ZfDN (2020) 26:157–172
https://doi.org/10.1007/s40573-020-00118-7

ORIGINAL PAPER

Maschinelles Lernen mit Aussagen zur Modellkompetenz

Dirk Krüger[1] · **Moritz Krell**[1]

Eingegangen: 11. Februar 2020 / Angenommen: 12. September 2020 / Online publiziert: 7. Oktober 2020

Zusammenfassung
Verfahren des maschinellen Lernens können dazu beitragen, Aussagen in Aufgaben im offenen Format in großen Stichproben zu analysieren. Am Beispiel von Aussagen von Biologielehrkräften, Biologie-Lehramtsstudierenden und Fachdidaktiker*innen zu den fünf Teilkompetenzen von Modellkompetenz ($N_{\text{Training}} = 456$; $N_{\text{Klassifikation}} = 260$) wird die Qualität maschinellen Lernens mit vier Algorithmen (*naïve Bayes, logistic regression, support vector machines* und *decision trees*) untersucht. Evidenz für die Validität der Interpretation der Kodierungen einzelner Algorithmen liegt mit zufriedenstellender bis guter Übereinstimmung zwischen menschlicher und computerbasierter Kodierung beim Training (345–607 Aussagen je nach Teilkompetenz) vor, bei der Klassifikation (157–260 Aussagen je nach Teilkompetenz) reduziert sich dies auf eine moderate Übereinstimmung. Positive Korrelationen zwischen dem kodierten Niveau und dem externen Kriterium Antwortlänge weisen darauf hin, dass die Kodierung mit *naïve Bayes* keine gültigen Ergebnisse liefert. Bedeutsame Attribute, die die Algorithmen bei der Klassifikation nutzen, entsprechen relevanten Begriffen der Niveaufestlegungen im zugrunde liegenden Kodierleitfaden. Abschließend wird diskutiert, inwieweit maschinelles Lernen mit den eingesetzten Algorithmen bei Aussagen zur Modellkompetenz die Qualität einer menschlichen Kodierung erreicht und damit für Zweitkodierungen oder in Vermittlungssituationen genutzt werden könnte.

Schlüsselwörter Kompetenzerfassung · Offenes Antwortformat · Computerbasierte Kodierung · Algorithmen · Evidenz für Validität

Machine Learning with Responses Related to Modeling Competence

Abstract
Computer-automated scoring can help to analyse responses to open-ended questions in large samples. For such purposes, the quality of computer-automated scoring must be evaluated. Using responses related to five aspects of modeling competence from biology teachers, pre-service biology teachers and science education experts ($N_{\text{training}} = 456$; $N_{\text{classification}} = 260$) as an exemplary case, the quality of computer-automated scoring is examined with four algorithms (*naïve Bayes, logistic regression, support vector machines* and *decision trees*). Evidence for the validity of the interpretation of the coding with these algorithms is available with satisfactory to good agreement between human and computer-automated scoring in training (345–607 statements depending on the aspect of modeling competence). In classification (157–260 statements depending on the aspect of modeling competence) this agreement is only moderate. Positive correlations between the identified level of modeling competence and the external criterion 'response length' indicates that coding with *naïve Bayes* does not provide valid results. Significant attributes that were used by the algorithms for classification correspond to relevant terms of the specifications for each level in the coding system. Finally, it is discussed to what extent computer-automated scoring with the algorithms used for modeling competence achieves the quality of a human coding and could thus be used as second human coder or in educational practice.

✉ Dirk Krüger
dirk.krueger@fu-berlin.de

[1] Institut für Biologie, Didaktik der Biologie, Fachbereich Biologie, Chemie und Pharmazie, Freie Universität Berlin, Berlin, Deutschland

Einleitung

Eine Reaktion auf das schlechte Abschneiden der deutschen Schüler*innen bei den ersten Erhebungen des *Programme for International Student Assessment* (PISA) war unter anderem ein neuer Modus der output-orientierten, bildungspolitischen Steuerung (Slepcevic-Zach und Tafner 2012), wodurch der Kompetenzerfassung eine besondere bildungspolitische Relevanz zukommt; „the assessment of competencies plays a key role in optimizing educational processes and improving the effectiveness of educational systems" (Leutner et al. 2017, p. 5). Dementsprechend wurden die Konzeption von Kompetenzmodellen, die Entwicklung von Verfahren zur Kompetenzerfassung sowie die Interpretation und Anwendung von Ergebnissen der Kompetenzerfassung als zentrale Aufgaben der empirischen Bildungsforschung identifiziert (Klieme et al. 2008; Leutner et al. 2017).

Ein theoretisch fundiertes und empirisch gut untersuchtes Gebiet der Naturwissenschaftsdidaktik und ein zentrales Ziel des naturwissenschaftlichen Unterrichts ist die Entwicklung prozeduraler und epistemologischer Facetten fachmethodischer Kompetenzen (Kind und Osborne 2017; KMK 2005, 2019). Zu den fachmethodischen Kompetenzen im Bereich Erkenntnisgewinnung (KMK 2005) wird Modellkompetenz gezählt (Chiu und Lin 2019; Gilbert und Justi 2016; Nicolaou und Constantinou 2014; Nielsen und Nielsen 2019; Upmeier zu Belzen und Krüger 2010; Upmeier zu Belzen et al. 2019). Modellkompetenz kann definiert werden als „die Fähigkeiten, mit Modellen zweckbezogen Erkenntnisse gewinnen zu können und über Modelle mit Bezug auf ihren Zweck urteilen zu können, die Fähigkeiten, über den Prozess der Erkenntnisgewinnung durch Modelle und Modellierungen in der Biologie zu reflektieren sowie die Bereitschaft, diese Fähigkeiten in problemhaltigen Situationen anzuwenden" (Upmeier zu Belzen und Krüger 2010, S. 49).

Um diese Kompetenzen zu erfassen, werden in der Naturwissenschaftsdidaktik unterschiedliche Aufgabenformate eingesetzt (z. B. *single best choice, forced choice,* offene Aufgaben; Grünkorn et al. 2014; Krell 2013; Terzer et al. 2013). Dabei haben offene Aufgabenformate das Potenzial, ein breiteres Spektrum an komplexen Kognitionen zu erfassen (Martinez 1999; Kuechler und Simkin 2010; Nehm et al. 2012), womit sie sich besser zur Erfassung von Kompetenzen als komplexe, kontextspezifische Dispositionen zur Problemlösung eignen (vgl. Hartig und Klieme 2006). Offene Aufgabenformate sind allerdings vergleichsweise zeitaufwendig in der Auswertung, dadurch kostenintensiv und fehleranfälliger in der Interpretation der Aussagen, insbesondere wenn subjektive Aspekte bei der Kodierung die Reliabilität und Objektivität gefährden (Yang et al. 2002). Dies kann ein Grund sein, weshalb offene Aufgabenformate seltener in Studien mit großen Stichproben und zur Individualdiagnose mit schneller Rückmeldung (z. B. im Unterricht) eingesetzt werden (Nehm und Haertig 2012).

Um größere Datenmengen schnell und zuverlässig auswerten und sie kontinuierlich zu Diagnosezwecken nutzen zu können, gibt es Bestrebungen, die Analyse von Aussagen auf Fragen im offenen Format schrittweise zu automatisieren (Liu et al. 2014; Moharreri et al. 2014; Williamson et al. 2012). Beim sogenannten maschinellen Lernen (*ML*) werden Algorithmen mit Hilfe von Aussagen, die von Menschen kodiert wurden, trainiert, um dann neue Aussagen durch die trainierten Algorithmen kodieren zu lassen (Mayfield und Rosé 2013; Williamson et al. 2012; Yang et al. 2002).

In einem aktuellen Review zur Forschung zu *ML* (Zhai et al. 2020) wird mit Blick auf nationale Standards (NGSS Lead States 2013) gefordert: „... future studies should explore using ML results to support other scientific practices such as modelling ...", und damit *ML* über Themengebiete wie naturwissenschaftliches Erklären (Linn et al. 2014) und Argumentieren (Zhu et al. 2017) auszuweiten. Ein in verschiedenen Studien (z. B. Göhner und Krell 2018; Krell und Krüger 2016) intensiv genutztes Instrument zur Erfassung der kognitiven Facetten von Modellkompetenz besteht aus fünf Fragen im offenen Antwortformat, die es bezogen auf fünf Teilkompetenzen (Upmeier zu Belzen und Krüger 2010) erlauben, Aussagen in drei Niveaus einzuteilen und damit das Elaborationsniveau einer antwortenden Person in Bezug auf das untersuchte Konstrukt valide einzuschätzen. In bisherigen Anwendungen umfasst die Auswertung der mit diesem Instrument gewonnenen Daten in der Regel eine unabhängige Kodierung der Aussagen durch zwei Personen sowie eine anschließende Kodierkonferenz zur Diskussion nicht übereinstimmend kodierter Aussagen (Göhner und Krell 2018; Krell und Krüger 2016). Für den Einsatz des Instruments in größeren Stichproben erscheint *ML* vor diesem Hintergrund als ein zeitsparendes Verfahren vielversprechend.

Ziel der vorliegenden Studie zu Kompetenzen von angehenden und praktizierenden Lehrkräften ist es, durch *ML* die Kodierung von Aussagen auf offene Fragen zur Modellkompetenz durch unterschiedliche Algorithmen zu evaluieren. Der Einsatz von *ML* in dieser Studie geht über Studien mit Schüler*innen (Zhai et al. 2020) hinaus und bezieht sich auf den Hochschulbereich (vgl. Zawacki-Richter et al. 2019). Zusätzlich werden Erfahrungen mit einer frei verfügbaren Software (*LightSide*; Mayfield und Rosé 2013) bereitgestellt, die für die Analyse englischer Sprache ausgelegt ist und in der naturwissenschaftsdidaktischen Forschung noch nicht mit Datensätzen in deutscher Sprache untersucht wurde. Die Validität der Interpretation von Ergebnissen mit *ML* wird einerseits durch die Übereinstimmung zwischen menschlicher und computerbasierter Kodierung

und andererseits durch Überprüfung mit einer externen Variable evaluiert (Williamson et al. 2012; Yang et al. 2002). Konkret wird die Eignung von unterschiedlichen, trainierten Algorithmen geprüft, es werden bedeutsame Attribute für die Klassifikation in den Aussagen identifiziert und die Nützlichkeit des *ML* für die fachdidaktische Forschung und praktische Anwendung diskutiert.

