Entwicklung und Reflexion einer Unterrichtssequenz zum Maschinellen Lernen als Aspekt von Data Science in der Sekundarstufe II

Simone Opel, Michael Schlichtig, Carsten Schulte[1], Rolf Biehler, Daniel Frischemeier, Susanne Podworny, Thomas Wassong[2]

Abstract: Die Bereiche „Data Science“ und „Big Data“ sowie ihre technischen, ethischen und gesellschaftlichen Auswirkungen werden zunehmend nicht nur in der Wissenschaft, sondern auch in diversen Medien diskutiert und somit verstärkt auch zu einem wichtigen Thema für alle. Um den Schülerinnen und Schülern der Sekundarstufe II einen theoretisch und fachwissenschaftlich fundierten Einstieg in diesen Themenbereich zu ermöglichen, wurde ein erster Entwurf eines interdisziplinären Curriculums entwickelt, das neben fachlichen Aspekten von Data Science einen Fokus auf sich hieraus ergebende gesellschaftliche Fragestellungen legt. Es werden neben der Konzeption des Kurses die bisherigen Erfahrungen aus der Durchführung – insbesondere in Hinsicht der darin enthaltenen Unterrichtseinheit zum Maschinellen Lernen - berichtet, sowie die sich hieraus ergebenden Implikationen für die Weiterentwicklung dargestellt und diskutiert.

Keywords: Data Science; Maschinelles Lernen; KI; Künstliche Neuronale Netze; Entscheidungsbäume; Big Data; Curriculum

1 Einleitung

Data Science sowie Fragen zu Maschinellem Lernen (ML) und Künstlicher Intelligenz (KI) sind inzwischen in großem Maß in verschiedene Systeme implementiert und werden in allen Bereichen intensiv diskutiert. Auch die gesellschaftliche Diskussion über Funktion, Nutzen und Gefahren dieser Systeme nimmt inzwischen viel Raum ein. Es erscheint uns wichtig, dass diese Fragestellungen in den Unterricht integriert werden. Daher entwickelten wir auf Basis curricularer Ideen aus der Informatik und Mathematik einen ersten Entwurf eines Curriculums für die Sekundarstufe II und führten dies in ein konkretes Unterrichtskonzept über, das im Rahmen des durch die Deutsche Telekom Stiftung ermöglichten Projekts *„ProDaBi[3]“* mit einem Oberstufenkurs erprobt wird. Im Rahmen dieses Artikels stellen wir neben diesem Kurs auch erste Erfahrungen der Umsetzung – insbesondere der Einheiten aus KI und ML – sowie die hinter dem Entwicklungsprozess liegenden Ideen vor.

[1] Universität Paderborn, Didaktik der Informatik, Fürstenallee 11, 33102 Paderborn, vorname.nachname@uni-paderborn.de
[2] Universität Paderborn, Didaktik der Mathematik, Warburger Str. 100, 33098 Paderborn, nachname@math.uni-paderborn.de
[3] ProDaBi – Projekt Data Science und Big Data in der Schule, Projektwebseite: https://www.prodabi.de

2 Data Science und Maschinelles Lernen – Aspekte für die Bildung

Data Science, ML und der Umgang mit Big Data geht weit über technische und wissenschaftliche Aspekte unterschiedlicher Disziplinen hinaus und trägt auch ethische, gesellschaftliche und soziale Auswirkungen in sich – daher werden hier nicht umsonst sehr verschiedene Kontexte diskutiert. Um möglichst viele Aspekte zu verstehen und ein gemeinsames Verständnis von *Data Science* zu entwickeln, wurden im Rahmen eines interdisziplinären, internationalen Symposiums[4] [Pa18] diese Aspekte diskutiert und zwei für uns curricular relevante Bereiche wurden auf Basis der dort geführten Diskussionen identifiziert: Für die *Informatikdidaktik* sind das neben der Entwicklung von Computational Thinking [TD16] insbesondere die Auswirkungen verschiedener Bereiche der Mensch-Maschine-Interaktion auf die Gesellschaft – und damit auch unser Umgang mit Big Data sowie den Methoden und Auswirkungen von Data Science [SBS18]. Im Bereich der *Statistikdidaktik* sollten die fundamentalen Ideen der Statistik um verschiedene Aspekte der *statistischen Kompetenz* [Ri16] erweitert werden. Daher stehen diese Ideen und Aspekte im Zentrum der Entwicklungen.

