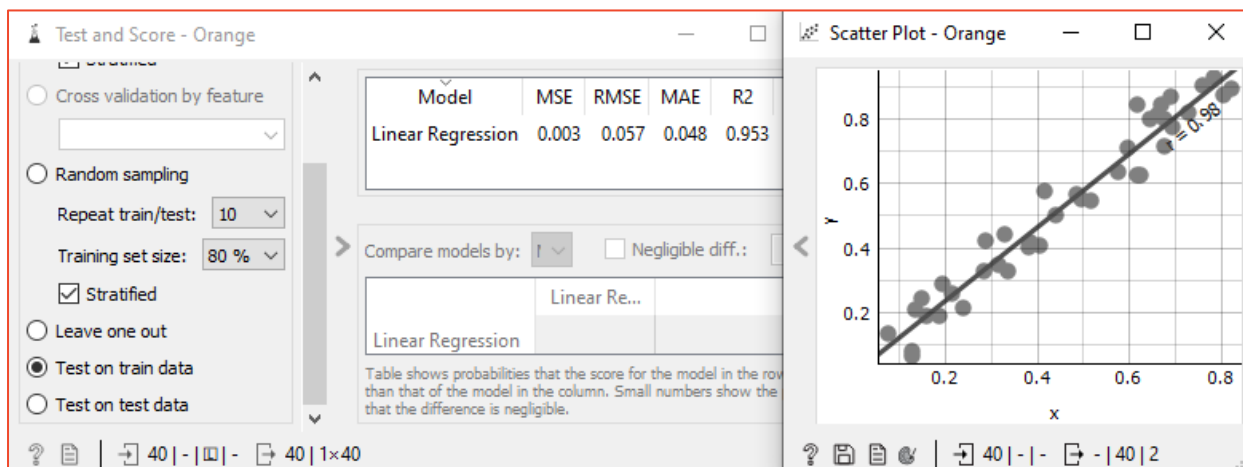


Abb. 38 Eigene Screenshots aus Orange Datamining

Variante 2: Starke negative Korrelation von ca.-1 (-0,98)

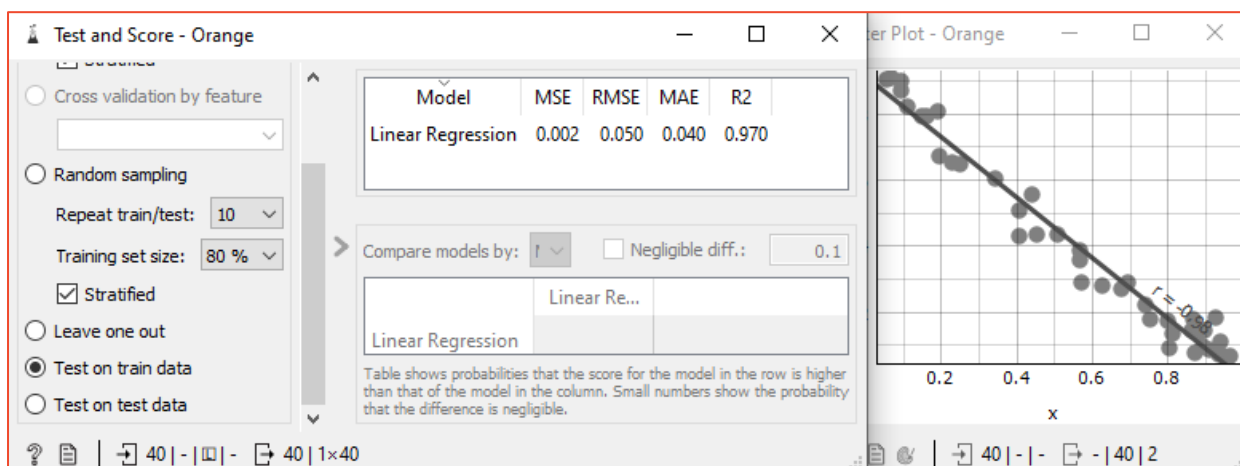


Abb. 39 Eigene Screenshots aus Orange Datamining

Variante 3: Keine Korrelation von ca. 0

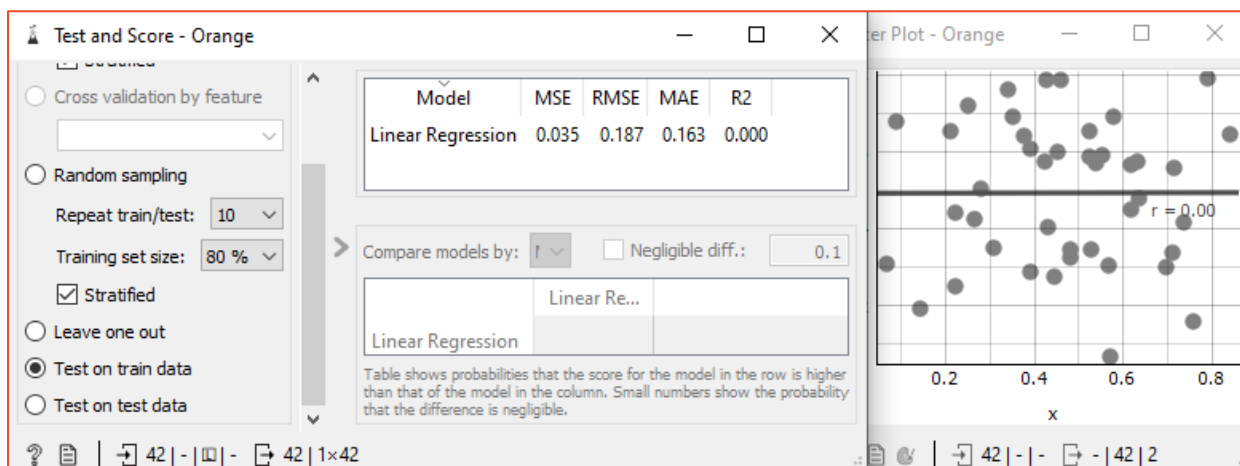
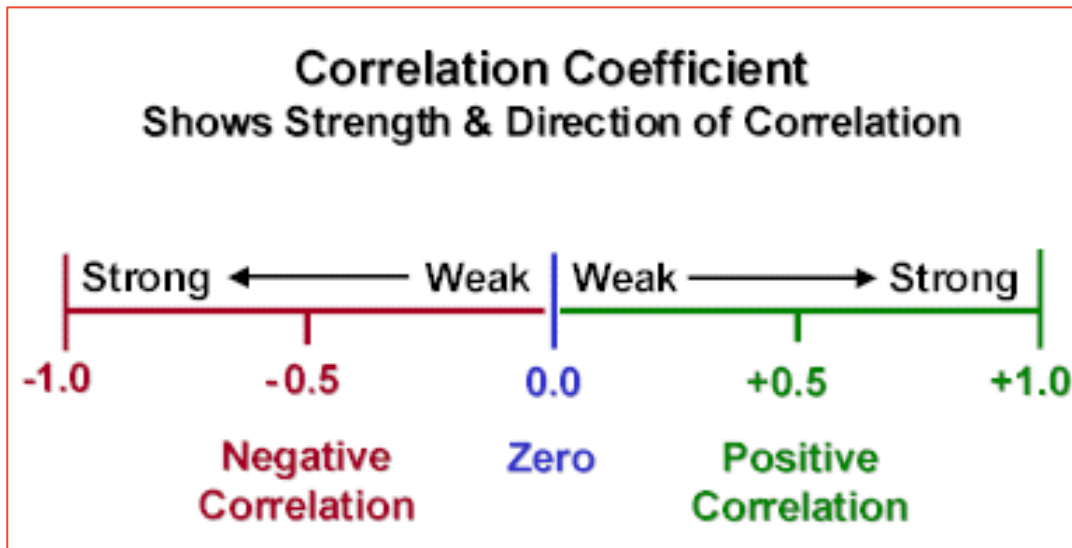


Abb. 40 Eigene Screenshots aus Orange Datamining

Zusammenfassung:

Die Korrelationskoeffizienten können zwischen den Werten -1 und +1 variieren. Ein Wert, der betragsmäßig nahe bei 1 liegt, bedeutet eine starke Korrelation zwischen x- und y-Werten. Ein positives Vorzeichen bedeutet hier eine positive Steigung, entsprechend komplementär ist das negative Vorzeichen. Ein Wert in der Nähe von 0 bedeutet, dass die Variablen nicht miteinander korrelieren.



Lektion 12: Unterscheiden Sie zwischen Korrelation und Kausalität!

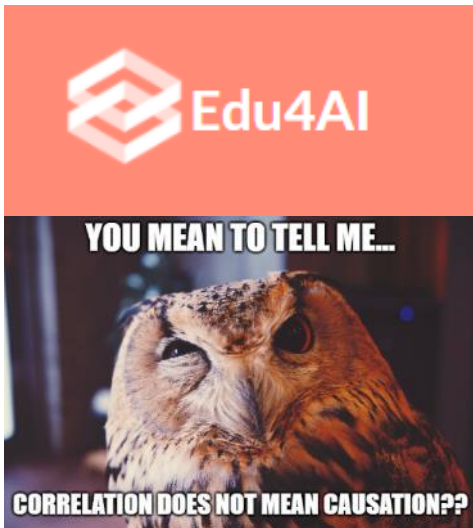


Abb. 41 Photo by [Joe Green](#) on [Unsplash](#)

Was die Schüler lernen sollten:

Manchmal sieht man Verbindungen, wo es keine gibt, und das kann irreführend sein. Hier sind einige Beispiele: Gibt es einen Zusammenhang zwischen dem Wetter und dem Flusspegel? Ja, sicher, denn es ist zu erwarten, dass bei stärkeren Regenfällen die Flusspegel steigen! Gibt es einen Zusammenhang zwischen dem Wetter und der Größe Ihrer Schuhe? Wahrscheinlich nicht!

Natürlich könnte es sein, dass es einen offensichtlichen Zusammenhang zwischen Ihrer Schuhgröße und zum Beispiel der durchschnittlichen Jahrestemperatur gibt. Schließlich leben wir in Zeiten des Klimawandels, und wenn Ihre Füße wachsen, kann auch die Jahrestemperatur steigen. Aber das ist nur eine so genannte "Korrelation": Die Korrelation ist ein Maß für die Beziehung zwischen Variablen. Sie sagt zunächst einmal nichts darüber aus, ob es sich um ein Ursache-Wirkungs-Prinzip handelt.

Mögliche Aktivitäten und Aufgaben der Schüler:

Lassen Sie die Schüler jeweils ein Beispiel von der Website auswählen und beschreiben, warum man sich von den hier gezeigten Diagrammen täuschen lassen könnte. In diesem Zusammenhang kann auch der Begriff "**Cum hoc ergo propter hoc**" recherchiert und einbezogen werden. Gefälschte Kausalitäten können bis in den Bereich des Aberglaubens reichen. Wie kommt es zum Aberglauben?

Wenn verschiedene Variablen in einem Abhängigkeitsverhältnis zueinander stehen, spricht man von "Kausalität". Die Schuhgröße zum Beispiel steht in einem kausalen Zusammenhang mit der Körpergröße - und das ist leicht zu erkennen.

Werfen wir einen Blick auf die Website von Tyler Vigen:

<https://tylervigen.com/spurious-correlations>

Er ist ein Mathematikprofessor mit einem lustigen Hobby: Er findet Korrelationen - also Merkmale, die scheinbar zusammenhängen, aber keinen begründbaren Zusammenhang haben.

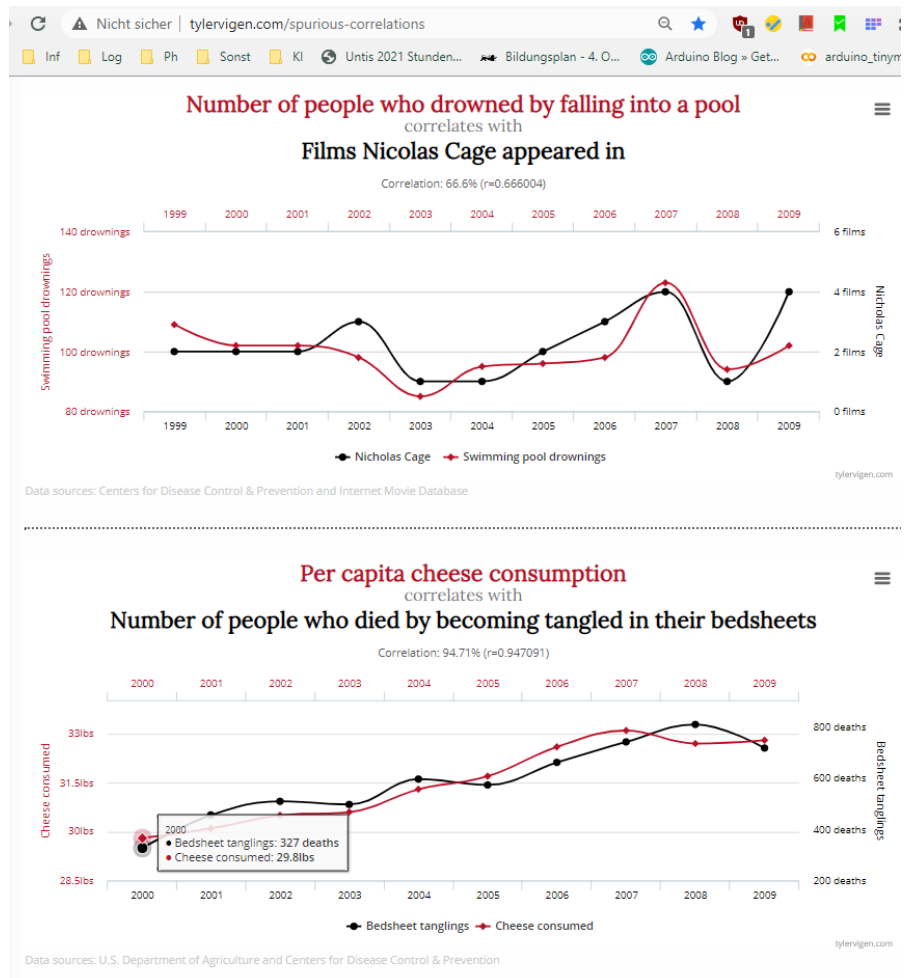


Abb. 42 Eigene Screenshots vom Webbrowser

Lektion 13: k-Means-Clustering.

Oder: Wie geht man mit nicht gelabelten Daten um?

Was die Schüler lernen sollten:



Abb. 43 r/datascience [postal_dude] K-means be like: Mine ! MINE ! MINE ! 2019

Wenn wir mit ungelabelten Daten umgehen müssen, sind unüberwachte Algorithmen manchmal in der Lage, Zusammenballungen von Datenpunkten zu finden und können daher unsere Daten durch die Bildung von Clustern strukturieren. Der k-means-Clustering-Algorithmus ist ein bekanntes Beispiel für diese Lernalgorithmen. Aber wie kann man bestimmen, wie viele Cluster angemessen sind - wie bestimmt man den 'k'-Wert?

Wir können ein Ellbogen-Diagramm verwenden und den charakteristischen Knick in der Diagrammlinie finden!

Mögliche Aktivitäten und Aufgaben der Schüler:

Anhand eines intuitiven Problems, bei dem die Lösung auf der Hand liegt, wird das Konzept des Clustering eingeführt. Der Algorithmus wird durch Beobachtung einer schrittweisen Simulation ausgearbeitet.

Anhand eines eindimensionalen Essensszenarios wird der erarbeitete Algorithmus von den Schülern angewendet und das Clustering einmal von Hand berechnet. Ein kleines Projekt in Orange Data Mining mit synthetischen, selbst gemalten Daten rundet die Lektion ab.

Als praxisbezogene Zusatzaufgabe kann der ‚Mall-Customers‘-Datensatz von Kaggle verwendet werden. Mit ihm kann man die Besucher eines Einkaufszentrums in verschiedene Kategorien einordnen. Und das ist für das Management wichtig, um gezieltes Marketing zu betreiben.

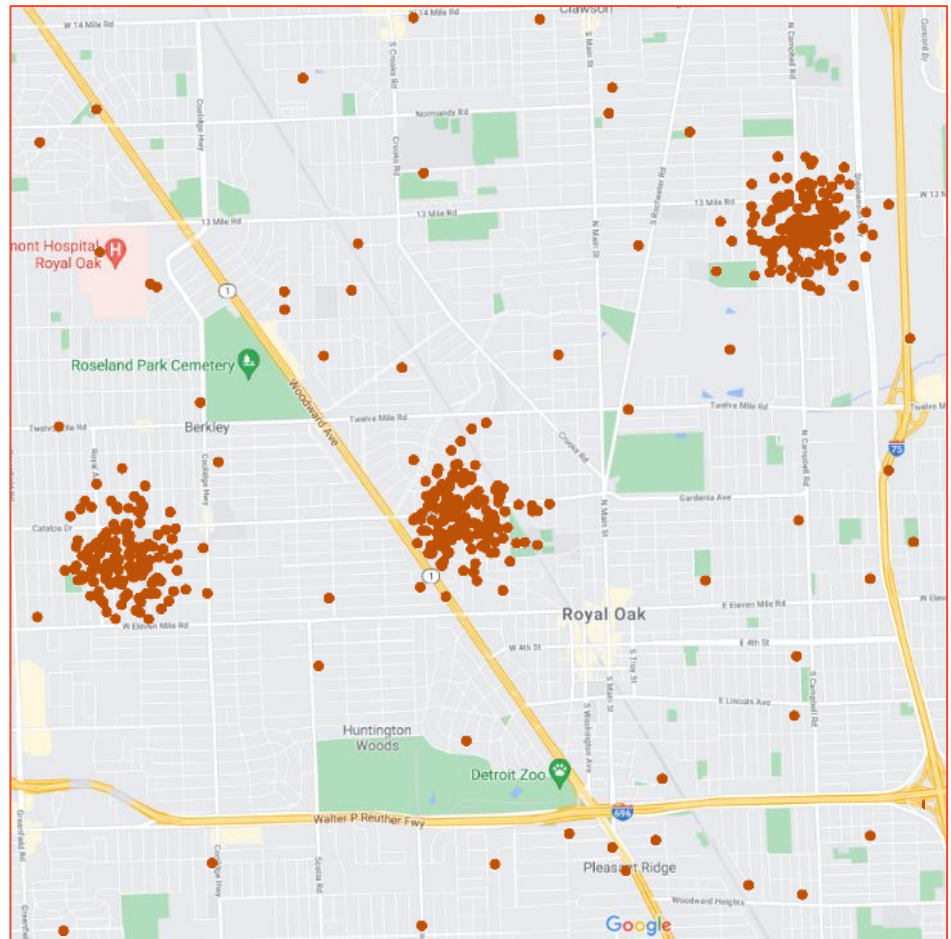
Teil I: Die Geschichte

Stellen Sie sich vor, Sie wollen einen Pizzalieferdienst mit einer kleinen Anzahl von Filialen in Ihrer Heimatstadt aufbauen. Da Sie Ihre Pizzen ausliefern müssen, versuchen Sie, die Wege für die Pizzaboten so kurz wie möglich zu halten.