Theoretischer Hintergrund

Modellkompetenz

Modellkompetenz ist ein in der Naturwissenschaftsdidaktik theoretisch und empirisch intensiv untersuchtes Konstrukt (Chiu und Lin 2019; Krüger et al. 2018; Nicolaou und Constantinou 2014; Nielsen und Nielsen 2019; Schwarz et al. 2012; Schwarz und White 2005). Modellbildung ist in der Biologie die Grundlage wissenschaftlichen Arbeitens (Bailer-Jones 2003; Giere et al. 2006; Odenbaugh 2005). Naturwissenschaftliche Modelle können sowohl aktuelles Wissen repräsentieren als auch hypothetisch Strukturen und Funktionszusammenhänge konstruieren und einer empirischen Überprüfung zugänglich machen (Gouvea und Passmore 2017; Reinisch und Krüger 2018). Grundsätzlich starten naturwissenschaftliche Arbeitsweisen mit dem Modellieren oder laufen auf Prozesse beim Modellieren hinaus (Lehrer und Schauble 2006; Upmeier zu Belzen et al. 2019). Als authentische Praxis in den Naturwissenschaften ist die vertiefte Reflexion über das Modellieren ein Ziel naturwissenschaftlichen Unterrichts (KMK 2019; NGSS Lead States 2013; Upmeier zu Belzen und Krüger 2019a, 2019b). Nach einem etablierten Strukturmodell können fünf Teilkompetenzen der Modellkompetenz unterschieden werden (Tab. 1), die jeweils in drei Niveaus mit ansteigender Komplexität differenziert sind (Krüger et al. 2018). Die fünf Teilkompetenzen beziehen sich auf kognitive Wissensfacetten, die ein Denken über Modelle und das Modellieren berücksichtigen (Upmeier zu Belzen und Krüger 2010): Dazu gehört ein Nachdenken über die ontologische Frage nach der Ähnlichkeit des Modells zu dem naturwissenschaftlichen Phänomen, das es repräsentieren soll (*Eigenschaften von Modellen*; *EvM*) und ein Verständnis dazu, dass und vor allem aus welchen theoretischen Erwägungen heraus es verschiedene Modelle zu einem naturwissenschaftlichen Phänomen geben kann (*Alternative Modelle; AM*). Elementar und handlungsleitend für die Reflexion über die Nutzung von Modellen ist es, den Nutzen von Modellen über die Repräsentation von Phänomenen hinaus auch als Forschungswerkzeug zu erkennen (*Zweck von Modellen*; *ZvM*), entsprechende Maßnahmen zur Überprüfung des Modells umsetzen zu können (*Testen von Modellen; TvM*) und nötigenfalls das Modell zu überarbeiten (*Ändern von Modellen*; *ÄvM*). Diese im Strukturmodell beschriebenen Teilkompetenzen entsprechen in weiten Teilen den auch in anderen naturwissenschaftsdidaktischen Arbeiten berücksichtigten Dimensionen (z. B. Crawford und Cullin 2005; Grosslight et al. 1991; Schwarz und White 2005; vgl. Krell et al. 2015).

Die einzelnen Teilkompetenzen in den verschiedenen Niveauausprägungen sind Tab. 1 zu entnehmen. Während die Niveaus I und II die Nutzung von Modellen im Sinne von Repräsentationen beschreiben, beschreibt das Niveau III die Nutzung dieser im Sinne von Forschungswerkzeugen.

Zur Modellkompetenz stehen umfangreiche durch Expert*innen kodierte Datensätze aus aktuellen Forschungsprojekten bereit, welche für die Evaluation von Kodierungen durch *ML* genutzt werden können (z. B. Göhner und Krell 2018; Günther et al. 2019; Krell und Krüger 2016). Mit den Daten wird die kognitive Facette von Modellkompetenz, also ein Denken über Modelle, und keine Tätigkeiten beim Modellieren erfasst (vgl. Göhner und Krell 2018). Das Denken über Modelle wird als zentrale Dimension von Modellkompetenz und Voraussetzung für erfolgreiche Problemlösungen in realen Situationen beim Modellieren betrachtet (Schwarz und White 2005).

Tab. 1 Kompetenzmodell der Modellkompetenz. (Krell et al. 2016)

	Niveau I	Niveau II	Niveau III
Eigenschaften von Modellen	Modelle sind Kopien von etwas	Modelle sind idealisierte Repräsentationen von etwas	Modelle sind theoretische Rekonstruktionen von etwas
Alternative Modelle	Unterschiede zwischen den Modellobjekten	Ausgangsobjekt ermöglicht Herstellung verschiedener Modelle von etwas	Modelle für verschiedene Hypothesen
Zweck von Modellen	Modellobjekt zur Beschreibung von etwas einsetzen	Bekannte Zusammenhänge und Korrelationen von Variablen im Ausgangsobjekt erklären	Zusammenhänge von Variablen für zukünftige neue Erkenntnisse voraussagen
Testen von Modellen	Modellobjekt überprüfen	Parallelisieren mit dem Ausgangsobjekt; Modell von etwas testen	Überprüfen von Hypothesen bei der Anwendung; Modell für etwas testen
Ändern von Modellen	Mängel am Modellobjekt beheben	Modell als Modell von etwas durch neue Erkenntnisse oder zusätzliche Perspektiven revidieren	Modell für etwas aufgrund falsifizierter Hypothesen revidieren

Modellkompetenz im Biologieunterricht erfassen

Zur Erfassung von Modellkompetenz wurden unterschiedliche Instrumente entwickelt (vgl. Mathesius und Krell 2019; Nicolaou und Constantinou 2014). Diese können bezüglich der genutzten Aufgabenformate unterschieden werden. Etablierte Aufgabenformate zur Erfassung von Modellkompetenz im deutschsprachigen Raum sind unter anderem *single best choice*-Aufgaben (Terzer et al. 2013), *forced choice*-Aufgaben (Gogolin und Krüger 2018; Krell 2013) sowie Aufgaben im offenen Format (Grünkorn et al. 2014).

Der Vorteil geschlossener Aufgabenformate (z. B. *single best choice; forced choice*) liegt in der zeiteffizienten Auswertung, während offene Aufgabenformate die Erfassung eines breiteren Spektrums an Kognitionen erlauben, aber in der Regel mit einem erhöhten Aufwand in der Auswertung einhergehen (Martinez 1999; Neuhaus und Braun 2007). Gleichzeitig transportieren geschlossene und offene Aufgabenformate spezifische, oftmals konstrukt-irrelevante Anforderungen (z. B. sprachliche Fähigkeiten, spezifische Auswahlstrategien; Martinez 1999; Prenzel et al. 2002; Thoma und Köller 2018). Vor dem Hintergrund der Definition von Kompetenzen als komplexe, kontextspezifische Dispositionen zur Problemlösung (Hartig und Klieme 2006) liegt das Potenzial offener Aufgaben zum Beispiel in der Fähigkeit, divergentes Denkens zu erfassen (z. B. die Fähigkeit, Hypothesen zu generieren; Martinez 1999). Abb. 1 illustriert das grundsätzliche Potenzial offener Aufgabenformate, ein breites Spektrum auch komplexerer Kognitionen zu erfassen. Da Modellkompetenz bisher in der fachdidaktischen Forschung vorwiegend mit geschlossenen Formaten erfasst wird (Mathesius und Krell 2019; Nicolaou und Constantinou 2014), eröffnen Verfahren des *ML* die Option, auch Antworten auf Fragen im offenen Format zeiteffizient auszuwerten (Lee et al. 2019; Liu et al. 2014; Moharreri et al. 2014).

Verfahren des maschinellen Lernens

Grundsätzlich lassen sich drei Arten des *ML* unterscheiden: überwachtes, bestärkendes und unüberwachtes *ML* (Buxmann und Schmidt 2018; Russel und Novig 2010). Bei dem hier realisierten überwachten Verfahren nehmen die Algorithmen beim *ML* auf der Grundlage der Häufigkeit und Verteilung von Attributen (d. h. Worten und Buchstaben in den Dokumenten) in Aussagen Gewichtungen für die einzelnen Attribute vor, um damit einer vorgegebenen menschlichen Kodierung möglichst nahe zu kommen (Trai-

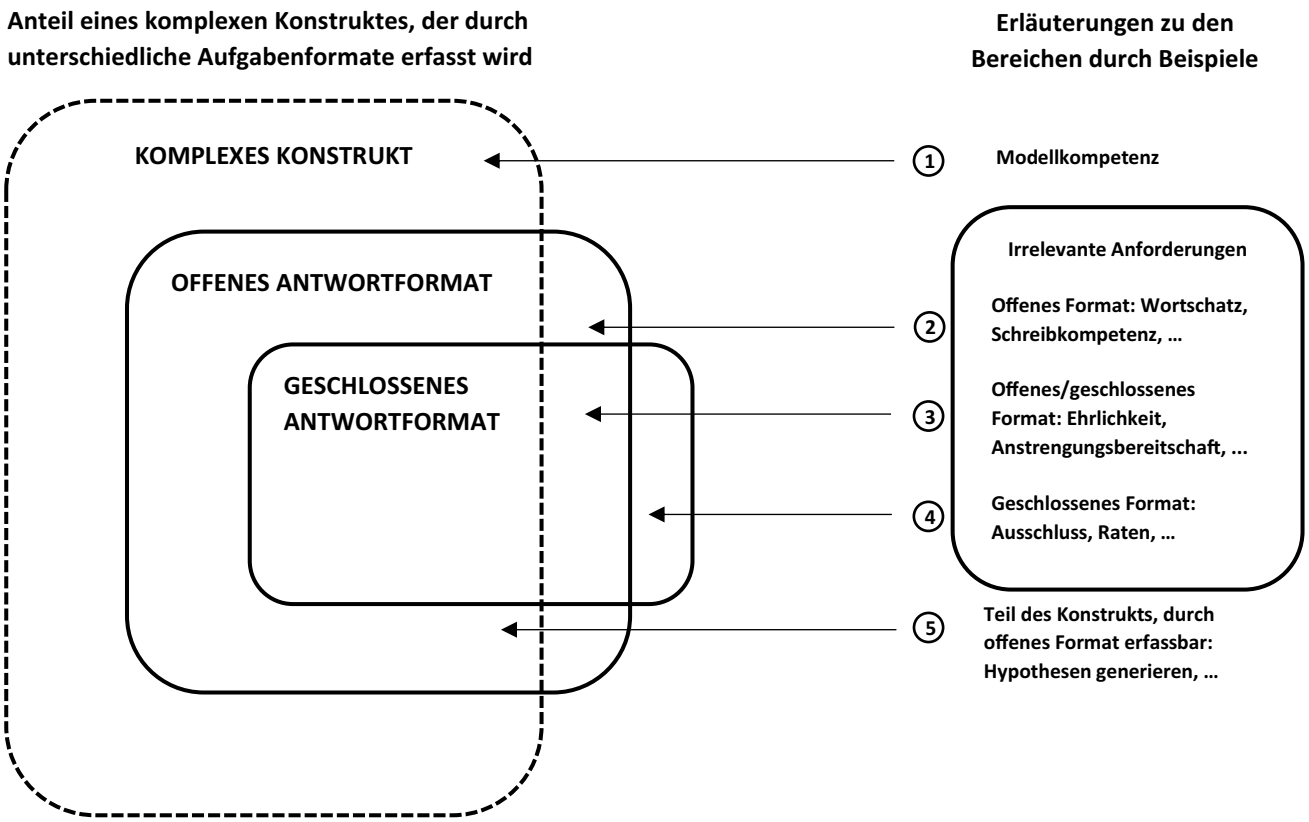

Abb. 1 Illustration des Potenzials von Aufgaben in verschiedenen Formaten zur Erfassung unterschiedlicher Kognitionen. (Verändert nach Martinez 1999)

ningsprozess). Diese Gewichtungen, das heißt Zahlenwerte, die die Bedeutung des Attributs für die Klassifikation angeben, werden anschließend genutzt, um eine Klassifikation von neuen Aussagen ohne vorgegebene menschliche Kodierung zu ermöglichen. Dies ist einerseits vom unüberwachten Verfahren zu unterscheiden, bei dem Muster in Daten, zum Beispiel Bildern, gesucht werden und in Kategorien ohne menschliche Vorgabe sortiert werden, und andererseits von bestärkenden Verfahren zu differenzieren, die zum Beispiel in Brettspielen wie Go und Schach umgesetzt werden können, bei denen eine zielerreichende Strategie gelernt werden soll und die Umorganisation der Bewertungen autonom ohne menschliches Zutun erfolgt (Buxmann und Schmidt 2018). Ein solcher selbstlernender Optimierungsprozess kann in diesem Projekt nicht realisiert werden, weil die Zielvorgabe (Niveau einer Aussage) nur durch die menschliche Kodierung bereitgestellt werden kann.

