Yannik FLEISCHER, Paderborn & Rolf BIEHLER, Paderborn

Automatisierte Entscheidungsverfahren als Thema im allgemeinbildenden Mathematikunterricht

Einleitung

Mathematik ist Teil unserer Welt und zugleich in ihr verborgen (Heymann, 1996). In der heutigen Lebenswelt ist solch verborgene Mathematik konstitutiv für zahlreiche Anwendungen von automatisierten Entscheidungsverfahren. Auf Onlineplattformen werden Inhalte oder Werbung automatisiert vorgeschlagen, Ärzte werden automatisiert bei Diagnosen unterstützt und mancherorts werden automatisiert juristische Fragestellungen bearbeitet (Dressel & Farid, 2018) oder gar Wahlkämpfe bestritten (Issenberg, 2012). Einher mit diesen Technologien geht ein gesellschaftlicher Diskurs über ihren Einsatz, der bisweilen einerseits von euphorischer Überhöhung und andererseits von ängstlicher Ablehnung so genannter „Frankenstein Algorithmen“ (sueddeutsche.de, 5.9.18) geprägt ist.

Entscheidungsmodelle, die dabei zum Einsatz kommen, werden oft nicht explizit programmiert, sondern durch selbstlernende Algorithmen aus einer Datengrundlage gelernt. Maschinelles Lernen liegt im Schnittbereich von Informatik und Mathematik, deren Anteile dabei gleichermaßen elementar sind (Dhar, 2013). Selbstlernende Algorithmen werden durch effiziente computergestützte Implementationen umgesetzt, deren Lernverfahren und Gütemaßstäbe auf mathematischen Konzepten aus der Statistik, Numerik, linearen Algebra oder Informationstheorie basieren. Mit steigender Relevanz wird vermehrt gefordert, dass datengetriebene Verfahren in der Schulbildung aufgegriffen werden (Engel, 2017; Ridgeway, 2015). Automatisierte Entscheidungsmodelle sind mit Alltagserfahrungen von Schüler*innen verbunden und das Behandeln der zugrundeliegenden Mathematik ermöglicht eine Neubewertung und fundierte Reflexion über deren Möglichkeiten und Grenzen. Dies entspricht dem Motiv der Weltorientierung in allgemeinbildendem Mathematikunterricht (Heymann, 1996) und kann einer mündigen Teilnahme an einem Diskurs über maschinelles Lernen zuträglich sein.

Das Projekt Data Science und Big Data in der Schule (ProDaBi)

Das Projekt ProDaBi (www.prodabi.de) ist ein interdisziplinäres Pilot-Projekt an der Universität Paderborn mit Beteiligung der Didaktiken der Mathematik und der Informatik, das von der Deutsche Telekom Stiftung initiiert und seit 2017 gefördert wird. Ein Ziel des Projektes ist die Entwicklung und Erprobung eines Data Science Curriculums zunächst in der Sekundarstufe II,

Siller, H.-S., Weigel, W. & Wörler, J. F. (Hrsg.). *Beiträge zum Mathematikunterricht 2020.* Münster: WTM-Verlag, 2020.
doi: 10.37626/GA9783959871402.0

dann auch für die Sekundarstufe I in Form eines Bausteinsystems. Das Projekt begann mit der Ausrichtung eines internationalen Symposiums, auf dessen Grundlage eine theoretische Basis für curriculare Ziele und Inhalte entwickelt wurde (Biehler & Schulte 2018). Die praktische Erprobung und Evaluation findet in einem Projektkurs statt, der im Schuljahr 2019/20 zum zweiten Mal in Kooperation mit Paderborner Gymnasien durchgeführt wird. In NRW können Schulen abiturrelevante Projektkurse ohne curriculare Vorgaben anbieten, die in der Oberstufe über ein Schuljahr belegt werden können. Unser Kurs besteht aus drei Modulen. Das erste Modul thematisiert Grundlagen der Statistik und der Datenexploration, ferner Grundkenntnisse der Programmiersprache Python und zugehöriger Bibliotheken für Data Science. Das zweite Modul thematisiert maschinelles Lernen in Form von zwei Verfahren: einerseits Entscheidungsbäume als Vertreter transparenter Modelle und andererseits künstliche neuronale Netze, deren Resultate „Black Boxes" darstellen. Das dritte Modul ist eine Projektarbeit, in der die Schüler*innen ihre bisher erworbenen Kompetenzen bei einer authentischen Fragestellung mit Realdaten anwenden (Vorhersage freier Parkplätze).