2.1 Daten und Datenprozesse als strukturgebende Komponenten

Im Gegensatz zur praktischen Informatik stehen im Bereich der *Data Science* nicht die Entwicklung von algorithmischen Strukturen sowie deren (algorithmische) Modellierung und Implementierung im Vordergrund, sondern der Umgang mit *Daten*. Das heißt, die Komplexität der Fragestellungen wird nicht nur durch die eingesetzten Algorithmen bestimmt, sondern wird in hohem Maße durch die Daten und die in ihnen implizit und explizit enthaltenen Informationen beschrieben. Nach der Definition der Empfehlungen GI zu den Bildungsstandards in der Sekundarstufe sind „Daten eine Darstellung von Information in formalisierter Art [...]. Daten werden wieder zu Information, wenn sie in einem Bedeutungskontext interpretiert werden“ ([Rö16], S. 9). Ein Informatiksystem verarbeitet somit nur Daten, die darin enthaltene Information wird durch Interpretation durch den Menschen gewonnen. Daten sind also „nicht nur Zahlen, sie sind Zahlen mit einem Kontext“ [CM97]. Es stellt sich die Frage, wie Schülerinnen und Schüler *Datenkompetenz* erwerben können, die Voraussetzung für einen kompetenten Umgang mit fehlerbehafteten oder unterschiedlich strukturierten Daten innerhalb eines Kontextes ist. Ridsdale et al. definieren *Datenkompetenz* als die prozessorientierte „Fähigkeit, Daten kritisch zu sammeln, zu analysieren, zu bewerten und anzuwenden“ ([Ri15], S. 3). Damit liegt der Schwerpunkt der Handlungen der Lernenden im Umgang mit verschiedenen Daten, so dass „Prozessmodelle zur Datenanalyse“ einen sinnvollen Ausgangspunkt zur Planung von unterrichtlichem Handeln bieten. Der Definition von Ridsdale et al. ähnelt das CRISP-DM-Modell[5] [Ch00], das ein vollständiges Verfahren zum Umgang mit Daten

[4] „Perspectives for data science education at school level – Educational contributions from statistics, computer science and sociocultural studies“; `http://go.upb.de/SymposiumProDaBi`

[5] CRISP-DM = **Cr**oss-**Industry S**tandard **P**rocess for **Data Mining**, Phasen: Verstehen der Fragestellung – Verstehen der Daten – Aufbereitung der Daten – Modellbildung – Bewertung des Modells – Einsatz

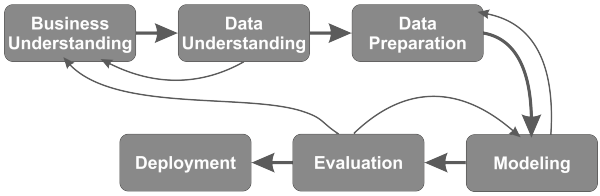

Abb. 1: Das CRISP-DM-Modell als Standardprozessmodell für Data Mining

beschreibt (vgl. Abb. 1). Dieses Prozessmodell erscheint uns einen sinnvollen Rahmen zur Erarbeitung eines Curriculums und entsprechenden Unterrichtsmaterials zu bieten, so dass entschieden wurde, das CRISP-DM-Modell als Basis weiterer Entwicklungen zu verwenden, da dieses Modell den Datenprozess umfassender darstellt als zum Beispiel der ebenfalls häufig verwendete PPDAC-Zyklus[6].

2.2 Relevanz sozialer und gesellschaftlicher Aspekte

Gesellschaftliche und soziale Implikationen besitzen eine große Relevanz im Bereich der *Data Science* (vgl. Kapitel 1). Diskussionen über gesellschaftliche Aspekte im Informatikunterricht werden – wenn überhaupt – meist nur entkoppelt von technologischen Fragen der Unterrichtsinhalte geführt und sind nicht fest in die Arbeit in Softwareprojekten und Lernaufgaben integriert. Auch eines der wenigen Data Science-Curricula und Kompetenzmodell für das schulische Umfeld [GR] stellt die verschiedenen Aspekte von Data Science zwar umfassend, aber im Wesentlichen aus fachwissenschaftlicher Sicht dar. Daher stellt die Frage, wie diese Aspekte integraler Bestandteil des Unterrichtsmaterials und des Curriculums werden können, eine wichtige, schrittweise zu lösende Herausforderung dar.