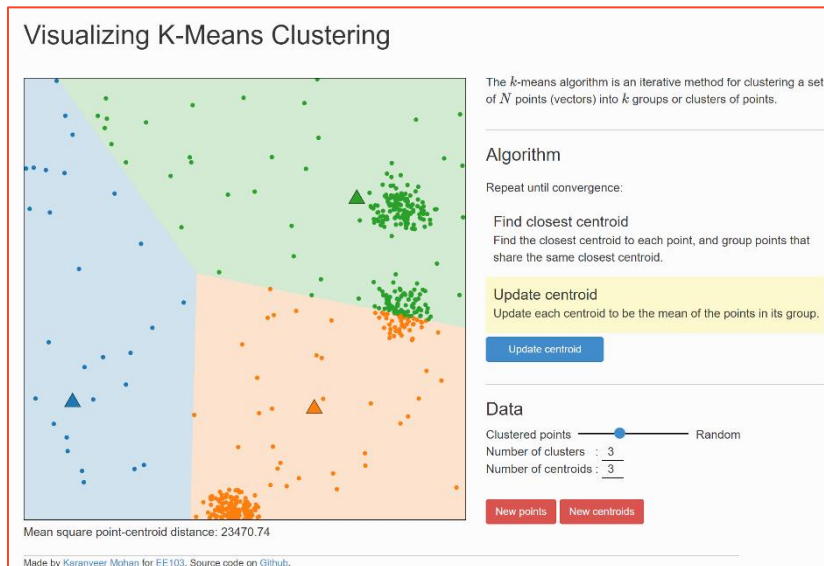
Wie viele Filialen brauchen Sie und wo sollen diese platziert werden, um Ihre Ziele zu erreichen? Alles, was Sie haben, ist eine Karte mit Ihren Kunden und den Standorten, an denen sie Pizzen bestellt haben.



Figure 13: Eine vorstellbare Verteilung von allen Kunden, die jemals bei der Pizzeria angerufen haben.

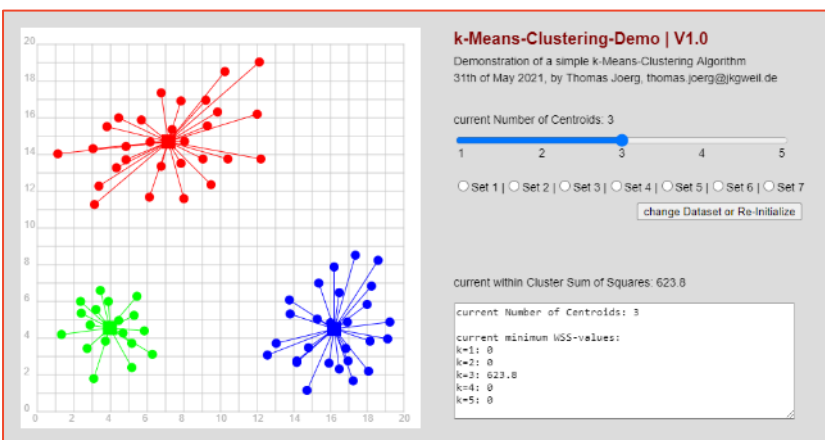


Teil II: Visualisierung des Algorithmus



<https://github.com/karanveerm/kmeans>

Diese Simulation wird auf github gehostet, da sie nicht mehr auf einer Internet-Webseite zu finden ist. Aber Sie können die App herunterladen und sie über die index.html starten



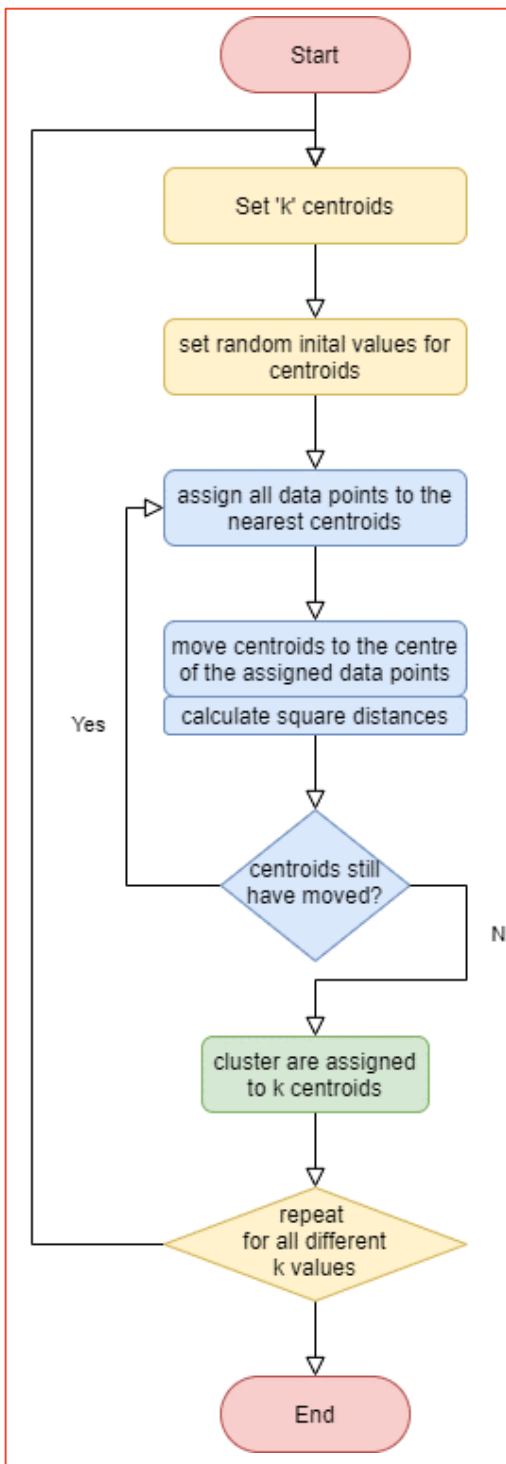
<https://iludis.de/kMeansClustering/index.html>

Diese Simulation enthält verschiedene Datensätze und berechnet die Summe der Quadrate für jede Konfiguration. Diese Werte können in eine Excel-Tabelle kopiert werden, um den Ellbogenplot zu visualisieren.

Abb. 44 Eigene Screenshots vom Webbrowser

Schritte für die Simulation

- Nehmen Sie als Ausgangszahl für die Cluster 3.
- Beginnen Sie mit 3 Centroiden und versuchen Sie, diese zum Zentrum der Datenpunktblöcke zu konvergieren.
- Wie bei vielen anderen Algorithmen des maschinellen Lernens müssen Sie die Prozedur viele Male mit verschiedenen, zufällig festgelegten Startwerten wiederholen.
- Der kleinste quadratische Mittelwert, den Sie finden, ist der beste:
 - a) Schreiben Sie den Algorithmus in Form eines Flussdiagramms oder einfach in Textform auf
 - b) Schreiben Sie den besten Wert des mittleren quadratischen Abstands auf.



Teil III: Formulating the algorithm

The flowchart diagram for the k-means clustering algorithm. Linguistically, it can be formulated a bit simpler like this:

a) *Beginnen Sie mit einer beliebigen Anzahl von Centroiden.*

b) *Verteile die Centroide zufällig zwischen den Punkten.*

c) *Berechne alle Abstände zwischen allen Punkten und den Centroiden.*

d) *Weisen Sie die Punkte dem nächstgelegenen Centroid zu.*

e) *Verschiebe nun die Centroide in die Mitte der ihnen zugeordneten Punkte.*

f) *Wiederhole den Vorgang ab Punkt c), bis sich die Centroids nicht mehr bewegen.*

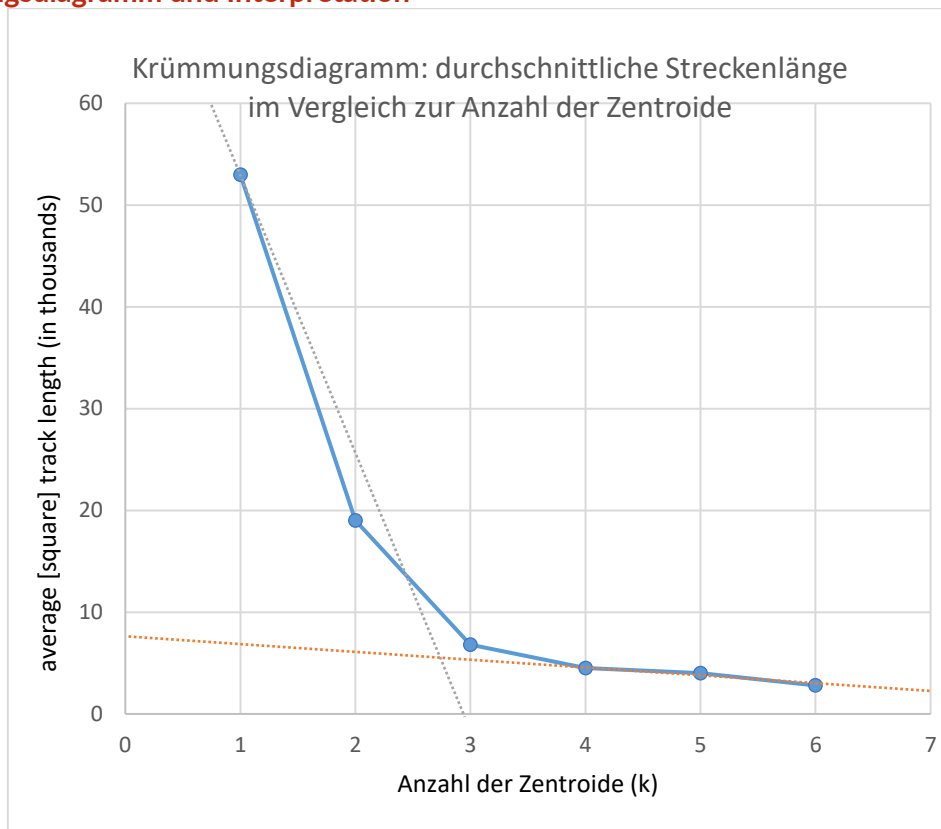
e) *Beginne wieder bei Punkt b), indem du die Anzahl der Centroids ändert.*

Teil IV: Schritte zur Ableitung des Ellbogenplots

k' Mittelpunkte	Mittlerer quadratischer Punkt-Schwerpunkt-Abstand
1	53000
2	19000
3	6800
4	4500
5	4000
6	2800

Wenn Sie alle optimalen Abstände für alle verschiedenen Schwerpunktnummern sammeln, erhalten Sie eine Tabelle ähnlich dieser. Erstellen Sie eine Excel-Tabelle mit diesen Werten, um ein Diagramm zu erstellen:

Teil V: Krümmungsdiagramm und Interpretation



an der charakteristischen Biegung des Graphen findet man die optimale Anzahl von Centroiden, da sich die mittleren quadratischen Abstände nicht mehr wesentlich ändern. Und wenn sich diese Werte nicht mehr wesentlich ändern, dann kann man auf weitere Centroiden verzichten.

Teil VI: k-Means-Clustering unplugged.

In der ersten Lerneinheit haben die Schüler eine Vorstellung von der Funktionsweise des Algorithmus entwickelt. Um diese Modellvorstellung zu vertiefen, werden sie nun ein einfaches Rechenbeispiel durchspielen. Dazu wählt die Lehrkraft 3 Schüler aus und befragt sie zu ihren Gewohnheiten:

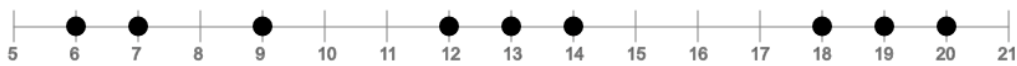
- Um wie viel Uhr frühstücken Sie?
- um wie viel Uhr ist Ihr Mittagessen,
- und um wie viel Uhr ist Ihr Abendessen?

Es wird eine Tabelle wie diese erstellt, die unser neuer Datensatz für das Clustering sein wird:

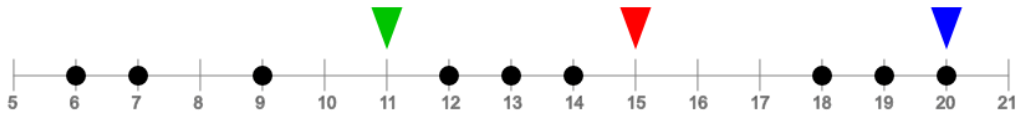
Studenten	Frühstückszeit	Mittagspause	Zeit zum Abendessen
1	9	14	19
2	7	13	20
3	6	12	18

Vollständige Visualisierung der Iterationen:

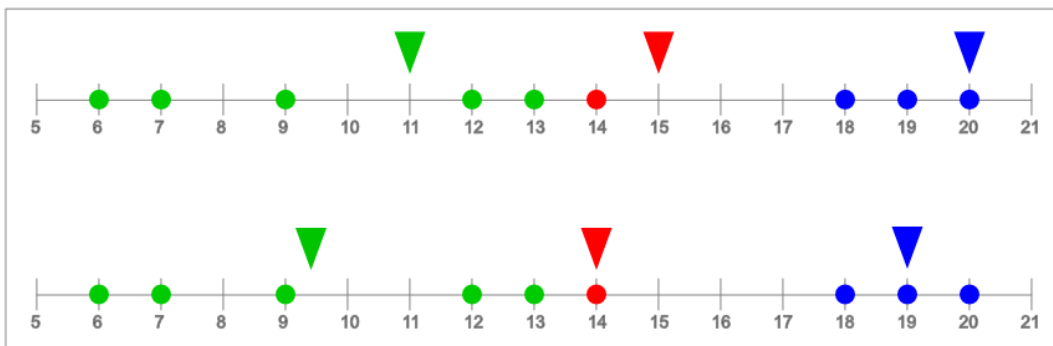
Alle Daten werden in eine Zahlenreihe übertragen:



Zahlenreihe mit Datenpunkten



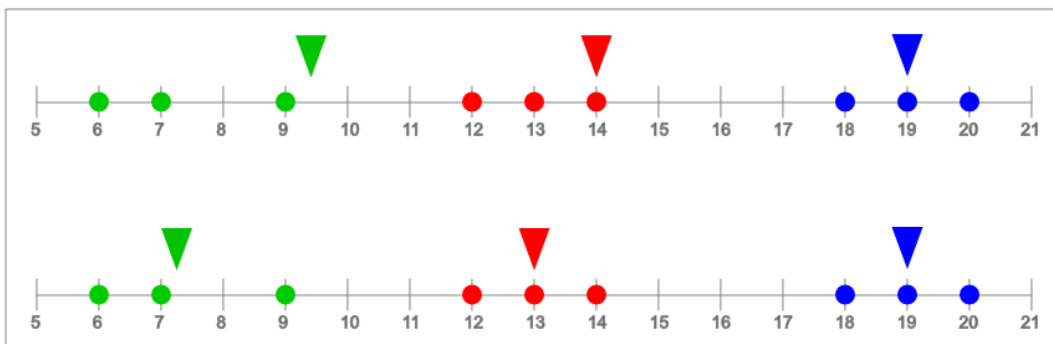
In diesem Beispiel werden drei Zentren nach dem Zufallsprinzip festgelegt.



Iteration 1:

Die Datenpunkte werden dem nächstgelegenen Schwerpunkt zugeordnet.

Die Centroiden werden in die Mitte der Datenpunkte verschoben.



Iteration 2:

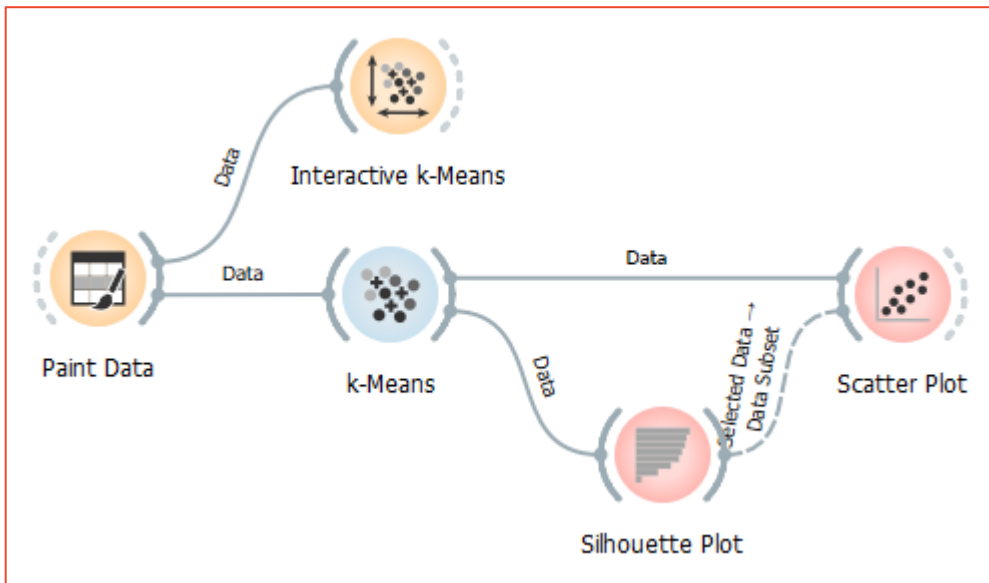
Die Datenpunkte werden dem nächstgelegenen Schwerpunkt zugeordnet.

Die Zentroide werden in die Mitte der Datenpunkte verschoben.

In Iteration 1 wird der Datenpunkt mit dem Wert "13" auf grün gesetzt. Diese Zuordnung wurde zufällig gewählt - sie hätte auch rot sein können, da die Abstände zu den beiden Mittelpunkten gleich sind.

