DATA SCIENCE & LAW: JURA TRIFFT INFORMATIK UND STATISTIK

Daniel Blochinger / Deniz Calhan / Steffen Hain

Prof. Dr. Daniel Blochinger, Professor für BWL/VWL im Studiengang Medien & Kommunikation, Duale Hochschule Baden-Württemberg Ravensburg, Oberamteigasse 4, 88214 Ravensburg, DE; E-Mail: blochinger@dhbw-ravensburg.de

Deniz Calhan, Legal Designerin & Innovation Consultant, DE; E-Mail: den.calhan@gmail.com

Dr. Steffen Hain, Legal Counsel, Merckle GmbH, Graf-Arco-Str. 3, 89079 Ulm, DE; E-Mail: steffen-hain@gmx.de

Schlagworte: ***Juristische Informatiksysteme, Legal Tech, Künstliche Intelligenz & Recht, Rechtsinformatik als Methodenwissenschaft, Ausbildung***

Abstract: *Der Projektkurs Data Science & Law, den die Universität Ulm gemeinsam mit dem Arbeitgeberverband Südwestmetall im Wintersemester 2020/21 ins Leben gerufen hat, ist eine interdisziplinäre Lehrveranstaltung an der Schnittstelle zwischen Recht, Statistik und Informatik. Der Kurs vermittelt den Studierenden der Universität Ulm mit wirtschaftswissenschaftlichem Hintergrund nicht nur Grundkenntnisse im Bereich der Digitalisierung des Rechts sowie auf dem Gebiet des kollektiven Arbeitsrechts, sondern auch Fähigkeiten bei der Aufbereitung und Analyse größerer Datensätze mit Hilfe der Programmiersprache R sowie deren Entwicklungsumgebung RStudio.*

1. Motivation des Kurses[1]

Inmitten des Zeitalters der digitalen Revolution, die nach der industriellen Revolution, der Einführung der Massenproduktion und der Entwicklung von Computertechnologie, als Industrie 4.0 gilt, finden Begriffe wie Künstliche Intelligenz, Machine Learning und Data Science zunehmend Eingang in die Welt des Rechts. In unmittelbarem Zusammenhang mit diesen Begriffen tritt häufig die schillernde Bezeichnung Legal Tech in Erscheinung. Nach der Definition von Oliver Goodenough[2] findet eine Aufteilung in Legal Tech 1.0, 2.0 und 3.0 statt, wobei es sich bei Legal Tech 1.0 insbesondere um Software handelt, die den Alltag eines Juristen erleichtert und ihn beispielweise bei der Dokumentenerstellung (Microsoft Word etc.) unterstützt. Programme im Sinne von Legal Tech 2.0 sollen hingegen klassische und unkomplizierte Aufgaben eines Juristen selbständig übernehmen können, ohne dass es einer menschlichen Überwachung bedarf. Schließlich sieht Goodenough in Systemen im Sinne von Legal Tech 3.0 die vollständige Ersetzung des Juristen unter Zuhilfenahme künstlicher Intelligenz und autonom handelnder Assistenten.

Digitale Rechtsberatungsstellen wie etwa Flightright zeigen, dass die Entwicklung softwarebasierter Technologien selbst vor dem Berufsstand der Juristen keinen Halt und die juristische Expertise bereits teilweise entbehrlich macht. Von alltäglichen Textverarbeitungsprogrammen über die Recherche mit Online-Datenbanken, die automatisierte Einleitung eines Mahnverfahrens bis hin zu Verfahren der computerisierten Dokumentenerstellung sowie dem Umgang mit Online-Streitbeilegungsplattformen[3] setzt der Beruf des Juristen gegenwärtig längst vielfältige Kenntnisse im Bereich digitalisierter Anwendungen voraus und dokumentiert den Wandel der Anforderungen an den Beruf des Juristen.

---

1 Dieser Beitrag wurde am 19. Dezember 2022 finalisiert. Aus konferenztechnischen Gründen konnte dieser Beitrag erst im Tagungsband 2024 erscheinen.