Maschinelles Lernen mit Entscheidungsbäumen

Entscheidungsbäume sind Klassifikationsmodelle, die eine Zielvariable aus anderen Variablen vorhersagen (z. B. Vorhersage einer Erkrankung aus medizinischen Merkmalen). Die gestuften Entscheidungsregeln sind in einer gerichteten Baumstruktur darstellbar. Diese können durch maschinelles Lernen aus einer Datengrundlage von Einzelfallbeispielen erzeugt werden. Selbstlernende Algorithmen, wie der ID3 (Quinlan 1986) testen anhand dieser Trainingsdaten, welche der Variablen sich am besten zum Vorhersagen der Zielvariable eignen. Als Gütemaßstab dienen z. B. das statistische Gütemaß der Fehlklassifikationsrate oder die aus der Informationstheorie stammende Entropiefunktion (Topsoe, 1974). Die Variable mit der besten Bewertung wird an die Spitze des Baumes gesetzt und auf gleiche Art werden weitere Variablen für die nächsten Stufen der Baumstruktur ausgewählt, bis der Datensatz perfekt klassifiziert oder keine Variable mehr übrig ist.

Im Rahmen des Projektes ProDaBi haben wir den Unterrichtsbaustein zu maschinellem Lernen mit Entscheidungsbäumen entwickelt. Dazu gehört ein Basisteil ohne Programmieren und ein Vertiefungsteil, der Grundkenntnisse im Programmieren voraussetzt. Im Basisteil wird mit dem in Abb. 1 dargestellten Plug-In des webbasierten Datenanalysetools Codap (Finzer, 2017; Engel et al., 2018) gearbeitet. Mit diesem Tool ist es möglich, zu einem Datensatz manuell einen Entscheidungsbaum zu erstellen und dynamisch die Gütekriterien des Baums zu erfassen. Im Unterricht wird u. A. ein selbst erhobener Datensatz zur Mediennutzung von Jugendlichen genutzt, dessen

Fragebogen sich an der JIM-Studie (Behrens, 2017) orientiert und in didaktisch reduzierter Form die persönlichen Daten zur Mediennutzung von 53 Personen mit je 15 Merkmalen (siehe Abb. 1) enthält. Es soll ein Entscheidungsbaum erstellt werden, der die Ausprägung (häufig/selten) der Zielvariable „Spielen_Onlinespiele“ vorhersagt. Ein Anwendungskontext eines solchen Baums ist z. B. das gezielte Schalten von Werbung in Onlineplattformen. Im Codap Tool können die gelisteten Merkmale per „Drag & Drop“ für die Vorhersage ausgewählt werden. Dann wird automatisch ein Datensplit durchgeführt, der den Datensatz nach den Ausprägungen des Merkmals teilt und die Teildatensätze hinsichtlich der Zielvariable auswertet.