Die Schüler können auch versuchen, 2 oder 4 verschiedene Zentroide zu verwenden. Die Anzahl von drei für die anfängliche Berechnung wurde gewählt, da das erwartete Ergebnis für die Schüler offensichtlich ist: Es ist deutlich zu sehen, dass der Algorithmus vernünftige und robuste Ergebnisse liefert.

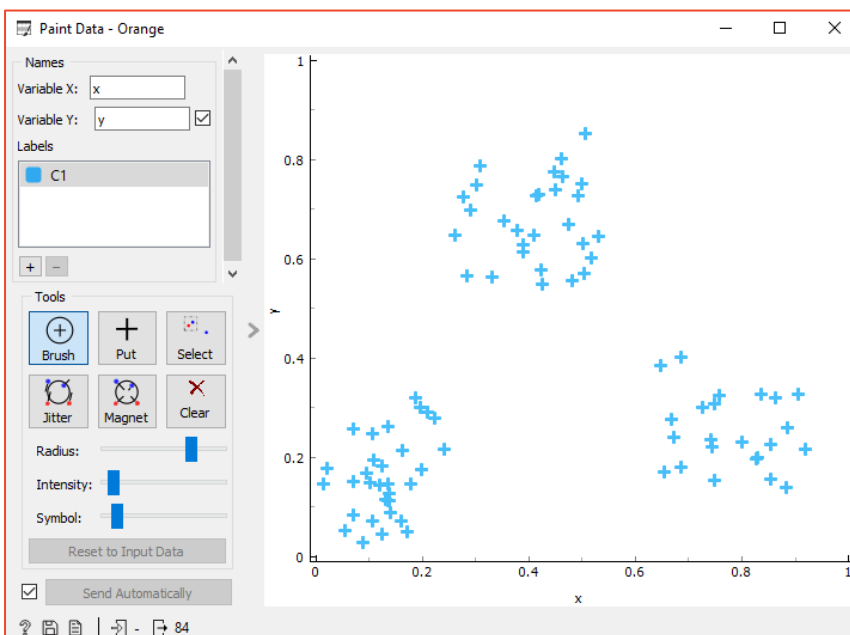
Teil VII: k-Means-Clustering mit Orange Data Mining



Die komplette Einrichtung des Projekts: Einfach und geradlinig kann es für eine Einführungslektion für Orange Data Mining verwendet werden.

Abb. 45 Eigene Screenshots aus Orange Datamining

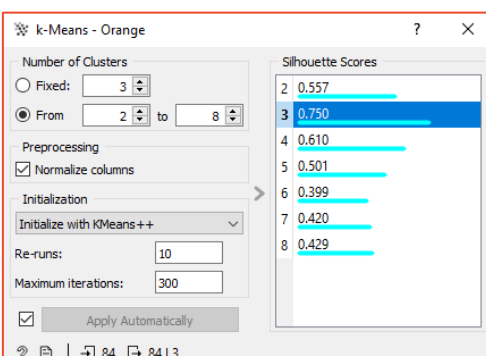
Kurze Erläuterungen zu den Knotenpunkten:



Daten malen:

Datenpunkte können mit dem Pinsel zufällig gezeichnet werden. Diese Datenpunkte werden als Werte in einer zugehörigen Datentabelle gespeichert und können für Berechnungen und Manipulationen verwendet werden.

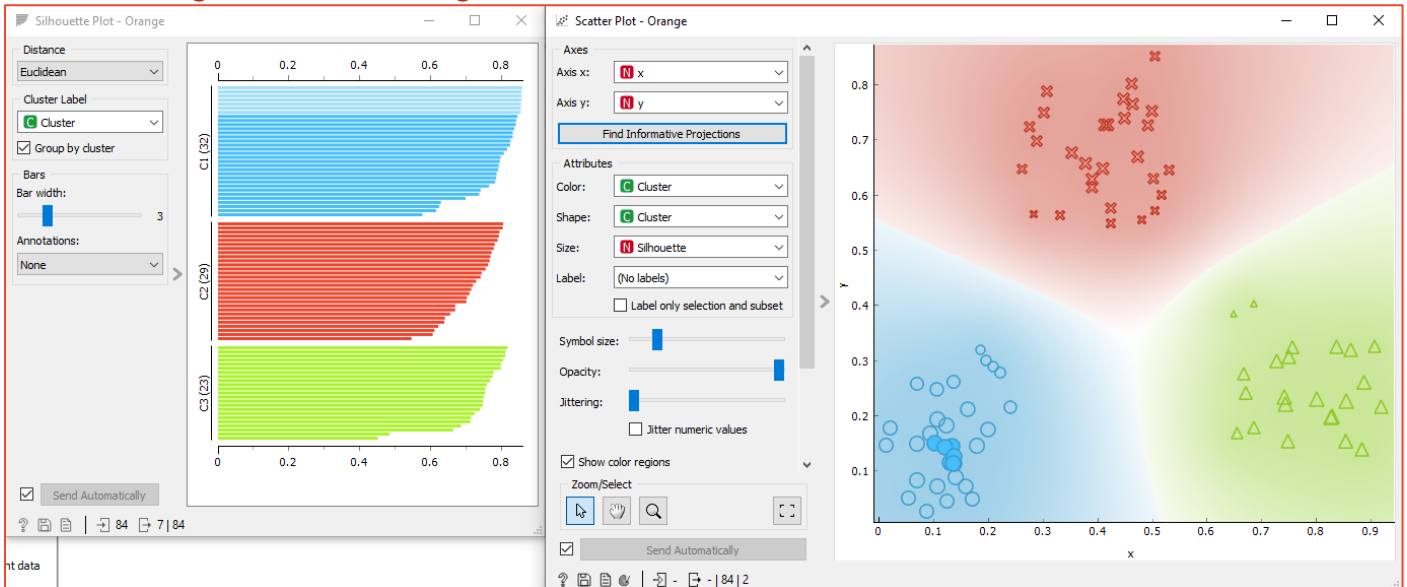
Es ist ein schnelles, unkompliziertes Verfahren zur Erstellung synthetischer Daten.



k-Means:

Um durch einen Bereich von k-Werten zu iterieren, wählen Sie die zweite Option "von". Es wird eine Berechnung der Silhouette-Scores durchgeführt. Kurz gesagt: Je höher der Silhouette-Score, desto besser passen die Daten. In diesem Beispiel ist es offensichtlich, dass drei Cluster am besten passen und daher auch den höchsten Silhouette-Score erhalten.

Shilouette-Diagramm und Streudiagramm



Im Scatter Plot können die klassifizierten Daten visualisiert werden. Er korrespondiert eng mit dem Shilouette Plot. Dieser Plot zeigt die Korrelation zwischen dem (unsichtbaren) Zentroid und dem einzelnen Datenpunkt: Je größer die Korrelation, desto höher der Score - und desto größer die Balkenlänge in der Balkengrafik des Datenpunktes. Wenn ein Datenpunkt in der Shilouette-Darstellung ausgewählt ist, wird er auch in der Streuungsdarstellung ausgewählt!

Add-On, „nice-to-have“: Der interaktive k-Means-Knoten

Name	Version	Action
<input type="checkbox"/> Associate	1.1.9	
<input type="checkbox"/> Bioinformatics	4.5.0	
<input checked="" type="checkbox"/> Educational	0.5.0	

Eine sehr schöne Funktion finden Sie unter Optionen-Add-ons:

Mit dem optionalen Installer kann das 'Educational'-Add-on installiert werden. Es enthält einen interaktiven k-Means-Knoten, der den Simulationen am Anfang des Kapitels sehr ähnlich ist.

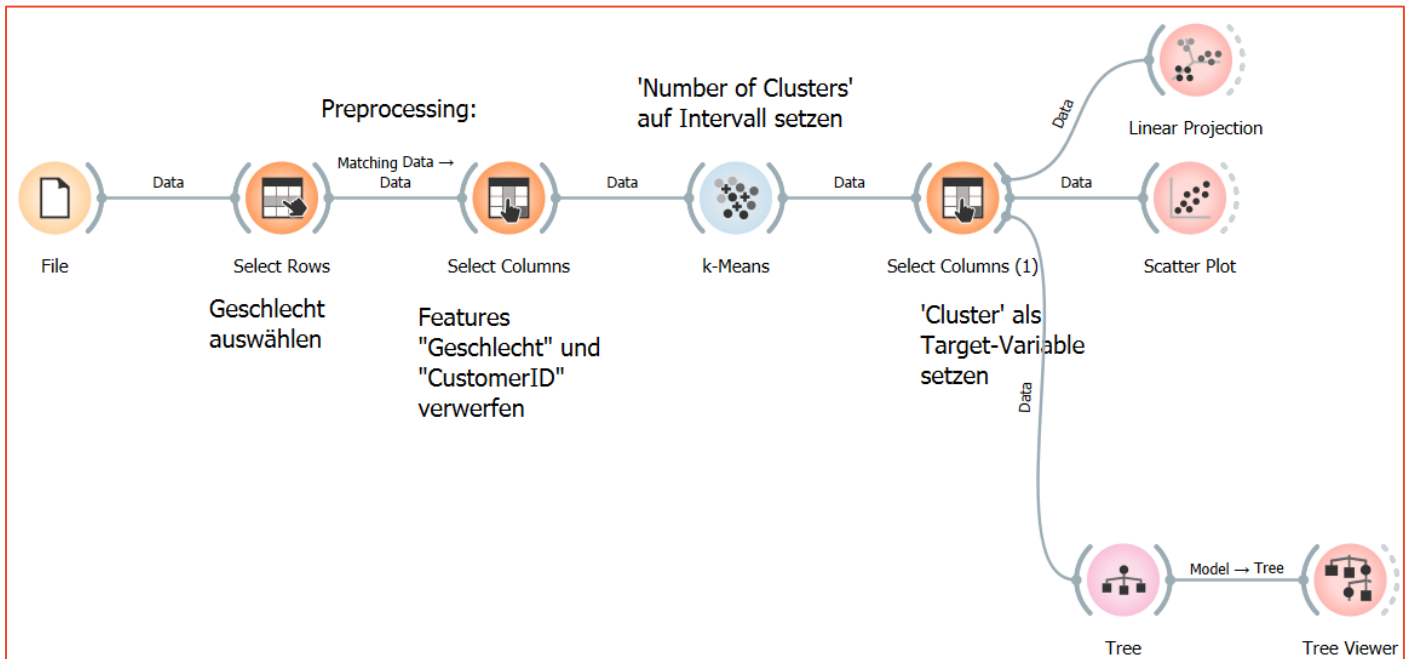
Abb. 46 Eigene Screenshots aus Orange Data Mining

Teil VIII: Mall-Customers-Datensatz

Wir beschäftigen uns mit Daten, die über Kunden in einer Shopping-Mall erhoben wurden. Diese sind hier zu finden und werden ab- und an aktualisiert. Die Ergebnisse können daher eventuell leicht anders aussehen:

<https://www.kaggle.com/datasets/vimalpillai/mailcustomers>

Der Datensatz wird in Orange Data Mining geladen. Der vollständige Musterworkflow sieht so aus:



- Im ersten Preprocessing-Schritt sollte man zwischen ‚männlich‘ und ‚weiblich‘ unterscheiden, weil sich die beiden Geschlechter unterschiedlich verhalten. Hier wurde ‚male‘ – also männlich – gewählt, weil sich die Daten so leichter strukturieren lassen.
- Im zweiten Preprocessing-Schritt werden die beiden Features „Geschlecht“ und „CustomerID“ verwerfen, da sie keinerlei Aussagekraft besitzen.
- Im eigentlichen Clustering-Schritt lässt man eine Reihe von unterschiedlichen Cluster-Anzahlen durchrechnen und wählt sich denjenigen mit dem höchsten Silhouette-Wert aus.
- Im Postprocessing-Schritt setzt man das neu hinzugekommene Feature „Cluster“ als Target, damit die weitere Verarbeitung reibungslos läuft.

Basierend auf den Clustern kann nun zum Beispiel ein Entscheidungsbaum trainiert werden, der sich interpretieren lässt:

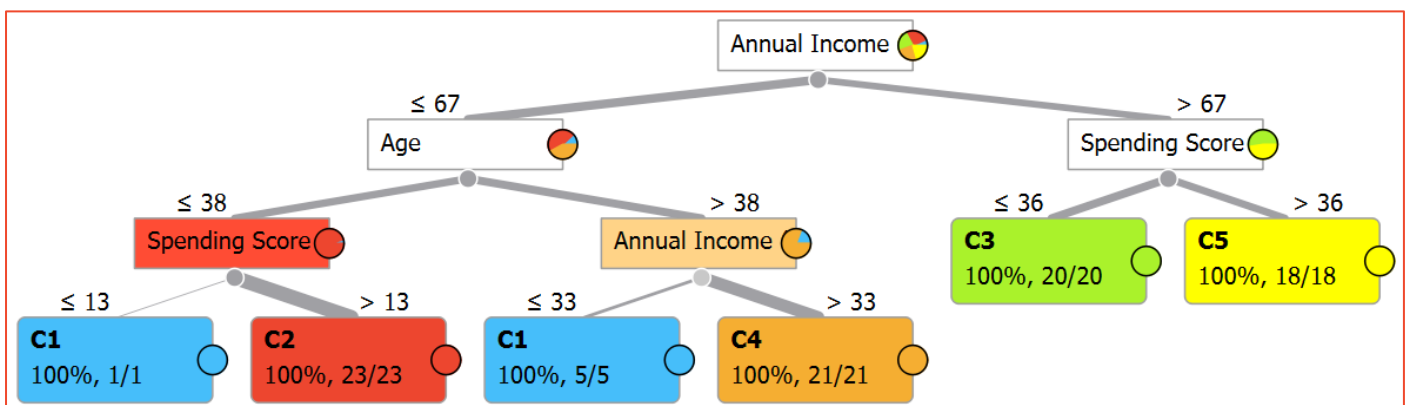


Abb. 47 Eigene Screenshots aus Orange Datamining

Lektion 14: k-Nächster Nachbar



Abb. 48 Bild [Gemeinfrei] erzeugt mit DALL-E; Prompt „several small owls sitting in a row on a branch“ von Jörg [CC BY-SA 4.0 International]

Oder: Können wir ein kNN von Grund auf neu entwickeln?

Was Schüler lernen sollten

Wie sieht ein implementierter ML-Algorithmus "unter der Haube" aus? Bislang wurden fertige Bibliotheken verwendet, die wie die berühmte "Black Box" aussehen. In dieser Lektion wird einer der einfachsten ML-Algorithmen implementiert - hier in Python, so dass die bekannten Plotting-Bibliotheken (z.B. Matplotlib) verwendet werden können. Es sind aber auch viele andere Programmiersprachen denkbar.

Mögliche Aktivitäten der Schüler

Die Schüler lernen das algorithmische Prinzip des kNN-Klassifikators anhand einer Visualisierung kennen. Anschließend formulieren sie den Algorithmus zunächst sprachlich und schreiben ihn auf. Anhand eines sehr einfachen Beispiels von drei Essenszeiten pro Tag wenden sie das Prinzip auf ein eindimensionales Problem an. Erst jetzt können sie schrittweise die Implementierung aufbauen: Die Datenstruktur des Arrays wird eingeführt und an einfachen Beispielen geübt. Dann wird die programmierbare Kontrollstruktur der Schleife eingeführt, mit der man auf alle Elemente eines Arrays zugreifen kann. Wenn beide Grundkonzepte verstanden sind, wird die Minimumsuche ausgearbeitet und als implementierbarer Algorithmus programmiert. Erst jetzt kann das 2D-Array als geeignete Datenstruktur für das bereits erarbeitete Essens-Modell konstruiert und geübt werden. Ein zufällig gewählter Zeitpunkt für ein unklassifiziertes Essen wird ermittelt und mit dem bereits angedeuteten Algorithmus klassifiziert. Dabei stoßen die Schülerinnen und Schüler auf verschiedene Lösungsmöglichkeiten, die alle - sofern sie zu richtigen Ergebnissen führen - gewürdigt und diskutiert werden. Die Schüler erkennen so, dass es nicht nur einen Weg gibt, ein Programmierproblem zu lösen.

Teil I: Interaktive Demo von kNN

<https://iludis.de/kNNDemo/index.html>

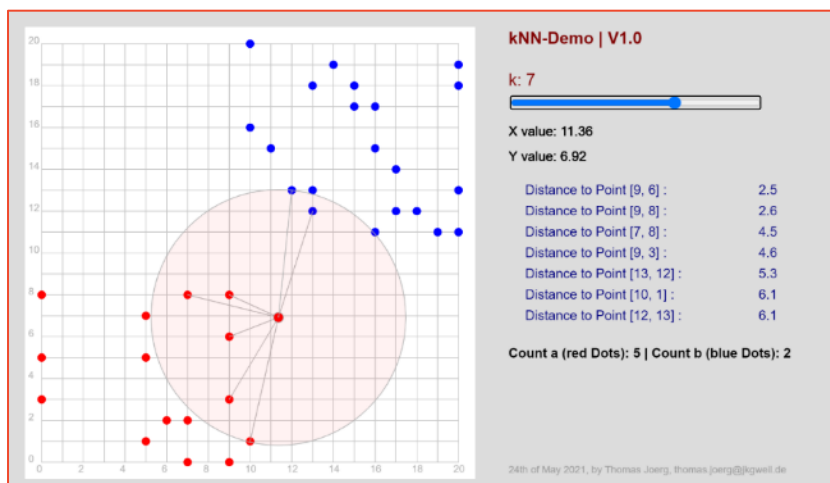


Abb. 49 Eigene Screenshots vom Webbrowser

Teil II: Eine knifflige Frage für die Schüler: Was bedeutet Training eines kNN-Modells?

Antwort: Es gibt keine wirkliches Training.

Lediglich die Parameter k und später eventuell die Methode zur Abstandsbestimmung können angepasst werden. Die Überwachung und Validierung erfolgt mit den üblichen Leistungsmetriken Accuracy, Präzision und Recall.

Natürlich kann auch die Merkmalsauswahl hinzugefügt werden. Aber die Optimierung von Hyperparametern wie zum Beispiel bei der linearen Regression oder dem Entscheidungsbaum findet nicht statt.

