

Datenanalyse und Vorhersage mit Klassifikationsbäumen

Ein Unterrichtskonzept für die Sekundarstufe II

Lehrerversion / Gesamtkonzept

Stand: Oktober 2020

Dr. rer. nat. Andreas Grillenberger

Kontakt: andreas@grillenberger.ch

Erstellt mit Unterstützung von *StRin RS Anne-Katrin Jäger*

Inhaltsverzeichnis

Lehrerversion / Konzept	3
Einführung für die Lehrkraft	3
Arbeitsblatt 1: Logik oder scharfes Hinsehen?	5
Arbeitsblatt 2: Händische Datenanalyse (Teil I)	9
Arbeitsblatt 3: Händische Datenanalyse (Teil 2)	13
Arbeitsblatt 4: Datenanalyse am Computer	15
Arbeitsblatt 5: Datenanalyse am Computer (Teil 2)	19
Arbeitsblatt 6: Diskussion der Ergebnisse	23
 Schüler-Arbeitsblätter	 27
Arbeitsblatt 1: Logik oder scharfes Hinsehen?	29
Arbeitsblatt 2: Händische Datenanalyse (Teil I)	31
Arbeitsblatt 3: Händische Datenanalyse (Teil 2)	33
Arbeitsblatt 4: Datenanalyse am Computer	35
Arbeitsblatt 5: Datenanalyse am Computer (Teil 2)	37
Arbeitsblatt 6: Diskussion der Ergebnisse	39

Einführung für die Lehrkraft

Ziele

Im Rahmen des Unterrichts soll den Schülern die Möglichkeit gegeben werden, zu erkennen wie die heute allpräsenten korrelationsbasierten Datenanalysen funktionieren. Diese versuchen, aus einem großen Berg an Daten Informationen zu gewinnen, ohne dass der konkrete Weg der Analyse vorher klar ist.

Es wird dabei angestrebt, dass die Schüler einen kritischen Blick auf Datenanalysen entwickeln und sich der Grenzen dieser Analysen bewusst werden.

Folgende Lernziele werden daher angestrebt: Die Schüler...

- erklären anhand eines Beispiels den Unterschied zwischen Kausalität und Korrelation bezogen auf Datenanalysen.
- beschreiben den Ablauf einer typischen korrelationsbasierten Datenanalyse (ggf. unter Zuhilfenahme eines Diagramms).
- beschreiben das Konzept „Klassifikationsbaum“ und erstellen einen solchen für gegebene Regeln.
- erstellen anhand eines „Klassifikationsbaums“ eine Prognose für einen Datensatz.
- beurteilen Analysen hinsichtlich ihrer Qualität anhand der auftretenden Fehl-Zuordnungen.
- führen einfache korrelationsbasierte Datenanalysen mit einem geeigneten Werkzeug am Computer selbst durch.
- beurteilen reale und fiktive Beispiele von korrelationsbasierten Datenanalysen hinsichtlich ihres Nutzens und ihrer Gefahren.

Zum Material

Alle Materialien stehen unter der CreativeCommons-Lizenz *CC BY-NC-SA 4.0*¹ und können unter Wahrung dieser Lizenz weiterverbreitet werden. Für den Einsatz im Unterricht an öffentlichen bzw. öffentlich anerkannten Schulen und die Weitergabe an die dortigen Schülerinnen und Schüler darf auf die im Rahmen der Lizenz geforderte Namensnennung explizit verzichtet werden. Alle im Folgenden genannten Aufgaben sind am Ende des Konzepts auch in einer Schülerversion zu Arbeitsblättern zusammengefasst.

Überblick

Zeitansatz: je nach Detailgrad werden ca. zwei bis vier Doppelstunden veranschlagt.

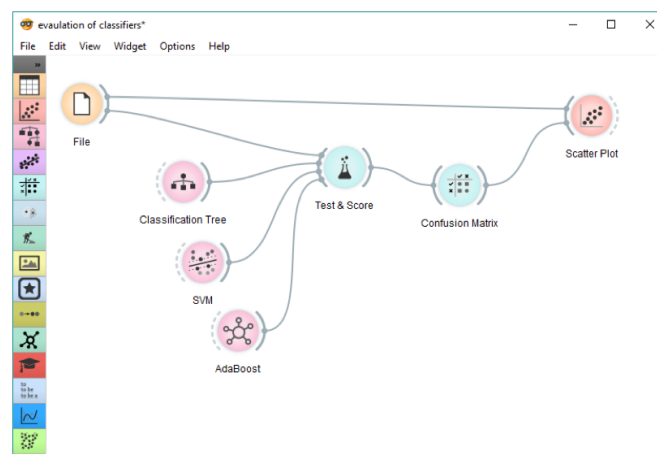
- Einblick in die Nutzung von Datenanalysen anhand eines Realweltbeispiels; Versuch dieses zu erklären, indem mögliche Kausalzusammenhänge diskutiert werden
- Begrifflichkeit: Kausalität vs. Korrelation als „mit gesundem Menschenverstand erklärbar“ vs. „unerklärbar, aber anhand von Daten richtig erscheinend“; Diskussion der damit einhergehenden Gefahr durch Fehleinstufungen
- Überblick über den Datenanalyseprozess und Feststellung wo Kausalität vs. Korrelation dabei zentral ist

¹<https://creativecommons.org/licenses/by-nc-sa/4.0/>

- Händische Durchführung einer einfachen korrelationsbasierten Analyse zum Erlernen der Grundsätze
 - Finden von Regeln im Datensatz
 - Darstellung der Regeln als Klassifikationsbaum
 - Nutzung des Klassifikationsbaums zur Prognose von Attributen weiterer Datensätze
 - Diskussion der Qualität
- Durchführung am Computer zum Erkennen des Potentials und der Grenzen bei Verwendung größerer Datensätze
 - Aufstellen einer Vermutung, welche Attribute des Datensatzes für die Bestimmung des vorherzusagenden Attributs relevant sind
 - Nutzung eines Datenanalysewerkzeugs zur Erstellung eines Klassifikationsbaums; Verifikation der vorherigen Hypothesen über relevante Attribute
 - Erstellen einer automatisierten Vorhersage anhand des erstellten Klassifikationsbaums; erste Einschätzung der Ergebnisse
 - Nutzung einer Confusion Matrix um Fehler in den Vorhersagen systematisch zu erkennen; Beurteilung der Analysequalität
 - Überprüfung von Möglichkeiten zur Verbesserung der Analysequalität
- Diskussion verschiedener fiktiver und realer Beispiele zur Verwendung von Datenanalysen hinsichtlich ihres Nutzens und ihrer Gefahren

Verwendetes Datenanalysewerkzeug

Es wird das Datenanalysetool Orange3 genutzt, das an der Universität Ljubljana entwickelt wird und für alle gängigen Desktop-Betriebssysteme zur Verfügung steht. Dieses Werkzeug erlaubt einen einfachen Zugang, da die Datenanalysen nicht durch textbasierte Programmierung durchgeführt, sondern in einer Art Datenflussdiagramm aufgebaut werden, wodurch ein guter Überblick über die Analyse gewonnen wird. Ursprünglich wurde eine reduzierte Version des Werkzeugs verwendet, die nicht nötige Funktionen ausblendete. Im Unterricht hat sich jedoch gezeigt, dass Schülerinnen und Schüler auch mit der nicht reduzierten Version problemlos zurechtkommen, sodass aus Aufwandsgründen auf diese verzichtet und auf die Originalversion von <https://orange.biolab.si> zurückgegriffen wurde.



Arbeitsblatt 1 (L): Logik oder scharfes Hinsehen?

Zum Einstieg wird die Nutzung eines für die Schüler interessanten und gleichzeitig Fragen aufwerfenden Realweltbeispiels vorgeschlagen. Es bietet sich hier beispielsweise die folgende Geschichte an, die sich in den USA ereignet haben soll (z. B. präsentieren mit Beamer o. Ä.):

Dem US-Einzelhandelsriesen Target gelang es durch die Analyse herauszufinden, welche Kundinnen schwanger sind. Duhigg schreibt, dies sei für das Unternehmen sehr wichtig gewesen, denn werdende Eltern seien so etwas wie der „Heilige Gral“ für Unternehmen wie Target. In einer Schwangerschaft änderten sich die Gewohnheiten, und wer vorher keine gute Kundin des Einzelhändlers gewesen sei, könne es danach werden - wenn man ihr zu richtigen Zeit die richtige Werbung zusendet.

Die Statistiker von Target, so berichtet es Duhigg, identifizierten etwa 25 Produkte, die darauf hinweisen, dass Kundinnen schwanger sind. Genauer gesagt, wenn sie sich im zweiten Trimester ihrer Schwangerschaft befinden. Denn zu diesem Zeitpunkt fingen sie an, sich neue Sachen zu kaufen, und Target schickte ihnen dann schon Werbung. Zu den identifizierten Produkten gehörten parfümfreie Körperlotion, große Mengen an Watte und Nahrungsergänzungsmittel wie Kalzium, Magnesium und Zink. Target habe in der Kundendatenbank gesucht und Zehntausende Frauen gefunden, die mit großer Wahrscheinlichkeit bald Mutter würden.

Der Autor Duhigg berichtet darüber, wie die Werbung für Schwangerschaftsprodukte den Vater einer Tochter in Rage versetzte. Er beschwerte sich in einem Target-Markt in der Nähe von Minneapolis darüber, dass seine Tochter - noch ein Teenager - Werbung für Babykleidung erhalten habe. Ob man sie dazu animieren wolle, schwanger zu werden, fragte er den Manager des Ladens. Dieser entschuldigte sich, doch als er später noch einmal sein Bedauern zum Ausdruck bringen wollte und den Vater anrief, stellte sich heraus, dass die Tochter wirklich schwanger war. Target hatte es nur vor dem Vater der jungen Frau gewusst.

