Datenflussorientierte Big-Data-Analyse

mit *Orange 3*

von Andreas Grillenberger

Die Erfassung und Verarbeitung von Daten von und über uns ist heute allgegenwärtig, sodass die Menschheit mittlerweile über riesige Datenmengen verfügt, die zum Teil öffentlich und frei verfügbar sind. Entsprechende Datenquellen sind auch für die Betrachtung von Big Data und Data Science im Informatikunterricht spannend, denn durch geeignete Wahl der Daten kann eine Betroffenheit bei den Schülerinnen und Schülern sichergestellt und damit eine hohe Motivation erreicht werden.

Neben der Auswahl von sinnstiftenden Datensätzen stellt jedoch die Wahl eines geeigneten Werkzeugs für den Informatikunterricht eine wichtige Entscheidung der Lehrperson dar, die auch von den Vorkenntnissen der Lernenden abhängt: Soll die Datenanalyse eher spielerisch erfolgen (z. B. mit SNAP!) oder eher durch professionelle Programmierung (z. B. mit PYTHON und entsprechenden Bibliotheken)? Einen interessanten Mittelweg stellt die Arbeit mit Werkzeugen dar, die zwar für die professionelle Nutzung, aber für Nicht-Informatiker entworfen worden sind: Solche Werkzeuge kombinieren oft Anteile des visuellen Zugangs mit den umfassenden Möglichkeiten professioneller Programmierung.

Aus didaktischer Sicht scheinen dabei insbesondere Werkzeuge spannend, deren Zugang auf einer datenflussorientierten Modellierung des Analyseprozesses basieren: Durch Modellierung des Datenflusses zwischen Eingabe-, Berechnungs- und Ausgabeknoten kann so relativ einfach eine erste Analyse selbst gestaltet werden. Durch Diskussion und Optimierung der Analysequalität können dabei, ohne dass tiefgehende Programmierkenntnisse vorhanden sein müssen, ein Einblick in die Herausforderungen und Möglichkeiten der Datenanalyse gewonnen und verschiedene Parameter, die solche Modelle beeinflussen, kennengelernt werden.

In diesem Beitrag wird daher eine kurze Unterrichtssequenz vorgestellt, deren Ziel es ist, einen ersten Einblick in die Big-Data-Analyse, ihre Grundlagen und ihre Möglichkeiten zu bieten und die beliebig – je nach Zielen der Lehrperson und des Unterrichts – erweitert werden kann.

Überblick über die Unterrichtsreihe

Für den im Folgenden beschriebenen Unterrichtsverlauf wurden vier Doppelstunden vorgesehen, durch Anpassungen ist jedoch auch eine kürzere oder längere Dauer möglich, da an verschiedenen Stellen eine deutlich vertieftere oder etwas oberflächlichere Betrachtung möglich ist. Die Unterrichtssequenz wurde so gestaltet, dass auf keinerlei Vorwissen aufgebaut werden muss, sodass diese flexibel in den Unterricht integrierbar ist. Das Thema ist sowohl für die Sekundarstufe I als auch für die Sekundarstufe II geeignet. In der Unterrichtssequenz werden insbesondere folgende Kompetenzen angestrebt, die sich in das im Beitrag *Big Data aus Perspektive der Informatikdidaktik* in diesem Heft dargestellte Kompetenzmodell einordnen lassen (siehe Seite 18 ff.):

Die Schülerinnen und Schüler …

- ▷ erläutern, warum und wie aus gespeicherten Daten verschiedene und ggf. neue Informationen gewonnen werden können (C1/P3).
- ▷ charakterisieren den Unterschied zwischen korrelations- und kausalitätsbasierten Zusammenhängen in Daten sowie der jeweiligen Aussagekraft (C1/P3, z. T. C4/P3).
- ▷ skizzieren den Ablauf einer (korrelationsbasierten) Datenanalyse (C3/P3).
- ▷ charakterisieren eine typische Analysemethode und erläutern das zugrunde liegende Prinzip an einem geeigneten Beispiel (C3/P3).
- ▷ führen einfache Datenanalysen unter Nutzung einer üblichen Methode durch, und zwar manuell sowie unter Nutzung eines geeigneten Softwarewerkzeugs (C3/P3).
- ▷ prognostizieren fehlende Attribute eines Datensatzes unter Rückgriff auf eine selbst durchgeführte Datenanalyse (C3/P3).
- ▷ bewerten das Ergebnis der Vorhersage und erläutern Ideen zur Verbesserung (C3/P3).