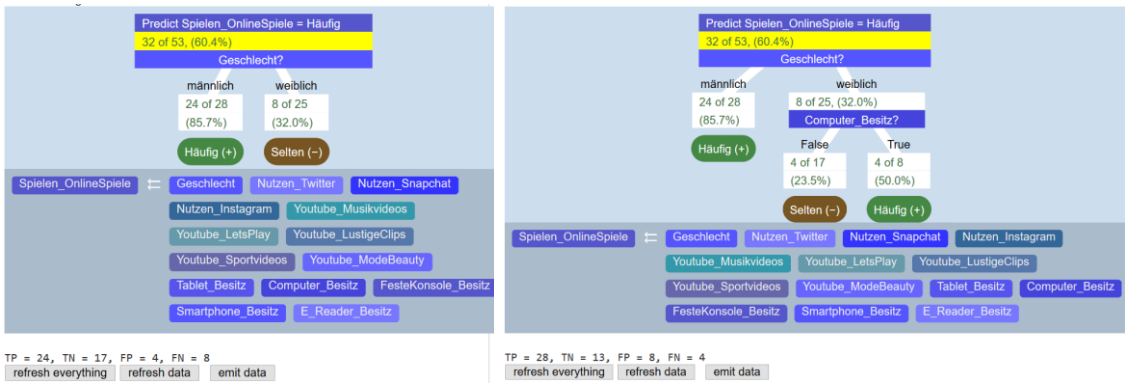

Abb. 1: Codap Entscheidungsbaum Tool

In Abb.1 (links) wurde das Merkmal „Geschlecht“ ausgewählt. Man kann ablesen, dass von 28 männlichen Personen 24 häufig Onlinespiele spielen. Von 25 weiblichen Personen spielen acht häufig. Der resultierende Entscheidungsbaum klassifiziert nach dem Mehrheitsprinzip und prognostiziert für männliche Personen, dass sie häufig Onlinespiele spielen und für weibliche, dass sie selten spielen. Um den Baum zu verbessern, können weitere Merkmale am Ende der Äste angefügt werden (siehe Abb. 1 rechts). Wie gut der aktuelle Baum den Datensatz klassifiziert, wird durch die angezeigten Anzahlen der True Positives (TP), True Negatives (TN), False Positives (FP) und False Negatives (FN) ausgedrückt. Aus diesen Werten können die Fehlklassifikationsrate und weitere statistische Gütekriterien, wie z. B. Sensitivität oder Spezifität berechnet werden. Die Schüler*innen erarbeiten sich mit Hilfe des Tools und gezielter Aufgabenstellungen verschiedene Aspekte zum Bewerten und Vergleichen der Bäume. Damit sind die Grundlagen vorhanden, um einen Entscheidungsbaum systematisch zu erstellen und als zentrales Ergebnis einen Algorithmus als Pseudocode zu formulieren.

Im Vertiefungsteil des Unterrichtsbausteins implementieren die Schüler*innen Teile eines Algorithmus, der aus Datensätzen automatisiert Entscheidungsbäume erstellt. Dafür verwenden wir die Programmiersprache Python in für Anfänger didaktisch optimierten Jupyter Notebooks (Toomey 2017).

Als weiteres leistungsfähiges Gütemaß für Datensplits wird die Entropiefunktion eingeführt. Vertiefend wird auf „overfitting“ eingegangen: die Überanpassung eines Modells an Trainingsdaten, die durch retrospektives beschneiden (pruning) des Modells anhand von Testdaten abgemildert wird. Nach der Implementation haben die Schüler*innen mit unserer eigens für den Unterricht entwickelten Bibliothek ein Werkzeug, das automatisch Modelle erstellt und optimiert. Im Unterricht lieferte der Algorithmus z. B. für die Vorhersage des Geschlechts aus dem Medienverhalten (94 Merkmale) ein Modell, das Personen als männlich klassifiziert, wenn sie häufig Onlinespiele spielen und Sportvideos schauen, und als weiblich, wenn sie selten Onlinespiele spielen und häufig Modevideos schauen. Testdaten wurden damit zu 85 Prozent korrekt klassifiziert. Solch ein Beispiel ermöglicht eine Reflexion über automatisierte Entscheidungsverfahren und einen Diskurs darüber, in welchen Anwendungen sie gesellschaftlich wünschenswert sind.

Ausblick

Im laufenden Projektkurs werden das Thema elementarisierende Unterrichtbausteine zu Entscheidungsbäumen weiterentwickelt und die erreichten Schülerkompetenzen und -einstellungen in einer Begleitstudie untersucht.

Literatur

Biehler, R. & Schulte, C. (2018). Perspectives for an interdisciplinary data science curriculum at German secondary schools. In R. Biehler, L. Budde, D. Frischemeier, B. Heinemann, S. Podworny, C. Schulte & T. Wassong (Hrsg.), *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts* (S. 2-14). Paderborn: Universitätsbibliothek Paderborn.

Behrens, P. & Rathgeb, T. (2017). *JIM-Studie 2017 – Jugend, Information, (Multi-)Media, Basisstudie zum Medienumgang 12- bis 19-jähriger in Deutschland*. Stuttgart.

Dhar, V. (2013): Data science and prediction. *Communications of the ACM* 56.

Dressel J. & Farid, H. (2018). The Accuracy, Fairness, and Limits of Predicting Recidivism. Science Advances.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, 16(1), 44-49.

Engel, J., Erickson, T. & Martignon, L. (2018). Teaching and learning about tree-based methods for exploratory data analysis. *ICOTS* 10.

Finzer, W. (2017). *Common Online Data Analysis Platform*. (www.codap.concord.org).

Issenberg, S. (2012). How President Obama’s campaign used big data to rally individual voters. *MIT Technology Review*.

Quinlan J. R. (1986). Induction of Decision Trees. *Machine Learning*, (1), 81–106.

Toomey, D. (2017). *Jupyter for Data Science – Exploratory analysis, statistical modeling, machine learning, and data visualization with Jupyter.*

Topsoe F. (1974). *Informationstheorie.* B. G. Teubner, Stuttgart.