2 Siehe https://www.huffpost.com/entry/legal-technology-30_b_6603658 (aufgerufen am 19. Dezember 2022).

3 Detailliert Anzinger, ZKM 2021, 53 ff.

Der Projektkurs Data Science & Law, den die Universität Ulm gemeinsam mit dem Arbeitgeberverband Südwestmetall im Wintersemester 2020/21 erstmalig ins Leben gerufen hat, vermittelt den Studierenden nicht nur Grundkenntnisse im Bereich der Digitalisierung des Rechts, sondern auch Fähigkeiten bei der Aufbereitung und Analyse größerer Datensätzen mit Hilfe der Programmiersprache R sowie deren Entwicklungsumgebung R-Studio. Der Kurs soll als Blaupause dienen, nicht nur an naturwissenschaftlich orientierten Universitäten, wie etwa der Universität Ulm, sondern insbesondere auch an den juristischen Fakultäten der europäischen Universitäten die Ausbildung dem Zeitgeist entsprechend anzupassen und dabei verstärkt auf den Einsatz statistischer Methoden und denjenigen des Machine Learnings zu setzen.

Neben der Universität Ulm bietet unter anderem die University of Ottawa eine ähnliche interdisziplinäre Veranstaltung an, die die Bereiche des Rechts und der Informatik verbindet. Die Existenz dieses Kurses unterstreicht die internationale Bedeutung von Data Science auf dem Gebiet des Rechts.

Machine Learning gilt als Teilbereich der künstlichen Intelligenz und trägt im Rahmen des Projektkurses dazu bei, Verbindungen und Zusammenhänge innerhalb eines Datensatzes zu ermitteln. Mit Hilfe der von Südwestmetall zur Verfügung gestellten arbeitsrechtlichen Datensätze lernt die Maschine und verbessert die Leistungsfähigkeit des Algorithmus durch ständiges Training.

Im Gegensatz zu regelbasierten Expertensystemen, die als Vorreiter des Maschinellen Lernens gelten und mit Wenn-Dann-Beziehungen arbeiten, handelt es sich beim Machine Learning um Algorithmen, die ihr Wissen mithilfe von Trainingsdaten selbstständig und autonom erweitern und sich insbesondere durch Anpassungsfähigkeit und Flexibilität auszeichnen. Insofern vermittelt der Kurs innovative Methoden bei der Erkenntniserlangung und Mustererkennung rechtlicher Vertragstexte.

2. Datensätze

Getreu dem Motto „The more data the better" hängt die Effektivität der einzelnen Maschine-Learning-Methoden ganz wesentlich von der Anzahl, aber auch von der Qualität, der verfügbaren Daten ab, anhand derer die Maschine angelernt wird. Neuronale Netze und andere Methoden des Maschinellen Lernens benötigen große Datensätze, um Erkenntnisse aus heterogenen Texten zu gewinnen und Muster zu erkennen, die den Alltag eines Juristen erleichtern sollen.

Vor diesem Hintergrund anonymisierte der „tarif- und sozialpolitische Arbeitgeberverband" Südwestmetall, der gegenwärtig 671 Unternehmen (887 Betriebe) aus der Metall- und Elektroindustrie in Baden-Württemberg vertritt, 112 „Betriebsvereinbarungen", welche die Studierenden in der Projektphase verwendeten. Als interdisziplinärer Kurs an der Schnittstelle zwischen Recht, Statistik und Informatik vermittelt er den Studierenden in der Grundausbildung neben Lerninhalten der Programmiersprache R ebenfalls solche des kollektiven Arbeitsrechts. Dies dient den Studierenden, die überwiegend aus Studiengängen mit wirtschaftswissenschaftlichem Hintergrund stammen, als Vorbereitung auf die Projektphase, in der sie auf den bereitgestellten Datensatz Methoden des Maschinellen Lernens anwenden, um Zusammenhänge und Muster zu erkennen. Konkret handelt es sich bei dem Datensatz um „Betriebsvereinbarungen" in Form von Interessensausgleichen und Sozialplänen von Unternehmen aus der Metall- und Elektroindustrie. Üblicherweise ziehen Betriebsänderungen, etwa eine größere Zahl betriebsbedingter Kündigungen oder Verlagerungen einzelner Betriebe ins Ausland, ähnliche Vertragsgestaltungen zwischen Betriebsrat und Arbeitgeber nach sich. Insbesondere der Sozialplan gewährleistet den Arbeitnehmern einen konkreten Ausgleich, z.B. in Gestalt einer Abfindung. Daneben kann beispielsweise die Errichtung einer Transfergesellschaft mitsamt der Vereinbarung von Transferkurzarbeitergeld Gegenstand eines Sozialplans sein. Um die Anonymität der Arbeitgeber zu wahren, die aufgrund der Sensibilität der Daten erforderlich ist, und um eine bessere Einlesbarkeit und Kompatibilität mit den Maschine-Learning-Methoden zu gewährleisten, pseudonymisierte die Südwestmetall-Niederlassung in Ulm die „Betriebsvereinbarungen" und wandelte sie in Word-Dateien um. Zum Zwecke der Übersichtlichkeit