2.3 Auf dem Weg zum Data Science-Kurs – Didaktische Ansätze

Eine wichtige Erkenntnis der Vorarbeiten war, dass – mehr als in der praktischen Informatik – eine gemeinsame Sicht auf Daten sowie die Entwicklung eines gemeinsamen Grundverständnisses aller Begriffe und Verfahrennotwendig ist (vgl. [Tv09]) – ein auch während der ersten Durchführung des Projektkurses nicht abgeschlossener Schritt. Weiterhin wurden schon existierende Materialien und Curricula auf ihre Einsatzmöglichkeit hin evaluiert. Allerdings sind sämtliche, meist hochschulische Curricula gut strukturiert, aber nur auf technologische Aspekte fokussiert. Daher ergibt sich die Notwendigkeit, diese Materialien selbst zu entwickeln und die relevanten Inhalte aus den Beiträgen des in Kap. 1 erwähnten Symposiums zu generieren. Das Auffinden relevanter Information aus Data Mining und Statistik auf der einen Seite, sowie die notwendigen Kompetenzen aus dem Bereich des ML

[6] PPDAC = **P**roblem – **P**lan – **D**aten – **A**nalyse – K(**C**)onclusion

sowie dem Design Künstlicher Neuronaler Netzwerke (KNN) auf der anderen Seite, ist dank der breiten Basis aller analysierten Unterlagen relativ einfach. Schwieriger gestaltet sich das Einbinden der gesellschaftlichen, sozialen und interdisziplinären Fragestellungen (vgl. Abschnitt 2.2). Daher entschieden wir uns, in diesem Projekt den Ansatz des „*Design-Based Research*“ (DBR) [Co03] zu verfolgen: Ausgehend von einem ersten Entwurf eines Data Science-Curriculums wird ein hierauf basierender Kurs entwickelt, durchgeführt und evaluiert. Aus den Erkenntnissen der Durchführung wird der Kurs und damit auch das anfangs noch sehr skizzenhafte Curriculum in mehreren Zyklen weiterentwickelt. Die im folgenden beschriebene Version des Data Science-Kurses wurde unter Verwendung dieses Ansatzes entwickelt und wird als sog. „Projektkurs“ in Kooperation mit einem Gymnasium vor Ort während des SJ 2018/19 erprobt und evaluiert.

3 Der Data Science-Kurs

Der so entstandene Kurs ist modular angelegt, wobei die beiden Bereiche „Data Mining und Statistik“ sowie „Künstliche Intelligenz und Maschinelles Lernen“ klar abgegrenzt sind und dem CRISP-DM-Modell (vgl. Abschnitt 2.1) folgend aufeinander aufbauend gestaltet werden. Zum Erwerben von Kompetenz, zumindest einfache Data Science Projekte selbst durchzuführen, wird als drittes Modul ein Projektmodul entwickelt und durchgeführt, so dass der Kurs in seiner ersten Version aus drei Modulen besteht:

1. *Von Daten zu Informationen:* Dieses Modul ist eine Einführung in Data Science und den Umgang mit Big Data und zielt darauf ab, das statistische Denken zu verbessern und Datenkompetenz zu entwickeln. Es werden dabei statistische Methoden auf Daten zur Informationsgewinnung angewendet, die Erkenntnisse reflektiert und ihre Aussagekraft kriterial diskutiert.

2. *Künstliche Intelligenz und Maschinelles Lernen:* Im Rahmen dieses Moduls lernen die Schülerinnen und Schüler zwei unterschiedliche Methoden kennen, Erkenntnisse aus Daten zu gewinnen, indem sie exemplarisch Entscheidungsbäume als Vertreter einer Symbolischen KI und KNN (hier Back-Probagation-Netze) als typischen Vertreter überwachten Lernens kennenlernen, analysieren und auf eigene Beispiele anwenden. Ziel ist, nicht nur die informatischen Aspekte von KI kennenzulernen, sondern auch ihre Erkenntnisse auf vorhandene Systeme anzuwenden und deren Grenzen, Chancen und Risiken zu diskutieren.