— Frankfurter Neue Presse, 13.09.2014

Verfügbar unter: <http://www.fnp.de/art673,1029989>

Dieser Unterrichtseinstieg wirft die Frage auf, wie der Supermarkt Target die entsprechenden Produkte erraten konnte und wie solche Analysen im Allgemeinen funktionieren. Dies kann im Unterrichtsgespräch diskutiert werden.

Während bei diesem Beispiel noch vermutet werden kann, dass findige Personen sich die Kriterien zur Erkennung einer Schwangerschaft überlegt und anhand der Kundendaten überprüft haben, zeigt sich spätestens im Folgenden zweiten Beispiel, dass dies nicht immer auf diese Weise funktionieren kann:

Für Schüler/-innen, z. B. auf Arbeitsblatt

Im Unterricht hast du bereits einen Artikel darüber gesehen, wie Daten heute im Einzelhandel verwendet werden, um Kunden auf sie zugeschnittene Werbung präsentieren zu können. Onlineshops gehen heute jedoch schon weiter und versuchen, ihren Kunden viele Produkte möglichst schnell liefern zu können:

Noch bevor ein Kunde überhaupt den Button "Kaufen" anklickt, soll die für ihn passende Ware schon auf dem Weg in Richtung seiner Wohnung sein. Dem Versandhändler Amazon wurde ein Patent zugesprochen, das einen „vorausschauenden Versand“ („anticipatory shipping“) ermöglichen soll. Das heißt: Bestimmte Waren werden schon einmal an ein Versandzentrum geschickt, in dessen Nähe sich ein oder mehrere Kunden höchstwahrscheinlich für das Produkt interessieren. Wird es dann schließlich bestellt, ist es umso schneller beim Empfänger.

— Spiegel Online, 18.01.2014

Verfügbar unter: <http://www.spiegel.de/netzwelt/web/a-944252.html>

Eine weitere Kontextualisierung kann durch Beispiele wie Same-Day-Delivery, wie sie verschiedene Onlinehändler anbieten, geschehen.

Den Schülern kann nun eine erste Aufgabe gegeben werden, in der das Ziel sein sollte, zu erkennen, dass diese Art der „Voraussage“ von Kundenverhalten nicht mehr rein auf logischen Schlüssen geschehen kann. Die Aufgabe könnte daher wie folgt lauten:

Aufgabe 1

Um herauszufinden, was ein Kunde als nächstes bestellen könnte, müssen die Versandhändler umfangreiche Daten über ihre Kunden sammeln und analysieren.

a) Was wissen Onlinehändler über ihre Kunden? Woher haben diese die jeweilige Information?

Information über den Kunden	Quelle
Vor- und Nachname	Registrierung
Adresse	Registrierung
Geburtsdatum	Registrierung
Beliebte Artikel	Einkäufe
Bekannte	Versandadressen
Aufenthaltsorte	Versand- & IP-Adressen
...	...

b) Wahrscheinlich sind nicht alle Informationen, die ein Onlinehändler über seine Kunden hat auch wichtig, wenn er herausfinden möchte, welchen Artikel der Kunde als nächstes bestellen könnte. Markiere in der Tabelle oben die Zeilen, von denen du denkst, dass sie für diese Zweck wichtig sind, indem du ein + neben die wichtigen Zeilen machst.

Die Ergebnisse können nicht auf Korrektheit überprüft werden, geben den Schülern aber die Gelegenheit sich Gedanken über das Konzept zu machen und zu erkennen, dass es eigentlich nur

wenige Daten gibt, die dafür wirklich hilfreich erscheinen. Erst bei Betrachtung der Gesamtmenge an Daten können aus diesen anscheinend relevante Schlüsse gezogen werden. Ziel ist es, über mögliche Attribute zu diskutieren. Das ist gut, denn es zeigt, dass keine logischen Schlüsse gezogen werden können und leitet damit hin zu korrelativen Analysen. Auf dieser Basis können dann korrelative Analysen besprochen werden. Die Ergebnisse können im Lückentext gesichert werden:

Aufgabe 2

Fülle folgenden Lückentext aus: Es gibt bei der Datenanalyse zwei Möglichkeiten, wie wir Vorhersagen treffen können:

1. Wenn wir bereits etwas über die zu analysierenden Daten wissen, dann können wir uns erklären wie etwas funktioniert und damit Schlussfolgerungen ziehen. Es gibt also logische Zusammenhänge, sog. Kausalzusammenhänge, die wir zur Vorhersage nutzen können.

Beispiel:

WENN ein Kunde in den letzten 5 Einkäufen Chips gekauft hat, DANN wird er auch beim nächsten Mal welche kaufen.

2. In anderen Bereichen erkennen wir keinerlei logische Zusammenhänge. Stattdessen können wir nach Mustern in den Daten suchen. Diese liefern uns auch Zusammenhänge, wir können sie uns aber oft nicht erklären. Solche Zusammenhänge bezeichnen wir als korrelative Zusammenhänge.

Beispiel:

WENN ein Kunde den Artikel X gekauft hat und er in Y wohnt und mindestens 35 Jahre alt ist, DANN wird er auch Z kaufen.

Kausalzusammenhänge helfen uns zwar dabei Dinge zu verstehen, sie sind aber für Datenanalysen oft relativ wenig interessant: Sie sind oft offensichtlich und bekannt, sodass sie nur wenig neue Informationen hervorbringen. Wir können uns aber logisch erklären, dass sie richtig/wahr sind. Die korrelativen Zusammenhänge sind daher oft spannender, da sie neue Informationen eröffnen. Sie haben aber den Nachteil, dass sie nicht unbedingt logisch nachvollziehbar sind: Wie genau Wohnort und Alter das Kaufverhalten prägen, können wir uns meist nicht logisch erklären. Außerdem müssen wir sie erst finden, was relativ schwierig ist.

Am Ende dieser Unterrichtsstunde haben die Schülerinnen und Schüler bereits einen ersten Einblick in die Ziele und Problematik der Datenanalyse erhalten. Auf dieser Basis wird in der nächsten Unterrichtsstunde eine erste händische Datenanalyse durchgeführt.

Arbeitsblatt 2 (L): Händische Datenanalyse (Teil I)

Auf dieser Basis kann mit den Schülern der Prozess der Datenanalyse thematisiert werden, dabei bietet es sich an, ein Modell des Prozesses vorzustellen, das auch als Art Advance Organizer dienen kann und den weiteren Lernprozess strukturiert. Dieser kann an der Tafel gemeinsam erarbeitet werden, beispielsweise unter Nutzung von Karteikarten für die vier Teilprozesse:

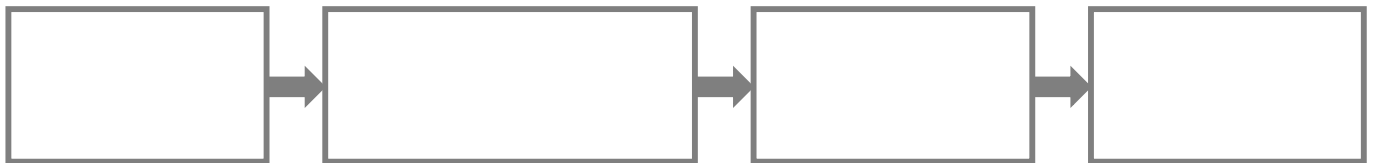


Zur Erlangung eines grundlegenden Verständnisses über die Datenanalysen wird ein erster Musterdatensatz zur Verfügung gestellt, der in einem Durchgang durch den gesamten Prozess händisch analysiert wird. Es wird dabei im Bereich der Erstellung des Vorhersagemodells versucht, sowohl die kausalitäts- als auch die korrelationsbasierte Datenanalyse zu thematisieren und deren Unterschiede herauszustellen.

Aufgabe 1

Im Unterricht wurde bereits gemeinsam erarbeitet, wie eine Datenanalyse abläuft. Vervollständige den folgenden Lückentext und das folgende Ablaufmodell:

Als erster Schritt der Datenanalyse, müssen die Daten erfasst/gewonnen und gespeichert werden. Aus diesen Daten wählt man sich üblicherweise eine kleine Teilmenge aus, um aus dieser das Vorhersagemodell zu erstellen, d. h. um Regeln zu finden, die die Vorhersage der gesuchten Eigenschaften ermöglichen. Diese Regeln können dann genutzt werden, um die Vorhersagen zu erstellen. Als letzter Schritt jeder Datenanalyse sollte die Bewertung der Ergebnisse erfolgen, mit dem Ziel eine möglichst gute Qualität der Ergebnisse sicherzustellen.



Der erste Schritt für die Schüler ist die Erstellung eines geeigneten Vorhersagemodells, dies erfolgt in der Folgenden Aufgabe:

Aufgabe 2

Ein Onlineshop hat über seine Kunden verschiedene Daten gesammelt und möchte nun seine Kunden durch den Versand von individuellen Gutscheinen zu weiteren Käufen anregen. Dazu will er herauszufinden, welche Produktkategorie für jeden Kunden jeweils besonders interessant ist. Der Shop hat bereits folgende Daten über jeden seiner Kunden gesammelt: Alter, Familienstand, Anzahl der Kinder, präferierte Zahlungsart, Kategorien der letzten vier eingekauften Produkte (Film, Sport, Software, Elektronik, Kleidung, Musik, Bücher oder Auto).