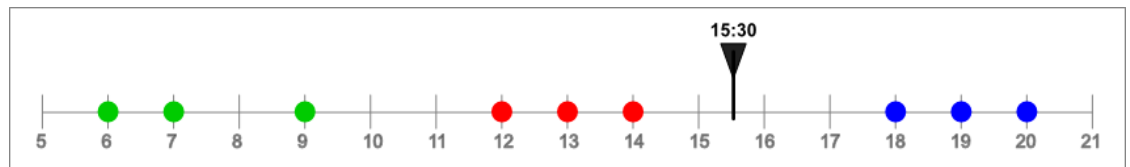
Stattdessen benötigt die Vorhersage IMMER den gesamten Datensatz im Speicher, da es keine Abstraktion von den Daten gibt.

Teil III: kNN unplugged

A) kNN 1D

Wir beginnen mit den vollständig klassifizierten Essensdaten aus der k-means-Lektion. Die grüne Farbe steht für das Frühstück, die rote für das Mittagessen und die blauen Punkte stehen für die Essenszeiten der verschiedenen Schüler. Nun wollen wir die zufällig festgelegte Zeit für eine Mahlzeit von 15:30 Uhr klassifizieren. Jeder wird sofort erkennen, dass wir diese Mahlzeit höchstwahrscheinlich als Mittagessen einstufen werden.

Zeit	Art der Mahlzeit
6	"B"
7	"B"
9	"B"
12	"L"
13	"L"
14	"L"
18	"D"
19	"D"
29	"D"



Da bei dieser Aufgabe alles klar ist, können wir uns auf die Konstruktion des Algorithmus konzentrieren:

1	Bestimmen Sie die Entfernung unseres zufälligen Essenszeitpunkts zu allen anderen Punkten
2	Bestimmen Sie den kleinsten Abstand
3	Ermitteln Sie die Klasse der Mahlzeit mit dem kleinsten Abstand und geben Sie diese als Ergebnis zurück.

Teil IV: Wie sollten wir den Algorithmus in Python implementieren?

Es wird empfohlen, Jupyter-Notebooks zu verwenden, da die Programmierung so leicht aufgeteilt werden kann.

Die Meal-Liste kann als zweidimensionale Python-Liste (2D-Array) geschrieben werden. Wir müssen einfach jede Zeile in eine einzelne Liste mit 2 Elementen packen und alle Zeilen in einer Containerliste zusammenfassen. Dieser Container wird 'allMeals' genannt:

```
## Variablen deklarieren und initialisieren
allMeals =
[[6, "B"], [7, "B"], [9, "B"], [12, "L"], [13, "L"], [14, "L"], [18, "D"], [19, "D"], [20, "D"]]
someMeal = 15.5
print(alleMahlzeiten)
print(alleMahlzeiten[0])
print(alleMahlzeiten[0][0])
```

OUTPUT:

```
[[6, 'B'], [7, 'B'], [9, 'B'], [12, 'L'], [13, 'L'], [14, 'L'], [18, 'D'], [19, 'D'], [20, 'D']]

[6, 'B']
6
```

Ein Beispiel für einige Übungen mit der 2D-Liste, um die Schüler mit den Berechnungen vertraut zu machen:

```
print(alleMahlzeiten[0][0] + alleMahlzeiten[4][0]) #Arithmetische Summe
print(alleMahlzeiten[0][1] + alleMahlzeiten[4][1]) #String-Verkettung
```

AUSGABE:

```
19
BL
```

Jetzt können wir testen, ob wir einen zufälligen Abstand berechnen können. Aber ein Problem wird auftauchen: Wie geht man mit negativen Werten um? Antwort: Es gibt kein absolutes Richtig oder Falsch; viele Wege sind möglich und wenn diese verschiedenen Methoden korrekte Ergebnisse liefern, sind diese Methoden korrekt!

```
## Berechnen Sie einen zufällig gewählten Abstand:
```

```
Abstand = alleMahlzeiten[0][0] - einigeMahlzeit
print(Abstand)
```

```
## Fallunterschied
```

```
wenn Abstand < 0:
    Entfernung = -Entfernung
print(Abstand)
```

```
## absoluter Wert
```

```
Entfernung = abs(Entfernung)
print(Abstand)
```

```
## Quadratwurzel aus Quadratzahl
```

```
Entfernung = (Entfernung**2)**0,5
print(Abstand)
```

AUSGABE:

```
-9,5
9.5
9.5
9.5
```

Der Hauptalgorithmus: Minimale Suche

```
# minimum-search
### mit der Annahme, dass der erste Wert der niedrigste sein könnte:
indexLowest = 0
niedrigste = abs(allMeals[0][0]-someMeal)

### iteriert alle anderen Werte und vergleicht die
for i in range(0, len(allMeals)):
    Abstand = abs(alleMahlzeiten[i][0]-eineMahlzeit)
    if (Abstand < niedrigster):
        niedrigste = Entfernung
        indexLowest = i

print("der niedrigste Abstand ist: ", niedrigster)
print("es hat den Index: ", indexLowest)
```

OUTPUT

```
der geringste Abstand ist: 1,5
es hat den Index: 5
```

Das Prinzip des Minimum-Search-Algorithmus

wird es an vielen Stellen im Internet beschrieben. Kurz und gut:

- Wir beginnen mit dem ersten Listenelement, das den Index 0 hat: Wir nehmen zunächst an, dass der Wert dieses Listenelements der niedrigste ist - solange wir es nicht besser wissen - und weisen diesen Wert daher der Variablen "lowest" zu.
- Wir durchlaufen alle Listenelemente mit der Kontrollstruktur einer Schleife.
- Bei jedem Schleifendurchlauf wird dieser "niedrigste" Wert immer wieder mit dem aktuellen Listenwert verglichen:
- Ist der aktuelle Listenwert kleiner als der "niedrigste" Wert, wird "niedrigster" mit diesem Listenwert überschrieben.
- Wenn Sie alle Listenelemente durchgearbeitet haben, haben Sie schließlich den kleinsten Wert gefunden.

```
exampleArray = [7, 12, 2, 5, 9]
niedrigste = exampleArray[0]

### iteriert alle anderen Werte und vergleicht die
for i in range(0, len(exampleArray)):
    wenn exampleArray[i] < niedrigster:
        niedrigste = exampleArray[i]

print("der niedrigste Wert ist: ", niedrigster)
```

OUTPUT

```
der niedrigste Wert ist: 2
```

Die Vorhersage ist einfach: Lesen Sie einfach das Klassen-Label des Elements mit dem niedrigsten Index und geben Sie es zurück:

```
print("Das Mahlzeiten-Element in der Liste ist: ", allMeals[indexLowest])  
print("Die vorhergesagte Klasse von someMeal ist: ", allMeals[indexLowest][1])
```

OUTPUT

```
Das Mahlzeiten-Element in der Liste ist: [14, 'L']  
Die vorhergesagte Klasse von someMeal ist: L
```

Teil V: Interne Differenzierung

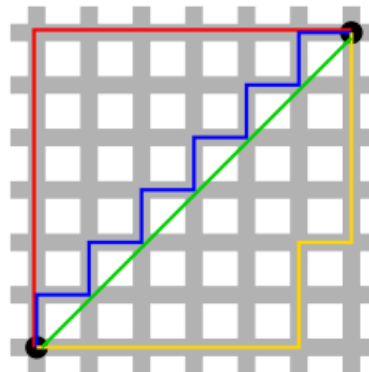
Die Erfahrung zeigt, dass sich das Feld der Schülerinnen und Schüler in Bezug auf Programmieraufgaben erweitert. Manche sind schneller, manche langsamer. Den schnelleren Schülern können zusätzliche Aufgaben gestellt werden:

- Bestimmen Sie die Klassenzugehörigkeit für $k = 2$ oder $k = 3$.
- Bestimmen Sie die Klassenzugehörigkeit in einem zweidimensionalen Fall.

Hier kann der euklidische Abstand mit dem Satz des Pythagoras verwendet werden oder der sogenannte Manhattan-Abstand, der sich aus der einfachen Subtraktion von x- und y-Werten ergibt.

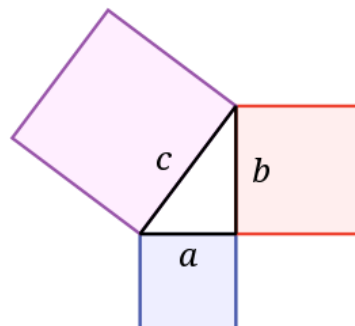
Manhattan-Entfernung:

https://en.wikipedia.org/wiki/Taxicab_geometry



Satz des Pythagoras

https://en.wikipedia.org/wiki/Pythagorean_theorem



Teil IV: kNN mit Orange Data Mining

Einführung in den Iris-Datensatz:

https://en.wikipedia.org/wiki/Iris_flower_data_set

Im Jahr 1936 sammelte der Mathematiker und Genetiker Sir Ronald Aylmer Fisher Daten über drei verschiedene Klassen von Irisblüten. Diese Daten wurden für die Klassifizierung der verschiedenen Arten verwendet. Es handelt sich um einen typischen Testdatensatz für ML-Algorithmen.

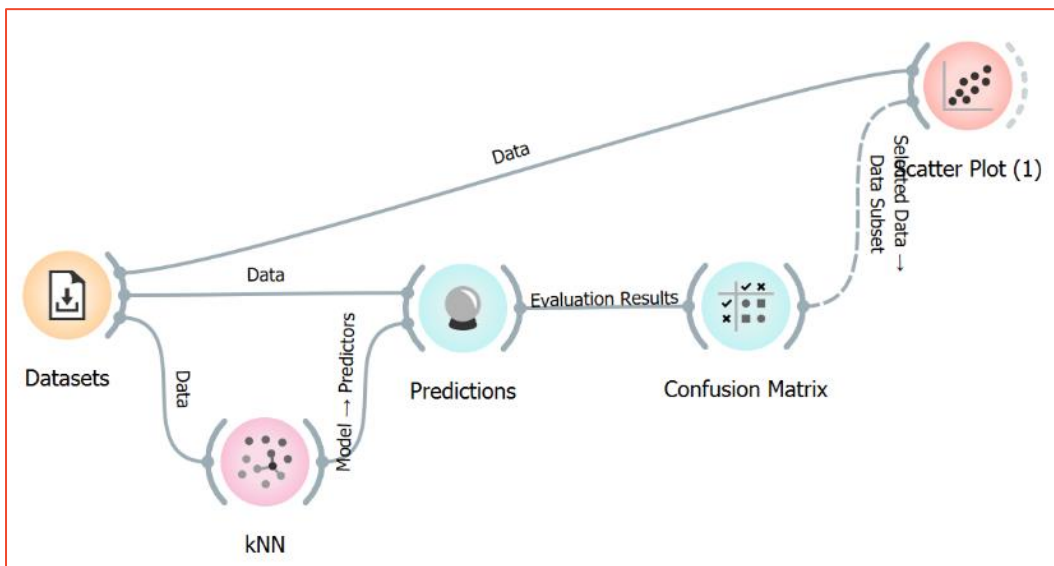
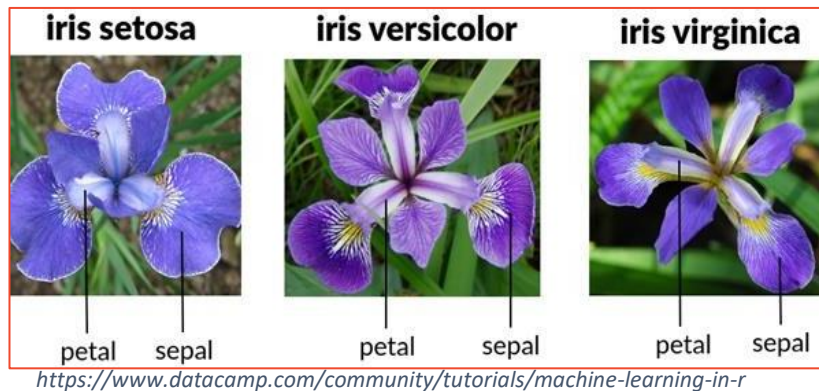
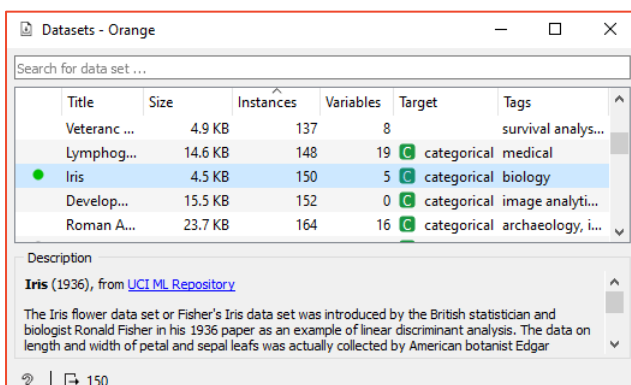


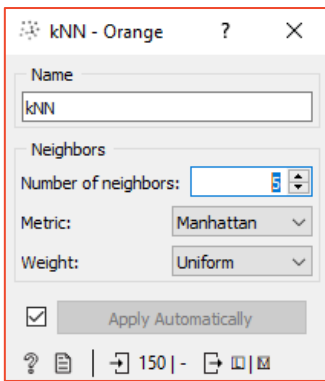
Abb. 50 Eigene Screenshots vom Webbrowser

Kurze Erläuterungen zu den Knotenpunkten:



Datensätze:

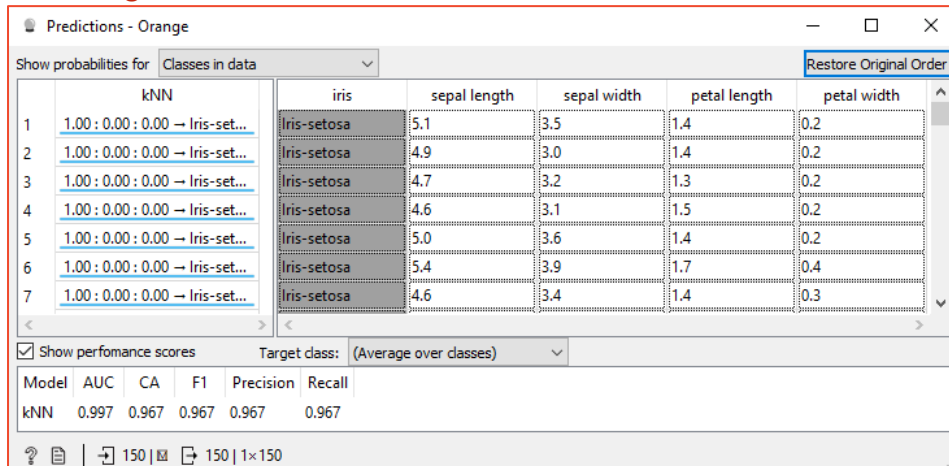
Orange Data Mining wird mit einer Reihe von vorgefertigten Datensätzen geliefert, die viele verschiedene Aspekte des Data Mining abdecken. In diesem Tutorial werden wir den Iris-Blumen-Datensatz verwenden.



kNN

Es können verschiedene Parameter gewählt werden, um eine höhere Genauigkeit zu erreichen. Für die Schüler sollte es ausreichend sein, die Methoden des "euklidischen" und des "Manhattan"-Abstands zu kennen.

Vorhersagen:

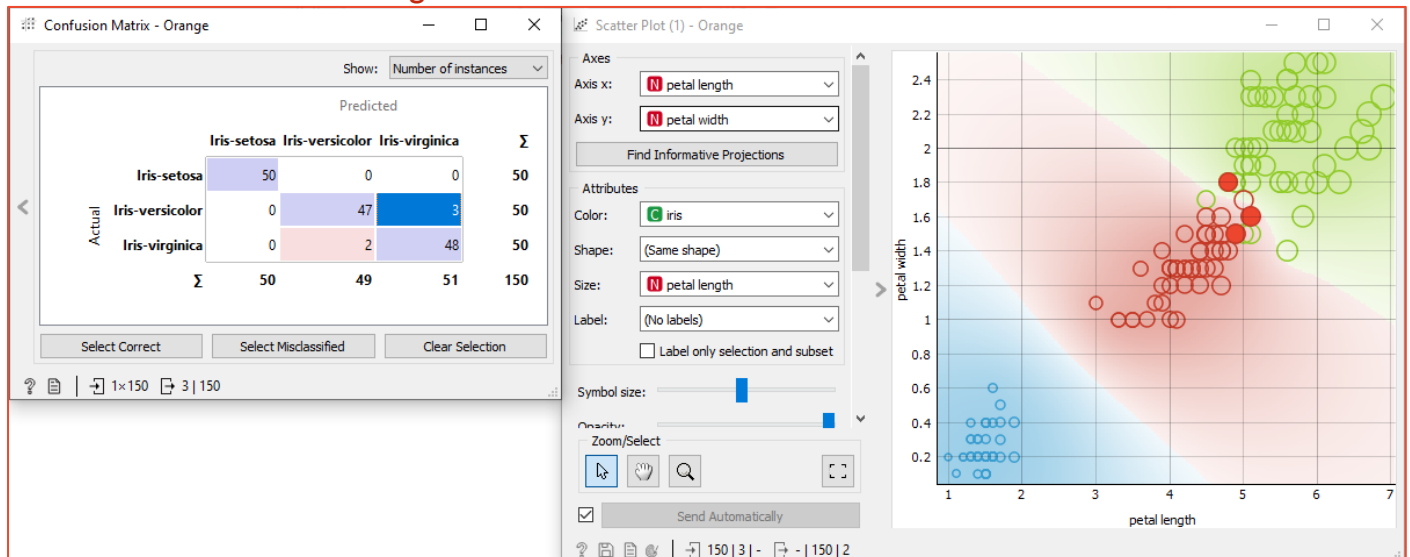


In diesem Prozessschritt werden die ursprünglichen Datenbeschriftungen verworfen und der kNN-Algorithmus wird auf die Datenbank angewendet. Nützlich ist die Anzeige der Leistungsmetrik im unteren Viertel des Fensters: CA - das heißt Accuracy - liegt bei 0,967, ebenso wie Recall und Precision. Dies zeigt, dass das Ergebnis mit den gewählten kNN-Parametern bereits recht gut ist.

Es gibt jedoch eine Situation, in der die Vorhersagen zu 100 % richtig sind. Welche sind das?

