Yannik FLEISCHER, Paderborn

Ein Unterrichtsmodul für Data Science und maschinelles Lernen mit Entscheidungsbäumen

Einleitung

„All models are wrong, but some are useful" (Box & Draper, 1987, S. 424). Dieser Aphorismus erinnert uns daran, dass selbst ausgefeilte datenbasierte Entscheidungsmodelle, sogenannte künstliche Intelligenz, an viele Stellen nützlich, aber auch begrenzt sind. Da wir tagtäglich mit solchen Entscheidungsmodellen in Kontakt treten ist es eine zentrale Kompetenz, diese kritisch zu beurteilen und ein Gespür dafür zu haben, welche Aufgaben mit ihnen angemessen gelöst werden können. Diese hohe Relevanz für Einzelne sowie die Gesellschaft erfordert eine verstärkte Behandlung von Data Science Inhalten auf Schulebene (Engel, 2017; Ridgway, 2016). Entscheidungsmodelle werden oft nicht explizit programmiert, sondern durch selbstlernende Algorithmen basierend auf Daten trainiert. Das sogenannte maschinelle Lernen (ML) liegt im Schnittbereich von Informatik und Mathematik, deren Anteile gleichermaßen elementar sind (Dhar, 2013). In neueren internationalen Rahmenlehrplänen für Statistik und Data Science wird ML anhand von Entscheidungsbäumen explizit ausfgenommen (Bargagliotti, 2020; IDSSP Curriculum team, 2019). Im Projekt Data Science und Big Data in der Schule (ProDaBi, www.prodabi.de) haben wir eine Unterrichtsmodul konzipiert, damit Schüler*innen ab Klasse 9 die ML-Methode der Entscheidungsbäume verstehen können und begründen, was an bestimmten Modellen falsch ist und warum sie nützlich sein könnten (oder auch nicht).

Unterrichten von maschinellem Lernen und Entscheidungsbäumen

Das in der Einleitung skizzierte Ziel ist nicht leicht zu erreichen. Sulmont et al. (2019) stellten fest, dass es möglich ist Studierende auch ohne fundierte Vorkenntnisse in mathematik und Informatik in ML zu unterrichten. Sie führen aus, dass Hindernisse für Studierende nicht primär das Verständnis von ML-Algorithmen, sondern eher die Evaluation eines Modells betreffen. In Bezug auf die Lehrkräfteausbildung fanden Zieffler et al.(2021) heraus, dass ein Entscheidungsbaum Algorithmus als solcher in einer Lehrkräftefortbildung gut gelehrt werden kann, aber auch hier die Bewertung von Bäumen eine Herausforderung darstellt. Als Problem wird unter anderem festgestellt, dass Lernende ML für nicht ‚accessible' halten (Sulmont et al., 2019). Dies könnte dafür sprechen den Modellerstellungsprozess durch Offenlegung zu demystifizieren. Hitron et al. (2019) arbeiteten mit 10-13-jährigen Schüler*innen und stellten fest, dass auch diese in der Lage waren, grundlegende

In: IDMI-Primar Goethe-Universität Frankfurt (Hrsg.),
Beiträge zum Mathematikunterricht 2022.
*56. Jahrestagung der Gesellschaft für Didaktik der Mathematik.*WTM.
https://doi.org/10.37626/GA9783959872089.0

ML-Konzepte zu verstehen. Allerdings wurde dort ausschließlich mit Black-Box Modellen gearbeitet und Hitron et al. (2019) regen an in künftiger Forschung zu untersuchen, welchen Effekt das Aufdecken zugrundeliegender Prozesse hat. Um dies adäquat zu unterrichten, wird ein Bedarf für mehr geeignete Visualisierungen und Tools formuliert (Sulmont et al., 2019).