3. *Datenprojekte:* Im Rahmen der Durchführung von Datenprojekten können die Schülerinnen und Schüler ihre Kompetenzen einsetzen, um reale Fragestellungen zu bearbeiten. Dabei werden sie motiviert, ihr Vorgehen im Sinne des CRISP-DM-Modells zu planen, um so zu für sie optimalen Ergebnissen kommen zu können und diese auch am Ende zu präsentieren und deren gesellschaftlichen Implikationen diskutieren zu können.

Es wurde bei der Entwicklung und Umsetzung immer auf eine enge Verflechtung der fachlichen Inhalte mit gesellschaftlichen und sozialen Aspekten geachtet, um eine mehrdimensionale und interdisziplinäre Sicht auf alle Aspekte der Themen zu erhalten.

3.1 Modul 1: Von Daten zu Informationen – Datendetektive

Folgend dem in Abschnitt 2.1 beschriebenen CRISP-DM-Modell ist für ein Vorhaben aus dem Bereich Data Science zunächst wichtig, die Fragestellung zu verstehen („Business Understanding“), entweder sinnvolle Daten zu erheben oder schon gesammelte Daten zu verstehen („Data Understanding“), und diese anschließend zu analysieren und aufzubereiten („Data Preparation“). Inspiriert von diesem Zyklus wird in diesem Modul (insg. 7 Wo.) im ersten Baustein zunächst die Verwendung von großen und offen verfügbaren Datenmengen diskutiert und anschließend an Hand von „Lärmdaten“ (vgl. Abb. 2) erprobt, derartige Daten selbst zu analysieren, unter Verwendung von Jupyter Notebooks[7] aufzubereiten und zu visualisieren. Im zweiten Baustein führen die Schülerinnen und Schüler explorative Datenanalysen unter Verwendung des multivariaten JIM-Datensatzes mit Hilfe des Online-Tools CODAP[8] durch und präsentieren und diskutieren ihre Erkenntnisse.

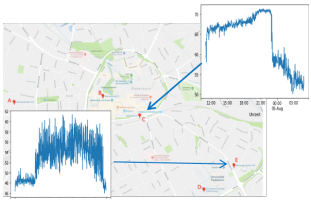

Abb. 2: Von den Schülerinnen und Schülern unter Verwendung von Python aufgearbeitete und verschiedenen Orten zugeordnete Lärmprofile

3.2 Modul 2: Künstliche Intelligenz und Maschinelles Lernen

Betrachtet man den CRISP-DM-Zyklus weiter, so folgt als nächster Schritt das Entwickeln eines Modells („Modeling“). Hier erkennen die Lernenden die Unterschiede zwischen klassischen algorithmischen Problemlösungsverfahren und datengetriebenen Prozessabläufen am Beispiel des ML (vgl. Abb. 3) und können die damit verbundene Rolle des Menschen innerhalb dieser Mensch-Maschine-Interaktion diskutieren und reflektieren (insg. 7 Wo.). Aus der Menge von Verfahren wurden im Vorfeld wichtige exemplarisch ausgewählt und in zwei Bausteinen mit den Schülerinnen und Schülern bearbeitet. Im ersten Baustein erwerben die Schülerinnen und Schüler unter Verwendung des „Sweet Learning Computers“ [Cu16b], einer Unplugged-Aktivität, ein grundsätzliches Verständnis von ML und diskutieren auf Basis dieses so erarbeiteten Wissens über aktuelle und zukünftige Chancen und Risiken dieser Technologien sowie ihren vielfältigen Einsatz. Dies wird vertieft durch die Einführung von *Entscheidungsbäumen*, die in relativ kurzer Zeit verstehbar sind und bei denen die wesentlichen Verfahren und Parameter zumindest im Grundsatz für die Lernenden transparent und erkennbar sind. Als Werkzeug wird hier wieder CODAP mit einem zusätzlichen Plug-In zum Darstellen von Entscheidungsbäumen

[7] Jupyter Notebook = interaktive, browserbasierte Umgebung zur Programmierung; `https://jupyter.org/`
[8] CODAP = Didaktisches Onlinetool zur Datenanalyse; `https://codap.concord.org`