Abb. 51 Eigene Screenshots vom Webbrowser

Konfusionsmatrix und Streudiagramm



Der Scatterplot visualisiert die Klassifizierungen - und einige weitere Aspekte: Wenn innerhalb der Konfusionsmatrix nichts ausgewählt ist, sind alle Datenpunkte nicht ausgefüllt. Wenn Sie eine Zelle innerhalb der Konfusionsmatrix auswählen, z.B. die Off-Diagonal-Elemente, die falsch klassifiziert werden, können Sie diese als gefüllte Punkte im Scatterplot sehen. Nun können Sie Ihr Modell so lange verfeinern, bis alle Off-Diagonal-Elemente der Matrix einen Nullwert haben und somit alle Datenpunkte gut klassifiziert sind.

Lektion 15: Support-Vektor-Maschinen

Was Schüler lernen sollten

Der große Widersacher der neuronalen Netze: Support Vector Machines waren die zentralen und leistungsfähigsten Algorithmen während des sogenannten "neuronalen Winters" in den 1970er und 1980er Jahren. Sie sind auch heute noch wichtig. Sie arbeiten nach dem Prinzip des sogenannten "Large Margin Classifier": Das Ziel des Algorithmus ist eine möglichst breite Trennung der einzelnen Klassen. Dabei wird die Trennebene als "Hyperebene" bezeichnet.

Die didaktische Reduktion in dieser Unterrichtseinheit besteht darin, das grundsätzliche Funktionsprinzip der Stützvektoren zu vermitteln, aber höhere Konzepte wie den "Kernel-Trick" lediglich als Idee anzusprechen, ohne sie zu vertiefen. Ebenso werden wir uns auf linear trennbare Klassen beschränken. Polynomielle Trennfunktionen können durch eine Simulation veranschaulicht werden, werden aber nicht vertieft.

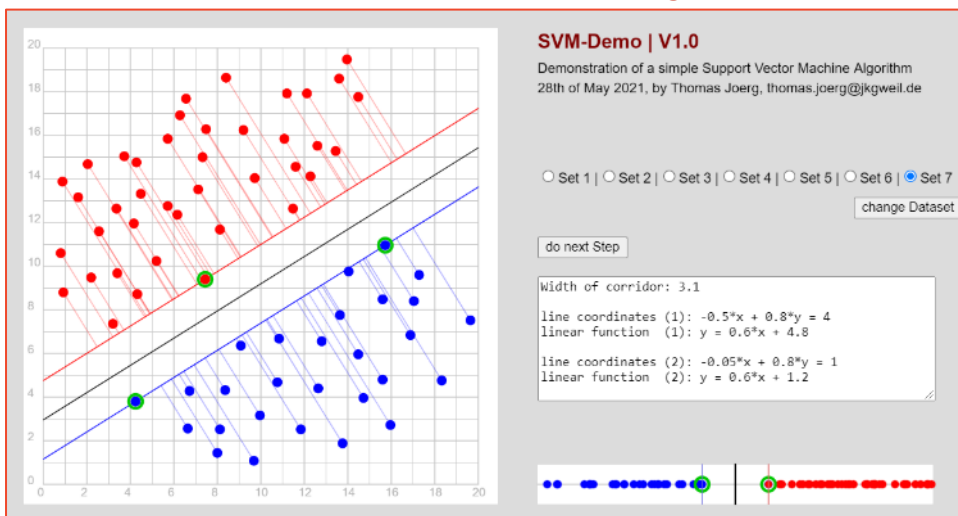


Abb. 52 [Uncle Sam](#) [Public Domain]

Mögliche Aktivitäten der Schüler

Was die Schüler herausfinden müssen: Das Prinzip der größtmöglichen Trennung: die Verwendung von so genannten mathematischen Stützvektoren, um eine bestmögliche Aufteilung der Klassen zu erreichen. Dazu analysieren sie zunächst eine Simulation und formulieren den Algorithmus sprachlich aus. Anschließend bearbeiten sie ein 1D-Problem, bei dem die Schülerinnen und Schüler verschiedene Einschränkungen und deren jeweilige Vor- und Nachteile diskutieren. Eine klare Grenze wird zum sogenannten "Kernel-Trick" gezogen. Üblicherweise werden Datenbanken mathematisch so modifiziert, dass Hyperebenen eingefügt werden können, um eine Trennung oder Klassifizierung zu erzwingen. Dies ist viel zu anspruchsvoll und wird hier bewusst weggelassen.

Teil I: "Bauen Sie die Straße so breit wie möglich"



Es werden verschiedene Datensätze untersucht und diskutiert, wie die Grenze durch die Verwendung eines möglichst breiten Kanals so dick wie möglich gemacht werden kann.

Dazu muss man zunächst genau die zwei Punkte aus jeder der verschiedenen Klassen finden, die am nächsten beieinander liegen. Denn: Der Kanal muss durch diese beiden Punkte verlaufen.

Abb. 53 Eigene Screenshots vom Webbrowser

Sie können an der Verbindungslinie zwischen diesen beiden Punkten beginnen: Die Diskriminatorlinie wird zunächst rechtwinklig zu dieser Verbindungslinie gesetzt. Aber jetzt werden Sie wahrscheinlich einige Punkte sehen, die innerhalb des Kanals liegen.

In weiteren Schritten wird der Kanal optimiert, bis kein Punkt mehr innerhalb des Kanals liegt. Die entsprechenden Diskriminatorlinien ergeben sich mathematisch (bitte nur diskutieren, nicht berechnen!) orthogonal zu den Stützvektoren der Datenpunkte

Wiederholen Sie den Vorgang für verschiedene Datensätze in der App!

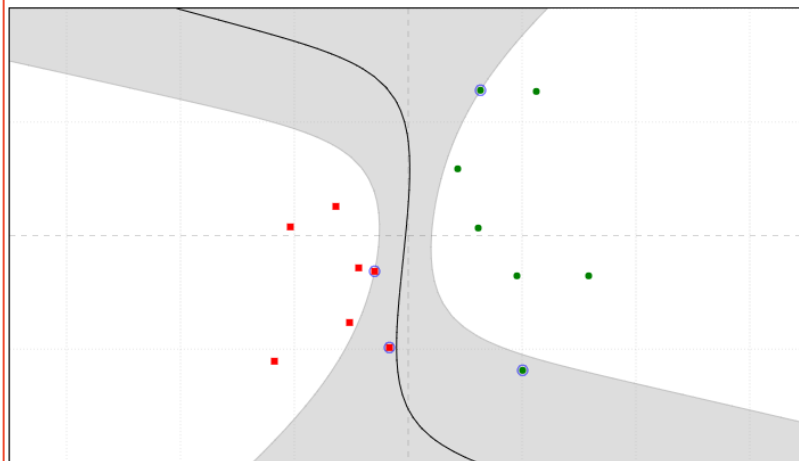
Teil II: SVM kann sehr kompliziert werden

Interactive demo of Support Vector Machines (SVM)

February 12, 2018

tags: [c++](#), [machine-learning](#), [svm](#), [wasm](#)

Note: you may have to **disable your adblocker** for this demo to work.



Toggle $\nu = 0.15$

Kernel: $\gamma = 1.0$ $c_0 = 0.0$

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\gamma \mathbf{x} \cdot \mathbf{y} + c_0)$$

<https://jgreitemann.github.io/svm-demo>

SVMs können auch dazu verwendet werden, nicht linear trennbare Daten voneinander zu trennen, um eine korrekte Klassifizierung zu erreichen. Dies wird jedoch mathematisch sehr schnell sehr anspruchsvoll und soll hier nicht weiter diskutiert werden. Das Simulationsbeispiel soll nur als Ausblick dienen.

Abb. 54 Eigene Screenshots vom Webbrowser

Teil III: Erweiterung des kNN-Projektes in Orange Data Mining

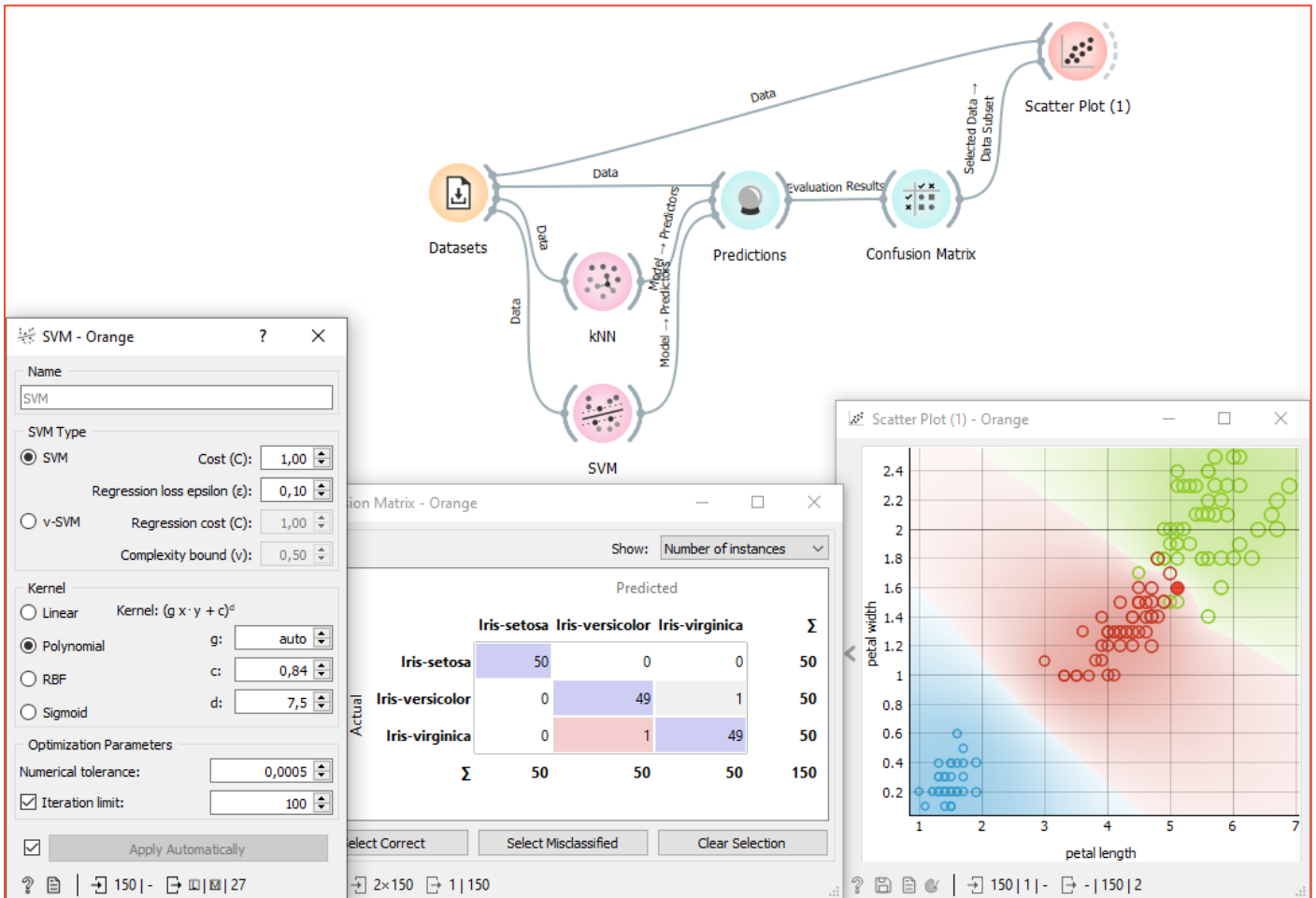


Abb. 55 Eigene Screenshots aus Orange Data Mining

Erweitern Sie das kNN-Projekt aus dem letzten Kapitel um einen SVM-Knoten. Experimentieren Sie mit den Hyperparametern, um das bestmögliche Modell zu erhalten. Dabei können Sie ein wenig mit den Parametern ‚herumspielen‘ um das beste Modell zu finden.

Teil IV: SVM 1D unplugged

Stellen Sie sich vor, wie Sie Kulturheidelbeeren von Erdbeeren unterscheiden können. Kulturheidelbeeren haben ein Durchschnittsgewicht von 8 Gramm, während das Durchschnittsgewicht von Erdbeeren etwa 15 Gramm beträgt. Größe und Gewicht dieser Früchte können jedoch sehr unterschiedlich sein. Es stellt sich also die Frage: Wie kann man sie zuverlässig trennen?

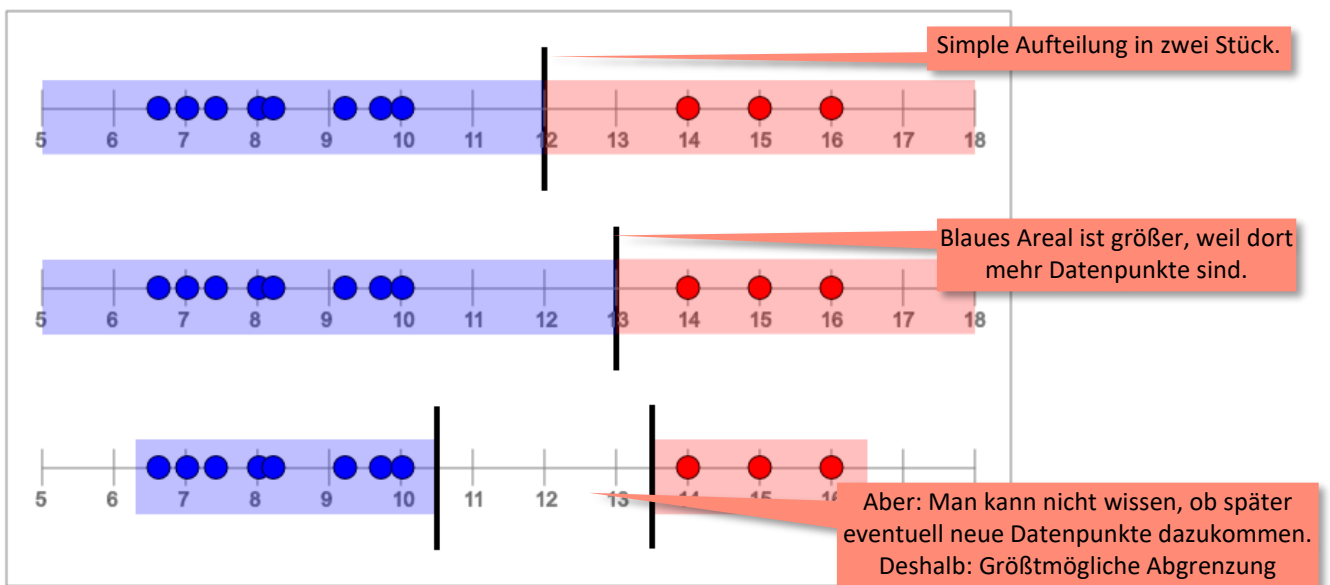


Abb. 56 Bild [\[Gemeinfrei\]](#) erzeugt mit [DALL-E](#); Prompt „Some cultivated blueberries / a bunch of strawberries, professional food photography“ von Jörg [\[CC BY-SA 4.0 International\]](#)

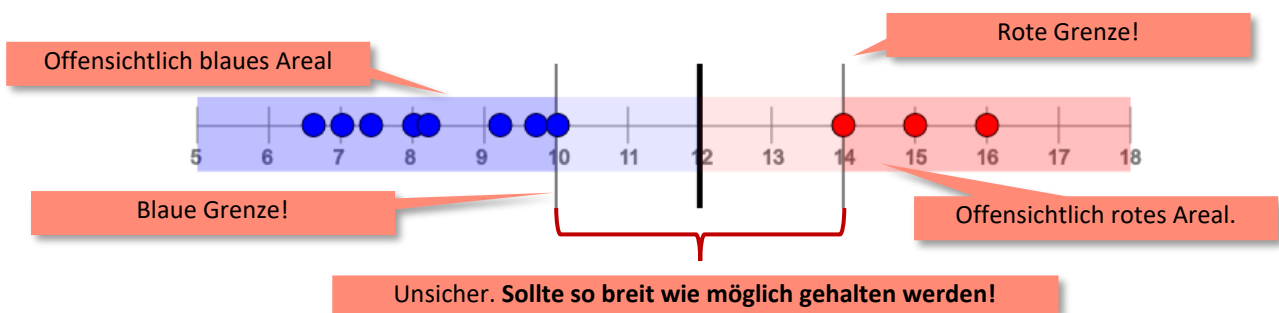
Wir erstellen einen kleinen Datensatz mit einigen Gewichten verschiedener Früchte und stellen sie auf einem Zahlenstrahl dar:



Wie kann man am geschicktesten eine Trennlinie ziehen, um eine Fläche für die blauen und eine Fläche für die roten Punkte zu definieren? Die Schüler sollten über das Problem nachdenken und ihre Ergebnisse präsentieren. Sie könnten wie diese möglichen Lösungen aussehen:



Es scheint ein vernünftiger Ansatz zu sein, den "Bereich der unbekanntem Klassifizierung" zu maximieren und die Unsicherheit zwischen den Grenzpunkten der beiden Bereiche zu berücksichtigen:



Lektion 16: Orange Data Mining Projekt: Preisvorhersage für einen Gebrauchtwagen

Was sollen die Schüler lernen?

Die Studierenden sollen den Weg von der Datenaufbereitung über die Datenanalyse bis hin zur endgültigen Vorhersage selbständig durcharbeiten. Es ist wichtig, dass die Schüler erkennen, dass trotz aller hoch entwickelten Algorithmen viel manuelle Arbeit bei der Anpassung von Daten und Modellen notwendig ist. Als Werkzeug verwenden die Studierenden die kostenlose Software "Orange Data Mining". "Orange" ist knotenbasiert und kapselt daher den komplexen mathematischen Hintergrund, so dass sich die Studierenden voll und ganz auf die Datenpipeline konzentrieren können.