Um jedem Kunden einen Gutschein zu schicken, den dieser wahrscheinlich einlöst, möchte der Onlineshop herausfinden, welche Kategorie für den Käufer besonders interessant ist. Welche WENN-DANN-Regeln vermutest du, die dem Onlineshop dabei helfen könnten? *Hinweis: natürlich kannst du mehrere Bedingungen mit „und“ verknüpfen, z. B. „Kategorie 1 = Elektronik und Anzahl Kinder = 0“.*

- WENN XX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX
- WENN XX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX
- WENN XX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX

Die Schüler werden möglicherweise zu verschiedenen Regeln kommen, im Endeffekt wird sich aber herausstellen, dass zu wenig Information verfügbar ist, um sinnvolle und stichhaltige Kausalzusammenhänge zu erkennen. Als nächstes kann daher folgende Aufgabe gestellt werden:

Aufgabe 3

Nachdem die Datenwissenschaftler des Unternehmens erkannt haben, dass keine stichhaltigen Zusammenhänge in den bisher vorliegenden Informationen erkennbar sind, wurde entschieden, es anders zu versuchen: Der Onlineshop hat daher einige seiner Kunden befragt, was für sie als nächstes interessant ist. Dabei sind folgende Daten herausgekommen. Welche Zusammenhänge erkennst du in der unten dargestellten Tabelle?

Beispiel:

WENN Kauf 1 ein Film ist und mit Girokarte bezahlt wurde, DANN interessiert der Kunde sich als nächstes für Artikel der Kategorie Auto.

Kurz: „Kauf1“=„Film“ und „bezahlt mit“=„Giro“ \Rightarrow „Interesse“=„Auto“

Alter	Kauf 1	Kauf 2	Kauf 3	Kauf 4	verheiratet	Kinder	bezahlt mit	Interesse
25-50	Film	Software	Film	Sport	Ja	1	Giro	Auto
25-50	Elektronik	Musik	Film	Software	Nein	1	VISA	Bücher
<18	Film	Elektronik	Sport	Sport	Nein	0	Giro	Auto
<18	Film	Musik	Kleidung	Sport	Nein	0	Giro	Auto
18-25	Bücher	Musik	Film	Haushalt	Nein	1	Giro	Bücher
<18	Bücher	Film	Film	Bücher	Nein	0	VISA	Bücher
25-50	Film	Film	Sport	Sport	Ja	1	Giro	Auto
25-50	Musik	Film	Film	Spielzeug	Nein	1	Giro	Bücher
25-50	Musik	Musik	Film	Haushalt	Nein	1	Giro	Bücher
25-50	Elektronik	Musik	Bücher	Software	Ja	1	Master	Elektronik
25-50	Software	Elektronik	Film	Spielzeug	Nein	1	Master	Bücher
25-50	Film	Film	Sport	Sport	Ja	0	Master	Elektronik
25-50	Musik	Elektronik	Bücher	Elektronik	Ja	1	Master	Elektronik

- WENN XX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX
- WENN XX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX
- WENN XX DANN ist der nächste Kauf aus der Kategorie XXXXXXXXXXXXXXXXXXXX

Die Schüler können hier verschiedene Regeln manuell aus den Daten herauslesen, die willkürlich wirken (und sind). Diese sind daher nicht logisch erklärbar, wie es für korrelationsbasierte Analysen oft üblich ist.

An dieser Stelle kann diskutiert werden, ob Regeln wie „kein Kind“ \Rightarrow „Interesse“ = „Elektronik“ aufgenommen werden sollten – aus den Daten ergibt sich das, es kann jedoch vermutet werden, dass diese auf nur einem Datensatz basierende Regel wenig stichhaltig ist.

Damit diese Regeln besser nutzbar sind, werden sie üblicherweise als Klassifikationsbaum dargestellt. Fachlich handelt es sich dabei um einen (nicht zwingend binären) Entscheidungsbaum, dessen Blättern die getroffenen Entscheidungen darstellen. Eine Aufgabe zur Überführung der Regeln in einen Baum könnte wie folgt aussehen:

Arbeitsblatt 3 (L): Händische Datenanalyse (Teil 2)

Aufgabe 1

Wenn der Onlineshop nun Vorhersagen treffen will, dann sortiert er die Kunden in verschiedene Kategorien ein - dies nennt man „Klassifikation“.

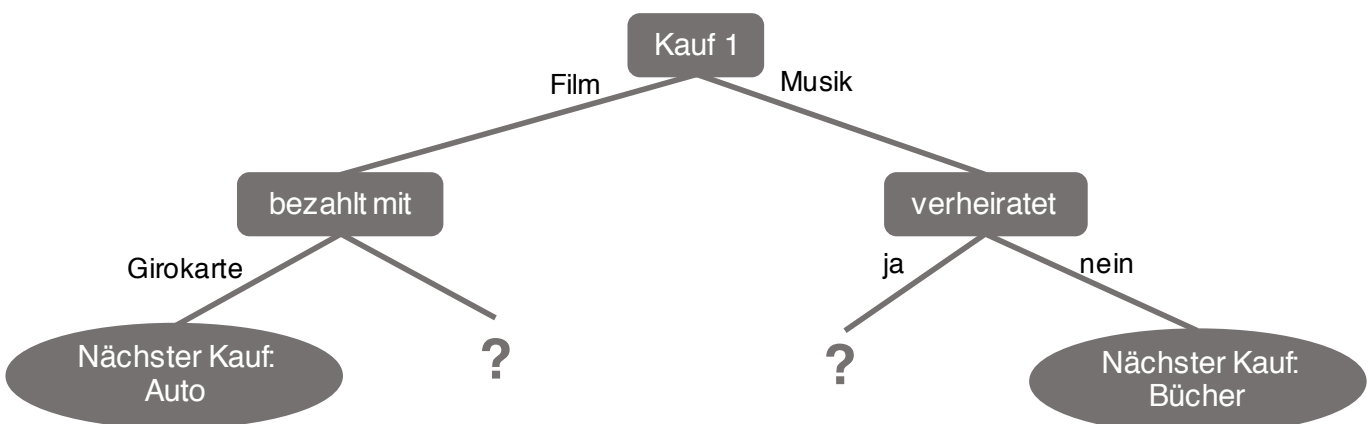
Um eine größere Menge an Regeln einfacher durchschaubar und anwendbar zu machen, stellt man diese als „Klassifikationsbaum“ dar. Dieser Baum symbolisiert die Entscheidungen, die anhand der Regeln getroffen werden.

Beispiel: Die Regeln

1. „Kauf 1“ = „Film“ und „bezahlt mit“ = „Giro“ \Rightarrow nächster Kauf = „Auto“
2. „Kauf 1“ = „Musik“ und nicht verheiratet \Rightarrow nächster Kauf = „Bücher“

können als Klassifikationsbaum wie unten abgebildet dargestellt werden.

- Markiere im Baum den Weg, den du gehen musst, wenn du herausfinden möchtest, was ein Kunde als nächstes gekauft hat, der als „Kauf 1“ einen Film gekauft und mit einer Girokarte bezahlt hat.
- Wir kennen zusätzlich die folgende Regel: Kauf 1 = „Film“ und „bezahlt mit“ ist „Mastercard“ und „verheiratet“ = „ja“ \Rightarrow nächster Kauf = „Elektronik“
Um diese im Baum zu berücksichtigen, musst du eine weitere Entscheidung ergänzen. Überlege dir, wo das sinnvoll ist und ergänze die Entscheidung.
- Überprüfe deine Ergänzung, indem du den Weg farbig markierst, den du durch den Baum gehen musst, um herauszufinden, was ein Kunde als nächstes kauft, der als „Kauf 1“ einen Film gekauft hat und mit „Mastercard“ bezahlt hat.



Mit diesem Klassifikationsbaum kann nun gut beschrieben werden, welche Regeln in den bisher bekannten Daten vorherrschen. Um eine Prognose zu treffen, müssen diese bekannten Informationen genutzt werden, um damit Informationen über einen Kunden vorherzusagen (eine Prognose zu generieren), die erst im Nachgang (wenn überhaupt) überprüft werden kann. Dazu kann den Schülern folgende Aufgabenstellung vorgelegt werden:

Aufgabe 2

Verwende den vorherigen Klassifikationsbaum, um zu entscheiden, an welcher Produktkategorie die folgenden Kunden wahrscheinlich als nächstes interessiert sind. Wenn diese Entscheidung anhand der beiden Regeln bzw. anhand des Baums nicht getroffen werden kann, schreibe ein ? in das Feld „vsl. interessiert an“.

Alter	Kauf 1	Kauf 2	Kauf 3	Kauf 4	verheiratet	Kinder	bezahlt mit	vsl. interessiert an
25-50	Film	Film	Sport	Sport	Ja	1	Giro	<i>Lösung: Auto</i>
25-50	Musik	Elektronik	Sport	Sport	Nein	1	Giro	<i>Lösung: Bücher</i>
>50	Film	Musik	Kleidung	Sport	Nein	1	Giro	<i>Lösung: Auto</i>
<18	Film	Musik	Film	Haushalt	Ja	1	Master	<i>Lösung: Elektronik</i>
>50	Bücher	Software	Film	Sport	Ja	1	VISA	<i>Lösung: ?</i>

Im Beispiel wurde absichtlich eine Stelle eingebaut, an der die Zuordnung mehrdeutig ist: Bei diesem Kunden ist es nicht möglich, eine eindeutige Vorhersage zu treffen. Dies kann genutzt werden, um eine Diskussion einzuleiten, was in solchen Fällen geschehen soll – und wie gut diese Analyse überhaupt sein kann: Was verursacht ein einzelner konträrer Datensatz der noch dazu kommt? Wie können wir die Analyse verbessern? Wie wichtig ist in diesem Fall eine möglichst gute Analyse?...