▷ reflektieren die Ergebnisse unter Einbeziehung ethischer und gesellschaftlicher Gesichtspunkte (C4/P3).

Die vier Doppelstunden wurden folgenden Themen gewidmet:

1. *Grundzüge des Analyseprozesses & Motivation:*
   Anhand eines Zeitungsartikels wird versucht, das Interesse der Schülerinnen und Schüler am Thema *Datenanalyse* zu wecken, und eine Diskussion darüber anzustoßen, wie diese funktionieren könnten. Am Beispiel werden daraufhin grundlegende Begriffe wie *Kausalität* und *Korrelation* eingeführt und ein Ablaufmodell der Datenanalyse erstellt.
2. *Der Weg von den Daten*
   und der Fragestellung zur Prognose:
   Basierend auf einem fiktiven Datensatz wird nachvollzogen, wie eine gegebene Fragestellung durch Analyse von Daten, Erzeugung eines Modells und darauf basierend der Prognose eines fehlenden Attributs des Datensatzes stattfinden kann. Dabei wird insbesondere der binäre Entscheidungsbaum in Form eines Klassifikationsbaums eingeführt.
3. *Nutzung eines echten Datensatzes –*
   Prognose von Schulnoten:
   Während zum Einstieg bislang fiktive Datensätze verwendet wurden, soll nun die Mächtigkeit von Datenanalysen thematisiert werden. Dazu kann ein frei verfügbarer Datensatz von Schülerdaten genutzt werden, um die Fragestellung zu beantworten, ob und wie gut aus den vorliegenden Daten eine Schulnote der Schülerinnen und Schüler vorhergesagt werden kann. Aufgrund des relativ guten Analyseergebnisses und der Betroffenheit der Schülerinnen und Schüler durch das Thema kann eine Diskussion der ethischen Aspekte dieser Analyse stattfinden und Probleme wie Stigmatisierung, die allgemein bei Datenanalysen vorkommen können, durch die Schülerinnen und Schüler selbst erkannt werden.
4. *Übertragung auf weitere Kontexte:*
   Für die letzte Doppelstunde wurde geplant, die bisher erworbenen Kompetenzen auf weitere Kontexte zu übertragen und somit beispielsweise die Datennutzung in der Medizin, durch Versicherungen und Banken zu hinterfragen und im Rahmen eines Gruppenpuzzles rechtliche, ethische und moralische Aspekte dieser Analysen zu diskutieren.

Im Folgenden wird der Fokus auf die dritte Unterrichtsdoppelstunde gelegt, die sich direkt mit der Analyse der Daten befasst, da diese den größten Einarbeitungsaufwand für die Lehrperson bedeutet. Alle Informationen zu den anderen Stunden können dem am Ende des Beitrags aufgeführten Unterrichtskonzept in den Internetquellen entnommen werden (vgl. Grillenberger, 2020). Dieses Konzept und alle zugehörigen Dateien stehen kostenfrei zur Verfügung.

Werkzeugauswahl: Das Data-Mining-Tool *Orange*

Um die Mächtigkeit und das Potenzial automatisierter Datenanalysen im Unterricht für die Schülerinnen und Schüler erlebbar zu machen, wurde ein Werkzeug für den Unterricht gesucht, das eine intuitive Nutzung ohne detailliertes Vorwissen (weder im Bereich Datenanalyse noch in der Programmierung) erlaubt. Entsprechend war für diesen Kontext die Verwendung beispielsweise einer klassischen Programmiersprache wie PYTHON nicht möglich. Besonders spannend für diesen Zweck schienen jedoch grafisch orientierte Analysewerkzeuge, bei denen die Anwender die Analyse als Datenflussmodell beschreiben und die alle für den Schulunterricht notwendigen Funktionalitäten bereitstellen. Ein bekannter Vertreter dieser Werkzeuge ist das an einer slowenischen Universität entwickelte und für die Nutzung durch Nicht-Informatiker, ursprünglich insbesondere im Bereich der Biologie, konzipierte und frei unter Open-Source-Lizenz verfügbare Werkzeug *Orange* (vgl. Orange, 1996 ff.; siehe auch Bild 1).