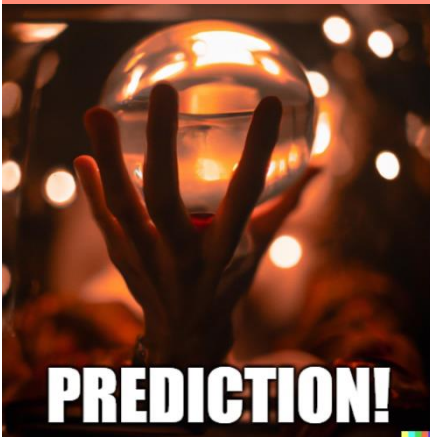


Abb. 57 Bild [\[Gemeinfrei\]](#) erzeugt mit [DALL-E](#); Prompt „a glass ball in the hand, human and spot lights, dark background “ von Jörg [\[CC BY-SA 4.0 International\]](#)

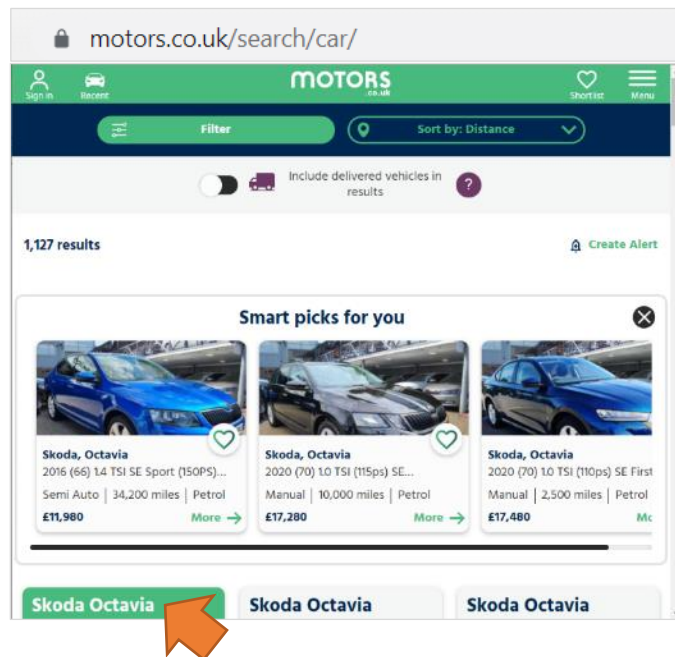
Anmerkung: Dieses Auto wird verwendet, da sich die multivariate Regressionsanalyse sehr gut auf diesen Datensatz anwenden lässt. Die Regression ergibt einen Korrelationskoeffizienten von etwa 0,9, was den Schülern ein schönes Erfolgserlebnis verschafft.

Mögliche Aktivitäten der Schüler

Führen Sie ein komplettes DataMining-Projekt von Anfang bis Ende durch: Ausgehend von einem (hoffentlich) interessanten realen Datensatz wird die gesamte Datenaufbereitungspipeline eingerichtet und verschiedene ML-Modelle damit trainiert. Dies geht bis hin zur Feinabstimmung der Parameter. So lernen die Studierenden spielerisch den Umgang mit datengetriebenen Projekten.

Teil I: Die Geschichte

Werfen wir einen Blick auf eine sehr beliebte Website, die Gebrauchtwagen anbietet. Wir interessieren uns für einen "Skoda Octavia". Ein Klick auf die Schaltfläche "Wie ist dieser Preis im Vergleich" liefert uns das Ergebnis: "Für dieses Auto sind keine Informationen verfügbar".



Hinweis: Die folgenden Grundelemente des Unterrichts können methodisch auf unterschiedliche Weise umgesetzt werden.

Was können wir tun, um nicht einen zu hohen Preis zu zahlen?

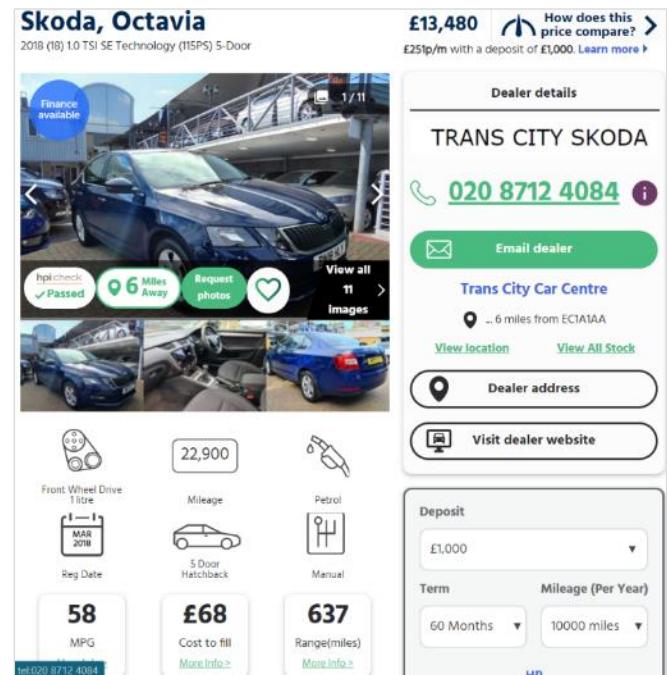


Abb. 58 Eigene Screenshots vom Webbrowser

Teil II: Auswahl der Merkmale.

Welche Merkmale sind für die Bildung des Preises für dieses Auto relevant?

- **"Alter"**: Je neuer, desto besser und daher höher der Preis,
- **"Kilometerstand"**: weniger Kilometerstand bedeutet höheren Preis,
- **"Kraftstofftyp"**: Autos mit Benzinmotor sind heutzutage teurer als Autos mit Dieselmotor,
- **"Getriebetyp"**: Automatik oder manuell, in der Regel ist ein manuell geschaltetes Auto etwas billiger,
- **"Meilen pro Gallone"**: je höher, desto besser und teurer,
- **"Motorgröße/Pferdestärke"**: Ein größerer Motor bedeutet in der Regel mehr PS und ist teurer.
- **"Zubehör"**: bestimmte Farben, Klimaanlage, bestimmte Felgen, Stereoanlagen ...

Teil III: Daten sind notwendig

Keine Vorhersage ist möglich ohne Anhaltspunkte - d.h. zuverlässige Daten. Man braucht so viele Daten wie möglich aus der Vergangenheit.

Woher bekommen wir diese Art von Daten? Wir werden Kaggle nutzen, eine sehr beliebte Website für Datenwissenschaftler und Spezialisten für maschinelles Lernen:

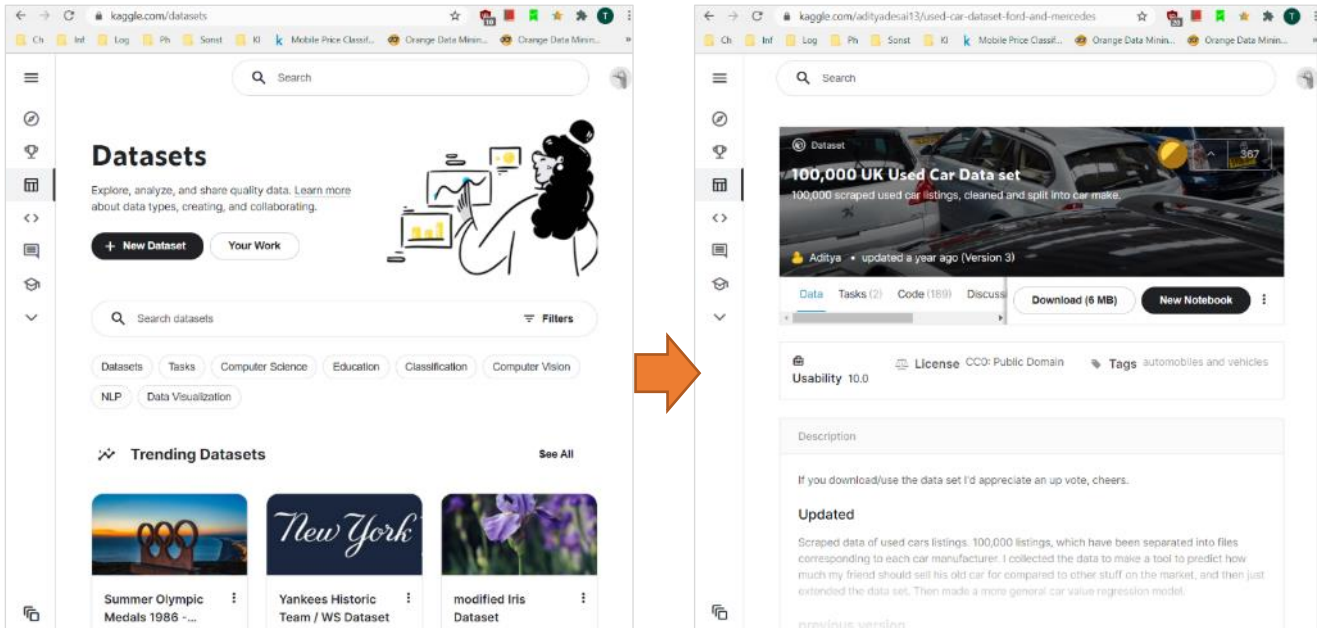
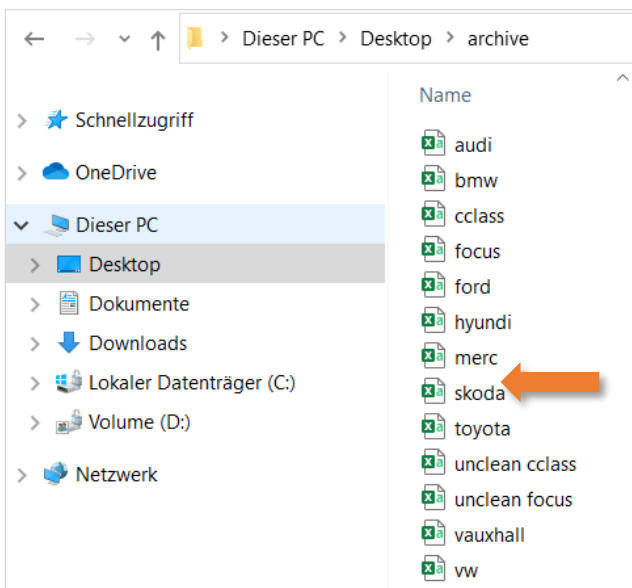


Abb. 59 Eigene Screenshots vom Webbrowser

Laden Sie das Archiv herunter und entpacken Sie alle Dateien:

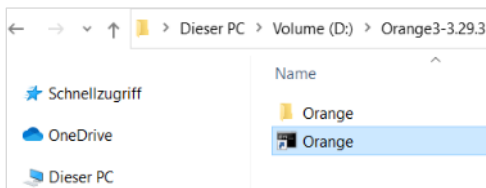


Jetzt können wir mit der Arbeit beginnen.

Wir benötigen diesen speziellen Datensatz:

Teil IV: Einführung in das "Orange Data Mining"

Das kostenlose Software-Tool "Orange Data Mining" kann von <https://orangedatamining.com/> heruntergeladen werden.



Wenn die portable Version verwendet wird, ist es nicht notwendig, sie zu installieren. Einfach auspacken und auf das schwarze "Orange"-Symbol doppelklicken.

Sehr nützliche Informationen finden Sie auf dem YouTube-Kanal <https://www.youtube.com/c/OrangeDataMining/videos>

Arbeitsablauf

Click this button

Double click opens dialog

search the file, load it and set "price" as the target variable

Name	Type	Role	Values
1 model	categorical	feature	Citigo, Fabia, ...
2 year	numeric	feature	
3 price	numeric	feature	
4 transmission	categorical	feature	Automatic, Manual, ...
5 mileage	numeric	meta	skip
6 fuelType	categorical	feature	Diesel, Hybrid, ...
7 tax	numeric	feature	
8 mpg	numeric	feature	
9 engineSize	numeric	feature	

Create a "Data Table"

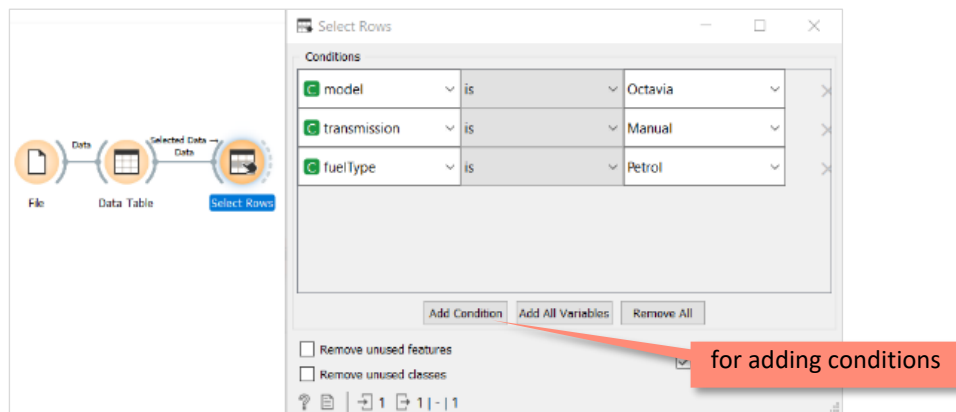
Connect it to "File"

and double click "Data Table"

Achten Sie darauf, NICHT bestimmte Zeilen auszuwählen, da Orange alle folgenden Schritte nur für diese ausgewählten Zeilen durchführt!

	model	year	price	transmission	mileage	fuelType
1	Octavia	2017	10550	Manual	25250	Petrol
2	Citigo	2018	8200	Manual	1264	Petrol
3	Octavia	2019	15650	Automatic	6825	Diesel
4	Yeti Outdoor	2015	14000	Automatic	28431	Diesel
5	Superb	2019	18350	Manual	10912	Petrol
6	Yeti Outdoor	2017	13250	Automatic	47005	Diesel
7	Superb	2019	15250	Manual	14850	Petrol
8	Octavia	2019	18950	Automatic	5850	Diesel
9	Kodiaq	2019	29900	Automatic	2633	Petrol
10	Octavia	2017	18990	Manual	20000	Petrol
11	Octavia	2016	8500	Automatic	91280	Diesel
12	Octavia	2017	8695	Manual	84450	Diesel
13	Superb	2016	7000	Manual	109060	Diesel
14	Rapid	2016	7950	Manual	33013	Petrol
15	Karoq	2018	19770	Manual	12884	Petrol

Es gibt zu viele Daten, die wir nicht brauchen, also müssen wir die relevanten Werte **filtern**. Platzieren Sie einen "Select Rows"-Knoten, verbinden Sie ihn mit der "Datentabelle", doppelklicken Sie darauf und legen Sie die entsprechenden Filterbedingungen fest:



Schauen wir uns also an, was wir bis jetzt haben. Um **Informationen zu erhalten**, können wir "Verteilungen" und "Korrelationen" verwenden:

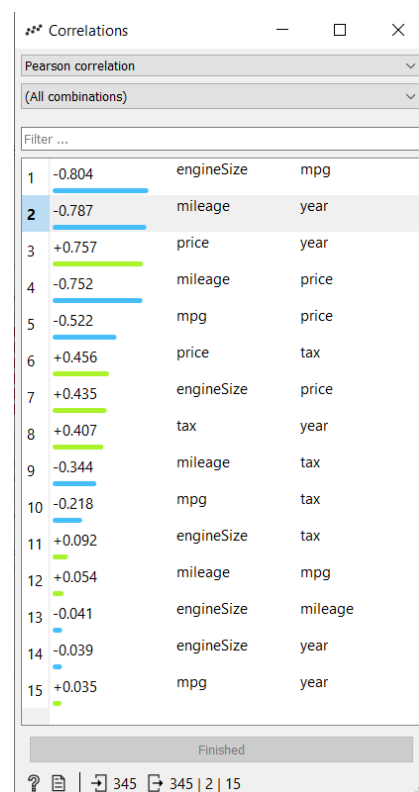
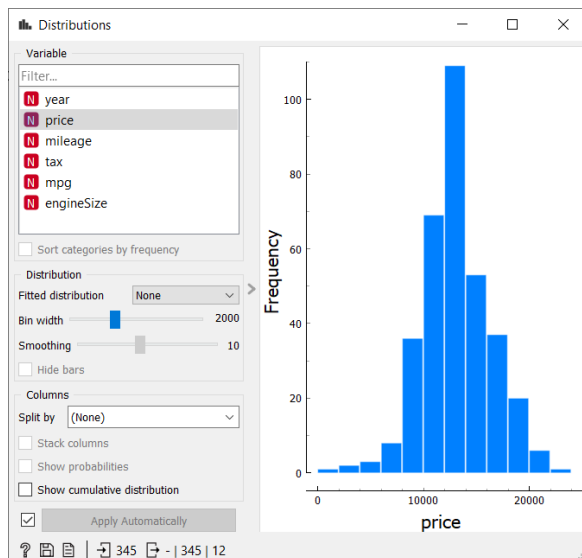
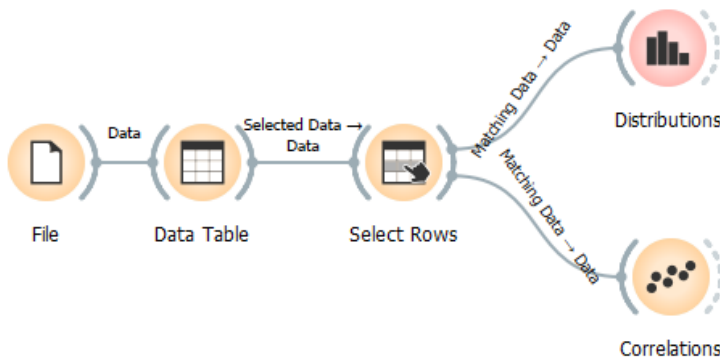


Abb. 60 Eigene Screenshots aus Orange Data Mining

- Der Preis kann als normalverteilt angenommen werden.
- Es besteht eine mittlere bis starke negative Korrelation zwischen Preis und Jahr.

Diskutieren Sie, was das bedeutet.

4. "Nice to have", aber nützlich: Funktionen schaffen.

Das "Jahr" eines Autos ist nicht sehr sinnvoll - die Arbeit mit dem "Alter" eines Autos wäre wahrscheinlich nützlicher, da wir von dem spezifischen Zeitpunkt, zu dem der Datensatz gesammelt wurde, verallgemeinern können. Nehmen wir eine einfache Umrechnung von "Jahr" in "Alter" vor. Der Datensatz wurde im Jahr 2020 erfasst, also müssen wir beide Werte mit dem **"Feature Constructor"** subtrahieren:

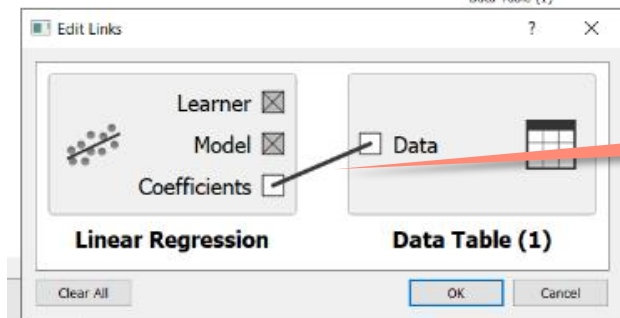
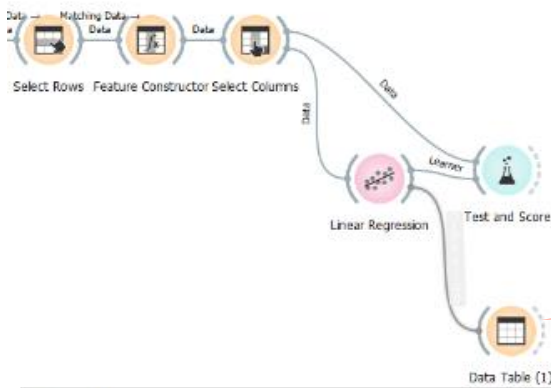
Teil V: Jetzt können wir mit dem maschinellen Lernen beginnen.