Arbeitsblatt 4 (L): Datenanalyse am Computer

Natürlich werden solche Datenanalysen in real nicht händisch, sondern rechnergestützt durchgeführt. Es soll den Schülern daher auch die Möglichkeit gegeben werden, das übliche Vorgehen auszuprobieren und damit auch mit etwas mehr Daten zu arbeiten, als bei der händischen Analyse vorher.

Als Datenbasis kann beispielsweise ein Datensatz gewählt werden, der die Schülerdaten von über 600 Schülern aus Portugal enthält, und in dem anhand dieser die Endnote/-punktzahl der Schüler prognostiziert werden soll. Diese Analyse kann beispielsweise wie folgt kontextualisiert werden. Es wurde hier absichtlich ein Kontext gewählt, der die Schüler direkt betrifft und der bei diesen sehr umstritten sein dürfte, um zu demonstrieren, dass auch die Schüler selbst von solchen Analysen prinzipiell direkt betroffen sein könnten.

Für Schüler/-innen, z. B. auf Arbeitsblatt

Es wäre für eure Lehrerin sicherlich eine sehr praktische Sache, die Idee der Onlineshops zu nutzen, um eure Schulnoten vorherzusagen: Dann würde es ausreichen jedes Mal nur ein paar Arbeiten zu korrigieren und die Noten aller anderen „vorherzusagen“. Doch wie (gut) funktioniert das wirklich?

Diese Fragestellung soll im Folgenden mit den Schülern ausführlich thematisiert werden, indem anhand des vorliegenden Datensatzes eine Analyse durchgeführt wird, die es erlauben wird, grundlegende Rückschlüsse auf die Qualität und die Möglichkeiten solcher Datenanalysen zu ziehen. Es bietet sich dabei an, den gleichen Prozess wie vorher händisch durchlaufen, auch an dieser Stelle wieder aufzugreifen. Als erstes steht die Erstellung von Klassifikationsregeln und eines Klassifikationsbaumes an. Um sich in den vorgegebenen Datensatz hineinzudenken, wird den Schülern zuerst folgende Aufgabe an die Hand gegeben:

Aufgabe 1

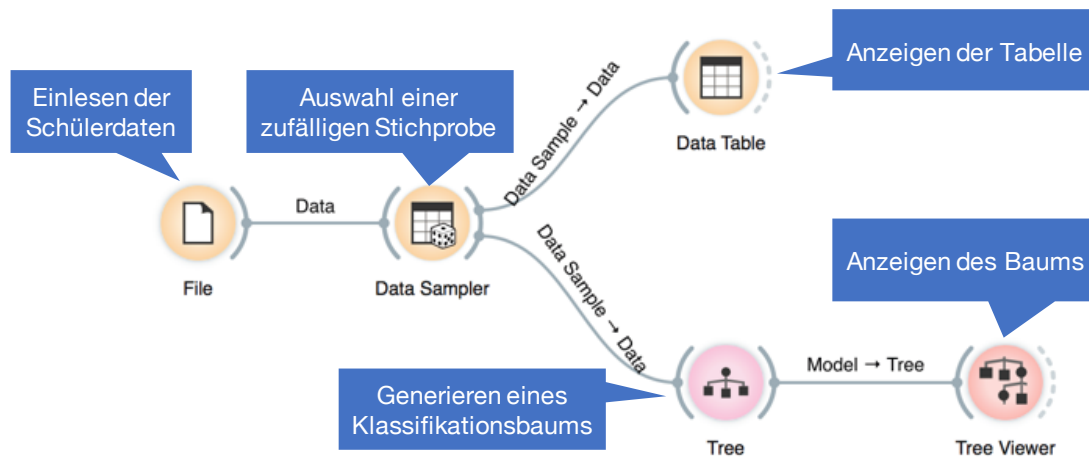
Zunächst ohne Computer: Von welchen der in der Tabelle unten dargestellten Eigenschaften der Schülerinnen/Schüler vermutest du, dass sie für das „Erraten“ bzw. „Berechnen“ der Endnoten wichtig sind? Markiere diese in der folgenden Tabelle.

Attribut	Beschreibung	relevant?
Schule	Kürzel der Schule: "GP" oder "MS"	
Geschlecht	"M" oder "W"	
Alter	Zahlenwert	
Wohnumfeld	"urban" oder "ländlich"	
Familiengröße	"≤3" oder ">3"	
Bildung Mutter	Grundschule; Hauptschule; Realschule/Gymnasium; Universität; keine	
Bildung Vater	vgl. Bildung Mutter	
Beruf Mutter	Gesundheitsbranche; Lehrerin; Hausfrau; Öffentlicher Dienst; sonstige	
Beruf Vater	vgl. Beruf Mutter	
Fahrzeit	Tägliche Fahrzeit des Schülers zur Schule: "<15min"; "15-30min"; "30-60min"; ">60min"	
Lernzeit	Zeit die der Schüler außerhalb des Unterrichts pro Woche zum Lernen aufwendet: "<2h"; "2-5h"; "5-10h"; ">10h"	
Wiederholte Schuljahre	Anzahl der Schuljahre die der Schüler wiederholt hat: 0; 1; 2 oder 3	
Unterstützung Familie	Ob der Schüler durch die Familie Unterstützung bei seinen schulischen Pflichten erhält: "Ja" oder "Nein"	
Nachhilfe	Ob der Schüler Nachhilfeunterricht nimmt: "Ja" oder "Nein"	
Außerunterrichtliche Schulaktivitäten	Nimmt der Schüler an Aktivitäten in der Schule außerhalb des Unterrichts teil: "Ja" oder "Nein"	
Internetzugang	Ob Zuhause ein Internetzugang für den Schüler nutzbar ist: "Ja" oder "Nein"	
familiäre Beziehungen	Als wie gut schätzt der Schüler seine familiären Beziehungen ein: "sehr schlecht"; "schlecht"; "mittelmäßig"; "gut"; "sehr gut"	
Freizeit	Menge an Freizeit: "sehr wenig"; "wenig"; "mittel"; "viel"; "sehr viel"	
Weggehen	Wie wichtig ist es dem Schüler; mit Freunden wegzugehen: "sehr wichtig"; "wichtig"; "mittel"; "unwichtig"; "sehr unwichtig"	
Gesundheit	Die Gesundheit des Schülers: "sehr schlecht"; "schlecht"; "mittelmäßig"; "gut"; "sehr gut"	
Abwesenheiten	Wie oft war der Schüler im aktuellen Schuljahr abwesend vom Unterricht: Zahlenwert	
Punkte 1	Punkte im ersten Test: 0 bis 20	
Punkte 2	Punkte im zweiten Test: 0 bis 20	
Punkte 3	Punkte im dritten Test (zu schätzen): 0 bis 20	

Nachdem die Schüler anhand der Aufgabe ihre Vermutungen geäußert und den Datensatz kennengelernt haben, kann nun eine Analyse am PC durchgeführt werden. Zur Reduzierung des Zeitaufwandes bietet es sich an, ein vorgefertigtes Analyseschema an die Hand zu geben, auf dem die Schüler aufbauen können. Eine erste Aufgabe kann dann die Arbeit mit dem Klassifikationsbaum sein:

Aufgabe 2

Nun werden wir das Ganze am Computer ausprobieren. Starte dazu am Computer das Programm „Orange3“ und lade das Projekt „Schulnoten“. Ein Teil der Analyse ist dort bereits vorbereitet:



Das Programm lädt also die Datei mit den Schülerdaten (*File*). Daraus wird ein kleiner Anteil der Daten (Standard: 30 %) ausgewählt (*Data Sampler*), der sozusagen den „korrigierten Arbeiten“ entspricht. Aus dieser Stichprobe werden automatisch Regeln gesucht und als Klassifikationsbaum gespeichert (*Tree*). Damit dieser angezeigt werden kann, wird er an den *Tree Viewer* übergeben.

Lasse dir nun den Klassifikationsbaum mit Hilfe des *Tree Viewer* anzeigen. Der Baum sieht etwas komplizierter aus, als der im letzten Arbeitsblatt. Kannst du Unterschiede zu den von dir erwarteten Attributen feststellen? Erscheinen die Kriterien, nach denen ein Schüler eine bestimmte Note bekommt, für dich logisch und sinnvoll?

Hinweis: Der Baum sieht bei dir möglicherweise anders aus als bei deinem Nachbarn. Das liegt daran, dass die 30 % der Schülerdaten auf jedem Computer getrennt zufällig ausgewählt werden. Du kannst auch bei dir eine neue Stichprobe auswählen, indem du im Data Sampler den Befehl „Sample Data“ nutzt. Es wird dann automatisch auch ein neuer Baum erzeugt.

Anhand der Aufgabe lernen die Schüler das verwendete Programm grundsätzlich kennen und können den Aufbau der Analyse sowie den entstehenden Klassifikationsbaum verstehen, was eine wichtige Grundlage für das weitere Vorgehen bildet, in dem die Schüler die Analyse nun erweitert und schlussendlich bewerten sollen.

Daher steht als nächstes die Verwendung des generierten Modells zur Vorhersage der Noten aller Schüler an:

Arbeitsblatt 5 (L): Datenanalyse am Computer (Teil 2)

Aufgabe 1

Natürlich wollen wir den Klassifikationsbaum verwenden um die Punkte der Schüler automatisch vorhersagen zu können. Dies kannst du machen, indem du von links das *Prediction*-Symbol nach rechts ziehst. Diese Funktion benötigt zwei Eingaben: Den *Baum*, anhand dessen es die Vorhersagen treffen soll, sowie die *Daten*, zu denen es etwas vorhersagen soll. Ziehe daher eine Verbindung vom Halbkreis rechts neben dem *Tree* (dieser Halbkreis entspricht dem Ausgang/Rückgabewert dieser Funktion) zum Eingang der *Prediction*-Funktion sowie vom Ausgang des *File* zum Eingang der *Prediction*.