sowie der Zusammenhangserkennung erfolgte in einem weiteren Schritt eine Vorklassifizierung des gesamten Datensatzes im Rahmen einer Excel-Tabelle nach neunundzwanzig Kriterien. Dazu gehören etwa Merkmale wie die Arbeitnehmerzahl, die Konzernzugehörigkeit, die konkrete Betriebsänderung, die Existenz und Berechnungsformel einer Abfindungsregelung, die Errichtung einer Transfergesellschaft und die Zusammensetzung einzelner Schlussbestimmungen. Aufgrund der Heterogenität und beschränkten Anzahl der einzelnen Datensätze und der daraus resultierenden Schwierigkeit mit Hilfe eines Algorithmus gewinnbringende Erkenntnisse zu erlangen, sollen die „Betriebsvereinbarungen" in den künftigen Semestern durch standardisierte Arbeitsverträge ersetzt werden. Mit Hilfe von Pseudonymisierungsprogrammen können homogene und standardisierte Arbeitsverträge, im Gegensatz zu den eher heterogen ausgestalteten „Betriebsvereinbarungen", schneller anonymisiert und dadurch größere Datenmengen generiert werden.

3. Kursinhalte

Der Projektkurs Data Science & Law besteht aus den drei Abschnitten „Basiskompetenzen in R/R-Studio", „Data Science in R/R-Studio" und „Projektphase". Die ersten beiden Abschnitte sind dabei ihrerseits in 15 Themenblöcke unterteilt.

Basiskompetenzen in R/R-Studio Im ersten Abschnitt „Basiskompetenzen in R/R-Studio" werden grundlegende Kenntnisse in der Programmiersprache R und der Entwicklungsumgebung R-Studio vermittelt. Die Inhalte überschneiden sich dabei vor allem in den Themenblöcken 1-5 willentlich mit denen anderer Lehrveranstaltungen, welche die Studierenden teilweise bereits abgeschlossen haben. Ziel des Abschnittes ist es die Studierenden auf ein gemeinsames Kompetenzniveau in der Programmiersprache R und der Entwicklungsumgebung R-Studio zu heben und ihnen eine solide Basis für das Verständnis der weiterführenden Inhalte in Abschnitt 2 zu vermitteln.

Themenblock Nr. 1 gibt eine Übersicht über die Benutzeroberfläche der Entwicklungsumgebung R-Studio sowie eine Anleitung für elementare Arbeitsabläufe wie z.B. das Schreiben und Ausführen von Skripten, die Verwendung der Konsole oder das Debuggen von Skripten mithilfe des Environment-Viewers.

Themenblock Nr. 2 behandelt die mehrdimensionale Datenstruktur „Vektor". Die Studierenden lernen leere oder auch mit spezifischen Daten vorbefüllte Vektoren zu erstellen, auf ausgewählte Elemente eines Vektors zuzugreifen und schließlich Vektoren in Berechnungen einzusetzen.

Themenblock Nr. 3 behandelt die komplexeren Datenstrukturen „Liste" und „Data-Frame", welche beide auf die im vorherigen Themenblock behandelten Vektoren aufbauen. Die Studierenden lernen mit diesen Datenstrukturen zu arbeiten, sie ineinander umzuwandeln und ihre jeweiligen Vor- und Nachteile zu verstehen.

Themenblock Nr. 4 behandelt die wichtigsten Programmstrukturen (Bedingungen und Schleifen) sowie das Entwickeln von parametrisch aufrufbaren Funktionen.

Themenblock Nr. 5 markiert den Übergang von reinen Grundlagen der Programmierung hin zum Arbeiten mit Daten. Zunächst lernen die Studierenden wie sie in R tabellarische Daten wie z.B. Excel-Dateien einlesen, filtern, sortieren und analysieren können.

Themenblock Nr. 6 ist der erste von drei Themenblöcken der sich mit dem Arbeiten mit textuellen Daten wie z.B. Word- oder PDF-Dateien befasst. Die Studierenden lernen insbesondere das automatisierte Strukturieren von Fließtexten mithilfe von regulären Ausdrücken.

Themenblock Nr. 7 behandelt die automatisierte Erfassung von textuellen Inhalten aus dem Internet (sogenanntes Webscraping) und geht dabei auch auf die rechtlichen Probleme beim Einsatz dieser Technik ein.