Wegen der hier verwendeten nominalen Daten (Wörter- und Buchstabenfolgen) und wegen der Einteilung in drei Niveaus werden Klassifikationstechniken verwendet, die diskrete Ausgänge voraussagen und in Kategorien einstufen. Für (überwachtes) *ML* wurden in der fachdidaktischen Forschung bereits unterschiedliche Algorithmen erfolgreich zur Kodierung von Aussagen genutzt (Moharreri et al. 2014; Zawacki-Richter et al. 2019; Zhai et al. 2020). Dazu gehören *naïve Bayes (NBayes), logistic regression* (*LogReg*; logistische Regression), *support vector machines* (*SVM*; Stützvektormethode) und *decision trees* (*DTrees*; Entscheidungsbäume; Abikoye et al. 2018; Mayfield und Rosé 2013). Diese Algorithmen unterscheiden sich in der Art des Umgangs mit den zur Verfügung stehenden Attributen, also Wort- oder Buchstabenfolgen und/oder deren Häufigkeiten, die für die Kodierung der Aussagen genutzt werden. Während beim *NBayes* jedes Attribut Bedeutung hat und keine Abhängigkeit zwischen den Attributen angenommen wird, wird bei *LogReg* der kleinstmögliche Merkmalsraum genutzt. *SVM* ist für dichotome ja/nein-Entscheidungen optimiert und *DTrees* variieren zwischen den Gewichtungen der Attribute (Witten et al. 2011). Weitergehende Erläuterungen zu den vier Algorithmen finden sich im Anhang.

Die Kodierung von Antworten zu Fragen im offenen Format mit Hilfe der genannten Algorithmen ist ein zweischrittiger Prozess: Im Training (Lernen) gewichten die Algorithmen zunächst die für die Kodierung relevanten Attribute auf der Basis einer zur Verfügung gestellten menschlichen Kodierung. In der Klassifikation (Voraussagen) wird anschließend auf der Basis dieser Gewichtungen eine computerbasierte Kodierung an einem neuen Datensatz ohne Berücksichtigung menschlicher Kodierung vorgenommen (Mayfield und Rosé 2013).

Stand der Forschung zu maschinellem Lernen in der Naturwissenschaftsdidaktik

In einem aktuellen Review zu *ML* in der Naturwissenschaftsdidaktik (Zhai et al. 2020) wird der Stand der Forschung unter Berücksichtigung von 49 Studien mit Schüler*innen analysiert. Das Review von Zawacki-Richter et al. (2019) gibt einen Überblick über den Hochschulbereich. Beide Reviews weisen Studien aus, in denen überwiegend mit überwachtem *ML* große Datensätze offen formulierter Antworten aus Erhebungen ausgewertet werden oder andererseits *ML* angewendet wird, um in Vermittlungssituationen Lernenden unmittelbar Rückmeldung zu ihren Antworten zu geben. In beiden Fällen befreit *ML* (Lehr-)Personen von zeitaufwendigen oder nicht individuell zu leistenden Aufgaben. Für den Hochschulbereich wird ein breites Spektrum an Anwendungen von *ML* zur Unterstützung von Student*innen (z. B. Erreichen von Studienzielen), der Lehre (z. B. Erfassung von Kenntnissen) und der Verwaltung (z. B. Auswahl von Studierenden für Studiengänge; Zawacki-Richter et al. 2019) beschrieben. In den berücksichtigten Studien werden unterschiedliche Algorithmen eingesetzt, wobei die Auswahl eines Algorithmus ein iterativer Prozess ist, in dem zunächst die Performanz unterschiedlicher Algorithmen exploriert und dann derjenige, der für den gegebenen Anwendungskontext am besten geeignet ist, ausgewählt wird. Zu den am meisten genutzten Algorithmen gehören diejenigen, die in dieser Studie eingesetzt werden: *NBayes, LogReg, SVM* und *DTrees* (Zhai et al. 2020). Desiderate bezüglich des Einsatzes von *ML* sind die Entwicklung neuer Aufgabenformate auf der Basis der vorliegenden Erfahrungen mit *ML* und die Ausweitung der Forschung zur Eignung von *ML* auf weitere relevante Zielkonstrukte naturwissenschaftlicher Bildung (u. a. Modellkompetenz). Dabei ist es anzustreben, (i) auf der Basis gut trainierter Algorithmen Bewertungen vorzunehmen, (ii) unmittelbares Feedback zu ermöglichen sowie (iii) neue Möglichkeiten zu entwickeln, Aufgaben zur Überprüfung von Kompetenz zielführender zu entwickeln (Zhai et al. 2020).

Studien mit *ML* in der Naturwissenschaftsdidaktik untersuchen den Inhalt (z. B. Nehm et al. 2012) oder die Qualität in Texten (z. B. Bridgeman et al. 2012). Dies geschieht, indem durch *ML* unmittelbares Feedback zu Texten, die von Lernenden verfasst wurden, in den Themengebieten Photosynthese oder Mitose gegeben wird (Linn et al. 2014). Andere Studien mit *ML* von Interviewdaten oder schriftlichen Antworten auf offene Fragestellungen befassen sich mit dem naturwissenschaftlichen Erklären (Linn et al. 2014) und Argumentieren (Zhu et al. 2017). Diese Bereiche sind theoretisch gut fundiert und es liegen Sammlungen von Begründungen oder Erklärungen in längeren mündlichen oder schriftlichen Ausführungen vor (Lee et al. 2019); zum Beispiel im Inhaltsbereich Jahreszeiten,

hier auch unter dem Blick auf Schülervorstellungen (Dam und Kaufmann 2008). In Liu et al. (2016) werden elf Studien mit *ML* zitiert. Darunter befinden sich auch die Arbeiten unter Beteiligung von Nehm und Kolleg*innen, bei denen es um Argumentationsstrukturen von Schüler*innen zur Evolution und zur natürlichen Selektion geht (Nehm et al. 2012; Ha et al. 2011; Nehm und Haertig 2012; Moharreri et al. 2014; Beggrow et al. 2014). Aktueller liegen Arbeiten mit unmittelbarem Feedback zum Argumentieren im Bereich Ökologie und Nachhaltigkeit mit verschiedenen Aufgabeninhalten vor (Lee et al. 2019). Hier erweisen sich trotz sehr guter computerbasierter Kodierung insbesondere falsch positive Rückmeldungen als problematisch, weil sie den Lernvorgang behindern. Abhängig davon, wie frei *ML* eingesetzt wird, unterscheiden Williamson et al. (2012) unterschiedliche Einsatzmöglichkeiten computerbasierter Kodierung, zum Beispiel die computerbasierte Kodierung im Sinne einer Zweitkodierung (*automated quality control of human scoring*) oder die alleinige computerbasierte Kodierung (*automated scoring alone*). Je freier die Ergebnisse der computerbasierten Kodierung genutzt werden sollen, desto wichtiger ist es, deren Güte systematisch zu evaluieren.

Überprüfung der Güte des maschinellen Lernens

Verfolgt man langfristig das Ziel, Daten durch *ML* ohne weitere menschliche Kontrolle analysieren zu lassen, muss Evidenz für die Güte des *ML* im vorgesehenen Anwendungskontext vorliegen. Ein etabliertes Vorgehen zur Überprüfung der Güte der Kodierung durch *ML* besteht in der Überprüfung der Übereinstimmung zwischen menschlicher und computerbasierter Kodierung sowohl im Trainingsprozess als auch bei der Klassifikation neuer Daten (Williamson et al. 2012; Yang et al. 2002). Hierfür muss zunächst die Güte der menschlichen Kodierung geprüft werden (Williamson et al. 2012). Anschließend kann die Übereinstimmung zwischen menschlicher und computerbasierter Kodierung mittels prozentualer Übereinstimmung oder *Cohens Kappa (K)* (um zufällige Übereinstimmung zu berücksichtigen) evaluiert werden, wobei Werte von $K > 0{,}60$ (Wirtz und Caspar 2002) oder $K > 0{,}70$ (Williamson et al. 2012) als Maß für eine gute Übereinstimmung betrachtet werden. Zusätzlich können die von dem bei *ML* zur Kodierung jeweils eingesetzten Algorithmus stark gewichteten Attribute mit dem intendierten Konstrukt (z. B. dem Kodierleitfaden) verglichen werden, um auf diese Weise den automatisierten Auswertungsprozess inhaltlich zu prüfen (Williamson et al. 2012; Yang et al. 2002). Über Korrelationsanalysen lässt sich darüber hinaus ermitteln, ob die Kodierung durch die Algorithmen stärker durch die Antwortlänge beeinflusst wird als die menschliche Kodierung, was als Mangel für Validität betrachtet wird (Mao et al. 2018).

Zielsetzung und Fragestellungen

Trotz der hervorgehobenen Relevanz der Entwicklung von Modellkompetenz als Ziel naturwissenschaftlichen Unterrichts und der bereits langen Tradition zum Erfassen von Wissen und Fähigkeiten zu Modellen und dem Modellieren in der fachdidaktischen Forschung (z. B. Grosslight et al. 1991) werden für letzteres auf Kosten der Möglichkeit, ein breiteres Spektrum an Kognitionen zu erfassen, überwiegend geschlossene Antwortformate eingesetzt (Nicolaou und Constantinou 2014). Vor dem beschriebenen Hintergrund ist das Ziel der vorliegenden Studie, die Güte des *ML* für die Auswertung von Aussagen auf offene Fragen zur Modellkompetenz zu evaluieren (Williamson et al. 2012). Hierzu wird die Übereinstimmung zwischen menschlicher und computerbasierter Kodierung sowie der Prozess der automatisierten Kodierung untersucht (Yang et al. 2002). Gleichzeitig soll geprüft werden, inwiefern die Ergebnisse Rückschlüsse auf die Eignung des *ML* in der fachdidaktischen Forschung und Praxis zulassen (Williamson et al. 2012). Die ersten beiden Forschungsfragen untersuchen die Qualität des *ML* in Bezug auf normative Setzungen (Wirtz und Caspar 2002; Williamson et al. 2012) mit den eingesetzten Algorithmen (F1; Zhai et al. 2020) und auf die Datensätze zu den fünf Teilkompetenzen der Modellkompetenz (F2; Krüger et al. 2018).