Um nun anzuzeigen, welche Vorhersagen Orange3 getroffen hat, können wir auf die *Prediction* doppelklicken. Du siehst dann eine Tabelle, die wie folgt aussieht:

Tree		Note 3	Schule	Geschlecht	Alter	Wohnumfeld	Familiengroesse
1	4.0	4.0	GP	W	18.0	urban	>3
2	4.0	4.0	GP	W	17.0	urban	>3
		3.0	GP	W	15.0	urban	<=3
		3.0		W	15.0	urban	>3
5	3.0	3.0		W	16.0	urban	>3
6	3.0	3.0	GP	M	16.0	urban	<=3
7	3.0	3.0	GP	M	16.0	urban	<=3

Wie sieht es aus - war deine Vorhersage perfekt? Wie gut würdest du sie in Schulnoten einschätzen (ankreuzen)?

① — ② — ③ — ④ — ⑤ — ⑥

Wenn du die echten Punktzahlen und die vorhergesagten vergleichst: Wie stark ist die maximale Abweichung, die du findest?

Bonusfrage: Es wurde als Daten an dieser Stelle wieder das File verwendet, nicht wie vorher der Data Sampler. Warum wäre es hier sinnlos, den Data Sampler als Eingabe zu nehmen?

Da die Auswertung anhand der Tabelle relativ mühsam und inakkurat ist, bietet es sich an, den Schülern noch eine Möglichkeit zu präsentieren, mit der das Ganze systematischer stattfinden kann. Es bietet sich daher an, eine sog. Confusion Matrix als Hilfsmittel zu verwenden. Diese zweidimensionale Matrix hat als eine Dimension den eigentlichen Wert, als andere den vorhergesagten. Somit erlaubt sie es, einen Einblick in die Validität der Vorhersage zu gewinnen und Ausreißer und deren Ausmaß schnell und einfach zu erkennen.

Aufgabe 2

Wenn wir unsere Analyse beurteilen wollen, ist es sehr aufwändig, nur die Tabelle anzusehen. Stattdessen können wir eine *Confusion Matrix* nutzen, die uns zeigt, wie „verwirrt“ die Analyse war. Diese kannst du (nachdem du das Symbol von der Liste links in den Arbeitsbereich rechts gezogen hast) direkt mit der Prediction verbinden. Wenn du die Confusion Matrix doppelklickst zeigt sie dir eine Tabelle, an der links die echten Punktzahlen stehen und oben die vorhergesagten Punktzahlen. In der Tabelle steht für jede dieser Kombinationen, wie viele Noten dort einsortiert wurden:

- Markiere im Diagramm die perfekten Schätzungen. Wo findest du diese?
Auf der Diagonalen von links oben nach rechts unten
- Bei wie vielen Schülern hat die Analyse richtig geschätzt?
Abweichend je nach Schueler wegen Sampling
- Bei wie vielen Schülern war die Vorhersage nur wenig falsch, d. h. bei wie vielen hat sie sich maximal um zwei Punkte verschätzt?
Abweichend je nach Schueler wegen Sampling

Nachdem die Schüler jetzt wahrscheinlich gesehen haben, dass die Analyse keinesfalls als perfekt angenommen werden kann, ist es sinnvoll, sich zu überlegen, wie diese verbessert werden kann. Dazu wird den Schülern eine wichtige Stellschraube eröffnet: die Größe des für die Erzeugung des Analysemodells genutzten Datensatzes, d.h. der gewählten Stichprobe. Um möglichst gute Ergebnisse zu erreichen müsste diese natürlich möglichst groß sein (idealerweise 100%). Das ist aber oft nicht sinnvoll, da noch ein Teil der Daten benötigt wird, um das Analysemodell zu testen.

Aufgabe 3

Wir können die Analyse noch etwas verbessern. Dazu kann die Anzahl der Schüler, die für die Erstellung des Baums verwendet werden, angepasst werden. Doppelklicke dazu auf den Data Sampler und ändere die Prozentzahl der Schülerdaten ab.

- Die Analyse verbessert sich. . .
 - ☐ beim Erhöhen der Samplegröße
 - ☐ beim Verringern der Samplegröße
- Mit welchem Prozentsatz der Schülerdaten wird die Analyse am besten?
Mit möglichst vielen Daten, d. h. 100 %
- Ergibt es Sinn, diesen Prozentsatz an Daten für die Erstellung des Modells zu nutzen? Was wären dabei mögliche Probleme?
Nein, da dafür dann (im konkreten Beispiel) alle Klausuren korrigiert sein müssen, was den Sinn der Vorhersage zerstört.
- Würdest du dich dabei wohl fühlen, wenn deine Lehrerin diese Möglichkeit nutzt, um deine Arbeiten zu bewerten?
☐ Ja ☐ Nein
- Falls es deiner Lehrerin gelingen würde, die Qualität der Analyse zu steigern, sodass nur noch wenige Schülerinnen bzw. Schüler falsch (besser oder schlechter) bewertet werden, wäre das dann eine ausreichend faire Lösung für dich?
☐ Ja ☐ Nein

Als letzter Schritt bei der automatisierten Analyse bietet es sich an, zu diskutieren, was passiert, wenn wir gesamt nur eine Schulklasse mit 30 Schülern ansehen. Als Lehrerdemo kann daher schnell das Modell umgebaut werden, sodass statt der gesamten 600 Schüler nur eine Stichprobe von 30 Schülern ausgewählt wird (von denen wiederum nur ein kleiner Anteil zur Erstellung des Modells genutzt wird). Es zeigt sich damit für die Schüler noch stärker, dass große Datenmengen sinnvoll sind, wenn Vorhersagen getroffen werden sollen, während kleine Datenmengen teils enorm fehlerträchtige Analyseergebnisse nach sich ziehen. Dies zeigt, warum jegliche datenbasierte Geschäftsmodelle darauf angewiesen sind, viele Daten über ihre Kunden zu sammeln. An dieser Stelle bietet sich ggf. auch, je nach Vorwissen der Schülerinnen und Schüler, ein Vergleich mit der Wahrscheinlichkeitsrechnung bzw. dem Gesetz der großen Zahlen an.

Arbeitsblatt 6 (L): Diskussion der Ergebnisse

Als letzter Teil der Unterrichtssequenz sollte eine Diskussion der Bedeutung des Gelernten stehen. Es ist an dieser Stelle wichtig, dass die Schüler sich folgende Aspekte bezüglich Datenanalysen bewusstmachen:

- Mit zunehmender Anzahl an Datensätzen wird eine Vorhersage typischerweise genauer.
- An manchen Stellen sind selbst kleinste Fehler bei der Vorhersage unerwünscht, während selbst viele Fehler an anderen Stellen tolerierbar sind.
- Dadurch, dass wir die Regeln auf denen die Vorhersagen basieren, zum Teil nicht logisch nachvollziehen können, erscheinen uns diese Vorhersagen oft als gefährlich.
- Selbst wenn wir vermuten, dass wir von Datenanalysen nur profitieren können, kann es sein, dass wir (warum auch immer) anders eingestuft werden - es sollte also kritisch hinterfragt werden, wie wir zu konkreten Nutzungsszenarien stehen.

Zu diesem Zweck bietet sich ein Gruppenpuzzle an, das wie folgt aufgebaut wird:

Teil I: Auftrag an die Stammgruppen:

Euch stehen fünf Beispiele aus dem Bereich zu Datenanalysen zur Verfügung, die ihr im Rahmen dieser Unterrichtsphase kennenlernen sollt.

Wählt euch jede/-r genau eines der fünf Beispiele aus, mit dem ihr euch im Folgenden etwas intensiver beschäftigen wollt. Geht dann in die Expertengruppen, in denen alle zusammenkommen, die sich mit demselben Beispiel beschäftigen.

Beispiele

- Analyse von Kreditkartendaten/-nutzung:
<http://bit.ly/2FpcRZx> – <http://bit.ly/2HcxEMP> – <http://bit.ly/2FrBzZg>
- Datenanalysen und Smart Cars:
<http://bit.ly/2Fvso9Y> – <http://bit.ly/2oKN5oS> – <http://bit.ly/2oJa9EE>
- Datenanalysen durch KFZ-Versicherungen:
<http://bit.ly/2Fw4ag4> – <http://bit.ly/2FrrKKD> – <http://bit.ly/2Fg5wvV>
- Beurteilung von Personen anhand von Datenanalysen: <http://bit.ly/2ssGJcA> – <http://bit.ly/2I3MxT4> – <http://bit.ly/2FqDm0u>
- Datenanalysen im Smart Home:
<http://bit.ly/2FYpqbM> – <http://bit.ly/2uBQHd0> – <http://bit.ly/2m0eaCs>

Teil II: Auftrag an die Expertengruppen

Ihr erhaltet zu einem Kontext von Datenanalysen und -vorhersagen ein Beispiel wie dieses real oder fiktiv eingesetzt werden könnte.

Versucht dieses Beispiel gemeinsam nachzuvollziehen. Besprecht dazu miteinander folgende Fragen und macht euch dazu Notizen. Bereitet euch darauf vor, eure Ergebnisse in den Stammgruppen kurz vorzustellen und zu diskutieren.

- Welche Daten werden genutzt?

- Wie werden die Daten analysiert?

- Was ist das Ziel dieser Analyse?

- Ist das Beispiel überhaupt praktisch machbar? Warum bzw. warum nicht?

- Sind mögliche Fehler bei der Datenanalyse tolerierbar oder nicht?