Bild 1: Beispielhafte Analyse in *Orange 3*.

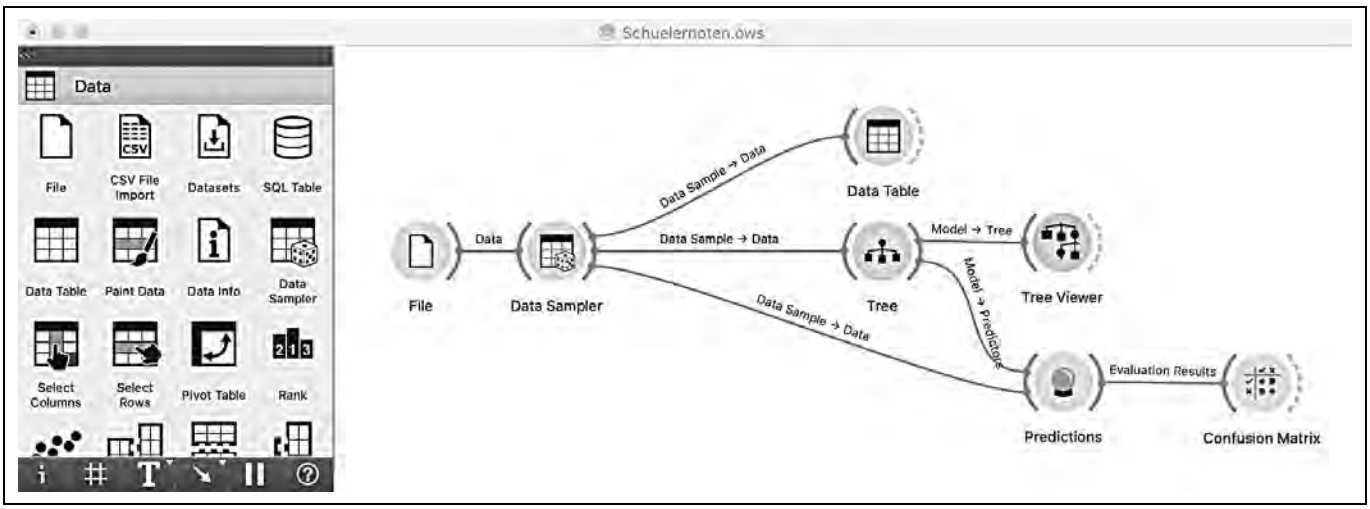


Bild 2: Analyseprozess.

Auswahl des Datensatzes: Klassifikationsaufgaben als einfacher Einstieg

Neben dem für den Informatikunterricht geeigneten Werkzeug ist eine wichtige Entscheidung für einen erfolgreichen Unterricht zum Thema *Big Data* die Auswahl geeigneter Datensätze, die das Interesse der Schülerinnen und Schüler wecken, die jeweils zu zeigenden Eigenschaften von Big Data deutlich klar machen können und gleichzeitig mit den zur Verfügung stehenden Mitteln analysierbar bleiben.

Trotz des oft erkennbaren Fokus von Big Data auf enorme Größe der Datensätze, die im Schulkontext nur mit hohem Aufwand analysierbar wären, können die Grundzüge von Big-Data-Analysen auch mit kleineren Datensätzen deutlich werden, da die hohe Datenmenge insbesondere zur Genauigkeit der Analyse beiträgt. Anhand von kleineren Datensätzen können damit zwar nur weniger valide Aussagen abgeleitet werden, die aber bei geeigneter Wahl des Datensatzes trotzdem spannend bleiben.

Für den Einstieg in die Big-Data-Analyse empfehlen sich dabei einfach zu verstehende Klassifikationsaufgaben: Hier ist das Ziel, in einem Datensatz ein Attribut „vorherzusagen", indem anhand der weiteren Attribute Klassen gebildet werden, die sich hinsichtlich des gesuchten Attributs gleich verhalten. Es handelt sich dabei üblicherweise um ein Attribut, das entweder aufwendig zu ermitteln ist oder das oft erst im Rückblick bestimmt werden kann – es muss jedoch zur Erzeugung eines Klassifikationsmodells ein gewisser Anteil an Daten vorhanden sein, bei denen dieses Attribut bekannt ist. Diese werden dann als Trainingsdaten zur Erzeugung des Modells genutzt, das dann auf weitere Daten angewandt werden kann.