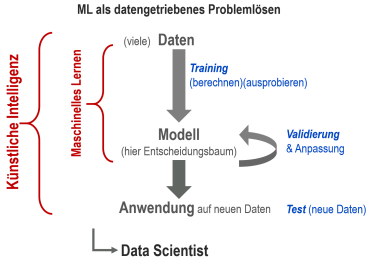

Abb. 3: Im Projektkurs erarbeitete Grafik zum Vorgehen zur Erstellung von KI-Modellen mittels ML

verwendet, das keine explizite Programmierung erfordert, sondern einen WYSIWYG-Editor bereitstellt. Allerdings stößt dieses Tool sehr bald an seine Grenzen. Insbesondere durch das manuelle Aufbauen des Entscheidungsbaumes kann kein tieferes Verständnis darüber erworben werden, wie Entscheidungsbäume algorithmisch erzeugt und zum automatischen Klassifizieren eingesetzt werden können. Daher werden im Anschluss mit Jupyter Notebook eigene Bäume zum schon bekannten JIM-Datensatz berechnet, wobei die Möglichkeit besteht, die Passung des entwickelten Baumes durch die Hinzunahme von Validierungsdaten („Evaluation“) zu überprüfen und anschließend zu optimieren. Während beim rein algorithmischen Problemlösen die Entwicklung eines stabilen und effizienten Algorithmus im Mittelpunkt stünde, ist hier die passende Wahl von Daten, Parametern und Optimierungsverfahren die größte Herausforderung für die Schülerinnen und Schüler, während die algorithmische Umsetzung der Bäume an sich durch die passende Wahl von Bibliotheken keine größere Herausforderung darstellt. Im folgenden zweiten Baustein werden zunächst (in Analogie zur Funktionsweise des Gehirns) durch die Unplugged-Aktivität „Brain in a Bag“ [Cu16a] die Grundbegriffe von *KNN* zusammen mit den Schülerinnen und Schülern erarbeitet. Anhand dieses durch weitere Erläuterungen und theoretische Inhalte ergänzte Spiel erkennen die Lernenden die grundlegenden Eigenschaften und Parameter von KNN ohne die Hürde eigener Programmierung. Um diese Erkenntnisse praktisch zu erproben und ein Gefühl für den Einfluss der verschiedenen Parameter zu erhalten, trainieren sie unter Verwendung des Online-Tools „Playground Tensorflow“[9] unterschiedliche Netze und beobachten und erfassen dabei direkt die Auswirkungen ihrer Änderungen. Für ein tieferes Verständnis modellieren, trainieren und validieren sie selbst unter Verwendung von Jupyter Notebook eigene KNN. Hier kommt der freie Datensatz mit handschriftlichen Ziffern aus der MNIST-Datenbank [LCB98] zum Einsatz. Da die Ziffern der amerikanischen Schreibweise entsprechen, führt eine Validierung durch handgeschriebene Zahlen der Schülerinnen und Schüler zu schlechten Ergebnissen, so dass hier ein Anlass geschaffen wird, die Grenzen und Möglichkeiten von verschiedenen Ansätzen von ML zu diskutieren. Auch kommen hier tagesaktuelle Artikel und Berichte zum Einsatz, die auch problematische oder kontrovers diskutierte Einsatzgebiete von KI und KNN zum Thema haben, so dass hier verstärkt gesellschaftliche Fragestellungen bearbeitet werden.

[9] Playground Tensorflow: Frei explorierbare Onlinevisualisierung von KNN; `https://playground.tensorflow.org`

3.3 Modul 3: Datenprojekte

Während in den ersten beiden Modulen die Erarbeitung neuen Wissens im Mittelpunkt stand, wird im dritten Modul ein gemeinsames Datenprojekt mit „realen“ Daten und Projektpartnern durchgeführt (insg. 10 Sitzungen). Dabei werden die in den vorherigen Modulen erworbenen Kompetenzen vertieft. Die Schülerinnen und Schüler erhalten den Projektauftrag, aus den Daten des örtlichen Parkleitsystems sowie aus den Bezahlsystemen eines Parkplatzes (Parkscheinautomat) und eines Parkhauses Vorhersagemodelle für die jeweilige Auslastung zu einem zukünftigen Zeitpunkt zu entwickeln und sich hierbei am Projektablauf des CRISP-DM zu orientieren. Zur Organisation und zur Unterstützung der arbeitsteiligen Arbeit an den Daten steht den Schülerinnen und Schülern ein Gitlab zur Verfügung, das sowohl zur Daten- als auch zur Aufgabenorganisation genutzt werden kann. Die betreuenden Personen fungieren im Rahmen dieses Projekts als Lernbegleiter und kümmern sich um den Kontakt mit den Projektpartnern sowie die interne Kommunikation und Organisation.