- Sehr ihr eine solche Analyse als hilfreich/sinnvoll/nützlich an oder eher als gefährlich? Sollte es zulässig sein, diese Art der Analyse zu nutzen?

- Würdet ihr eure Daten für diesen Zweck freiwillig hergeben?

- Wofür könnten die Daten zukünftig - wenn sie schon einmal da sind - noch genutzt werden?

Teil III: Auftrag an die Stammgruppen:

Nachdem ihr nun in den Expertengruppen die Beispiele diskutiert habt, sollt ihr diese nun gemeinsam vergleichen. Dazu habt ihr in eurer Gruppe nun einen Experten für jedes der fünf Beispiele. Damit alle über die Beispiele Bescheid wissen, erklärt ihr diese einander kurz. Geht dabei insbesondere auf die in den Expertengruppen diskutierten Aspekte ein.

Diskutiert nun die Gemeinsamkeiten und Unterschiede der Beispiele: Handelt es sich bei den Analysen um von euch gewünschte und sinnvolle Arten der Datennutzung? Wenn ja, welche Vorteile hat das für euch? Wenn nein, wie könnt ihr diesen möglicherweise entfliehen und verhindern, dass eure Daten so genutzt werden?

Schülerversionen der Arbeitsblätter

Arbeitsblatt 1: Logik oder scharfes Hinsehen?

Im Unterricht hast du bereits einen Artikel darüber gesehen, wie Daten heute im Einzelhandel verwendet werden, um Kunden auf sie zugeschnittene Werbung präsentieren zu können. Onlineshops gehen heute jedoch schon weiter und versuchen, ihren Kunden viele Produkte möglichst schnell liefern zu können:

Noch bevor ein Kunde überhaupt den Button "Kaufen" anklickt, soll die für ihn passende Ware schon auf dem Weg in Richtung seiner Wohnung sein. Dem Versandhändler Amazon wurde ein Patent zugesprochen, das einen „vorausschauenden Versand“ („anticipatory shipping“) ermöglichen soll. Das heißt: Bestimmte Waren werden schon einmal an ein Versandzentrum geschickt, in dessen Nähe sich ein oder mehrere Kunden höchstwahrscheinlich für das Produkt interessieren. Wird es dann schließlich bestellt, ist es umso schneller beim Empfänger.

— Spiegel Online, 18.01.2014

Aufgabe 1

Um herauszufinden, was ein Kunde als nächstes bestellen könnte, müssen die Versandhändler umfangreiche Daten über ihre Kunden sammeln und analysieren.

a) Was wissen Onlinehändler über ihre Kunden? Woher haben diese die jeweilige Information?

Information über den Kunden	Quelle

b) Wahrscheinlich sind nicht alle Informationen, die ein Onlinehändler über seine Kunden hat auch wichtig, wenn er herausfinden möchte, welchen Artikel der Kunde als nächstes bestellen könnte. Markiere in der Tabelle oben die Zeilen, von denen du denkst, dass sie für diese Zweck wichtig sind, indem du ein + neben die wichtigen Zeilen machst.

Aufgabe 2

Fülle folgenden Lückentext aus: Es gibt bei der Datenanalyse zwei Möglichkeiten, wie wir Vorhersagen treffen können:

1. Wenn wir bereits etwas über die zu analysierenden Daten wissen, dann können wir uns erklären wie etwas funktioniert und damit _____ ziehen. Es gibt also logische Zusammenhänge, sog. _____, die wir zur Vorhersage nutzen können.

Beispiel:

WENN ein Kunde in den letzten 5 Einkäufen Chips gekauft hat, DANN wird er auch beim nächsten Mal welche kaufen.

2. In anderen Bereichen erkennen wir keinerlei logische Zusammenhänge. Stattdessen können wir nach _____ in den Daten suchen. Diese liefern uns auch Zusammenhänge, wir können sie uns aber oft nicht erklären. Solche Zusammenhänge bezeichnen wir als _____ Zusammenhänge.

Beispiel:

WENN ein Kunde den Artikel X gekauft hat und er in Y wohnt und mindestens 35 Jahre alt ist, DANN wird er auch Z kaufen.

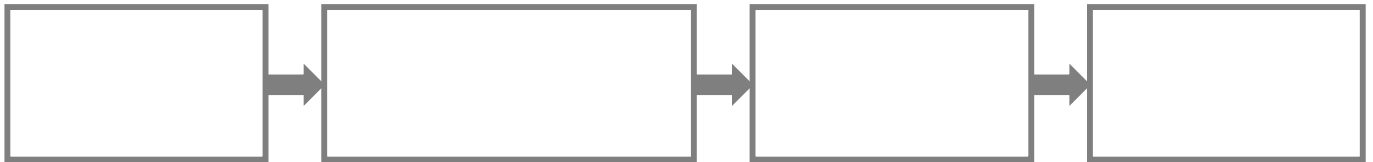
_____ helfen uns zwar dabei Dinge zu verstehen, sie sind aber für Datenanalysen oft relativ wenig interessant: Sie sind oft _____ und _____, sodass sie nur wenig neue Informationen hervorbringen. Wir können uns aber logisch erklären, dass sie _____ sind. Die korrelativen Zusammenhänge sind daher oft spannender, da sie neue Informationen eröffnen. Sie haben aber den Nachteil, dass sie nicht unbedingt _____ sind: Wie genau Wohnort und Alter das Kaufverhalten prägen, können wir uns meist nicht logisch erklären. Außerdem müssen wir sie erst finden, was relativ _____ ist.

Arbeitsblatt 2: Händische Datenanalyse (Teil I)

Aufgabe 1

Im Unterricht wurde bereits gemeinsam erarbeitet, wie eine Datenanalyse abläuft. Vervollständige den folgenden Lückentext und das folgende Ablaufmodell:

Als erster Schritt der Datenanalyse, müssen die Daten _____ und gespeichert werden. Aus diesen Daten wählt man sich üblicherweise eine _____ Teilmenge aus, um aus dieser das _____ zu erstellen, d. h. um Regeln zu finden, die die Vorhersage der gesuchten Eigenschaften ermöglichen. Diese Regeln können dann genutzt werden, um die _____ zu erstellen. Als letzter Schritt jeder Datenanalyse sollte die _____ der Ergebnisse erfolgen, mit dem Ziel eine möglichst gute _____ der Ergebnisse sicherzustellen.



Aufgabe 2

Ein Onlineshop hat über seine Kunden verschiedene Daten gesammelt und möchte nun seine Kunden durch den Versand von individuellen Gutscheinen zu weiteren Käufen anregen. Dazu will er herauszufinden, welche Produktkategorie für jeden Kunden jeweils besonders interessant ist.

Der Shop hat bereits folgende Daten über jeden seiner Kunden gesammelt: Alter, Familienstand, Anzahl der Kinder, präferierte Zahlungsart, Kategorien der letzten vier eingekauften Produkte (Film, Sport, Software, Elektronik, Kleidung, Musik, Bücher oder Auto).

Um jedem Kunden einen Gutschein zu schicken, den dieser wahrscheinlich einlöst, möchte der Onlineshop herausfinden, welche Kategorie für den Käufer besonders interessant ist. Welche WENN-DANN-Regeln vermutest du, die dem Onlineshop dabei helfen könnten? *Hinweis: natürlich kannst du mehrere Bedingungen mit „und“ verknüpfen, z. B. „Kategorie 1 = Elektronik und Anzahl Kinder = 0“.*

- WENN _____
DANN ist der nächste Kauf aus der Kategorie _____
- WENN _____
DANN ist der nächste Kauf aus der Kategorie _____
- WENN _____
DANN ist der nächste Kauf aus der Kategorie _____

Aufgabe 3

Nachdem die Datenwissenschaftler des Unternehmens erkannt haben, dass keine stichhaltigen Zusammenhänge in den bisher vorliegenden Informationen erkennbar sind, wurde entschieden, es anders zu versuchen: Der Onlineshop hat daher einige seiner Kunden befragt, was für sie als nächstes interessant ist. Dabei sind folgende Daten herausgekommen. Welche Zusammenhänge erkennt du in der unten dargestellten Tabelle?

Beispiel:

WENN Kauf 1 ein Film ist und mit Girokarte bezahlt wurde, DANN interessiert der Kunde sich als nächstes für Artikel der Kategorie Auto.

Kurz: „Kauf1“=„Film“ und „bezahlt mit“=„Giro“ \Rightarrow „Interesse“=„Auto“

Alter	Kauf 1	Kauf 2	Kauf 3	Kauf 4	verheiratet	Kinder	bezahlt mit	Interesse
25-50	Film	Software	Film	Sport	Ja	1	Giro	Auto
25-50	Elektronik	Musik	Film	Software	Nein	1	VISA	Bücher
<18	Film	Elektronik	Sport	Sport	Nein	0	Giro	Auto
<18	Film	Musik	Kleidung	Sport	Nein	0	Giro	Auto
18-25	Bücher	Musik	Film	Haushalt	Nein	1	Giro	Bücher
<18	Bücher	Film	Film	Bücher	Nein	0	VISA	Bücher
25-50	Film	Film	Sport	Sport	Ja	1	Giro	Auto
25-50	Musik	Film	Film	Spielzeug	Nein	1	Giro	Bücher
25-50	Musik	Musik	Film	Haushalt	Nein	1	Giro	Bücher
25-50	Elektronik	Musik	Bücher	Software	Ja	1	Master	Elektronik
25-50	Software	Elektronik	Film	Spielzeug	Nein	1	Master	Bücher
25-50	Film	Film	Sport	Sport	Ja	0	Master	Elektronik
25-50	Musik	Elektronik	Bücher	Elektronik	Ja	1	Master	Elektronik

• WENN _____

DANN ist der nächste Kauf aus der Kategorie _____

• WENN _____

DANN ist der nächste Kauf aus der Kategorie _____

• WENN _____

DANN ist der nächste Kauf aus der Kategorie _____

Arbeitsblatt 3: Händische Datenanalyse (Teil 2)

Aufgabe 1

Wenn der Onlineshop nun Vorhersagen treffen will, dann sortiert er die Kunden in verschiedene Kategorien ein - dies nennt man „Klassifikation“.