Datensätze für solche Klassifikationsaufgaben stehen in verschiedenen Datenportalen bereits vorgefertigt zur Verfügung, sodass völlig verschiedene Interessen bedient werden können: Beispielsweise existieren Datensätze aus dem Bereich der Gesundheit (Erkennen der Gutartigkeit oder Bösartigkeit von Krebszellen aus deren Form), der Lebensmittelqualität (Erkennen der Weinqualität anhand verschiedener chemischer Faktoren), der Finanzwelt (Vorhersage des Haushaltseinkommens anhand Wohnumfeld u. Ä.) und vieles mehr. Entsprechende Datensätze sind beispielsweise im *Machine Learning Repository* der University of California, Irvine (vgl. UCI, 1987 ff.) oder bei den *Datasets* von Kaggle (vgl. Kaggle, 2010 ff.) auffindbar.

Im hier vorgestellten Unterrichtsbeispiel wurde ein Datensatz aus dem genannten *Machine Learning Repository* gewählt: Ein Datensatz mit Daten über ca. 600 portugiesischer Schülerinnen und Schüler, der unter anderem Informationen über Alter, Wohnumfeld, Bildungsniveau und Berufe der Eltern, schulische und außerschulische Aktivitäten sowie je drei Noten der Schülerinnen und Schüler enthält. Der Datensatz wurde für die deutschen Schülerinnen und Schüler minimal angepasst, indem Attribute übersetzt und die Noten im Datensatz in deutsche Schulnoten umgerechnet wurden. Der vorbereitete Datensatz ist in der zip-Datei des bereits erwähnten Konzepts enthalten (vgl. Grillenberger, 2020). Ziel der Analyse im Unterricht war dann, die dritte Note aus den ersten beiden Noten sowie aus allen anderen Informationen vorherzusagen. Um eine direkte Betroffenheit zu betonen und entsprechende Diskussionen anzuregen, wurde die Analyse dann im Unterricht als „fairere, objektivere und schnellere Methode der Benotung" vorgestellt.

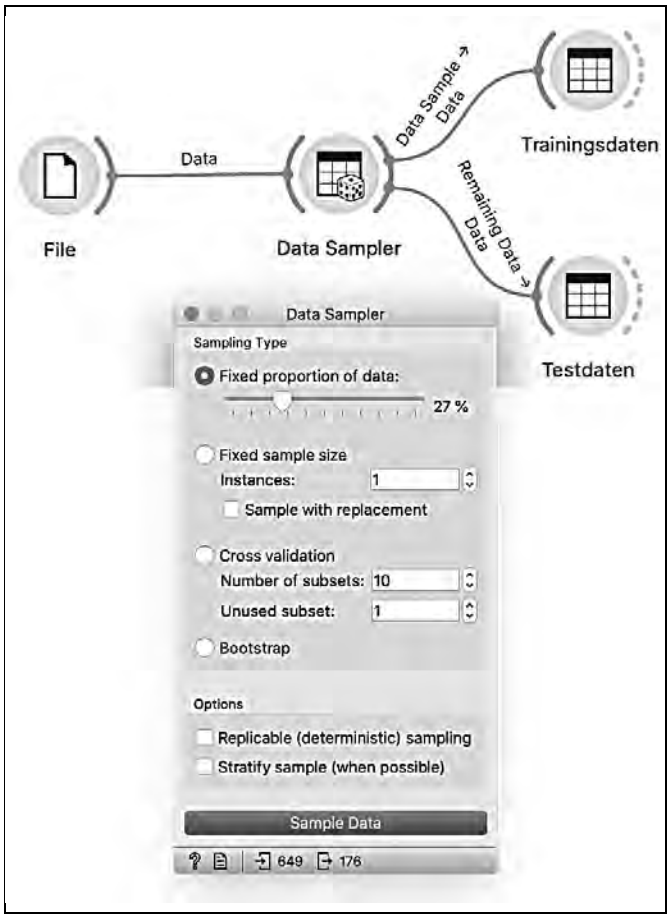


Bild 3: Nutzung des Data Samplers zur Selektion von Trainings- und Testdaten.