4 Erfahrungen der ersten Kursdurchführung

Der Projektkurs findet dreistündig mit 2 Schülerinnen und 17 Schülern der Jgst. 12, die alle Informatik belegt haben und daher über grundlegende Kenntnisse von Java verfügen, in einer Laborumgebung statt (insg. 24 Wo.). Die verwendete Programmiersprache Python sowie Jupyter Notebook waren für sie neu – der Einstieg war aber mit etwas Unterstützung gut zu bewältigen. Durch die Größe der Gruppe können die im Folgenden präsentierten Erkenntnisse nicht generalisiert werden, zeigen jedoch erste Hinweise, ob der Kurs an sich auch für andere Gruppen umsetzbar ist.

4.1 Erkenntnisse aus Modul 1: Daten und Informationen – Datendetektive

Das erste Modul dient zur Vermittlung von Kompetenzen zu statistischen Exploration, Verarbeitung und Darstellung von Daten (vgl. Abschnitt 3.1). Die Modulabschlusspräsentationen der Schülerinnen und Schüler zur statistischen Untersuchung von eigenen Fragestellungen anhand des JIM-Datensatzes zeigen ebenso wie verschiedene Diskussionsrunden, dass es ihnen gelungen ist, ein kritisches Verständnis zu Daten, Information und ihrer Visualisierung zu entwickeln. So wurde beispielsweise in den Präsentationen häufig die Größe des verwendeten JIM-Datensatzes kritisch gewürdigt und die damit verbundene geringe Aussagekraft der sich ergebenden kleinen Teilmengen eingeordnet. Im Rahmen der Feedbackrunde am Ende des Moduls wurde berichtetet, dass einerseits die Auswertung der Lärmdaten mit Python spannender als das Arbeiten mit dem JIM-Datensatz angesehen wurde, aber beide Bausteine wurden von den Lernenden nicht als besonders persönlich relevant befunden. Daher ist zu überlegen, wie im nächsten Durchlauf die subjektiv empfundene Relevanz und damit Motivation noch erhöht werden können.

4.2 Erkenntnisse aus Modul 2: Künstliche Intelligenz und Maschinelles Lernen

In Modul 2 liegt der Schwerpunkt auf der Vermittlung der Grundideen von ML und KI sowie dem Erwerb der Kompetenz zum eigenständigen Programmierung von KI-Modellen (vgl. Abschnitt 3.2). Die beiden Unplugged-Einheiten haben sich als guter thematischer Einstieg in den jeweiligen Baustein erwiesen und waren für die Lernenden hilfreich zur Bildung eines ersten Verständnisses des Themenbereichs. Das Feedback zeigt, dass das angemessene schrittweise Anheben des Schwierigkeitsniveaus bei dem komplexen Themenfeld von ML und KI eine Herausforderung darstellt: Während der „Sweet Learning Computer“durchweg positiv bewertet wurde, wurden die darauf folgenden Beispiele zum manuellen Erstellen von Entscheidungsbäumen mittels CODAP teils als monoton und wenig motivierend empfunden. Für die Erarbeitung der algorithmischen Darstellung von Entscheidungsbäumen wird der Begriff der Entropie nach Shannon benötigt. Die hierfür gestaltete Selbstlerneinhalt mit den zugehörigen Aspekten der Informationstheorie in Jupyter Notebook hat sich als zu schwierig selbst für diese sehr leistungsstarke Lerngruppe erwiesen, so dass dies stattdessen gemeinsam im Plenum erarbeitet wurde. Beim Feedback zu diesem Baustein bemängelten die Schülerinnen und Schüler, dass – ebenso wie in Modul 1 – in diesem Baustein aus didaktischen Gründen nur kleine Datensätze mit nur begrenzter Aussagekraft zum Einsatz kamen. Die den zweiten Baustein eröffnende Unplugged-Einheit „Brain in a Bag“ sowie die im ersten Baustein erarbeitete Darstellung (vgl. Abb. 3) zum datengetriebenen Problemlösen ermöglichte den Lernenden bereits eine sachlich fundierte Diskussion zur gesellschaftlichen Bedeutung der Verwendung von KI-Systemen, in der Probleme wie Vorurteile in Trainingsdaten („Diskriminierende Algorithmen“) divers betrachtet und diskutiert wurden. Im weiteren Verlauf des Moduls erarbeiteten sich die Schülerinnen und Schüler das eigene Erstellen, Trainieren und Validieren von KNN unter Verwendung von Python und Jupyter Notebook mit dem MNIST-Datensatz zur Ziffernerkennung. Dass es den Schülerinnen und Schülern damit gelang, ein grundlegendes Verständnis zu ML und KI sowie der Programmierung eines passendes Netzes zu entwickeln, zeigte sich auch in der Modulabschlusspräsentation, in der sie für den schon bekannten JIM-Datensatz Modelle zur Vorhersage des Geschlechts der Studienteilnehmer implementierten und optimierten. Sie haben damit wichtige Kompetenzen entwickelt, die im anschließenden Projektmodul (vgl. Abschnitt 3.3) gefestigt und weiterentwickelt wurden.