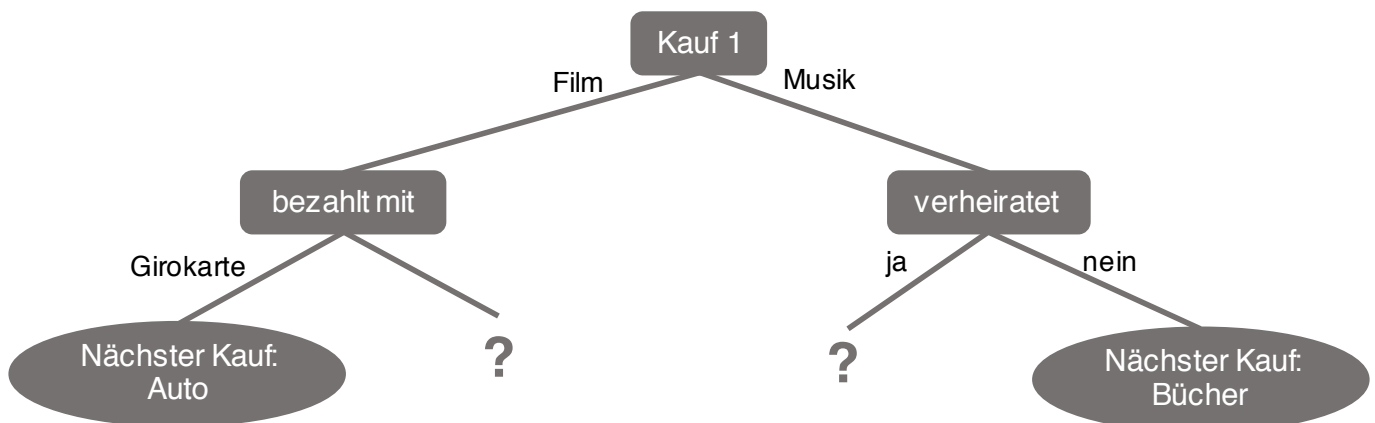
Um eine größere Menge an Regeln einfacher durchschaubar und anwendbar zu machen, stellt man diese als „Klassifikationsbaum“ dar. Dieser Baum symbolisiert die Entscheidungen, die anhand der Regeln getroffen werden.

Beispiel: Die Regeln

1. „Kauf 1“ = „Film“ und „bezahlt mit“ = „Giro“ \Rightarrow nächster Kauf = „Auto“
2. „Kauf 1“ = „Musik“ und nicht verheiratet \Rightarrow nächster Kauf = „Bücher“

können als Klassifikationsbaum wie unten abgebildet dargestellt werden.

- Markiere im Baum den Weg, den du gehen musst, wenn du herausfinden möchtest, was ein Kunde als nächstes gekauft hat, der als „Kauf 1“ einen Film gekauft und mit einer Girokarte bezahlt hat.
- Wir kennen zusätzlich die folgende Regel: *Kauf 1 = „Film“ und „bezahlt mit“ ist „Mastercard“ und „verheiratet“ = „ja“ \Rightarrow nächster Kauf = „Elektronik“*
Um diese im Baum zu berücksichtigen, musst du eine weitere Entscheidung ergänzen. Überlege dir, wo das sinnvoll ist und ergänze die Entscheidung.
- Überprüfe deine Ergänzung, indem du den Weg farbig markierst, den du durch den Baum gehen musst, um herauszufinden, was ein Kunde als nächstes kauft, der als „Kauf 1“ einen Film gekauft hat und mit „Mastercard“ bezahlt hat.



Aufgabe 2

Verwende den vorherigen Klassifikationsbaum, um zu entscheiden, an welcher Produktkategorie die folgenden Kunden wahrscheinlich als nächstes interessiert sind. Wenn diese Entscheidung anhand der beiden Regeln bzw. anhand des Baums nicht getroffen werden kann, schreibe ein ? in das Feld „vsl. interessiert an“.

Alter	Kauf 1	Kauf 2	Kauf 3	Kauf 4	verheiratet	Kinder	bezahlt mit	vsl. interessiert an
25-50	Film	Film	Sport	Sport	Ja	1	Giro	
25-50	Musik	Elektronik	Sport	Sport	Nein	1	Giro	
>50	Film	Musik	Kleidung	Sport	Nein	1	Giro	
<18	Film	Musik	Film	Haushalt	Ja	1	Master	
>50	Bücher	Software	Film	Sport	Ja	1	VISA	

Arbeitsblatt 4: Datenanalyse am Computer

Es wäre für eure Lehrerin sicherlich eine sehr praktische Sache, die Idee der Onlineshops zu nutzen, um eure Schulnoten vorherzusagen: Dann würde es ausreichen jedes Mal nur ein paar Arbeiten zu korrigieren und die Noten aller anderen „vorherzusagen“. Doch wie (gut) funktioniert das wirklich?

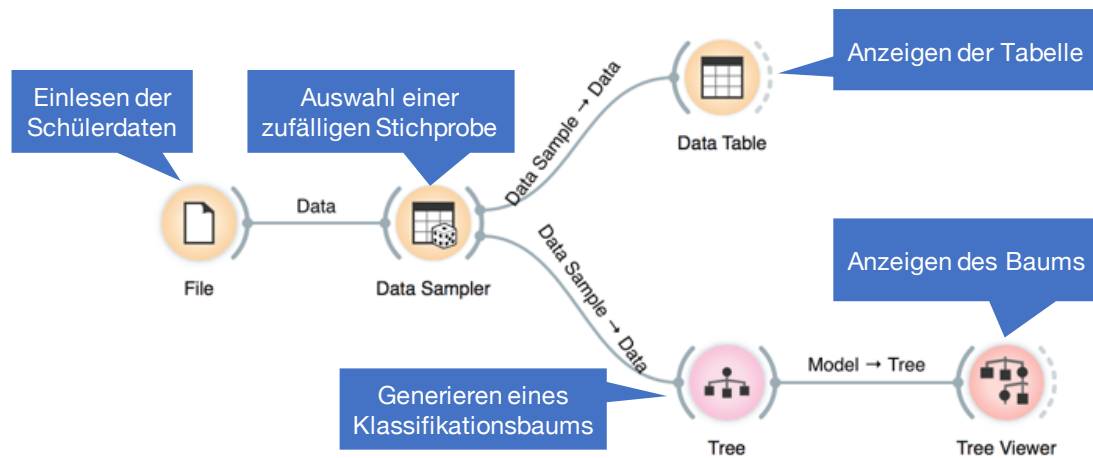
Aufgabe 1

Zunächst ohne Computer: Von welchen der in der Tabelle unten dargestellten Eigenschaften der Schülerinnen/Schüler vermutest du, dass sie für das „Erraten“ bzw. „Berechnen“ der Endnoten wichtig sind? Markiere diese in der folgenden Tabelle.

Attribut	Beschreibung	relevant?
Schule	Kürzel der Schule: „GP“ oder „MS“	
Geschlecht	„M“ oder „W“	
Alter	Zahlenwert	
Wohnumfeld	„urban“ oder „ländlich“	
Familiengröße	„≤3“ oder „>3“	
Bildung Mutter	Grundschule; Hauptschule; Realschule/Gymnasium; Universität; keine	
Bildung Vater	vgl. Bildung Mutter	
Beruf Mutter	Gesundheitsbranche; Lehrerin; Hausfrau; Öffentlicher Dienst; sonstige	
Beruf Vater	vgl. Beruf Mutter	
Fahrzeit	Tägliche Fahrzeit des Schülers zur Schule: „<15min“; „15-30min“; „30-60min“; „>60min“	
Lernzeit	Zeit die der Schüler außerhalb des Unterrichts pro Woche zum Lernen aufwendet: „<2h“; „2-5h“; „5-10h“; „>10h“	
Wiederholte Schuljahre	Anzahl der Schuljahre die der Schüler wiederholt hat: 0; 1; 2 oder 3	
Unterstützung Familie	Ob der Schüler durch die Familie Unterstützung bei seinen schulischen Pflichten erhält: „Ja“ oder „Nein“	
Nachhilfe	Ob der Schüler Nachhilfeunterricht nimmt: „Ja“ oder „Nein“	
Außerunterrichtliche Schulaktivitäten	Nimmt der Schüler an Aktivitäten in der Schule außerhalb des Unterrichts teil: „Ja“ oder „Nein“	
Internetzugang	Ob Zuhause ein Internetzugang für den Schüler nutzbar ist: „Ja“ oder „Nein“	
familiäre Beziehungen	Als wie gut schätzt der Schüler seine familiären Beziehungen ein: „sehr schlecht“; „schlecht“; „mittelmaessig“; „gut“; „sehr gut“	
Freizeit	Menge an Freizeit: „sehr wenig“; „wenig“; „mittel“; „viel“; „sehr viel“	
Weggehen	Wie wichtig ist es dem Schüler; mit Freunden wegzugehen: „sehr wichtig“; „wichtig“; „mittel“; „unwichtig“; „sehr unwichtig“	
Gesundheit	Die Gesundheit des Schülers: „sehr schlecht“; „schlecht“; „mittelmaessig“; „gut“; „sehr gut“	
Abwesenheiten	Wie oft war der Schüler im aktuellen Schuljahr abwesend vom Unterricht: Zahlenwert	
Punkte 1	Punkte im ersten Test: 0 bis 20	
Punkte 2	Punkte im zweiten Test: 0 bis 20	
Punkte 3	Punkte im dritten Test (zu schätzen): 0 bis 20	

Aufgabe 2

Nun werden wir das Ganze am Computer ausprobieren. Starte dazu am Computer das Programm „Orange3“ und lade das Projekt „Schulnoten“. Ein Teil der Analyse ist dort bereits vorbereitet:



Das Programm lädt also die Datei mit den Schülerdaten (*File*). Daraus wird ein kleiner Anteil der Daten (Standard: 30 %) ausgewählt (*Data Sampler*), der sozusagen den „korrigierten Arbeiten“ entspricht. Aus dieser Stichprobe werden automatisch Regeln gesucht und als Klassifikationsbaum gespeichert (*Tree*). Damit dieser angezeigt werden kann, wird er an den *Tree Viewer* übergeben.