Analyse der Daten mit *Orange 3*

Die Datenanalyse mit *Orange 3* mit dem Ziel, die dritte Note der Schülerinnen und Schüler vorherzusagen, folgt dem in Bild 2 (vorige Seite) dargestellten Ablauf.

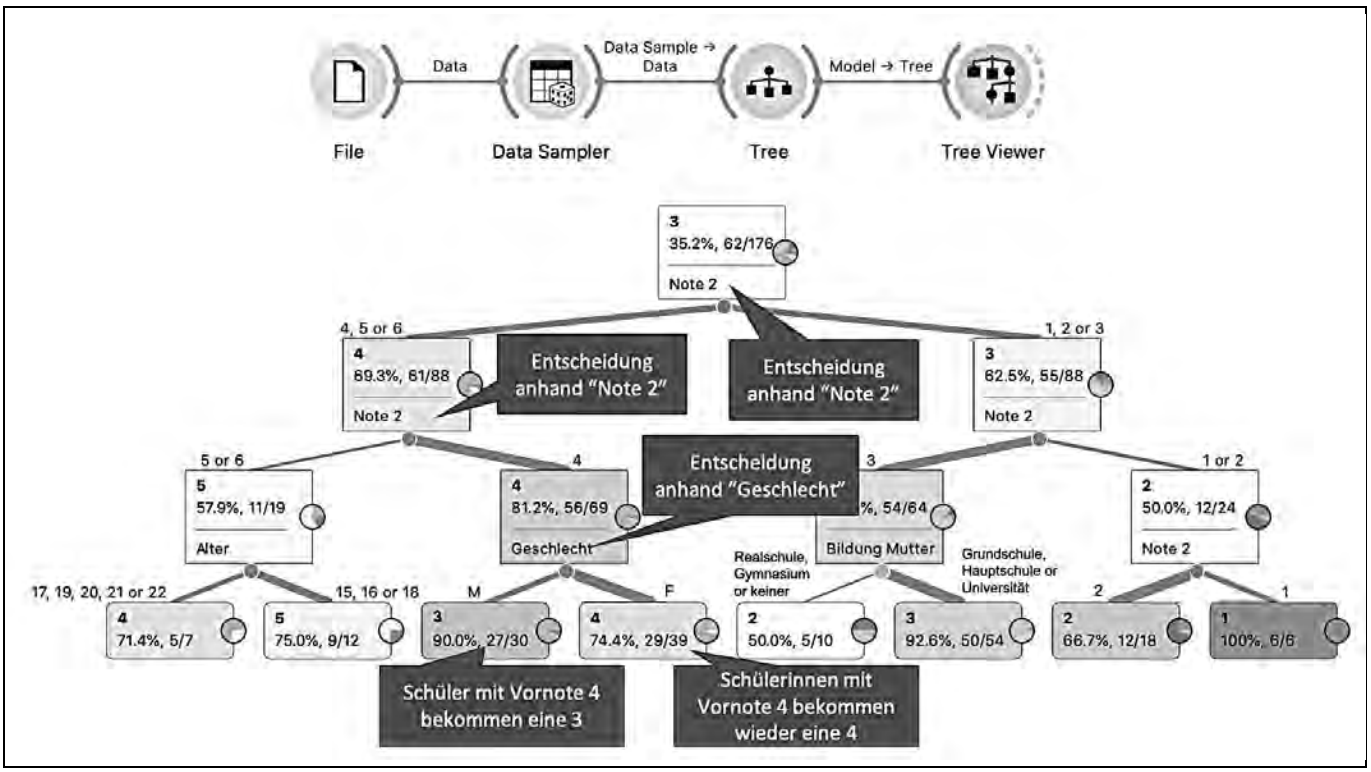
Da der Datensatz nicht bereits in Trainings- und Testdaten aufgeteilt war, wurde aus dem kompletten Datensatz ein gewisser Anteil an Daten (30 bis 50 Prozent, anpassbar) ausgewählt, die als Trainingsdaten fungierten; der Rest wurde als Testdaten genutzt. Diese Aufgabe wird durch die Komponente „Data Sampler" erledigt, die als Eingang den gesamten Datensatz bekommt und diesen entsprechend der Einstellungen aufteilt (siehe Bild 3, vorige Seite).