4.3 Erkenntnisse aus Modul 3: Datenprojekt

Im Rahmen einer Kooperation mit zwei lokalen Unternehmen (RTB, Bad Lippspringe sowie ASP, Paderborn[10]) erhielten die Schülerinnen und Schüler den Auftrag, aus Daten des örtlichen Parkleitsystems sowie aus den Bezahlsystemen eines Parkplatzes (Parkscheinautomat) und eines Parkhauses Vorhersagemodelle für die jeweilige Auslastung zu einem zukünftigen Zeitpunkt zu entwickeln. Da sich die Daten von Parkhäusern und Parkplätzen

[10] `https://www.rtb-bl.de` und `https://www.paderborn.de/microsite/asp/`

grundlegend unterscheiden, wurde die Aufgabe in zwei parallel bearbeitete Projekte aufgeteilt. Alle Daten lagen entweder als unbearbeitete Text-Dateien vor oder wurden direkt aus einer Datenschnittstelle entnommen. Die Schülerinnen und Schüler benötigten die ersten Wochen, um – folgend dem CRISP-DM-Prozess – die großen Datenmengen zu sichten, zu verstehen und aufzuarbeiten. Dazu wurden von den Gruppen Jupyter Notebooks entwickelt, die sie anschließend um die gewünschten Lerner zu erweiterten, wobei es den Gruppen freigestellt war, ob sie zur Modellierung Entscheidungsbäume oder KNN verwenden. Die größte Herausforderung stellte hier nicht die Implementierung der ML-Algorithmen dar, sondern die strategisch günstige Vorbereitung der Daten und die optimale Auswahl der Konfiguration des Lerners. Bei der Abschlusspräsentation konnten beide Gruppen aber nicht nur funktionierende Vorhersagemodelle einschließlich einer sinnvollen Web-GUI für ihre jeweiligen Parkmöglichkeiten an die Projektpartner übergeben, sondern überzeugten auch in der abschließenden Diskussion mit ihrem grundlegendes Verständnis nicht nur über fachliche Problemstellungen, sondern auch durch ihre Fähigkeit, die hierzu gehörenden gesellschaftlichen Fragen fundiert erörtern können.