Wir müssen unsere Überlegungen von Anfang an anwenden: welche Merkmale sind relevant und welche nicht. Daher müssen wir **die Merkmale** mit dem Knoten "Spalten auswählen" **auswählen**:

Nun können wir verschiedene Algorithmen darauf anwenden, um ein Modell zu erstellen. Beginnen wir mit einer linearen Regression:

Belassen Sie zunächst die "Lineare Regression" auf ihren Standardwerten. Machen Sie einen Doppelklick auf "Test and Score", um die Leistung des Regressionsalgorithmus zu überprüfen.

Um die Regressionskoeffizienten zu überprüfen, müssen wir einen "Data-Table"-Knoten mit dem Knoten der linearen Regression verbinden. Ein Doppelklick auf die Verbindungslinie stellt die korrekte Datenverknüpfung sicher:



Diese Verbindung muss ausgewählt sein.

... dann erhält man die Koeffizienten.

	name	coef
1	intercept	30750
2	age	-1016.29
3	engineSize	686.907
4	mpg	-274.135
5	mileage	-0.0553704

Ein erster Versuch einer Vorhersage "von Hand"

Das bedeutet, dass wir mit der folgenden (leicht gerundeten!) Formel einen mittleren Preis für unser Auto mit einem $R^2 = 0,907$ vorhersagen können, was ziemlich hoch ist:

$$\text{predicted price} = 30750 - 1016 \cdot \text{age} + 687 \cdot \text{engineSize} - 274 \cdot \text{mpg} - 0.0554 \cdot \text{mileage}$$

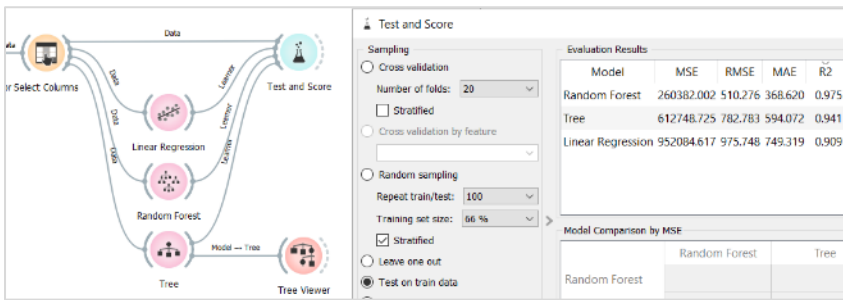
Unser Beispielfahrzeug hatte die folgenden Werte:

Alter: 2 (Jahr 2018)	Motorgröße: 1.0	Kilometerstand: 22900	Treibstoffverbrauch: 58	Preis: 13480
----------------------	-----------------	-----------------------	-------------------------	--------------

$$\text{predicted price} = 30750 - 1016 \cdot 2 + 687 \cdot 1 - 274 \cdot 58 - 0.0554 \cdot 22900 = 12244$$

Das bedeutet: Das Auto ist zu diesem Preis zu teuer!

Teil VI: Anwendung zusätzlicher Modelle



Versuchen wir nun, einige weitere Algorithmen für die Regression zu vergleichen: Entscheidungsbaum und Random Forest

Beide baumbasierten Algorithmen schneiden besser ab als die lineare Regression. Speziell mit dem "Tree"-Knoten können wir den "Tree Viewer"-Knoten zur Inspektion und Visualisierung verwenden:

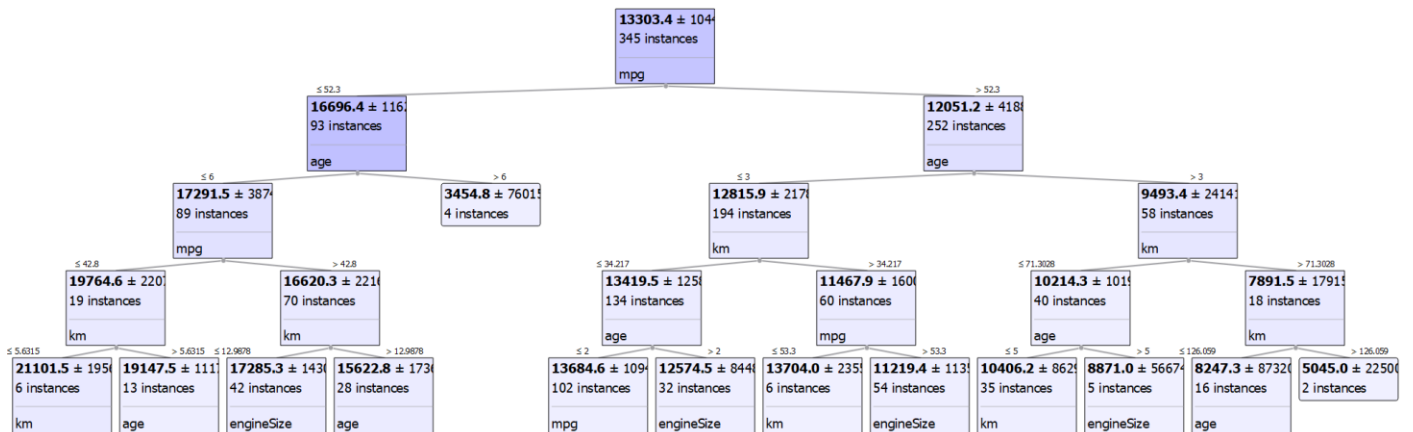


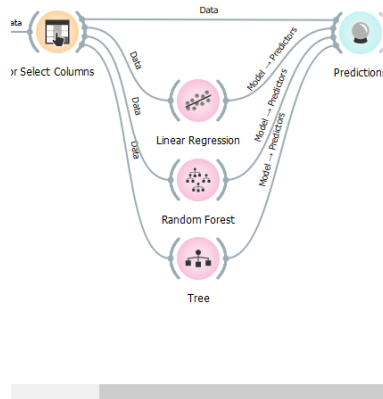
Abb. 61 Eigene Screenshots aus Orange Datamining

Vorhersage mit Orange

Ersetzen Sie den "Test and Score"-Knoten durch den "Prediction"-Knoten und doppelklicken Sie auf ihn. Hier sehen Sie alle von den Modellen vorhergesagten Werte im Vergleich zu den tatsächlichen Werten. Das "Random Forest"-Modell schneidet offensichtlich am besten ab:

Teil VI: Abschließende Gedanken und Vorschläge

Es gibt einige eingebaute Datensätze wie den "iris"-Datensatz in Orange. Und auf Kaggle gibt es viele weitere für eigene Projekte - sogar einige Wettbewerbe, z. B. die Vorhersage des Überlebens der Titanic-Katastrophe. Die Schüler sollten sich mit Klassifizierung und Regression beschäftigen.



	Linear Regression	Random Forest	Tree	price
16	11709	10139	10424	10000
288	11213	10090	10481	10200
41	10720	10236	10434	10250
307	10576	10832	10274	10254
189	10679	10329	10434	10295
293	10676	10329	10434	10380
294	10825	10331	10434	10390
73	10882	10567	10274	10490
175	10039	10214	10274	10490
143	11925	11462	11788	10492
205	12188	11067	11788	10495
142	10898	10603	10481	10498
96	10036	10214	10274	10500
62	10892	10422	10434	10519
1	12379	10883	11788	10550
82	10970	10812	10274	10698
282	10379	10558	10434	10700
334	11695	10915	10424	10700
110	10644	11319	11886	10746
313	12461	10971	11788	10790
117	11051	10997	10434	10795
146	10342	10791	10434	10795
147	11315	11106	11157	10795
47	11632	10840	10481	10798
183	12516	10920	11788	10900

Model	MSE	RMSE	MAE	R2
Linear Regression	952084.617	975.748	749.319	0.909
Random Forest	260938.983	510.822	368.354	0.975
Tree	612748.725	782.783	594.072	0.941

Lektion 17: Orange Data Mining Project, wie überlebt man die Titanic?



9: Hyperrealistisches Foto der Titanic vor der Skyline von New York, sonniges helles Wetter, leichte Bewölkung, professionelles Foto, vom Meer aus fotografiert, ruhige See, lebendige Beleuchtung, Meisterwerk

Oder: Maschinelles Lernen aus dem Desaster

Was sollen die Schüler lernen?

Die Schüler analysieren einen unvollkommenen Originaldatensatz. Nicht perfekt bedeutet, dass ein Datensatz unvollständig ist und einige Merkmale unbrauchbar sind. Dieses Szenario ist also sehr realitätsnah: Bevor ML-Modelle trainiert werden, sind die Studierenden gezwungen, sich mit dem Datensatz von Grund auf vertraut zu machen. Es müssen grundlegende Annahmen über die Relevanz von Merkmalen getroffen werden, bevor eine sinnvolle Auswahl von Daten getroffen werden kann. Die Ergebnisse sind nicht immer zufriedenstellend. Es ist mit einer durchschnittlichen Genauigkeit der Ergebnisse im Bereich von 0,8 zu rechnen. Aber: Es lassen sich immer wieder durchaus plausible und überraschende kausale Zusammenhänge identifizieren. Zum Beispiel ist die Überlebenswahrscheinlichkeit stark abhängig vom Geschlecht oder der Klasse, in der man eingebucht wurde.

Mögliche Aktivitäten der Schüler

Führen Sie ein komplettes DataMining-Projekt von Anfang bis Ende durch: Ausgehend von einem (hoffentlich) interessanten realen Datensatz wird die gesamte Datenaufbereitungspipeline eingerichtet und verschiedene ML-Modelle damit trainiert. Dies geht bis hin zur Feinabstimmung der Parameter. So lernen die Studierenden spielerisch den Umgang mit datengetriebenen Projekten.

Es war ein berühmter Wettbewerb auf Kaggle.com: Wie kann man die Überlebenswahrscheinlichkeit der Titanic-Katastrophe vorhersagen? Der Originaldatensatz kann hier gefunden werden:

<https://www.kaggle.com/competitions/titanic>

Teil I: Erläuterung der Attribute des Datensatzes

Es gibt 12 Merkmale, einige sind mehr, andere weniger relevant:

Attribut	Erläuterung
SURVIVED	Die Zielvariable, zwei verschiedene Klassen: 0: nicht überlebt, 1: überlebt
PCLASS	Erste Klasse (teuer), zweite und dritte Klasse: 1, 2 und 3
NAME	Der Name des Passagiers
GENDER	Männlich und weiblich
AGE	Alter in Jahren
SIBSP	Geschwister und Ehegatten: Anzahl der Verwandten plus Ehemann oder Ehefrau an Bord
PARCH	Anzahl der Eltern oder Kinder an Bord
TICKET	ID der Fahrkarte
FARE	Preis der Fahrkarte
CABIN	Kabinennummer
EMBARKED	Eingangshafen, C: Cherbourg, Q: Queenstown, S: Southampton
PASSENGERID	Fortlaufende Nummer für jede Zeile, "Primärschlüssel"-Variable

Teil II: Didaktische Hinweise für die Lehrkraft, was die Schüler lernen und tun sollen:

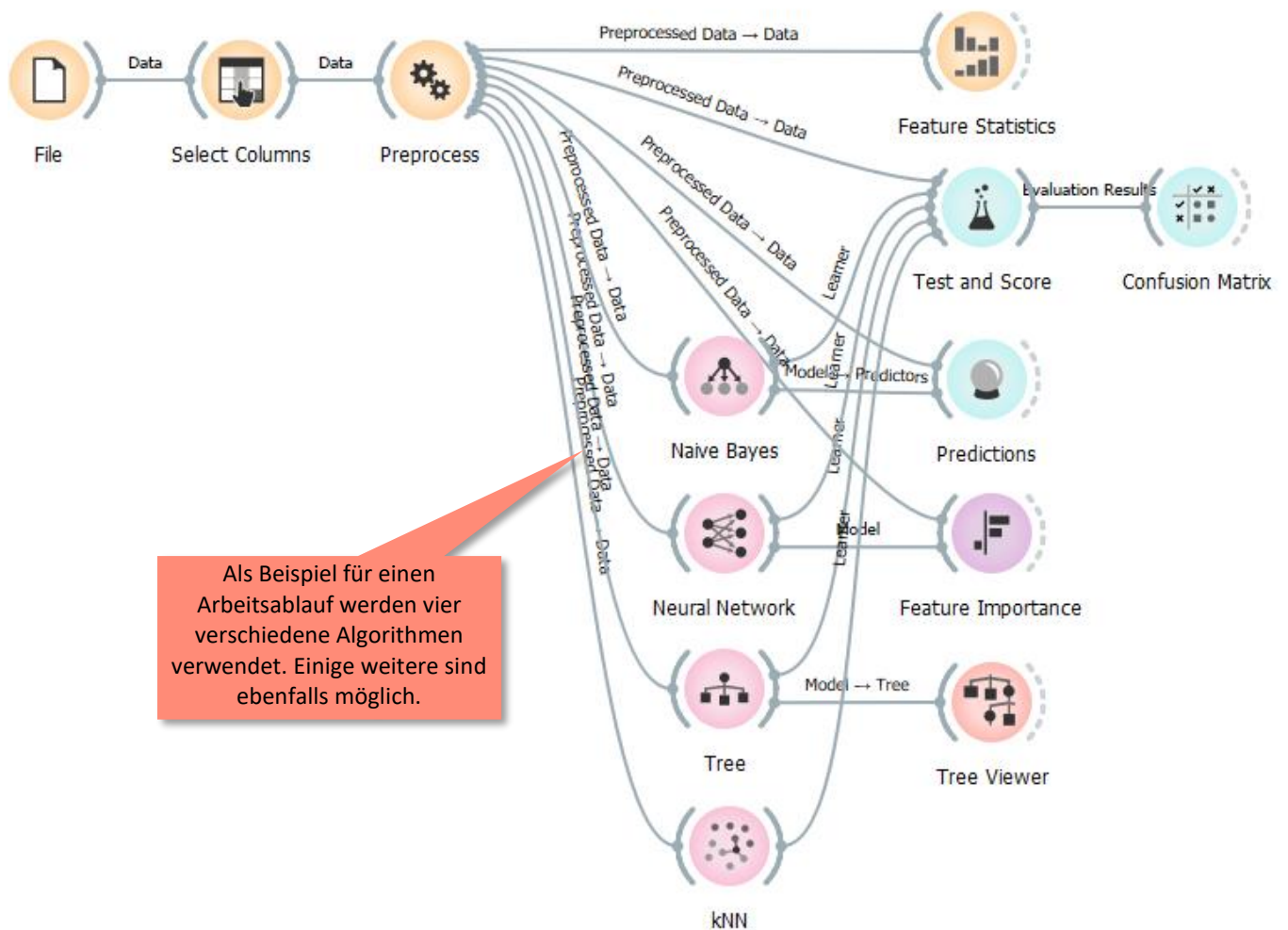
Bei diesem Datensatz ist der Aufwand etwas höher. Bevor ein Modell erstellt werden kann, ist eine Vorbereitung erforderlich:

1. Es gibt eine ganze Reihe von Einträgen, bei denen Werte fehlen. Normalerweise muss man diese Einträge zuerst entfernen. Dies geschieht mit dem Knoten "Preprocess".
2. Es gibt einige Merkmale, die für eine Vorhersage keine Rolle spielen. Dies sind die Kabinennummer, der Name, die "Passenger ID" und das Ticket. Diese werden mit dem Knoten "Select Columns" aus dem Datensatz entfernt.

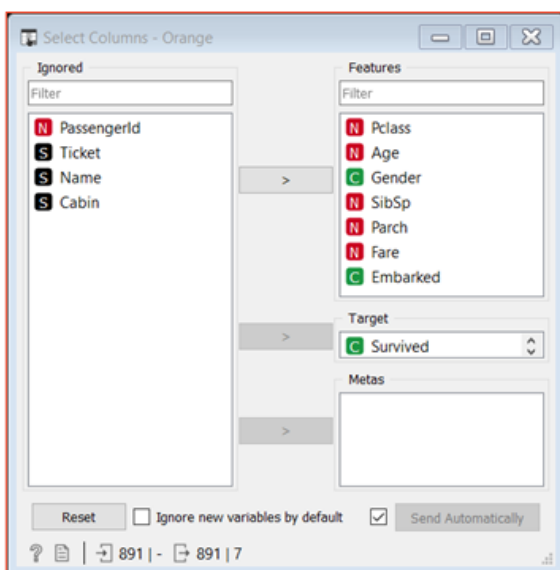
Sobald der Datensatz bereinigt ist, werden die Daten zunächst nach verschiedenen Aspekten gesichtet. Hier sticht der Knoten "Feature Statistics" hervor. Hier können Sie verschiedene bedingte Wahrscheinlichkeiten formulieren, was eine gute Möglichkeit ist, das Bayes-Theorem zu üben.

Der Knoten "Neuronales Netz" wird nicht ernsthaft für eine Vorhersage benötigt, denn hier wird mit "Kanonen auf Spatzen" geschossen. Aber der Knoten wird benötigt, um den Knoten "Merkmalsbedeutung" zu berechnen. Hier können sinnvolle Interpretationen gemacht werden, welches Merkmal besonders wichtig ist und welche Merkmale eher unwichtig sind.