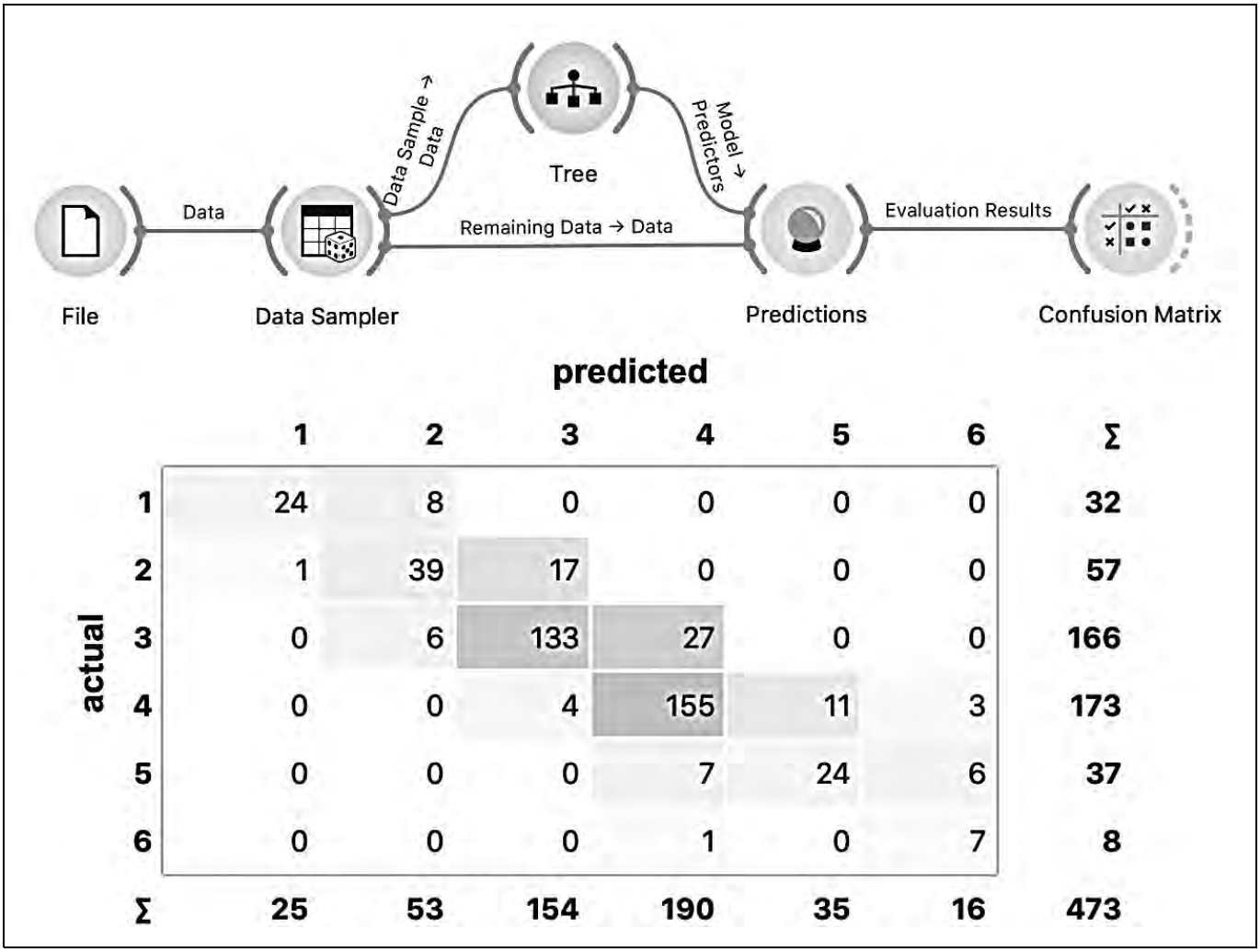
Der Trainingsdatensatz wird zur Erstellung eines Klassifikationsbaums durch die Komponente „Tree" genutzt. Die Gestalt dieses Baums kann durch verschiedene Einstellungen verändert werden, beispielsweise indem ein binärer Baum erzwungen wird, wodurch die Breite des Baums limitiert wird, aber die Tiefe steigt, oder stattdessen die Tiefe limitiert wird, was einerseits die Übersichtlichkeit steigern, aber auch zur Vermeidung einer Überanpassung an die Trainingsdaten („overfitting") beitragen kann. Der durch die Tree-Komponente automatisch erstellte Baum kann durch den „Tree Viewer" betrachtet werden (siehe Bild 4). Der Baum zeigt dabei die Entscheidungen, anhand derer ein Datensatz einer bestimmten Klasse zugeordnet (und entsprechend die dritte Note abgeleitet) wird. Dabei fallen zum Teil Entscheidungen, die ethisch sicherlich fragwürdig sind und eine ideale Diskussionsgrundlage bilden – beispielsweise wird manchmal (je nach zufällig gewählten Trainingsdaten) das Geschlecht miteinbezogen.