Themenblock Nr. 8 behandelt die Aufbereitung von textuellen Daten für Machine Learning Methoden und dabei insbesondere die Erzeugung von s.g. Term-Dokument-Matrizen (wie oft kommen bestimmte Worte oder auch Wortgruppen in bestimmten Dokumenten vor?) und die Berechnung von Ähnlichkeitsmaßen, welche die Ähnlichkeit von zwei Texten in einer einzigen Kennzahl ausdrücken.

Data Science in R/R-Studio Im zweiten Abschnitt „Data Science in R/R-Studio“ lernen die Studierenden ausgewählte Methoden des Machine Learnings kennen. Bei der Vermittlung der Methodik setzen wir vor allem auf Anschaulichkeit: die Studierenden sollen verstehen, wie die Methoden mit den Daten arbeiten und welche Voraussetzungen an die Daten, Vorteile und Einschränkungen diese deshalb mit sich bringen. Auf eine tiefgehende Analyse der Methoden auf mathematisch / statistischer Ebene wird bewusst verzichtet, da diese zum einen den Rahmen eines Projektkurses und zum anderen den Rahmen einer interdisziplinären Ausbildung sprengen würde. Viel Raum erhält dagegen die Vermittlung der praktischen Anwendung der Machine Learning Methoden in R und die Interpretation der Ergebnisse. Dabei kommt ein Datensatz aus dem Bereich des Steuerrechts mit 7760 Entscheidungen des Bundesfinanzhofs zum Einsatz, der mit den vorgestellten Methoden exemplarisch analysiert wird.

Themenblock Nr. 9 stellt diesen Datensatz vor und gibt auch eine kurze Übersicht über die Arbeit des Bundesfinanzhofs und dessen Organisation in unterschiedliche Senate.

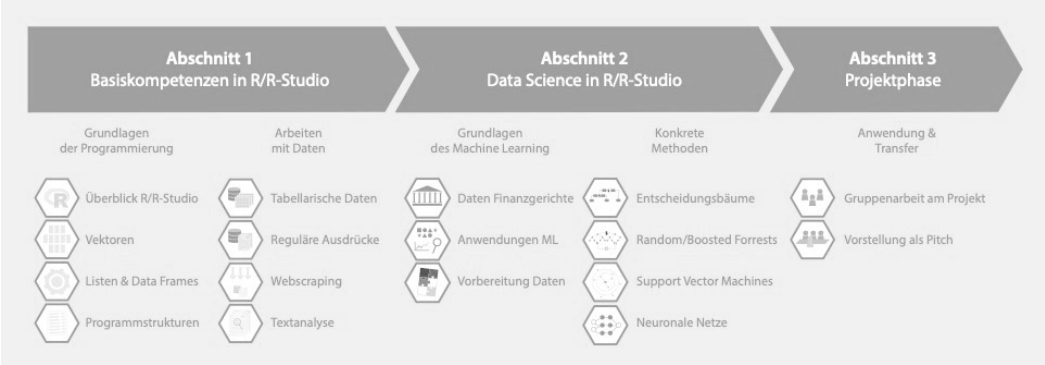
Themenblock Nr. 10 erklärt die drei Standardanwendungen von Machine Learning (Klassifizieren, Su-chen und Vorhersagen) sowie die gängigen Maße zur Beurteilung des Lernerfolgs der Maschine (Accuracy, Recall und Precision)

Themenblock Nr. 11 zeigt exemplarisch, wie der Datensatz „Bundesfinanzhof“ für die Anwendung von Machine Learning Methoden vorbereitet wird. Dabei wird das Wissen aus Themenblock 8 wieder-holt, vertieft und konkret angewendet.

Themenblock Nr. 12 stellt mit Entscheidungsbäumen die erste konkrete Machine Learning Methode vor. Dabei wird zunächst an stilisierten Beispielen anschaulich erklärt, wie die Methode arbeitet und anschließend konkret gezeigt wie die Methode auf den im vorherigen Themenblock aufbereiteten Datensatz „Bundesfinanzhof“ angewendet werden kann und wie die erhaltenen Ergebnisse zu interpretieren sind.

Themenblock Nr. 13 erweitert die Methode der Entscheidungsbäume zu Random Forrests und Gradi-ent Boosted Forrests. Die Inhalte werden mit demselben Schema vermittelt wie beim vorherigen Themenblock: Übersicht - Anwendung - Interpretation.

Themenblock Nr. 14 und Themenblock Nr. 15 stellen zwei weitere Methoden vor: Support Vector Ma-chines und Neuronale Netzwerke.