Teil III: Eine beispielhafte vollständige orangefarbene Einrichtung



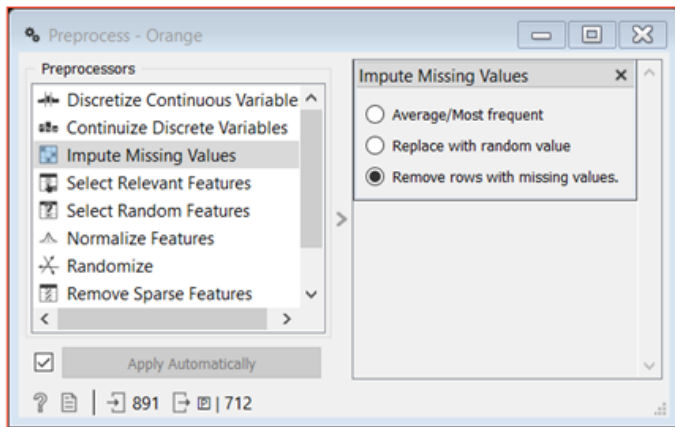
Teil IV: Kurze Erläuterung der Knotenpunkte



Spalten auswählen

Die irrelevanten Merkmale werden aus dem Datensatz entfernt, "Überlebt" wird als Zielmerkmal definiert.

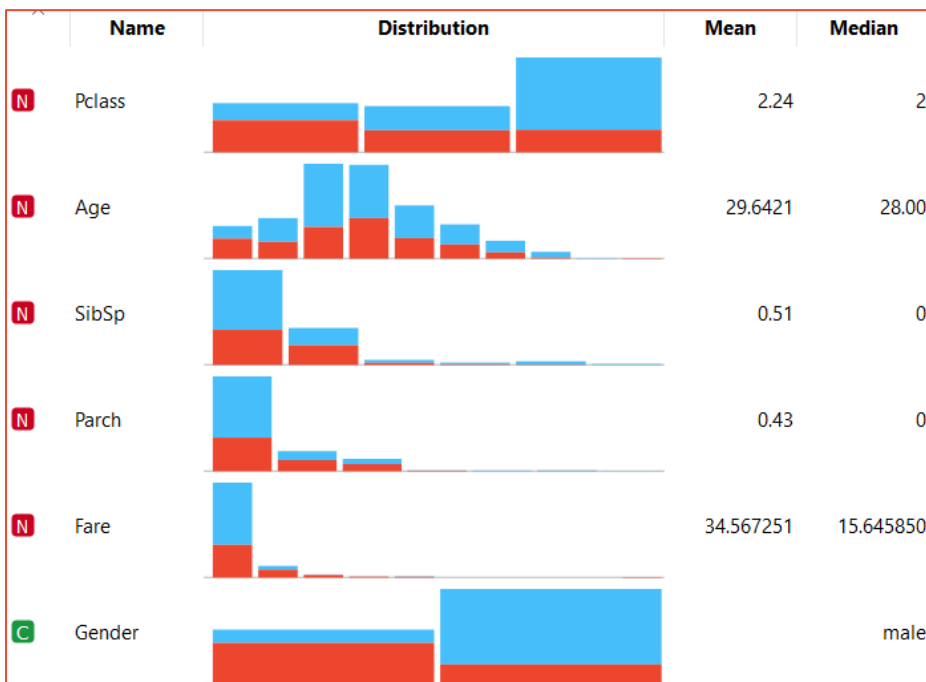
Abb. 62 Eigene Screenshots aus Orange Datamining



Vorverarbeiten

Es fehlen einige Werte. Dies kann auf verschiedene Weise behoben werden. Hier entfernen wir einfach die Zeilen mit fehlenden Werten.

22	1	2	34.00	male	0	0	13.0000	S
23	1	3	15.00	female	0	0	8.0292	Q
24	1	1	28.00	male	0	0	35.5000	S
25	0	3	8.00	female	3	1	21.0750	S
26	1	3	38.00	female	1	5	31.3875	S
27	0	3	?	male	0	0	7.2250	C
28	0	1	19.00	male	3	2	263.0000	S
29	1	?	?	female	0	0	7.0792	Q
30	0	3	?	male	0	0	7.8958	S
31	0	1	40.00	male	0	0	27.7208	C
32	1	1	?	female	1	0	146.5208	C
33	1	3	?	female	0	0	7.7500	Q
34	0	?	?	male	0	0	10.6000	S
35	0	1	28.00	male	1	0	82.1708	C



Merkmals-Statistik

Hier können Sie z. B. das Schätzen von bedingten Wahrscheinlichkeiten üben:

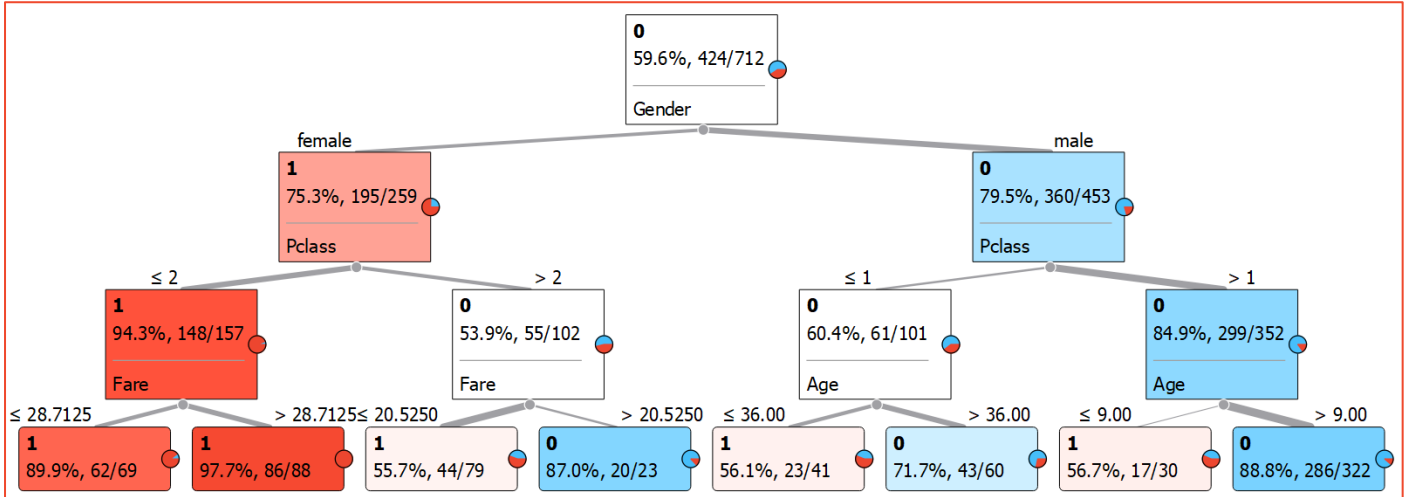
Wie hoch war die Überlebenswahrscheinlichkeit unter der Bedingung, weiblich / männlich zu sein?

Wie hoch war die Überlebenswahrscheinlichkeit unter der Bedingung, in der ersten Klasse eingeschrieben gewesen zu sein?

Achtung! Hier ist eine sensible Sprache gefragt. Zum einen, weil es um den Tod von Menschen geht und zum anderen, weil es sich um geschlechtersensible Überlegungen handelt.

Abb. 63 Eigene Screenshots aus Orange Datamining

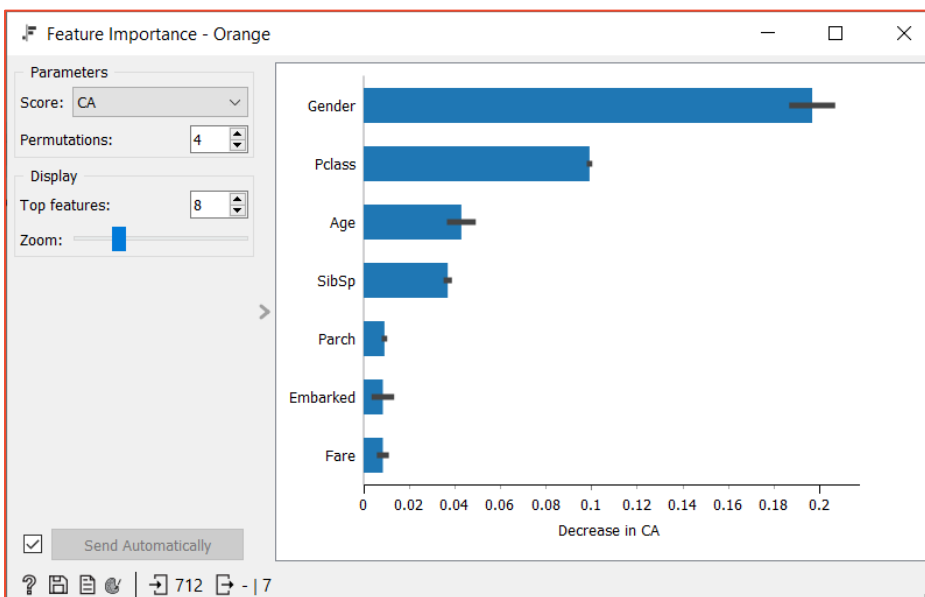
Entscheidungsbaum & Baumbetrachter



Laut dem Test- und Score-Knoten erreicht der Entscheidungsbaum eine Genauigkeit von 0,78, was recht gut ist. Außerdem gibt uns der Entscheidungsbaum die Möglichkeit, die Beziehungen zu interpretieren:

- Dementsprechend haben die Damen der ersten Klasse die höchste Überlebenswahrscheinlichkeit.
- Als männlicher Passagier in der dritten Klasse hatte man eine halbwegs vernünftige Überlebenswahrscheinlichkeit, wenn man jünger als 9 Jahre war. Kinder hatten also eine höhere Überlebenschance als Erwachsene.

Wahrscheinlich galt der berühmte Spruch "Frauen und Kinder zuerst".

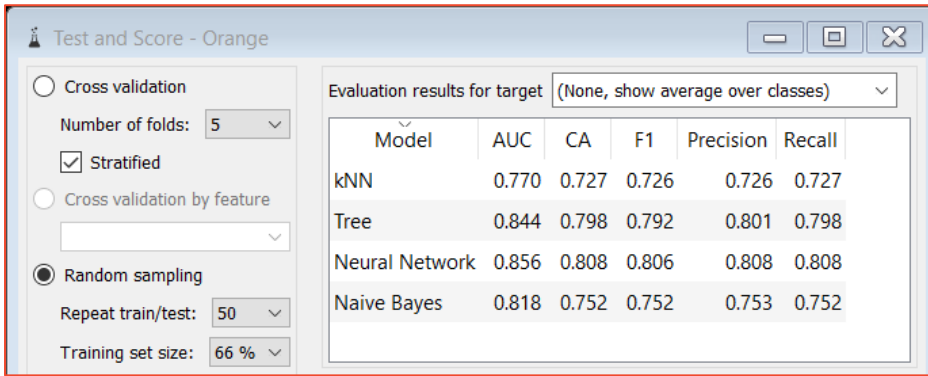


Bedeutung des Merkmals

Zusammen mit dem NN-Knoten, der hier als eine Art Analysewerkzeug dient, können die bisherigen Aussagen noch einmal bestätigt und verstärkt werden.

Um zu überleben, waren das Geschlecht, die Passagierklasse und das Alter am wichtigsten. Der Einreisehafen zum Beispiel spielte erwartungsgemäß kaum eine Rolle.

Abb. 64 Eigene Screenshots aus Orange Datamining



Test und Ergebnis

Mit einem Train-Test-Split von 66 % und 50 Epochen scheinen Baum und Neuronales Netz die besten Modelle für die Vorhersage der Überlebenswahrscheinlichkeit zu sein.

Abb. 65 Eigene Screenshots aus Orange Datamining

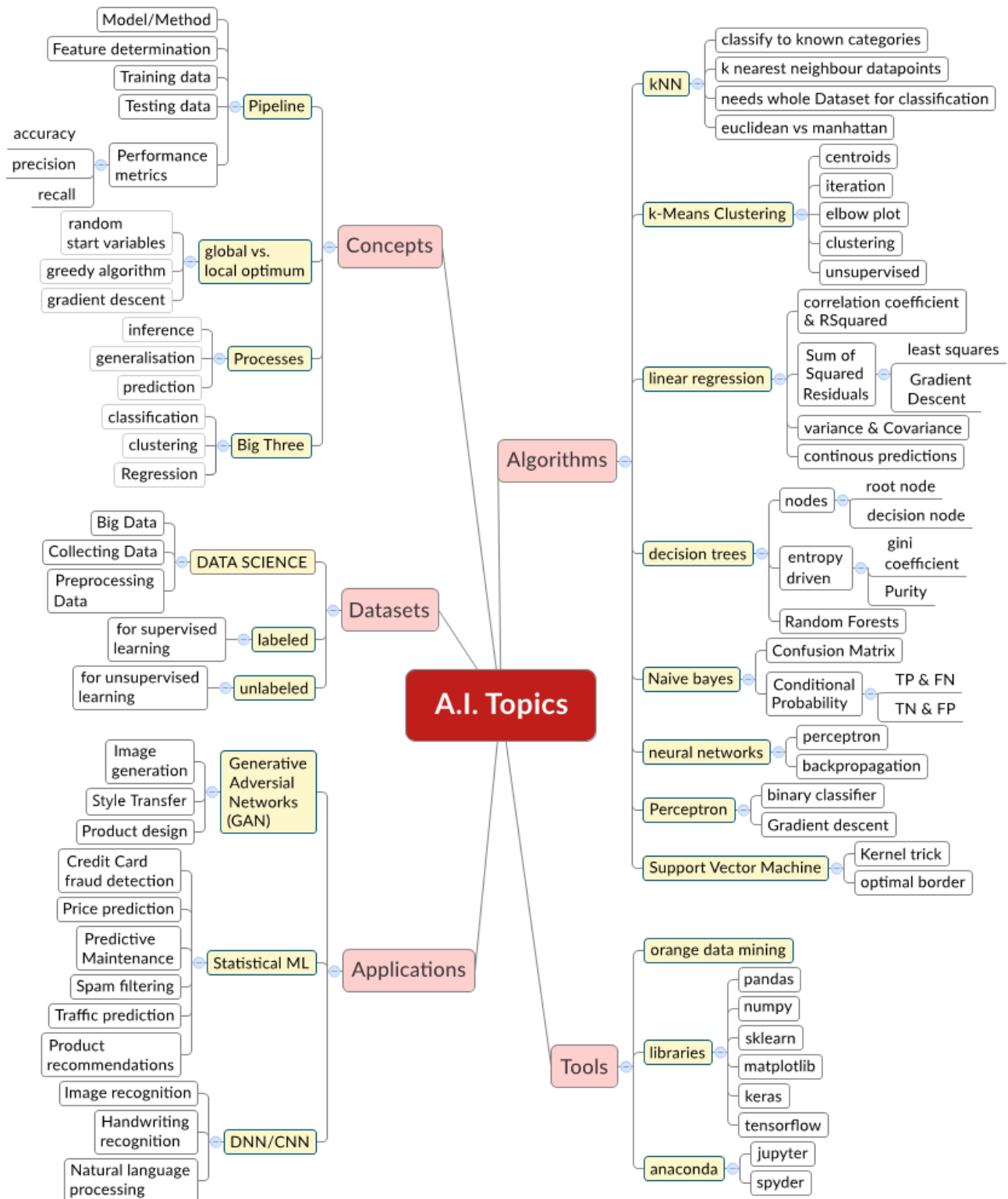
Lektion 18: Zusammenfassung der gesamten Lektion

Was sollen die Schüler lernen?

Die Schülerinnen und Schüler sollten in einer Gesamtübersicht die gelernten Unterrichtsinhalte visualisieren. Hier gibt es die Möglichkeit, die Lerninhalte miteinander zu vernetzen. Ein Gesamtüberblick sorgt für besseres Verständnis.

Mögliche Aktivitäten der Schüler

Die Schüler fassen den gesamten Inhalt der Unterrichtsstunde zusammen, indem sie sich in Gruppen zusammenfinden und darüber diskutieren. Das Arbeitsergebnis könnte z. B. eine Mindmap oder eine kurze Präsentation sein. Diese kann durch einen kurzen Vortrag oder ein Poster präsentiert werden.



Didaktische Überlegungen zum Schluss

Was ist das Wesentliche an den vielen verschiedenen ML-Algorithmen? Was sollten die Schülerinnen und Schüler am Ende der Unterrichtseinheit implizit über sie verstanden haben?

Es wurden viele verschiedene Algorithmen untersucht: Lineare und logistische Regression, Naive Bayes, Support Vector Machines, Entscheidungsbäume, kNN, und k-Means-Clustering. Im nächsten Kurs werden Neuronen und neuronale Netze hinzukommen. Gibt es ein gemeinsames Prinzip hinter all diesen Einzelalgorithmen?

Ja, es ist die Fähigkeit all dieser Algorithmen, sich an unbekannte neue Daten anzupassen. Dies geschieht dadurch, dass diese Algorithmen versuchen, sich so gut wie möglich die Fehler iterativ zu minimieren. Die dynamische Anpassungsfähigkeit durch Fehlerminimierung wird somit zum Herzstück der verschiedenen Methoden.

Streng genommen handelt es sich um Optimierungsalgorithmen. Im Falle des Entscheidungsbaums ist die treibende Kraft die Entropiereduktion, die durch einen ‚greedy‘ Algorithmus durchgeführt wird. Ein k-means-Algorithmus optimiert die Lage der Zentren in verschiedenen Iterationsschritten, und eine Support-Vector-Maschine optimiert die Breite des Korridors zwischen den Punkten. Eine logistische Regression wiederum optimiert nach der ‚Maximum Likelihood‘-Methode.

Der lineare Regressionsalgorithmus unterscheidet sich von dem in der statistischen Wissenschaft verwendeten Algorithmus durch seine Implementierung: Durch iterative Änderung der Steigung wird der Einfluss der Restquadrate minimiert. Dies wurde in diesem Kurs zwar nicht ausdrücklich gelehrt, da es sich um höhere Mathematik handelt, aber hier werkelt das Prinzip des Gradientenabstiegs.

Das zentrale Element des Gradientenabstiegs findet sich in neuronalen Netzen wieder: Ein Neuron lernt auch durch die Optimierung seiner Gewichte. Vermittelt werden diese durch den Backpropagation-Algorithmus. Und der arbeitet, ähnlich wie die lineare Regression, nach dem Prinzip des Gradientenabstiegs.

Dies ist der rote Faden, den die Studierenden mitnehmen sollten: Setzt man das Bild aus den einzelnen Mosaiksteinen zusammen, kommt man zu der Erkenntnis, dass das Wesen des maschinellen Lernens in der Optimierung auf unbekannte Daten liegt.



Abb. 66 Bild [[Gemeinfrei](#)] erzeugt mit [DALL-E](#); Prompt „single long red thread against white background, high quality photo, depth of field“ von Jörg [[CC BY-SA 4.0 International](#)]