F1 Inwiefern unterscheidet sich menschliche und computerbasierte Kodierung von Aussagen zur Modellkompetenz in Abhängigkeit von den vier Algorithmen *NBayes, LogReg, SVM* und *DTrees*?

F2 Inwiefern unterscheidet sich menschliche und computerbasierte Kodierung von Aussagen zur Modellkompetenz in Abhängigkeit von den fünf Teilkompetenzen der Modellkompetenz (Tab. 1)?

Eine hohe Übereinstimmung der durch die Algorithmen besonders stark gewichteten Attribute mit den Ausführungen im Kodierleitfaden (F3) wird als Evidenz für eine valide Interpretation der durch die Algorithmen vorgenommenen Kodierung betrachtet (Williamson et al. 2012). Ferner liefern Korrelationsanalysen zwischen den durch die Algorithmen kodierten Niveaus mit der externen Variable „Antwortlänge“ (F4) Hinweise darauf, inwieweit die Algorithmen von einem Oberflächenmerkmal beeinflusst werden (Mao et al. 2018).

F3 Inwiefern besteht eine Übereinstimmung zwischen besonders relevanten Attributen für das *ML* mit Aussagen zur Modellkompetenz und Schlüsselwörtern im zugrundeliegenden Kodierleitfaden (Krell und Krüger 2016; Tab. 1)?

F4 Inwiefern unterscheiden sich die Korrelationen zwischen kodierten Niveaus und Antwortlänge bei menschlicher und computerbasierter Kodierung?

Die abschließende Frage (F5) geht dem Untersuchungsinteresse nach, mit welcher Unsicherheit beziehungsweise welchem Nutzen das *ML* beim Einsatz in einer Zweitkodierung verbunden ist (Williamson et al. 2012).

F5 Zu welchen Fehlern (und Zeitersparnissen) führt die Nutzung von *ML* mit Aussagen zur Modellkompetenz beim Einsatz für Zweitkodierungen?

Methoden

Datenerhebung

Es wurde je eine offene Frage zu jeder der fünf Teilkompetenzen der Modellkompetenz eingesetzt (Krell und Krüger 2016). Die Daten stammen teilweise aus bereits durchgeführten Projekten (Göhner und Krell 2018; Günther et al. 2019; Krell und Krüger 2016). Zusätzlich wurden für die vorliegende Studie Bachelor- und Masterstudierende des Lehramts (mit Fach Biologie) sowie Fachdidaktiker*innen mit Expertise im Bereich Modellkompetenz befragt. Letztere wurden als Expert*innen explizit dazu aufgefordert, idealtypische Aussagen für alle drei Niveaus jeder Teilkompetenz zu formulieren.

Stichprobe, Design und Statistik

Für das Training der Algorithmen wurden Aussagen von Biologielehrkräften ($n = 148$), Biologie-Lehramtsstudierenden ($n = 217$) und Fachdidaktiker*innen ($n = 91$) genutzt ($N = 456$). Für die Klassifikation wurden 260 Biologie-Lehramtsstudierende befragt. Neu hinzugekommene Datensätze wurden von den Autoren dieses Beitrags vollständig und unabhängig voneinander mit Hilfe eines bestehenden Kodierleitfadens (Krell und Krüger 2016) mit guter bis sehr guter Beurteilerübereinstimmung kodiert (*Cohens K*; Tab. 2); bei Daten aus bereits durchgeführten Projekten wurde die bestehende Kodierung übernommen. Die Gleichverteilung der Antworten über die drei Niveaus zwischen Trainings- und Klassifikationssatz wurde mit dem X^2-Test untersucht (Döring und Bortz 2016). Die Klassifikationssätze weichen auf dem 1 %-Signifikanzniveau in der Verteilung auf die drei Kompetenzniveaus nicht von den Verteilungen der Trainingssätze ab (Tab. 2).

Tab. 2 Trainings-(Lernen) und Klassifikations-(Voraussagen) Datensätze: Häufigkeit von Aussagen in den Niveaus I, II, III (menschliche Kodierung; *Cohens K* berechnet mit 50 % der Klassifikationsdaten)

	Training				Klassifikation				
Teilkompetenz, Beurteilerübereinstimmung (*K*) und offene Frage	I	II	III	∑	I	II	III	∑	X^2-Test
Eigenschaften von Modellen ($K = 0{,}67$) Beschreiben Sie, inwieweit ein Modell Ihrer Meinung nach dem biologischen Phänomen entspricht?	75 Ø Wörter/Antwort: 24	399	51	525	15 Ø Wörter/Antwort: 28	161	16	192	$p > 0{,}01$
Alternative Modelle ($K = 0{,}85$) Aus welchen Gründen gibt es Ihrer Meinung nach zu einem biologischen Phänomen verschiedene Modelle?	53 Ø Wörter/Antwort: 23	367	130	607	4 Ø Wörter/Antwort: 32	111	42	157	$p > 0{,}01$
Zweck von Modellen ($K = 0{,}72$) Welchen Zweck erfüllen Ihrer Meinung nach Modelle in der Biologie?	287 Ø Wörter/Antwort: 14	156	149	592	109 Ø Wörter/Antwort: 19	88	63	260	$p > 0{,}01$
Testen von Modellen ($K = 0{,}66$) Wie lässt sich Ihrer Meinung nach überprüfen, ob ein biologisches Modell seinen Zweck erfüllt?	37 Ø Wörter/Antwort: 19	185	123	345	8 Ø Wörter/Antwort: 21	102	64	174	$p > 0{,}01$
Ändern von Modellen ($K = 0{,}62$) Aus welchen Gründen wird ein gegebenes biologisches Modell Ihrer Meinung nach verändert?	126 Ø Wörter/Antwort: 17	356	81	563	36 Ø Wörter/Antwort: 24	157	39	232	$p > 0{,}01$

X^2-Test auf Gleichverteilung zwischen den beiden Datensätzen

Maschinelles Lernen mit LightSide

Für *ML* wurde die Software *LightSide* eingesetzt (Mayfield und Rosé 2013). Die Aussagen der Proband*innen wurden zur Eingabe in die Software, die für englische Buchstabenerkennung entwickelt wurde, wie folgt überarbeitet: Großbuchstaben wurden in Kleinbuchstaben umgewandelt, Satzzeichen wurden entfernt, ß wurde in ss und Umlaute (ä, ö, ü) in Doppelvokale (ae, oe und ue) umgeschrieben.

Es wurden die vier Algorithmen *NBayes, LogReg, SVM* und *DTrees* in *LightSide* eingesetzt (Mayfield und Rosé 2013; Zhai et al. 2020). Beim Trainingsprozess wurden folgende allgemeine Einstellungen vorgenommen, um die Attribute zu erhalten: Kombinationen aus einem Wort und bis zu drei in Folge auftretende Worte (*basic features*; uni-, bi-, trigrams) und Buchstabenfolgen aus drei bis vier Buchstaben inklusive einer Lücke zwischen Worten (*character n-grams*). Eine Reduktion der ausgewählten Attribute, die sich durch Konjugation oder Deklinationen unterscheiden, lässt die für englische Sprache konstruierte Software nicht automatisch zu. Bei den Modellrechnungen mit den Algorithmen wurden alle Prozessdurchläufe manuell eingestellt und in Gruppen von zehn zufällig ausgewählten Aussagen durchgeführt (z. B.: Eigenschaften von Modellen: 52 Trainingsdurchläufe). Zur Optimierung der Prozesse wurden die Standardeinstellungen im Programm genutzt (*LogReg*: *L2 regularization*; *SVM*: *normalize* und LibLINEAR zur Vermeidung von Overfitting (irrelevante Attribute werden beibehalten) beim Reduktionsprozess der Attribute; *DTrees*: *prune tree* und minimal zwei Objekte in den Ästen zum Beschneiden des Baumes und Reduzierung der Komplexität). Tab. 3 gibt die Anzahl der von den Algorithmen genutzten Attribute in beiden Datensätzen an.

Evidenz für Validität

Für eine valide Interpretation der durch die Algorithmen vorgenommenen Kodierungen werden verschiedene Evidenzquellen evaluiert (AERA et al. 2014). Als Voraussetzung wird die Reliabilität durch die Übereinstimmung mit der menschlichen Kodierung (prozentuale Übereinstimmung, *Cohens K*) geprüft. Zur Prüfung von Validität wird die vergleichende Betrachtung der *Beziehung zu anderen Variablen* (hier: Antwortlänge) sowie der *Testinhalt* durch den Abgleich gewichteter Attribute mit dem Kodierleitfaden herangezogen. Auch die inhaltliche Analyse von Fehlkodierungen durch die Algorithmen trägt dazu bei, die Kodierung zu verstehen und im folgenden Schritt zu optimieren.

Jeder Algorithmus wurde in Abhängigkeit von der Gesamtzahl *N* an Aussagen je Teilkompetenz (Tab. 2) mit zufälliger Fallauswahl trainiert (1-1/*N*) % und jeweils gegenüber den nicht eingesetzten Aussagen getestet (1/*N*) %. Nach *N* Durchläufen wurde die prozentuale Übereinstimmung und *Cohens K* in diesem Trainingsprozess bestimmt. Der Klassifikationsprozess wurde mit einem Datensatz durchgeführt, der nicht im Trainingsprozess genutzt wurde (Tab. 2).

Als Beziehung zu anderen Variablen wurde die inhaltlich irrelevante Länge der Antworten herangezogen. Evidenz für Validität liegt vor, wenn die Korrelationen zwischen kodierten Niveaus und Antwortlänge bei computerbasierter

Tab. 3 Übereinstimmung (in % und Cohens *K*) beim Training (Lernen) und Klassifizieren (Voraussagen) in den fünf Teilkompetenzen und mit vier Algorithmen

Teilkompetenzen		*NBayes*		*LogReg*		*SVM*		*DTrees*	
	Attribute/versch. Wörter	%	Kappa	%	Kappa	%	Kappa	%	Kappa
Eigenschaften von Modellen	Training: 4861/380	73	0,44	87	0,64	88	0,67	76	0,36
	Klassifikation: 2740/180	80	0,31	86	0,34	84	0,35	**84**	**0,44**
Alternative Modelle	Training: 5088/376	76	0,57	86	0,70	86	0,71	80	0,58
	Klassifikation: 2650/156	71	0,41	80	0,50	**79**	**0,50**	76	0,46
Zweck von Modellen	Training: 3849/268	81	0,70	87	0,80	87	0,80	84	0,75
	Klassifikation: 2852/177	70	0,55	72	0,57	72	0,56	**75**	**0,62**
Testen von Modellen	Training: 3278/209	87	0,77	89	0,80	91	0,84	78	0,62
	Klassifikation: 2602/151	73	0,50	**78**	**0,57**	76	0,55	67	0,42
Ändern von Modellen	Training: 4027/270	83	0,70	92	0,84	93	0,86	89	0,80
	Klassifikation: 2790/173	75	0,54	86	0,71	**86**	**0,72**	81	0,62

Fett hervorgehobene Zahlen geben je Teilkompetenz den besten Algorithmus beim Klassifizieren an

Kodierung nicht höher ausfällt als bei menschlicher Kodierung.