Lasse dir nun den Klassifikationsbaum mit Hilfe des *Tree Viewer* anzeigen. Der Baum sieht etwas komplizierter aus, als der im letzten Arbeitsblatt. Kannst du Unterschiede zu den von dir erwarteten Attributen feststellen? Erscheinen die Kriterien, nach denen ein Schüler eine bestimmte Note bekommt, für dich logisch und sinnvoll?

Hinweis: Der Baum sieht bei dir möglicherweise anders aus als bei deinem Nachbarn. Das liegt daran, dass die 30 % der Schülerdaten auf jedem Computer getrennt zufällig ausgewählt werden. Du kannst auch bei dir eine neue Stichprobe auswählen, indem du im Data Sampler den Befehl „Sample Data“ nutzt. Es wird dann automatisch auch ein neuer Baum erzeugt.

Arbeitsblatt 5: Datenanalyse am Computer (Teil 2)

Aufgabe 1

Natürlich wollen wir den Klassifikationsbaum verwenden um die Punkte der Schüler automatisch vorhersagen zu können. Dies kannst du machen, indem du von links das *Prediction*-Symbol nach rechts ziehst. Diese Funktion benötigt zwei Eingaben: Den *Baum*, anhand dessen es die Vorhersagen treffen soll, sowie die *Daten*, zu denen es etwas vorhersagen soll. Ziehe daher eine Verbindung vom Halbkreis rechts neben dem *Tree* (dieser Halbkreis entspricht dem Ausgang/Rückgabewert dieser Funktion) zum Eingang der *Prediction*-Funktion sowie vom Ausgang des *File* zum Eingang der *Prediction*.

Um nun anzuzeigen, welche Vorhersagen Orange3 getroffen hat, können wir auf die *Prediction* doppelklicken. Du siehst dann eine Tabelle, die wie folgt aussieht:

Tree		Note 3	Schule	Geschlecht	Alter	Wohnumfeld	Familiengroesse
1	4.0	4.0	GP	W	18.0	urban	>3
2	4.0	4.0	GP	W	17.0	urban	>3
		3.0	GP	W	15.0	urban	<=3
		3.0		W	15.0	urban	>3
5	3.0	3.0		W	16.0	urban	>3
6	3.0	3.0	GP	M	16.0	urban	<=3
7	3.0	3.0	GP	M	16.0	urban	<=3

Wie sieht es aus - war deine Vorhersage perfekt? Wie gut würdest du sie in Schulnoten einschätzen (ankreuzen)?

① — ② — ③ — ④ — ⑤ — ⑥

Wenn du die echten Punktzahlen und die vorhergesagten vergleichst: Wie stark ist die maximale Abweichung, die du findest?

Bonusfrage: Es wurde als Daten an dieser Stelle wieder das File verwendet, nicht wie vorher der Data Sampler. Warum wäre es hier sinnlos, den Data Sampler als Eingabe zu nehmen?

Aufgabe 2

Wenn wir unsere Analyse beurteilen wollen, ist es sehr aufwändig, nur die Tabelle anzusehen. Stattdessen können wir eine *Confusion Matrix* nutzen, die uns zeigt, wie „verwirrt“ die Analyse war. Diese kannst du (nachdem du das Symbol von der Liste links in den Arbeitsbereich rechts gezogen hast) direkt mit der *Prediction* verbinden. Wenn du die *Confusion Matrix* doppelklickst zeigt sie dir eine Tabelle, an der links die echten Punktzahlen stehen und oben die vorhergesagten Punktzahlen. In der Tabelle steht für jede dieser Kombinationen, wie viele Noten dort einsortiert wurden:

- Markiere im Diagramm die perfekten Schätzungen. Wo findest du diese?

- Bei wie vielen Schülern hat die Analyse richtig geschätzt?

- Bei wie vielen Schülern war die Vorhersage nur wenig falsch, d. h. bei wie vielen hat sie sich maximal um zwei Punkte verschätzt?
-

Aufgabe 3

Wir können die Analyse noch etwas verbessern. Dazu kann die Anzahl der Schüler, die für die Erstellung des Baums verwendet werden, angepasst werden. Doppelklicke dazu auf den Data Sampler und ändere die Prozentzahl der Schülerdaten ab.

- Die Analyse verbessert sich. . .
 - ☐ beim Erhöhen der Samplegröße
 - ☐ beim Verringern der Samplegröße
- Mit welchem Prozentsatz der Schülerdaten wird die Analyse am besten?

- Ergibt es Sinn, diesen Prozentsatz an Daten für die Erstellung des Modells zu nutzen? Was wären dabei mögliche Probleme?

- Würdest du dich dabei wohl fühlen, wenn deine Lehrerin diese Möglichkeit nutzt, um deine Arbeiten zu bewerten?
 - ☐ Ja ☐ Nein
- Falls es deiner Lehrerin gelingen würde, die Qualität der Analyse zu steigern, sodass nur noch wenige Schülerinnen bzw. Schüler falsch (besser oder schlechter) bewertet werden, wäre das dann eine ausreichend faire Lösung für dich?
 - ☐ Ja ☐ Nein

Arbeitsblatt 6: Diskussion der Ergebnisse

Teil I: Auftrag an die Stammgruppen:

Euch stehen fünf Beispiele aus dem Bereich zu Datenanalysen zur Verfügung, die ihr im Rahmen dieser Unterrichtsphase kennenlernen sollt.

Wählt euch jede/-r genau eines der fünf Beispiele aus, mit dem ihr euch im Folgenden etwas intensiver beschäftigen wollt. Geht dann in die Expertengruppen, in denen alle zusammenkommen, die sich mit demselben Beispiel beschäftigen.

Beispiele

- Analyse von Kreditkartendaten/-nutzung:
<http://bit.ly/2FpcRZx> – <http://bit.ly/2HcxEMP> – <http://bit.ly/2FrBzZg>
- Datenanalysen und Smart Cars:
<http://bit.ly/2Fvso9Y> – <http://bit.ly/2oKN5oS> – <http://bit.ly/2oJa9EE>
- Datenanalysen durch KFZ-Versicherungen:
<http://bit.ly/2Fw4ag4> – <http://bit.ly/2FrrKKD> – <http://bit.ly/2Fg5wvV>
- Beurteilung von Personen anhand von Datenanalysen: <http://bit.ly/2ssGJcA> – <http://bit.ly/2I3MxT4> – <http://bit.ly/2FqDm0u>
- Datenanalysen im Smart Home:
<http://bit.ly/2FYpqbM> – <http://bit.ly/2uBQHd0> – <http://bit.ly/2m0eaCs>

Teil II: Auftrag an die Expertengruppen

Ihr erhaltet zu einem Kontext von Datenanalysen und -vorhersagen ein Beispiel wie dieses real oder fiktiv eingesetzt werden könnte.

Versucht dieses Beispiel gemeinsam nachzuvollziehen. Besprecht dazu miteinander folgende Fragen und macht euch dazu Notizen. Bereitet euch darauf vor, eure Ergebnisse in den Stammgruppen kurz vorzustellen und zu diskutieren.

- Welche Daten werden genutzt?

- Wie werden die Daten analysiert?

- Was ist das Ziel dieser Analyse?

- Ist das Beispiel überhaupt praktisch machbar? Warum bzw. warum nicht?

-
-
- Sind mögliche Fehler bei der Datenanalyse tolerierbar oder nicht?

-
-
- Sehr ihr eine solche Analyse als hilfreich/sinnvoll/nützlich an oder eher als gefährlich? Sollte es zulässig sein, diese Art der Analyse zu nutzen?

-
-
- Würdet ihr eure Daten für diesen Zweck freiwillig hergeben?

-
-
- Wofür könnten die Daten zukünftig - wenn sie schon einmal da sind - noch genutzt werden?
-
-

Teil III: Auftrag an die Stammgruppen:

Nachdem ihr nun in den Expertengruppen die Beispiele diskutiert habt, sollt ihr diese nun gemeinsam vergleichen. Dazu habt ihr in eurer Gruppe nun einen Experten für jedes der fünf Beispiele.

Damit alle über die Beispiele Bescheid wissen, erklärt ihr diese einander kurz. Geht dabei insbesondere auf die in den Expertengruppen diskutierten Aspekte ein.

Diskutiert nun die Gemeinsamkeiten und Unterschiede der Beispiele: Handelt es sich bei den Analysen um von euch gewünschte und sinnvolle Arten der Datennutzung? Wenn ja, welche Vorteile hat das für euch? Wenn nein, wie könnt ihr diesen möglicherweise entfliehen und verhindern, dass eure Daten so genutzt werden?