Projektphase Im letzten Abschnitt wenden die Studierenden ihre in den ersten beiden Abschnitten erworbenen Kenntnisse im Rahmen einer Projektarbeit auf einen Datensatz mit 110 Betriebsvereinbarungen an. Während in diesem Abschnitt keine neuen Inhalte vermittelt werden, werden bereits behandelte Inhalte je nach Bedarf für das konkrete Projekt selektiv vertieft.

4. Machine Learning Methoden

Im Kurs werden insgesamt fünf ausgewählte Machine Learning Methoden vermittelt: Entscheidungsbäume, Random Forrests, Boosted Forrests, Support Vector Machines und Neuronale Netzwerke. Die Auswahl soll einen breiten Überblick über verschiedene Methoden geben und gleichzeitig den Studierenden für die Projektarbeit geeignete Werkzeuge an die Hand geben. Bei allen fünf Methoden soll eine zu erklärende Variable des Datensatzes durch mehrere erklärende Variablen des Datensatzes vorhergesagt werden.

Um den Lernerfolg später beurteilen zu können müssen wir dazu zunächst den Datensatz aufteilen. Mit einem Teil trainieren wir die Maschine, mit dem anderen schauen wir, ob die Maschine tatsächlich gelernt hat, also ihre Erfahrung aus dem Training auch auf andere Daten übertragen kann oder ob sie lediglich ihren Trainingsdatensatz „auswendig gelernt“ hat. Im einfachsten Fall teilen wir den Datensatz 50:50 auf.

Entscheidungsbäume Die einfachste im Kurs vermittelte Methode ist der Entscheidungsbaum. Entscheidungsbäume sind eine Sammlung an miteinander verknüpften Wenn-Dann-Entscheidungen. Die Maschine „probiert“ verschiedene „Wenn-Dann“ Abfragen in unterschiedlichen Kombinationen durch und sucht nach einem Entscheidungsbaum, der den ihr gegebenen Trainingsdatensatz möglichst gut erklären kann.

Entscheidungsbäume sind einfach zu interpretieren und auch bei kleinen Datensätzen einigermaßen robust. Ihr Nachteil ist, dass sie komplexere Zusammenhänge nicht so gut erfassen können wir andere Methoden dies tun. Da unser Datensatz in der Projektphase eher klein ist stellen wir unseren Studierenden zwei Weiterentwicklungen der Entscheidungsbäume vor.

Random Forrests Die Grundidee hinter Random Forrests ist es den Trainingsdatensatz in eine Vielzahl an Teildatensätzen zufällig aufzuteilen und mit jedem dieser Teildatensätze einen separaten Entscheidungsbaum anzutrainieren. Die Vorhersage erfolgt dann über eine Abstimmung aller Entscheidungsbäume mit einfacher Mehrheit.

Boosted Forrests sind eine konsekutive Weiterentwicklung der Random Forrests. Statt einen Wald aus vielen zufällig erzeugten Bäumen zusammenzustellen wird bei der Erzeugung der Bäume ein iteratives Verfahren angewendet, dass nach und nach „bessere“ Bäume erzeugt.

Support Vector Machines ist eine Methode, die mit der klassischen linearen Regression verwandt ist. Die Idee ist aus den Datenpunkten sogenannte Stützvektoren auszuwählen, die eine Grenze zwischen den zu klassifizierenden Daten ziehen.

Die SVM versucht die Stützvektoren so zu wählen, dass der Kanal, der dadurch entsteht, so breit wie möglich ist. In diesem Kanal dürfen keine Datenpunkte liegen. Je breiter der Kanal und je weniger Datenpunkte auf der falschen Seite des Kanals liegen, umso besser gilt das Trainingsergebnis. SVMs können komplexere Zusammenhänge erklären, sind jedoch im Gegensatz zu den Entscheidungsbäumen weniger anschaulich.

Neuronale Netzwerke versuchen die Funktionsweise des Gehirns nachzubilden und funktionieren über ein Netzwerk an miteinander verbundenen Neuronen welche „Eingangssignale“ (Daten oder vorhergeschaltete Neuronen) verarbeiten und entweder an weitere Neuronen weiterleiten oder an Ausgangsneuronen weitergeben. Neuronale Netzwerke können sehr komplexe Zusammenhänge erfassen, sind jedoch gleichzeitig am schwierigsten zu interpretieren und am anfälligsten für das Problem Overfitting.

5. Kursablauf und Didaktik

Der Kursablauf ist analog zu den Kursinhalte in die drei Abschnitte „Basiskompetenzen in R/R-Studio“, „Data Science in R/R-Studio“ und „Projektphase“ gegliedert. Die eingesetzten Didaktischen Methoden führen die Studierenden schrittweise an das selbstständige Arbeiten mit Daten und Machine Learning Methoden heran.