Ergebnisse

Überprüfung der Übereinstimmung zwischen menschlicher und computerbasierter Kodierung

Die folgenden Ergebnisse beziehen sich auf den Einsatz der vier beschriebenen Algorithmen, weitere 39 Algorithmen, die in *LightSide* zur Verfügung stehen, erzielten in keinem Fall bessere Übereinstimmungswerte. Die Übereinstimmung zwischen menschlicher und computerbasierter Kodierung für die vier Algorithmen ist in Tab. 3 separat für die fünf Teilkompetenzen der Modellkompetenz dargestellt. Das Spektrum der Übereinstimmung beim Training rangiert zwischen 73–93 %, *Cohens K* zwischen $0{,}36 \leq K \leq 0{,}87$. Es gelingt mit den zur Verfügung stehenden Daten, bei allen Teilkompetenzen einen Algorithmus zu finden, der mit gutem *Cohens K* ($K > 0{,}60$; Wirtz und Caspar 2002) trainiert werden kann. Für jede Teilkompetenz konnte mit *SVM* am besten trainiert werden und kaum schlechter mit *LogReg*. *NBayes* schneidet in allen Teilkompetenzen bis auf *Testen von Modellen* am wenigsten zuverlässig ab (Tab. 3).

Bei der Klassifikation sinkt *Cohens K* auf akzeptable bis gute Werte ($0{,}44 \leq K \leq 0{,}72$). Dabei ist in drei Fällen (Tab. 3: *EvM, ZvM, TvM*) der am besten trainierte Algorithmus nicht derjenige, der am besten klassifiziert. Während *ÄvM* mit guten Werten klassifiziert wird, gelingt bei *EvM* das Training und die Klassifikation weniger gut. Es besteht ein negativer Zusammenhang zwischen der Anzahl eingesetzter Attribute in den Teilkompetenzen und *Cohens K* beim Training (Tab. 3): Je mehr Attribute die Antworten einer Teilkompetenz bereitstellen, umso schlechter lässt sich jeder Algorithmus trainieren ($-0{,}89 \leq r \leq -0{,}49$; Pearson Korrelation). Die Qualität der computerbasierten Kodierung je Teilkompetenz wird dabei maßgeblich negativ durch lange durchschnittliche Antworten und eine hohe Anzahl verschiedener Wörter beeinflusst (vgl. *EvM* und *AM* mit *ZvM, TvM* und *ÄvM*; Tab. 2 und 3).

Die Darstellung des jeweils kodierten Niveaus über alle Aussagen lässt sich als „Schwierigkeit" der Aufgaben bei der Klassifikation durch den Menschen und durch die vier Algorithmen interpretieren (Abb. 2). Schwierigkeit einer Aufgabe bedeutet hier, dass weniger (schwierig) oder mehr (leicht) Kodierungen in höheren Niveaus vorliegen. Im Vergleich zwischen menschlicher und computerbasierter Kodierung unterschätzen die Algorithmen tendenziell das Niveau der Aussagen mit maximal kleinem Effekt (*Cohens* $d < 0{,}22$), kodieren also seltener ein hohes Niveau (Abb. 2). Davon ist besonders oft die Kodierung von *SVM* und *LogReg* (Ausnahme: *TvM*) betroffen, während *NBayes* tendenziell überschätzt (Ausnahme: *EvM*).

Inhaltliche Prüfung des maschinellen Lernens

Tab. 4 zeigt besonders gewichtete Wörter, die beim Trainingsprozess von den Algorithmen zur Zuordnung zu einem Niveau genutzt werden. Die Begriffe decken sich weitgehend mit der theoretischen Grundlage (Tab. 1) und auch mit den Kodierhinweisen und Schlüsselwörtern im Kodierleitfaden. Einzelne Begriffe (Artikel, Personalpronomen, Konjunktionen) zeigen keine solche Übereinstimmung mit den

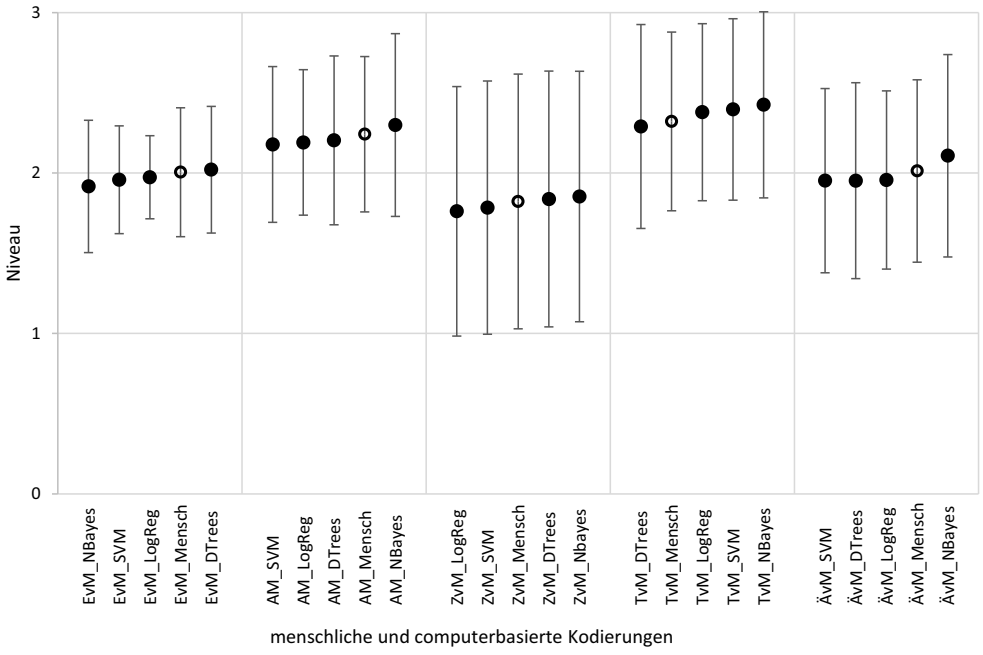

Abb. 2 Vergleich der Kodierungen von Mensch und Algorithmen bei der Klassifikation

Tab. 4 Bis zu zehn positiv wichtige Wörter (sofern vom Algorithmus angegeben) im besten Algorithmus je Teilkompetenz für eine Zuordnung zu einem Niveau

Teilkompetenz/ Algorithmus	Niveau I	Niveau II	Niveau III
Eigenschaften von Modellen/*DTrees*	**Sowe**it – möglichst genau – vergrö**ssert** – biologischen Ausgangsobjekt – vielleicht – *wenn* – das Modell ist	**Kons**truiert	Ver**mutu**ng – **H**ypothese –V**orst**ellung – **Theo**rie
Alternative Modelle/*SVM*	*Es* – Modelle – Größe – Hersteller – Vereinfachung – Materialien – *auch* – in – verschiedene – *stehen*	Neue – Erkenntnisse – veranschaulichen – *um* – *die* – Modell – Funktionen – Details – von Modellen – Fokus	Theorien – Vorstellungen – *über* – Hypothesen – unterschiedliche – Meinungen – *geben* – Erkenntnissen – *es kann* – *der*
Zweck von Modellen/*DTrees*	*Algorithmus ohne positiv ausgewiesene Attribute für Niveau I*	Erläutern – *sich* – Vergleich – Erkenntnisg**ewin**nung – V**erst**ändnis/**verst**ehen – **Erkl**aerung	Ableiten und – abzuleiten – Annahmen – dem Modell – *lassen* – Theorien – vorhersagen – **Ere**ignisse – **Zuk**unft/**zuk**ünftige – überpr**uefe**n
Testen von Modellen/*LogReg*	*Sein* – *werden* – funktioniert – Modell – *bei* – *ob* – *ist* – vereinfacht – prüfen – stabil	Modellkritik – *das* – das Modell – *wenn* – Vergleich – Original – Funktionsweise – Vergleich mit – Sachverhalt – *oder*	Hypothesen – *am* – dem Modell – *es* – *dem* – *aus* – *man* – *über* – vorhersagen – *eine*
Ändern von Modellen/*SVM*	Vereinfachung – falsch – *ist* – *zu* – Modell – *es* – *muss* – *auch* – *hat sich* – *sich nicht*	Neue Erkenntnisse – Original – neue Forschungsergebnisse – *dass* – *haben* – *es die* – Realität – darstellt – *an* – der Forschung – *so*	Eine neue – dem Modell – widerlegt – *in der* – Hypothese(n) – *aus* – *eine* – *lassen sich* – verändert

Fett Beispiele für gewichtete Buchstaben in Worten; *kursiv* Worte mit positiven Gewichtungen, die sich inhaltlich nicht erschließen

theoretischen Grundlagen (kursiv gesetzte Begriffe; Tab. 4). Die Analyse einer nicht direkt verständlichen positiven Gewichtung eines Wortes oder von Buchstabenfolgen erklärt sich in einigen Fällen durch die Betrachtung der entsprechenden vollständigen Aussagen.

In den Kreuztabellen aus menschlicher und computerbasierter Kodierung erkennt man beim Training und der Klassifikation Fehlkodierungen (Tab. 5). Die Analyse der Unter- und Überschätzungen in der computerbasierten Kodierung gibt Aufschluss über die Fehleranfälligkeit. In den folgenden Beispielen sind relevante Begriffe für das kodierte Niveau kursiv markiert: Die Niveau I-Aussage „bei der Bildung von Modellen kann es notwendig sein unterschiedliche Materialien zu verwenden, auch wenn die *Hypothese* identisch bleibt“ wird durch *SVM* in Niveau III kodiert (Tab. 5a; Kodierung Mensch I/*SVM* III). Die folgenden Aussagen kodiert *SVM* in Niveau I statt Niveau III: „ein *Modell* kann verändert werden, wenn es seinen Zweck nicht erfüllt und vom *Modell* abgeleitete Vorhersagen *falsch* sind“ (Tab. 5a; III/I); „ein *Modell muss* verändert werden, wenn die gemachten Aussagen nicht beobachtbar oder *falsch* sind, ein *Modell muss auch* verändert werden, wenn *falsche* Annahmen gemacht worden sind“ (Tab. 5b, III/I) und „Erklärung, Theorie des Phänomens ist *falsch*, Experiment widerspricht massiv dem *Modell, Modell* wird durch ein weiteres *Modell* erweitert“ (Tab. 5b; III/I).

Überprüfung mit der externen Variable „Antwortlänge“

Die Korrelation zwischen dem Niveau der Kodierung und der Antwortlänge liegt bei *LogReg*, *SVM* und *DTrees* zwischen $0{,}05 < r < 0{,}33$ und ist damit niedrig bis mittel (Döring und Bortz 2016; Tab. 6). Demgegenüber korreliert das kodierte Niveau bei *NBayes* mit einer Ausnahme (Training *TvM*) signifikant positiv mit der Antwortlänge und fällt signifikant höher aus als bei menschlicher Kodierung (kleine bis große Effekte; Döring und Bortz 2016; Tab. 6).