5 Ausblick und Fazit

Im Rahmens dieses Artikels beschreiben wir die Entwicklung eines Data Science-Kurses, der in der Sek II als Projektkurs unterrichtet werden kann. Nicht jede Lehrkraft, die gerne dieses Thema in den Unterricht einbinden möchte, hat diese Zeitressource. Daher planen wir, die Module auf Basis unserer bisherigen Erfahrungen weiterzuentwickeln, so dass es möglich wird, nur ausgewählte Teile durchzuführen. Im Moment werden die Erkenntnisse der aktuellen Durchführung sowie der Erprobungen einzelner Komponenten in die Kursmaterialien eingearbeitet. Im Anschluss daran werden die Materialien interessierten Lehrkräften unterschiedlicher Schularten zur Erprobung und Weiterentwicklung zur Verfügung gestellt. Zur Weiterentwicklung des Gesamtkurses im Sinne des DBR wird dieser im nächsten Schuljahr mit der bisherigen Partnerschule in modifizierter Form durchgeführt. Da die Schülerinnen und Schüler einerseits das durch die Modulstruktur bedingte „Vorratslernen“ bemängelten und andererseits an einigen Stellen Schwierigkeiten hatten, die erlernten Inhalte zielgerichtet auf die Projektfragestellungen anzuwenden, ist geplant, die Theoriephasen so mit der Arbeit an einem Datenprojekt zu verzahnen, dass die theoretischen Inhalte gezielt während des Projektablaufs eingebaut werden. Zudem werden die in den verschiedenen Modulen enthaltenen gesellschaftlichen, sozialen und ethischen Fragestellungen nochmals gebündelt in einem zusätzlichen Modul diskutiert und bearbeitet, um die Schülerinnen und Schüler noch tiefer mit den darin enthaltenen Fragestellungen vertraut zu machen. Dies alles erfordert noch einiges an konzeptioneller Arbeit, die sich jedoch durch einen höheren Kompetenzgewinn seitens der Lernenden bemerkbar machen wird. Zusammengefasst sind wir trotz aller noch zu leistenden Entwicklungsarbeit überzeugt, durch den vorgestellten *Data Science-Kurs* einen großen Schritt hin zu einem umfassenden Curriculum für diesen sehr komplexen Bereich getan zu haben, so dass wir gespannt auf die nächsten Durchführungen des Kurses und seiner Module sind.

Literaturverzeichnis

[Ch00] Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R.: Cross Industry Standard Process for Data Mining 1.0, Step-by-step Data Mining Guide. 2000.

[CM97] Cobb, George W.; Moore, David S.: Mathematics, statistics, and teaching. The American Mathematical Monthly, 104(9):801–823, 1997.

[Co03] Cobb, Paul; Confrey, Jere; diSessa, Andrea; Lehrer, Richard; Schauble, Leona: Design Experiments in Educational Research. Educational Researcher, 32(1):9–13, 2003.

[Cu16a] Curzon, Paul: , Brain in a Bag, A CS4FN Computing Activity. https://www.youtube.com/watch?v=lux_ybamClU, 2016. Accessed: 2018-06-15.

[Cu16b] Curzon, Paul: , The Sweet Learning Computer, A CS4FN Computing Activity. www.cs4fn.org/machinelearning, 2016. Accessed: 2019-02-10.

[GR] Grillenberger, Andreas; Romeike, Ralf: In: Proceedings of the 17th Koli Calling Conference on Computing Education Research. New York.

[LCB98] LeCun, Yann; Cortes, Corinna; Burges, Christopher J.C.: , MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 1998. Accessed: 2018-06-15.

[Pa18] Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts, 2018.

[Ri15] Ridsdale, Chantel; Rothwell, James; Smit, Michael; Ali-Hassan, Hossam; Bliemel, Michael; Irvine, Dean; Kelley, Daniel; Matwin, Stan; Wuetherick, Bradley: Strategies and best practices for data literacy education: knowledge synthesis report. 2015.

[Ri16] Ridgway, Jim: Implications of the Data Revolution for Statistics Education. International Statistical Review, 84(3):528–549, 2016.

[Rö16] Röhner, Gerhard; Brinda, Torsten; Denke, Volker; Hellmig, Lutz; Heußer, Theo; Pasternak, Arno; Schwill, Andreas; Seiffert, Monika: Bildungsstandards Informatik für die Sekundarstufe II. Beilage zu LOG IN, Heft, (183/184), 2016.

[SBS18] Sentance, Sue; Barendsen, Erik; Schulte, Carsten: Computer Science Educ.: Perspectives on Teaching + Learning in School. Bloomsbury Academic, London, 3 2018.

[TD16] Tedre, Matti; Denning, Peter J.: The Long Quest for Computational Thinking. In: Proceedings of the 16th Koli Calling International Conference on Computing Education Research. Koli Calling '16, ACM, New York, NY, USA, S. 120–129, 2016.

[Tv09] Thijs, Annette; van den Akker, Jan: Curriculum in development. SLO - Netherlands Institute for curriculum development, 2009.