Basiskompetenzen in R/R-Studio Im ersten Abschnitt „Basiskompetenzen in R/R-Studio“ werden die Studierenden noch am meisten an die Hand genommen. Zu jedem der acht Themenblöcke im ersten Abschnitt gibt es ein YouTube Video (überwiegend durch Einblendungen und Animationen ergänzte Screencasts), welches die Inhalte langsam und mit vielen Beispielen erklärt.

Aufbauend auf den Videos bearbeiten die Studierenden Übungsblätter in Form von R-Dateien mit zunächst sehr konkreten und kurzen Aufgaben. Ab dem fünften Themenblock werden die Aufgaben länger und offener. Bei Schwierigkeiten oder zur Lösungskontrolle können die Studierenden auf eine weitere R-Datei mit den fertig gelösten Aufgaben zurückgreifen.

Einmal pro Woche wird ein Live-Stream via Zoom angeboten, bei dem die Studierenden zum einen Fragen zu den Videos und den Übungen stellen können und zum anderen einen 10–30-minütigen Impuls zu den juristischen, technischen und praktischen Aspekten von Data Science & Law bekommen.

Um den Studierenden ein Maximum an Flexibilität zu ermöglichen werden alle Videos, Übungen und Lösungen von Anfang an zugänglich gemacht. Die Foliensätze der Impulse werden nach dem jeweiligen Livestream hochgeladen. Eine Aufzeichnung des Livestreams gibt es dagegen nicht.Das vorhandene asynchrone Material ist mehr als ausreichend für den Kurs und Aufzeichnungen stellen für einige Studierende eine zusätzliche Hemmschwelle bzgl. Fragen und aktiver Teilnahme dar.

Data Science in R/R-Studio Im zweiten Abschnitt wird weiter auf YouTube Videos als asynchrone Methode der Wissensvermittlung gesetzt. Die Videos zu den Themenblöcken 9 bis 15 arbeiten jedoch verstärkt mit Animationen zur anschaulichen Vermittlung der Machine Learning Methoden.

Anstelle von Übungsblättern dürfen die Studierenden ihre neu erworbenen Kompetenzen im Rahmen eines sogenannten Kaggle-Wettbewerbs einsetzen. Dabei werden in Einzel- oder Partnerarbeit 3 Aufgaben mithilfe von Machine Learning Methoden angegangen. Am Ende des Wettbewerbs reichen die Teilnehmer ihre Ergebnisse in Form von R-Skripten ein. Die Bewertung der Aufgaben ergibt sich dann objektiv über die Performance, welche diese Skripte in den drei gegebenen Aufgabenstellungen erreichen. Der Gewinner bzw. die Gewinner erhalten Amazongutscheine über jeweils 15€.

Die Live-Streams werden analog zum ersten Abschnitt weiter wöchentlich Angeboten: eine Gelegenheit Fragen zu den Videos zu stellen und ein 10–30-minütiger Impuls zu den juristischen, technischen und praktischen Aspekten von Data Science & Law bekommen.

Projektphase Für den dritten und letzten Abschnitt teilen sich die Studierenden in Gruppen von 4-6 Personen ein und wählen sich eines der im nächsten Unterkapitel vorgestellten Projekte aus.

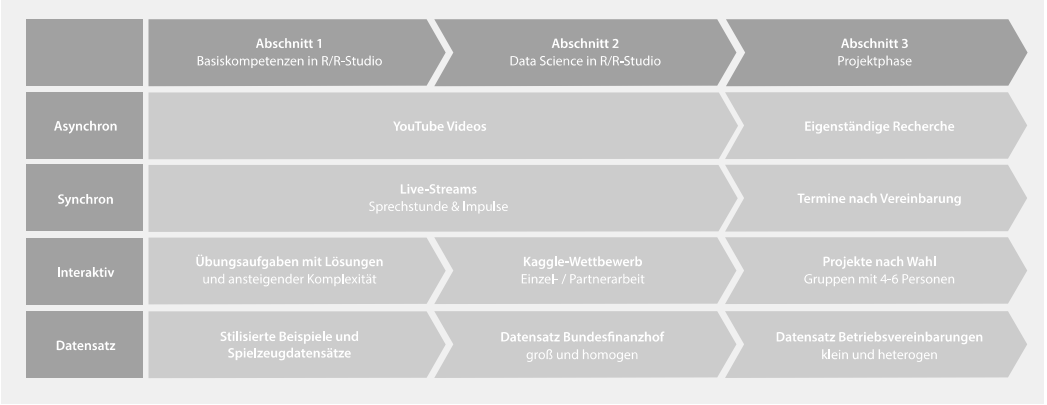
Ähnlich wie beim Kaggle-Wettbewerb sind diese Projekte mit Machine Learning Methoden anzugehen. Anders als beim Kaggle-Wettbewerb ist die Aufgabenstellung jedoch umfangreicher, offener und komplexer und als Datensatz kommt ein deutlich kleinerer und gleichzeitig heterogenerer Datensatz an Betriebsvereinbarungen zum Einsatz.