Überprüfung der Eignung zur Zweitkodierung

Tab. 7 illustriert die potenzielle Nützlichkeit des *ML* für eine Zweitkodierung. Für diejenigen Aussagen, in denen sich die computerbasierte von der menschlichen Kodierung unterscheidet (*ML* ≠ Mensch), wäre eine wiederholte Prüfung nötig, was in annähernd 32 % der Fälle (*ML* fehlerhaft) ke-

Tab. 5 Vergleich der Niveauzuordnungen I, II und III von menschlicher und *SVM*-Kodierung beim Training (a) und der Klassifikation (b) für die Teilkompetenz *Ändern von Modellen*

(a) Training	Algorithmus *SVM*			(b) Klassifikation	Algorithmus *SVM*		
Mensch	I	II	III	Mensch	I	II	III
I	116	9	1	I	31	5	–
II	13	339	4	II	11	141	5
III	1	14	66	III	2	9	28

Tab. 6 Korrelationen zwischen dem Niveau der Kodierung beim Training und der Klassifikation mit der Antwortlänge. Signifikanztest für Vergleich der Korrelationen der Algorithmen vs. Mensch. (Nach Eid et al. 2011, S. 548 f)

	Training					Klassifikation				
Teilkompetenz	Mensch	*NBayes*	*LogReg*	*SVM*	*DTrees*	Mensch	*NBayes*	*LogReg*	*SVM*	*DTrees*
EvM	0,19	**0,58*** **	0,17	0,16	0,03	0,11	**0,42*** **	0,03	–0,01	0,00
AM	0,18	**0,52*** **	0,18	0,20	0,14	0,16	**0,49*** **	0,14	0,17	0,11
ZvM	0,21	0,33***	0,17	0,17	0,20	0,19	**0,51*** **	0,33**	0,32**	0,23
TvM	0,14	0,10	0,07	0,09	0,07	0,05	0,15**	–0,00	–0,01	–0,06
ÄvM	0,30	**0,51*** **	0,30	0,31	0,29	0,29	**0,51*** **	0,28	0,26	0,30

Effektstärke fett hervorgehoben: mittel und groß
* $p < 0{,}05$; ** $p < 0{,}01$; *** $p < 0{,}001$

Tab. 7 Vergleich des menschlichen und maschinellen Lernens (*ML*) für eine Zweitkodierung (Angaben in %)

	ML ≠ Mensch		*ML* = Mensch	
	ML fehlerhaft	Mensch fehlerhaft	Zutreffende Kodierung	Unentdeckte menschliche Fehler
Eigenschaften von Modellen	44	5	47	3
Alternative Modelle	33	2	66	2
Zweck von Modellen	33	7	59	2
Testen von Modellen	28	2	66	4
Ändern von Modellen	20	9	66	5
Durchschnitt	32	5	60	3

ine Veränderung des menschlichen Urteils ergeben hätte. In 5 % dieser Fälle hätte durch *ML* eine menschliche Fehlkodierung revidiert und optimiert werden können. Für insgesamt circa 63 % der Aussagen (*ML* = Mensch) hätte eine zweite menschliche Kodierung eingespart werden können, wobei dabei in circa 3 % der Fälle ein menschlicher Fehler unentdeckt geblieben wäre (Tab. 7).

Diskussion

Das Ziel der vorliegenden Studie bestand in der Evaluation der Güte des *ML* für die Auswertung von Aussagen auf offene Fragen zur Modellkompetenz (Williamson et al. 2012). Hierzu wurde die Übereinstimmung zwischen menschlicher und computerbasierter Kodierung sowie der Prozess des *ML* untersucht (Yang et al. 2002). Diese Evaluation liefert Hinweise dazu, inwiefern bestimmte Algorithmen für Teilkompetenzen der Modellkompetenz bereits zur Zweitkodierung oder zum unmittelbaren Feedback in Vermittlungssituationen (Williamson et al. 2012; Zhai et al. 2020) geeignet sind. Damit trägt diese Studie dazu bei, dem von Zhai et al. (2020) identifizierten Desiderat zu begegnen und ein wichtiges Zielkonstrukt naturwissenschaftlicher Bildung, Modellkompetenz, der computerbasierten Kodierung durch *ML* zugänglich zu machen. Im Folgenden werden zunächst methodische Aspekte und Limitationen angemerkt und anschließend die formulierten Forschungsfragen auf Basis der Ergebnisse diskutiert.

Methodische Aspekte

Grundsätzlich muss angemerkt werden, dass es deutlich mehr als die vier eingesetzten Algorithmen gibt (Mayfield und Rosé 2013). Unter den in *LightSide* angebotenen 39 weiteren Algorithmen konnte keiner gefunden werden, der erfolgreicher als die hier besprochenen vier trainiert werden konnte.

Im Datensatz wurden Rechtschreibfehler korrigiert (vgl. Ha et al. 2011). Darüber hinaus wurde keine Korrektur vorgenommen, das heißt es wurden keine Wörter (z. B. solche ohne nachvollziehbare Bedeutung) aus dem Datensatz entfernt. Grundsätzlich kann das Streichen bedeutungsloser Wörter den Trainingsprozess optimieren (Mayfield und Rosé 2013). Auf eine solche Prozedur wurde verzichtet, weil für *LightSide* kein automatisierter und objektiver Prozess in deutscher Sprache vorlag, der zum Beispiel konjugierte Verben und deklinierte Wortarten als ein Attribut berücksichtigen konnte oder Artikel entfernt. Sollten die Trainingsdaten allerdings für den Einsatz im Unterricht oder zur Beantwortung weitergehender Forschungsfragen eingesetzt werden, empfiehlt es sich, solch eine Überarbeitung von Datensätzen vorzunehmen (Ha et al. 2011).

In dieser Studie wurde *Cohens K* und der von Wirtz und Caspar (2002) beziehungsweise Williamson et al. (2012) vorgeschlagene Grenzwert für eine gute Beurteilerübereinstimmung genutzt. Alternativ dazu wird vorgeschlagen, die Übereinstimmung zwischen menschlicher und computerbasierter Kodierung mit Hilfe des quadrierten gewichteten Kappas zu prüfen (Liu et al. 2016; Mao et al. 2018;

Williamson et al. 2012). Davon wurde bei den hier vorliegenden drei ordinalen Niveaus, deren Distanz sich nicht metrisch ausdrücken lässt, abgesehen. Es gibt keinen theoretisch überzeugenden Grund, eine abweichende Kodierung einer Aussage in Niveau I und III stärker zu gewichten als Kodierungsdifferenzen zwischen den Niveaus I und II beziehungsweise II und III.

Die Verteilung aller Datensätze in den Niveaus zur Klassifikation unterschied sich nicht signifikant von den Verteilungen im Training (Tab. 2). Damit sind die Werte für *Cohens K* nicht durch Über- oder Unterrepräsentation einzelner Niveaus beeinflusst, die der Algorithmus gegebenenfalls schlechter oder besser klassifizieren kann. Dennoch, bei der Klassifikation wurden nur Biologie-Lehramtsstudierende befragt, also eine Teilpopulation der Befragten im Training. Es ist nicht auszuschließen, dass trotz vergleichbar akademisch und didaktisch geprägter Sprache aller beteiligten Personen die Auswertung davon beeinflusst ist.

Schließlich ist der Datensatz zum Training mit 356–632 Aussagen zum Beispiel verglichen mit der Studie von Ha et al. (2011) mit bis zu 1056 Kurzantworten eher klein. Allerdings konnte bei Ha et al. (2011) durch eine Verdoppelung von 500 auf gut 1000 Aussagen nur eine geringe Verbesserung der Güte des *ML* erzielt werden. Aus unseren Ergebnisse wird deutlich, dass kurze Antworten mit einer kleinen Anzahl verschiedener Wörter (*ZvM, TvM, ÄvM*; Tab. 2 und 3) bessere Werte für *Cohens K* beim Training und der Klassifikation erzielen als lange Antworten mit einer großen Anzahl verschiedener Wörter (*EvM, AM*; Tab. 2 und 3). Offensichtlich ist in der Diversität möglicher Antworten noch keine Sättigung erreicht. Werden zufällig nur vereinzelt in einer Antwort zu einem Niveau auftretende Attribute stark gewichtet, führt die Nutzung dieser Attribute in anderen Niveaus zu Fehlkodierungen. Das wird für die Algorithmen zusätzlich durch die Art der menschlichen Kodierung erschwert, in der nur das höchste Niveau, das in einer Aussage angesprochen wird, kodiert wird (Krell und Krüger 2016). Es ist demnach möglich, dass eine Antwort Aussagen zu allen Niveaus enthält. Es ist zu prüfen, ob die Erhöhung der Antworten zu einer größeren Übereinstimmung zwischen menschlicher und computerbasierter Kodierung führt. Eine bloße Erweiterung des Aussagenpools führt jedenfalls nicht automatisch zu besserem Training oder erfolgreicherer Klassifikation (Williamson et al. 2012; Yang et al. 2002). So verbesserte zum Beispiel das Hinzufügen der Klassifikationsaussagen zu den Trainingsaussagen die Trainingssätze nicht in der gewünschten Weise.

Übereinstimmung zwischen menschlicher und computerbasierter Kodierung (F1, F2)

Beim Training gelingt es immer einen der vier eingesetzten Algorithmen zu finden, dessen *Cohens K* die mindestens geforderte Übereinstimmung ($K = 0{,}60$; Wirtz und Caspar 2002) zwischen menschlicher und computerbasierter Kodierung erreicht. Die gewünschte Qualität von $K > 0{,}70$ (Williamson et al. 2012) wurde im Training für vier der fünf Teilkompetenzen erreicht (Tab. 3). Für die Klassifikation ist dieser Wert nur in der Teilkompetenz *ÄvM* erreicht, für *ZvM* noch für $K > 0{,}60$ (Tab. 3). Dabei ist der beste Trainings-Algorithmus immer *SVM*. Dennoch produziert *SVM* nur in den Teilkompetenzen *AM* und *ÄvM* die besten Klassifikations-Ergebnisse, sonst gelingt es mit *DTrees* und *LogReg* besser zu klassifizieren. Es ist somit kein Automatismus, dass gutes Training in guter Klassifikation mündet (Ha et al. 2011).

Es wird deutlich, dass die Wahl des genutzten Algorithmus für die Güte des *ML* von Aussagen zur Modellkompetenz eine Rolle spielt, wobei die Anzahl der zur Verfügung gestellten Attribute für die Algorithmen ohne Bedeutung bleibt, wie sich an *DTrees* und *SVM* bei vielen als auch wenigen Attributen zeigt (Tab. 3).