Datenbasierte Entscheidungsbäume mit CODAP erstellen

Ein solches Tool für das manuelle Erstellen und Visualisieren von Entscheidungsbäumen ist CODAP (Finzer, 2017) mit einem speziellen Entscheidungsbaum PlugIn, dass kürzlich unter Beteiligung von ProDaBi weiterentwickelt wurde, um damit halbautomatisch Entscheidungsbäume erstellen zu können. Entscheidungsbäume sind Klassifikationsmodelle, die gestufte Entscheidungsregeln in einer gerichteten Baumstruktur dargestellen und die Ausprägung einer Zielvariable basierend auf anderen Variablen vorhersagen. Selbstlernende Algorithmen (Quinlan, 1993) testen anhand dieser Trainingsdaten welche der Variablen sich am besten zum Vorhersagen der Zielvariable eignen. Als Gütemaßstab dienen statistische Gütemaße (z. B. Fehlklassifikationsrate (FKR), Sensitivität). 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. Mit CODAP ist es möglich, zu gegebenen Daten manuell einen Entscheidungsbaum zu erstellen und dynamisch die Gütekriterien des Baums zu erfassen. Im Unterricht wird ein im Jahr 2021 erhobener Datensatz zur Mediennutzung von über 1200 Jugendlichen (160 Merkmale) genutzt (Podworny et al., 2022). Eine daraus gezogene Stichprobe mit 53 Fällen und 15 je binär codierten Merkmalen stellt eine didaktisch reduzierte Form des Datensatzes dar, die zum Einstieg in Entscheidungsbäume genutzt wird (siehe Abb. 1).

cases (53 Fälle)

Index	Spielen OnlineSpiele	Geschlecht	Tablet Besitz	Computer Besitz	FesteKonsole Besitz	Smartphone Besitz	E Reader Besitz	Nutzen Twitter	Nutzen Snapchat	Nutzen Instagram	Youtube Musikvideos	Youtube LetsPlay	Youtube LustigeClips	Youtube Sportvideos	Youtube ModeBeauty
1	Häufig	männlich	Ja	Nein	Ja	Ja	Nein	Selten	Selten	Häufig	Häufig	Häufig	Selten	Selten	Selten
2	Häufig	männlich	Nein	Ja	Nein	Ja	Nein	Selten	Häufig	Häufig	Häufig	Häufig	Häufig	Häufig	Selten
3	Häufig	männlich	Nein	Nein	Ja	Ja	Ja	Häufig	Häufig	Selten	Häufig	Häufig	Häufig	Häufig	Selten
4	Selten	weiblich	Nein	Nein	Nein	Ja	Nein	Selten	Selten	Häufig	Selten	Selten	Selten	Selten	Selten

Abb. 1: Ausschnitt des didaktisch reduzierten JIM-Paderborn (JIM-PB) Datensatzes

Im Unterricht soll ein Entscheidungsbaum erstellt werden, der die Ausprägung (Häufig/Selten) der Zielvariable ‚Spielen_Onlinespiele' vorhersagt. Als Anwendungskontext eines solchen Baums ist wird das gezielte Schalten von Werbung auf Onlineplattformen thematisiert. In CODAP können die gelisteten Merkmale per ‚Drag & Drop' für die Vorhersage ausgewählt werden. Dann wird automatisch ein sogenannter Datensplit durchgeführt, der den Datensatz nach den Ausprägungen des Merkmals teilt und die Teildatensätze hinsichtlich der Zielvariable auswertet.