Das nun in Form des Klassifikationsbaums vorliegende Vorhersagemodell kann jetzt durch die „Prediction"-Komponente einfach auf die Testdaten – also hier alle Nicht-Trainingsdaten – angewandt werden. Die entsprechenden Ergebnisse können (auch mit Vergleich mit den eigentlichen Noten, die ja in diesem Fall zuvor bekannt waren) durch Doppelklick auf die Prediction (deutsch = Vorhersage) tabellarisch eingesehen oder zur Analyse durch eine weitere Komponente betrachtet werden: Die Konfusionsmatrix („confusion matrix") stellt die jeweils durch die Vorhersage zugeordneten Noten den eigentlichen Noten der Schülerinnen und Schüler aus dem Testdatensatz gegenüber und hilft daher bei der Untersuchung der Analysequalität. Dabei zeigt sie im vorliegenden Fall, insbesondere bei Berücksichtigung der geringen Größe des Datensatzes (649 Tupel) und des noch kleineren Trainingsdatensatzes (27 % des Datensatzes), eine erstaunlich gute Analysequalität: In Bild 5 (nächste Seite) wurden von den 473 Schülerinnen und Schülern aus dem Testdatensatz ca. 80 Prozent korrekt vorhergesagt; diese liegen also auf der Hauptdiagonalen der Konfusionsmatrix. Nur bei vier Personen wäre eine Abweichung von mehr als einer Note eingetreten. Wenn statt der sehr groben Notenskala eine detailliertere Punkteskala verwendet wird, kann die Qualität der Prognose noch deutlich gesteigert werden.

Nachdem die Analyse im Unterricht durchgeführt wurde, bietet sich an, verschiedene Parameter der Analyse und deren Einfluss auf die Qualität des Ergebnisses zu untersuchen, insbesondere eine Limitierung des

Bild 4:
Beispiel eines kommentierten Klassifikationsbaums.





predicted

actual	1	2	3	4	5	6	Σ
1	24	8	0	0	0	0	32
2	1	39	17	0	0	0	57
3	0	6	133	27	0	0	166
4	0	0	4	155	11	3	173
5	0	0	0	7	24	6	37
6	0	0	0	1	0	7	8
Σ	25	53	154	190	35	16	473

Klassifikationsbaums (binär, Tiefe) oder die Größe des Trainingsdatensatzes.

Bild 5: Beispiel einer Konfusionsmatrix (erstellt mit 27 % Trainingsdaten-Anteil, binärem Klassifikationsbaum mit maximaler Tiefe 4).

Erfahrungen aus dem Unterricht

Das Unterrichtskonzept wurde bereits – meist in adaptierter Form – durch verschiedene Lehrpersonen an verschiedenen Schulen, Schularten (Realschule und Gymnasium) und in verschiedenen Klassenstufen (ab 9. Klasse) eingesetzt. Dabei wurden größtenteils positive Erfahrungen berichtet, da diese jedoch nicht systematisch erfasst wurden und daher hier nur kurz anekdotisch berichtet werden können. Insgesamt zeigte sich bei allen Durchführungen die Angemessenheit des Themas und der Aufbereitung für den Unterricht. Es zeigte sich dabei insbesondere, dass – nach einem zuerst eher gering erscheinenden Interesse am Thema – die Motivation der Schülerinnen und Schüler schon mit den ersten Beispielen (Zeitungsartikel zum Thema) schnell anstieg und Diskussionen nicht nur darüber entstanden, ob der Artikel so korrekt sein kann, sondern auch darüber, ob gewisse Analysen überhaupt aus ethischen Gründen stattfinden sollten oder nicht. Nach der eher trockenen Einführung in Grundzüge der Analyse in der zweiten Doppelstunde, die je nach Klasse noch etwas motivierender ausgebaut werden sollte, konnte ein starker Interessenzuwuchs in der dritten Doppelstunde bei der Durchführung der oben skizzierten Analyse festgestellt werden: Insbesondere das Thema der Notenvorhersage war für die Schülerinnen und Schüler wichtig; es schien gleichzeitig für sie real zu sein, dass man Schulnoten zukünftig vielleicht so berechnen könnte – und somit wurden auch wichtige Aspekte durch sie selbstständig erkannt, die Big-Data-Analysen oft vorgeworfen werden, wie beispielsweise:

- Stigmatisierung („Ich kann mich dann ja vielleicht gar nicht mehr verbessern, wenn ich zweimal schlecht war“).
- Unabhängigkeit des Resultats von der eigenen Leistung („Dann muss ich mich ja gar nicht mehr anstrengen!“ bis hin zu „Was passiert denn, wenn sich niemand mehr anstrengt?“).

- ▷ Angemessenheit des Trainingsdatensatzes („Wenn nur die schlechten in dem Trainingsdatensatz sind, werden wir dann alle schlecht benotet?“).
- ▷ Einbeziehung sachfremder Attribute („Was hat mein Wohnort mit meinen Noten zu tun?“).

Alles in allem konnte im Allgemeinen das Ziel der Unterrichtsreihe erreicht werden: Die Schülerinnen und Schüler bekamen einen ersten Einblick in die Big-Data-Analyse, konnten diese kritisch hinterfragen und deren Probleme verstehen. Weiterhin konnten, dem Eindruck der Lehrpersonen nach, das Funktionsprinzip und der Analyseprozess klar von den Schülerinnen und Schülern nachvollzogen und die Bedeutung einer guten Datenquelle von diesen erkannt werden, was sich insbesondere in den Diskussionen immer wieder zeigte. Gleichzeitig wurde aber auch deutlich, dass viel Potenzial für eine weitere Vertiefung vorhanden ist, beispielsweise indem weitere Datensätze untersucht und somit andere Kontexte mit einbezogen werden, weitere Analysemethoden thematisiert werden, ein Blick hinter die Kulissen des verwendeten Klassifikationsalgorithmus ermöglicht wird, das Thema der Kausalität im Vergleich zur Korrelation vertiefter betrachtet wird oder auch eine Verknüpfung mit weiteren verwandten Themen wie dem Maschinenlernen stattfindet (siehe auch den nachfolgenden Beitrag *AI Replugged* von Lennard Kerber, Seite 67 ff. in diesem Heft). Aus Sicht des Autors zeigte sich damit, dass das Thema den Unterricht stark bereichern kann – selbst wenn ggf. nur relativ beschränkte Unterrichtszeit dafür zur Verfügung steht. Das vorliegende Unterrichtskonzept bietet genau für diesen Fall eine erste Idee, die sich aber individuell gut anpassen lässt – sei es durch Wahl anderer Datensätze, Beispiele oder Werkzeuge oder durch entsprechende Vertiefung je nach Vorwissen der Lernenden oder der angestrebten Unterrichtsziele.

Dr. Andreas Grillenberger
Freie Universität Berlin
Didaktik der Informatik
Königin-Luise-Straße 24–26
14195 Berlin

E-Mail: andreas.grillenberger@fu-berlin.de

Internetquellen

Grillenberger, A. (unter Mitwirkung von A.-K. Jäger): Datenanalyse und Vorhersage mit Klassifikationsbäumen – Ein Unterrichtskonzept für die Sekundarstufe II. 2020.
https://dataliteracy.education/Unterrichtskonzept-Orange.zip

Kaggle – Your Home for Data Science: Datasets. 2010 ff.
https://www.kaggle.com/datasets

Orange – Data Mining. 1996 ff.
https://orange.biolab.si/

UCI – University of California, Irvine: Machine Learning Repository. 1987 ff.
http://archive.ics.uci.edu/ml/index.php

Alle Internetquellen wurden zuletzt am 18. Februar 2021 geprüft und können auch aus dem Service-Bereich des LOG IN Verlags (https://www.log-in-verlag.de/) heruntergeladen werden.