Übereinstimmung zwischen relevanten Attributen und dem Kodierleitfaden sowie der Antwortlänge (F3, F4)

Die identifizierten, stark von den Algorithmen gewichteten Wort-Attribute decken sich weitgehend mit dem Kodierleitfaden und sind auch Schlüsselwörter für die menschliche Kodierung (Krell und Krüger 2016). Die Auflistung (Tab. 4) macht deutlich, dass bisher weder Artikel, Personalpronomen noch Konjunktionen oder Deklinationen aus dem Trainings-Datensatz entfernt wurden und noch keine Reduktion auf eine grammatikalische Variante eines Begriffs erfolgen konnte. Hierfür müsste eine lexikonartige Datenbank genutzt werden, die diesen Prozess automatisch übernimmt. *LightSide* bietet eine solche Datenbank im Englischen an (Mayfield und Rosé 2013), für andere Sprachen fehlt sie. Hierzu werden zurzeit Folgestudien in spanischer und englischer Sprache durchgeführt, die einerseits den Einfluss der Sprachen auf die Auswertung mit *LightSide* prüfen und im Falle englischer Datensätze die Effekte einer Datenbereinigung mit den vorliegenden Datenbanken untersuchen.

Die Analyse der fehlerhaften Kodierungen gibt Aufschluss über deren Ursache. Sobald in Niveau I-Aussagen stark gewichtete Wörter (Tab. 4) wie „Hypothese“ oder „Theorie“ vorkommen, folgt durch die Algorithmen eine Kodierung in Niveau III. Da bei der menschlichen Kodierung immer das höchste angesprochene Niveau kodiert wird, erkennen die Algorithmen außerdem in Antworten auf Niveau I nicht kurze Antwortteile auf Niveau III, insbesondere dann, wenn zusätzlich hoch gewichtete Attribute niedriger Niveaus (z. B. das Attribut *falsch* im Niveau I) doppelt benutzt werden. Hier fehlt den Algorithmen, die auf der Basis von Wort- oder Buchstabenfolgen kodie-

ren, das semantische Verständnis insbesondere in längeren Textpassagen (Lintean et al. 2008).

Die zum Teil hohen Korrelationen zwischen dem kodierten Niveau und der Antwortlänge bei *NBayes* zeigen, dass die externe Variable diese Kodierung beeinflusst. Bei der unabhängigen Betrachtung der Attribute durch *NBayes* klärt die Antwortlänge bis zu 34 % der Varianz der Kodierung mit *NBayes* auf (Bestimmtheitsmaß r^2 bei $r = 0{,}58$; Tab. 6). Offensichtlich gelingt es bei langen Antworten nicht, durch die Auswahl weniger hoch gewichteter Attribute wiederholte Niveau I- oder II-Aussagen von Niveau III-Aussagen zu differenzieren. *NBayes* erweist sich in allen Datensätzen als nicht geeignet (Mao et al. 2018).

Schließlich ließe sich mit den Attributen, die die Algorithmen für die Niveaus nutzen, ein neuer Typus von Aufgaben entwickeln (Kim et al. 2017). Dieser Aufgabentyp stellt Wörter zur Verfügung, die zu sinnvollen Sätzen kombiniert werden müssen. Die Herausforderung der Aufgabenentwicklung ist, neben einer großen Kombinationsmöglichkeit der Wörter auch Antworten auf verschiedenen Niveaus gleichzeitig zu erlauben. Der Ansatz ist vielversprechend für unmittelbares Feedback in Vermittlungssituationen, weil er die Sättigungsproblematik des aktuellen Datensatzes umgeht. Das setzt voraus, dass mit diesen Aufgaben wie bei Kim et al. (2017) ein hohes *Cohens K* ($K > 0{,}80$) beim Vergleich der menschlichen und computerbasierten Kodierung erreicht wird.

Nützlichkeit des maschinellen Lernens (F5)

Es sollte die Nützlichkeit des *ML* für einen fachdidaktischen Einsatz geprüft werden (Williamson et al. 2012). Im Prinzip reicht der eingesetzte Datensatz aus, *ML* auf dem Niveau einer menschlichen Zweitkodierung zu entwickeln. Als Vorteil ergibt sich, dass circa 63 % der Aussagen kein zweites Mal kodiert werden müssten (Tab. 7). Zusätzlich führt die Überprüfung nicht übereinstimmender Kodierungen dazu, menschliche Fehler in durchschnittlich 5 % der Aussagen zu entdecken. Es bleibt allerdings auch das Risiko, dass 3 % menschliche Fehler unentdeckt bleiben, wenn sowohl *ML* als auch Mensch gleichzeitig fehlerhaft kodieren (Tab. 7); was allerdings auch bei menschlicher Zweitkodierung auftreten kann. In den vorliegenden Datensätzen besitzen die Werte für *Cohens K* bereits eine ausreichende Qualität, um die computerbasierte Kodierung in Forschungszusammenhängen einzusetzen. Dies ist auch in anderen Studien mit Kurzantworten, zum Beispiel im Bereich von Evolution, gelungen (Ha et al. 2011). Insbesondere für Differenzierungsmaßnahmen im Biologieunterricht ist die computergestützte Auswertung schnell zu erzielen und wäre damit beispielsweise für didaktisch motivierte Gruppeneinteilungen in Unterrichtssituationen zu nutzen. Während in Forschungszusammenhängen jede Fehlkodierung bedeutsam ist, wären im Unterricht Unterschätzungen für eine Benotung und Überschätzungen in einer Intervention besonders kritisch (Lee et al. 2019). Es kann allerdings vermutet werden, dass im Rahmen einer Intervention eine inhaltlich motivierte Differenzierung auf der Basis des hier durchgeführten *ML* zu einer begründeteren Gruppeneinteilung führt als eine Einteilung durch eine Lehrperson ohne Datenbasis.

Zusammenfassend lässt sich feststellen, dass die Güte des *ML* von Aussagen zur Modellkompetenz bereits ausreichend erscheint, um für eine Zweitkodierung eingesetzt zu werden. Dies ist vor dem Hintergrund der festgestellten Relevanz von Kompetenzerfassung und der entsprechend notwendigen Entwicklung von Verfahren zur Kompetenzerfassung sowie zur Interpretation von erzielten Ergebnissen (Klieme et al. 2008; Leutner et al. 2017) vielversprechend. Für Modellkompetenz ließen sich dementsprechend in Forschungsvorhaben die Vorteile offener Aufgabenformate (Abb. 1) nutzen und gleichzeitig die zeitaufwendige Zweitkodierung computerbasiert gestalten. Es muss einschränkend berücksichtigt werden, dass die Nutzung computerbasierter Verfahren für die interpretative Auswertung von Aussagen auf Fragen im offenen Format grundsätzlich ein Abtreten von Autonomie und Kontrolle transportiert, insbesondere wenn die Klassifikationsprozesse der genutzten Algorithmen nicht vollständig nachvollzogen werden können. Für die von Williamson et al. (2012) und Zhai et al. (2020) genannten weiteren Einsatzmöglichkeiten des *ML* zum Beispiel in Vermittlungssituationen muss daher umfassende Evidenz für die Güte des *ML* im vorgesehenen Anwendungskontext vorliegen. Hierzu gehört zunächst die Prüfung, inwieweit sich Schüleraussagen von den hier gesammelten Daten von (angehenden) Lehrpersonen und Expert*innen unterscheiden. Ferner gilt es, den Prozess der automatisierten Kodierung besser zu verstehen (Yang et al. 2002), wozu eine tiefere (d. h. qualitative) Betrachtung der nicht übereinstimmenden Kodierungen unter Berücksichtigung der durch die Algorithmen besonders stark gewichteten Attribute beitragen kann.

Funding Open Access funding enabled and organized by Projekt DEAL.

Anhang

Naïve Bayes (NBayes) wird typischerweise zur Kategorisierung von Texten genutzt (z. B. Spam-Erkennung). Er beruht auf der Annahme, dass die Attribute unabhängig voneinander auftauchen. Der Algorithmus berechnet die Wahrscheinlichkeit des Auftretens jedes Attributs, anschließend in jeder Klasse. Attribute mit der höchsten Wahrscheinlichkeit in einer Klasse werden ausgewählt, um Zuordnungen

vorzunehmen (Mayfield und Rosé 2013). Der Kern der Berechnungen basiert auf dem Bayes-Theorem, nach dem die Wahrscheinlichkeit von A ermittelt wird, wenn B aufgetreten ist. *NBayes* klassifiziert jedes Objekt in die Klasse (hier: Niveaus von Modellkompetenz), zu der es mit der größten Wahrscheinlichkeit gehört. Der größte Nachteil dieser Klassifikation besteht in der Annahme, dass die Attribute unabhängig sein müssen. In den meisten Fällen ist dies nicht der Fall, was die Leistung des Algorithmus beeinträchtigt (Han et al. 2011; Witten et al. 2011).

Logistic regression (*LogReg*) ist ein statistisches Verfahren, mit dem der Wert einer logarithmierten Zielvariable in Abhängigkeit von einer oder mehreren unabhängigen Variablen geschätzt wird. *LogReg* wird oft für binäre Prognosen genutzt, weil das Ergebnis der Regressionsfunktion nur Werte im Intervall [0,1] annehmen kann. Dazu wird jedes Attribut aus dem Trainingsdatensatz gewichtet. Ziel ist es, die Gewichtungen so zu wählen, dass der Fehler zwischen Schätzwert und echtem Wert aller Objekte (Aussagen) möglichst klein ist. Bei der Klassifikation werden die Aussagen aufgrund der festgelegten Gewichtungen der Attribute einem der drei Niveaus zugeteilt (Han et al. 2011; Witten et al. 2011).

Support vector machines (*SVM*) transformiert die Trainingsdatensätze in Vektoren und wählt aus einer kleinen Anzahl bedeutsame Grenzinstanzen zwischen jeweils zwei Klassen aus (Mayfield und Rosé 2013). Zwischen diesen wird eine lineare Diskriminanzfunktion bestimmt, die mit dem größtmöglichen Abstand in einem Hyperraum zwischen zwei Klassen trennt. Jeder Vektor wird in einer Hyperebene kartiert und der Algorithmus sucht nach den Vektoren einer Klasse, die am nächsten an den Vektoren der anderen Klasse liegen. Diese Vektoren sind die Stützvektoren und sie dienen zur Grenzziehung für die Hyperebene. Je größer der Abstand zwischen zwei Klassen, desto geringer ist die Fehlerrate. *SVM* sucht bei der Transformation der Daten nach der Gewichtung einzelner Attribute, so dass der Abstand zwischen den Klassen (hier: den drei Niveaus) maximiert wird. Der Algorithmus entscheidet bei neuen Daten, je nachdem auf welche Seite der Hyperebene der neue Datensatz liegt (Han et al. 2011; Witten et al. 2011).

Decision trees (*DTrees*) startet mit einem Wurzelknoten, an dem noch alle Datensätze gebündelt und unsortiert vorliegen. Um die Datensätze zu sortieren, müssen Entscheidungskriterien gefunden werden. Für jede Entscheidung durch ein Kriterium wird vom Knoten aus ein neuer Zweig gebildet. Von jedem neuen Knoten aus wächst durch ein neues Entscheidungskriterium der Entscheidungsbaum. Sollten an Zweigen neue Knoten entstehen, an denen keine weitere Aufspaltung möglich ist, weil entweder nur noch ein Datensatz vorhanden ist oder kein Kriterium gefunden werden kann, um die Datensätze weiter aufzuteilen, so endet an diesem Knoten das Wachstum in einem sogenannten Blatt. Diese Blätter repräsentieren die Klassen, hier die drei Niveaus (Han et al. 2011; Witten et al. 2011).