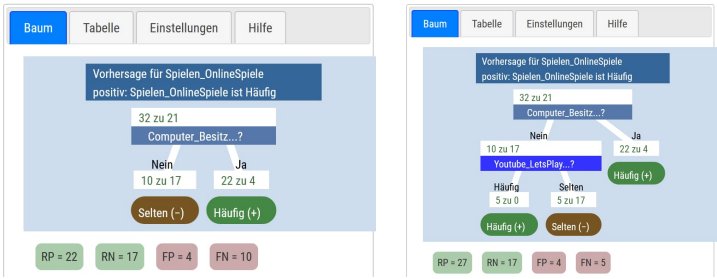

Abb. 2: Darstelllungen von Entscheidungsbäumen in CODAP

In Abb. 2 (links) wurde das Merkmal ‚Computer_Besitz' ausgewählt. Man kann ablesen, dass unter den Computer besitzenden Personen 22 häufig Onlinespiele spielen und vier selten (22 zu 4). Unter denjenigen die keinen Computer besitzen spielen zehn häufig und 17 selten (10 zu 17). Der resultierende Entscheidungsbaum klassifiziert nach dem Mehrheitsprinzip und prognostiziert z. B. für Computer besitzende Personen, dass sie häufig Onlinespiele spielen. Wie gut der aktuelle Baum den Datensatz klassifiziert, wird durch die angezeigten Anzahlen der Richtig Positiven (RP), Richtig Negativen (TN), Falsch Positiven (FP) und Falsch Negativen (FN) ausgedrückt, die auch als sogenannte Confusion Matrix dargestellt werden können (Abb. 3 links). Aus diesen Werten können die FKR und weitere statistische Gütekriterien (z. B. Sensitivität, Spezifität) berechnet werden. Der einstufige Baum liegt bei 14 Fällen falsch (FKR: 0.264). Um den Baum noch zu verbessern, können weitere Merkmale am Ende der Äste angefügt werden. Im zweistufigen Baum in Abb. 2 (rechts) ist in einem Ast das Abbruchkriterium eines ‚reinen' Teildatensatzes (5 zu 0) erreicht.

Abb. 3: Weitere Funktionen des CODAP Entscheidungsbaum PlugIns

Vertiefend lernen die Schüler*innen systematisch Entscheidungsbäume zu erstellen, indem sie ‚Maschine spielen' und einen Entscheidungsbaum auf gleiche Art erstellen wie ein Algorithmus. Eine neue durch ProDaBi mitgestaltete Funktion ermöglicht eine Tabelle anzulegen (Abb. 3 rechts) mit der übersichtlich auszuwerten ist, welches Merkmal (FokusKnoten) man an einem bestimmten Punkt des Baums nutzen muss um die FKR möglichst stark zu reduzieren. Ein weiterer wichtiger Aspekt im Unterrichtsmodul ist das Evaluieren eines erstellten Baums mit Testdaten, d. h. mit Daten die nicht zum Erstellen des Baums genutzt wurden. Auch dies ist in CODAP bequem umsetzbar. Damit werden Fehler des Modells aufgedeckt und es kann über die Einsetzbarkeit des Modells argumentiert werden.

Evaluation und Ausblick

Das Unterrichtsmodul wurde in einer 9 Klasse erprobt und mit einem Fragebogen evaluiert. Die Schüler*innen gaben dort u. A. mehrheitlich an, dass sie Spaß im Umgang mit CODAP hatten und dass Sie verstanden haben, wie man datenbasiert Entscheidungsbäume erstellt.

Abb. 4: Einblick in die Evaluation des Unterrichtsmoduls in einer 9. Klasse

Aktuell wird eine Interviewstudie in einem Oberstufenkurs durchgeführt, bei der im Anschluss an die Unterrichtsreihe das Verständnis der Schüler*innen sowie der Umgang mit dem Tool untersucht werden soll.

Literatur

Bargagliotti, A. (2020). *Pre-K-12 guidelines for assessment and instruction in statistics education II (GAISE II)* (Second edition). American Statistical Association.

Box, G. E. P. & Draper, N. R. (1987). *Empirical model-building and response surface*. Wiley. https://books.google.de/books?id=QO2dDRufJEAC

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73. https://doi.org/10/gdm28r

Engel, J. (2017). Statistical Literacy for Active Citizenship: A Call for Data Science Education. *Statistics Education Research Journal*, *16*(1), 44–49. https://doi.org/10.52041/serj.v16i1.213

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

IDSSP Curriculum team. (2019). *Curriculum Frameworks for Introductory Data Science*. http://idssp.org/files/IDSSP_Frameworks_1.0.pdf

Podworny, S., Fleischer, Y., Stroop, D. & Biehler, R. (2022). *An example of rich, real and multivariate survey data for use in school.* CERME 12, Bozen, Italy.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Ridgway, J. (2016). Implications of the Data Revolution for Statistics Education: The Data Revolution and Statistics Education. *International Statistical Review*, *84*(3), 528–549. https://doi.org/10/f3q6f6

Sulmont, E., Patitsas, E. & Cooperstock, J. R. (2019). What Is Hard about Teaching Machine Learning to Non-Majors? Insights from Classifying Instructors' Learning Goals. *ACM Transactions on Computing Education*, *19*(4), 1–16. https://doi.org/10/ghnpbm

Zieffler, A., Justice, N., delMas, R. & Huberty, M. D. (2021). The Use of Algorithmic Models to Develop Secondary Teachers' Understanding of the Statistical Modeling Process. *Journal of Statistics and Data Science Education*, *29*(1), 1–17. https://doi.org/10.1080/26939169.2021.1900759