Während der Gruppenarbeit haben die Studierenden jederzeit die Möglichkeit sich an ihre Dozenten zu wenden. Allerdings müssen sie in diesem Abschnitt dazu aktiv das Gespräch suchen und einen Termin ausmachen. Livestreams bzw. Sprechstunden gibt es keine mehr.

Am Ende der Projektphase reichen die Gruppen ihre Ergebnisse (Quellcode und Datensätze) ein, verfassen eine kurze Zusammenfassung ihres Projekts und stellen dieses vor den Dozenten und Vertretern von Südwestmetall in Form eines Pitch vor.

Der Pitch-Aspekt ist dabei nicht nur ein Aufhänger, sondern soll von den Studierenden ernst genommen werden: sie präsentieren ihr Projekt vor einem interdisziplinären Panel: Juristen, Volkswirte, Softwareentwickler und Manager. Der Vortrag muss die richtige Balance zwischen zu technisch detailversessen und zu oberflächlichem marketingorientiert finden und die wichtigsten Punkte prägnant rüberbringen: Warum ist unser Projekt interessant? Welches Problem löst es? Wie erreichen wir diesen Mehrwert? Wie könnte man das Projekt noch weiterentwickeln?

Als Endnote erhalten die Studierenden ein gewichtetes Mittel aus ihren schriftlichen und mündlichen Leistungen. Der eingereichte Quellcode und die Zusammenfassung bilden zusammen ein Drittel der Endnote. Der Pitch, beurteilt von allen anwesenden Prüfern, bildet die übrigen zwei Drittel der Endnote.



6. Projekte

Nachdem den Studierenden in den ersten zwei Monaten des Semesters Lerninhalte in den Bereichen der Programmiersprache R und deren Entwicklungsumgebung R-Studio sowie auf dem Gebiet des Legal Techs und kollektiven Arbeitsrechts vermittelt wurden, arbeiteten sie in der einmonatigen Projektphase in Kleingruppen von bis zu 6 Personen an vorgegebenen Projekten. Den Abschluss der Projektphase bildete die Präsentation der Ergebnisse und des erarbeiteten Quellcodes im Rahmen eines „Pitches“. Die Bewertung erfolgte durch den Geschäftsführer der Niederlassung des Arbeitgeberverbands Südwestmetall in Ulm, Professoren der Universität Ulm aus den Bereichen der Rechtswissenschaft und der Volkswirtschaftslehre, Universitätsmitarbeiter mit juristischem und volkswirtschaftlichem Hintergrund sowie durch einen bei einem Automobilhersteller angestellten Data Scientist als „Gastjuror“.

Bei den einzelnen Projekten handelte es sich um vorgegebene Themen, die nachfolgend überblicksartig vorgestellt werden:

Extraktion von Abfindungsregeln

Die Grundsatzfrage des ersten Projekts lautet: Können Abfindungsregeln mithilfe von Künstlicher Intelligenz und Machine Learning-Methoden zuverlässig aus den zur Verfügung gestellten „Betriebsvereinbarungen“, die den Studierenden als DOCX-Dokumente vorliegen, extrahiert werden?

In einem weiteren Schritt sollten sich die Studierenden mit der Frage befassen, wie die einzelnen Abfindungsformeln zu interpretieren sind und wie man sie in einem Legal-Tech-Tool verarbeiten könnte, das einen

Mehrwert für die Praxis, insbesondere für den Arbeitgeberverband Südwestmetall und dessen Mitglieder, darstellt. In Betracht käme beispielsweise ein Tool für den Arbeitgeber, welches die Kosten branchenüblicher „Betriebsvereinbarungen" vergleicht und Rückschlüsse darauf zulässt, inwiefern die einzelnen Abfindungsformeln für den Arbeitgeber, etwa aufgrund eines Faktors oder Divisors, bei der Berechnung der jeweiligen Abfindung besonders günstig oder teuer sind. Insbesondere aufgrund der Vielfalt einzelner Abfindungsberechnungsformeln in den jeweiligen Sozialplänen könnte ein solches Tool die Arbeitgeber bei der Gestaltung der Sozialpläne effizient unterstützen.