Literatur

Abikoye, Omokanye, & Aro (2018). Text classification using data mining techniques: a review. *Computing and Information Systems Journal*, *22*(2), 1–8.

American Educational Research Association, & American Psychological Association & National Council on Measurement in Education [AERA, APA & NCME] (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bailer-Jones, D. (2003). When scientific models represent. *International Studies in the Philosophy of Science*, *17*, 59–74.

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: how closely do they match clinical interview performance? *Journal of Science Education and Technology*, *23*(1), 160–182.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*, 1.

Buxmann, P., & Schmidt, H. (Hrsg.). (2018). *Künstliche Intelligenz. Mit Algorithmen zum wirtschaftlichen Erfolg*. Wiesbaden: Springer Gabler.

Chiu, M. H., & Lin, J. W. (2019). Modeling competence in science education. *Disciplinary and Interdisciplinary Science Education Research*, *1*(1), 1–11.

Crawford, B. A., & Cullin, M. J. (2005). Dynamic assessments of preservice teachers' knowledge of models and modelling. In K. Boersma, M. Goedhart, O. de Jong & H. Eijkelhof (Hrsg.), *Research and the quality of education* (S. 309–323). Dordrecht: Springer Netherlands.

Dam, & Kaufmann (2008). Computer assessment of interview data using latent semantic analysis. *Behavior Research Methods*, *40*, 8–20.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Eid, M., Gollwitzer, M., & Schmitt, M. (2011). *Statistik und Forschungsmethoden Lehrbuch*. Weinheim: Beltz.

Giere, R., Bickle, J., & Mauldin, R. (2006). *Understanding scientific reasoning*. London: Thomson Learning.

Gilbert, J. K., & Justi, R. (2016). *Modelling-based teaching in science education*. Cham: Springer.

Gogolin, S., & Krüger, D. (2018). Students' understanding of the nature and purpose of models. *Journal of Research in Science Teaching*, *55*(9), 1313–1338.

Göhner, M., & Krell, M. (2018). Modellierungsprozesse von Lehramtsstudierenden der Biologie. *Erkenntnisweg Biologiedidaktik*, *17*, 45–61.
Gouvea, J., & Passmore, C. (2017). 'Models of' versus 'Models for. *Science & Education*, *26*(1–2), 49–63.
Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: conceptions of middle and high school students and experts. *Journal of Research in Science Teaching*, *28*, 799–822.
Grünkorn, J., Upmeier zu Belzen, A., & Krüger, D. (2014). Assessing student's understandings of biological models and their use in science to evaluate theoretical framework. *International Journal of Science Education*, *34*(10), 1651–1684.
Günther, S. L., Fleige, J., Upmeier zu Belzen, A., & Krüger, D. (2019). Using the case method to foster preservice biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK) related to models and modeling. *Journal of Science Teacher Education*, *30*(4), 321–343.
Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *Life Sciences Education*, *10*, 379–393.
Han, J., Kamber, M., & Pei, J. (2011). *Data mining—concepts and techniques* (3. Aufl.). Heidelberg: Elsevier.
Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Heidelberg: Springer.
Kim, K. J., Pope, D. S., Wendel, D., & Meir, E. (2017). WordBytes: exploring an intermediate constraint format for rapid classification of student answers on constructed response assessments. *Journal of Educational Data Mining*, *9*(2), 45–71.
Kind, P., & Osborne, J. (2017). Styles of scientific reasoning: a cultural rationale for science education? *Science Education*, *101*, 8–31.
Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 3–22). Göttingen: Hogrefe.
KMK (Hrsg.). (2019). Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf. Zugegriffen: 23.09.2020.
KMK [Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der BRD] (Hrsg.). (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*. München: Wolters Kluwer. http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf
Krell, M. (2013). *Wie Schülerinnen und Schüler biologische Modelle verstehen: Erfassung und Beschreibung des Modellverstehens von Schülerinnen und Schülern der Sekundarstufe I*. Dissertation. Berlin: Logos.
Krell, M., & Krüger, D. (2016). Testing models: A key aspect to promote teaching-activities related to models and modelling in biology lessons? *Journal of Biological Education*, *50*, 160–173.
Krell, M., Reinisch, B., & Krüger, D. (2015). Analyzing students' understanding of models and modeling referring to the disciplines biology, chemistry, and physics. *Research in Science Education*, *45*, 367–393. https://doi.org/10.1007/s11165-014-9427-9.
Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2016). Modellkompetenz im Biologieunterricht. In A. Sandmann & P. Schmiemann (Hrsg.), *Biologiedidaktische Forschung: Schwerpunkte und Forschungsgegenstände* (S. 83–102). In: Logos.
Krüger, D., Kauertz, A., & Upmeier zu Belzen, A. (2018). Modelle und das Modellieren in den Naturwissenschaften. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 141–157). Berlin: Springer.
Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education*, *8*(1), 55–73. https://doi.org/10.1111/dsji.2010.8.issue-1.
Lead States, N. G. S. S. (Hrsg.). (2013). *Next generation science standards: for states, by states*. Washington, DC: The National Academies Press.
Lee, H.-S., Pallant, A., Pryputniewicz, S., & Lord, T. (2019). Automated text scoring and real-time adjustable feedback: supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*, 590–622.
Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy. In K. Renninger, W. Damon, I. Sigel & R. Lerner (Hrsg.), *Handbook of child psychology* (S. 153–196). Hoboken: Wiley.
Leutner, D., Fleischer, J., Grünkorn, J., & Klieme, E. (2017). Competence assessment in education: an introduction. In D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (Hrsg.), *Competence assessment in education* (S. 1–6). Cham: Springer.
Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. *Science*, *344*(6180), 155–156.
Lintean, M., Rus, V., & Azevedo, R. (2008). Automatic detection of student mental models based on natural language student input during metacognitive skill training. *International Journal of Artificial Intelligence in Education*, *21*(3), 169–190.
Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*, 19–28.
Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments: automated scoring of science assessment. *Journal of Research in Science Teaching*, *53*, 215–233.
Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121–138.
Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*, 207–218.
Mathesius, S., & Krell, M. (2019). Assessing model competence with questionnaires. In A. Upmeier zu Belzen, D. Krüger & J. H. van Driel (Hrsg.), *Towards a competence-based view on models and modeling in science education* (S. 117–129). Berlin: Springer.
Mayfield, E., & Rosé, C. P. (2013). LightSIDE: Open source machine learning for text. In M. Shermis & J. Burstein (Hrsg.), *Handbook of automated essay evaluation* (S. 146–157). New York,: Routledge.
Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, *7*, 15.
Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *Journal of Science Education and Technology*, *21*(1), 56–73.
Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183–196.
Neuhaus, B., & Braun, E. (2007). Testkonstruktion und Testanalyse – praktische Tipps für empirisch arbeitende Didaktiker und Schulpraktiker. In H. Bayrhuber, D. Elster, D. Krüger & H. Vollmer (Hrsg.), *Kompetenzmessung und Assessment* (S. 135–164). Innsbruck: Studienverlag.
Nicolaou, C. T., & Constantinou, C. P. (2014). Assessment of the modeling competence: a systematic review and synthesis of empirical research. *Educational Research Review*, *13*(3), 52–73.
Nielsen, S. S., & Nielsen, J. A. (2019). A competence-oriented approach to models and modelling in lower secondary science

education: practices and rationales among Danish teachers. *Research in Science Education*. https://doi.org/10.1007/s11165-019-09900-1.

Odenbaugh, J. (2005). Idealized, inaccurate but successful: a pragmatic approach to evaluating models in theoretical ecology. *Biology & Philosophy*, *20*, 231–255.

Prenzel, M., Häußler, P., Rost, J., & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, *30*, 120–135.

Reinisch, B., & Krüger, D. (2018). Preservice biology teachers' conceptions about the tentative nature of theories and models in biology. *Research in Science Education*, *48*, 71–103. https://doi.org/10.1007/s11165-016-9559-1.

Russel, S. J., & Novig, P. (2010). *Artificial intelligence. A modern approach* (3. Aufl.). Hoboken: Pearson.

Schwarz, C., & White, B. (2005). Metamodeling knowledge: developing students' understanding of scientific modeling. *Cognition and Instruction*, *23*(2), 165–205.

Schwarz, C., Reiser, B. J., Acher, A., Kenyon, L., & Fortus, D. (2012). MoDeLS: Challenges in defining a learning progression for scientific modeling. In A. C. Alonso & A. W. Gotwals (Hrsg.), *Learning progressions in science: current challenges and future directions* (S. 101–137). The Netherlands: Sense Publishers.

Slepcevic-Zach, P., & Tafner, G. (2012). Input – Output – Outcome: Alle reden von Kompetenzorientierung, aber meinen alle dasselbe? In M. Paechter, M. Stock, S. Schmölzer-Eibinger, P. Slepcevic-Zach & W. Weirer (Hrsg.), *Pädagogik. Handbuch Kompetenzorientierter Unterricht*. Weinheim, Basel: Beltz.

Terzer, E., Hartig, J., & Upmeier zu Belzen, A. (2013). Systematische Konstruktion eines Tests zu Modellkompetenz im Biologieunterricht unter Berücksichtigung von Gütekriterien. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 51–76.

Thoma, G.-B., & Köller, O. (2018). Test-wiseness: Ein unterschätztes Konstrukt? Empirische Befunde zur Überprüfung und Erlernbarkeit von test-wiseness. *Zeitschrift für Bildungsforschung*, *8*, 63–80.

Upmeier zu Belzen, A., & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 41–57.

Upmeier zu Belzen, A., & Krüger, D. (2019a). Modelle und Modellieren im Biologieunterricht: Ein Fall für Erkenntnisgewinnung. *Unterricht Chemie*, *171*, 38–41.

Upmeier zu Belzen, A., & Krüger, D. (2019b). Modelle als methodische Werkzeuge begreifen und nutzen: Empirische Befunde und Empfehlungen für die Praxis. In J. Groß, M. Hammann, P. Schmiemann & J. Zabel (Hrsg.), *Biologiedidaktische Forschung: Erträge für die Praxis* (S. 129–146). Berlin: Springer.

Upmeier zu Belzen, A., van Driel, J., & Krüger, D. (2019). Introducing a framework for modeling competence. In A. Upmeier zu Belzen, D. Krüger & J. van Driel (Hrsg.), *Towards as competence-based view on models and modeling in science education* (S. 3–19). Cham: Springer.

Williamson, D., Xi, X., & Breyer, F. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*, 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität [Interrater-accordance and interrater-reliability]*. Göttingen: Hogrefe & Huber.

Witten, I. H., Frank, E., & Hall, M. (2011). *Data mining: practical machine learning tools and techniques* (3. Aufl.). Heidelberg: Elsevier.

Yang, Y., Buckendahl, C., Juszkiewicz, P., & Bhola, D. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, *15*, 391–412. https://doi.org/10.1207/S15324818AME1504_04.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(39), 1–27.

Zhai, X., Yin, Y., Pellegrino. J. W., Haudek, K. C., Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, *56*(1), 111–151.

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648–1668.