Abfindung: Wie hoch fällt sie aus?

Das zweite Projekt befasst sich mit der Abfindung im Einzelnen und soll Rückschlüsse über die jeweilige Höhe einer Abfindung im Rahmen einer „Betriebsvereinbarung", bestehend aus Interessensausgleich und Sozialplan, ermöglichen.

Unter Anwendung von Künstlicher Intelligenz und Machine Learning-Methoden sollen die Studierenden Algorithmen trainieren, die aufgrund der Situationsbeschreibung innerhalb eines Dokuments Erkenntnisse über die jeweilige Abfindungshöhe liefern. Dabei sollen insbesondere auf den ersten Blick belanglos erscheinende Informationen, wie etwa die Präambel oder sonstige Standardtextbausteine eines Interessensausgleichs, betrachtet und deren Zusammenhänge mit der jeweiligen Abfindungshöhe analysiert werden. Ziel dieses Projekts ist es, zu erkennen, welche realen Zusammenhänge sich hinter der jeweiligen Abfindungshöhe verbergen.

Transfergesellschaft: ja oder nein?

Zahlreiche Betriebe aus der Metall- und Elektroindustrie sehen im Falle betriebsbedingter Kündigungen die Übernahme des gekündigten Arbeitnehmers in eine Transfergesellschaft vor. Um die drohende Arbeitslosigkeit entlassener Arbeitnehmer zu verhindern, ermöglicht § 111 SGB III grundsätzlich die Gewährung von Kurzarbeitergeld in Höhe von 60% des bisherigen Nettomonatsentgelts für eine Zeit von maximal zwölf Monate innerhalb einer Transfergesellschaft. Die Errichtung einer Transfergesellschaft dient ausschließlich der Vermittlung einer neuen Arbeitsstelle für den gekündigten Arbeitnehmer.

Gegenstand des dritten Projekts sind diejenigen „Betriebsvereinbarungen", die die Errichtung einer Transfergesellschaft vorsehen. Die Studierenden untersuchen mithilfe Künstlicher Intelligenz und Maschine-Learning-Methoden die Situationsbeschreibungen der „Betriebsvereinbarungen", etwa im Rahmen der Präambel oder anderer Textbausteine, und analysieren, ob bestimmte, sich wiederholende Hinweise innerhalb des Textes auf die Gründung einer Transfergesellschaft schließen. Darüber hinaus befassen sich die Studierenden mit der Frage, welche anderen realen Zusammenhänge die Errichtung einer Transfergesellschaft nach sich ziehen könnten.

Schlussbestimmungen: reine Formsache?

Als Bestandteil nahezu aller juristischer Vertragsgestaltungen enthalten die von Südwestmetall zur Verfügung gestellten „Betriebsvereinbarungen" ebenfalls eine Reihe unterschiedlicher Schlussbestimmungen. Der Kündigungsausschluss, die Mitteilung über die persönlichen Verhältnisse ebenso wie die Verankerung salvatorischer Klauseln sind nur einige Beispiele häufig auftretender Schlussbestimmungen in Interessensausgleichsregelungen sowie Sozialplänen.

Ein weiteres Projekt des Kurses beschäftigt sich mit ebenjenen Schlussbestimmungen. Im Rahmen des Projekts sollen die Studierenden die Schlussbestimmungen clustern und darauf aufbauend Rückschlüsse auf den weiteren Inhalt der „Betriebsvereinbarung" ziehen. Konkret verfolgt dieses Projekt das Ziel, zu überprüfen, ob unter Einsatz von Künstlicher Intelligenz und Maschinellem Lernen Zusammenhänge zwischen verschiedenen Varianten von Schlussbestimmungen sowie der Existenz einzelner Schlussklauseln mit anderen Eigenschaften der „Betriebsvereinbarungen", wie etwa einer konkreten Betriebsänderung oder einer für den Arbeitgeber besonders günstigen oder teuren Abfindungsregel, bestehen. Letzteres könnte beispielsweise

daran liegen, dass gewisse Rechtsanwaltskanzleien wiederholt bestimmte Muster an Schlussbestimmungen festlegen und dabei besonders gut oder schlecht für die jeweiligen Parteien verhandeln. Optimalerweise ist im Ergebnis ein bestimmtes Muster erkennbar, das die Aufmerksamkeit der Arbeitgeber bei der Existenz bestimmter Schlussbestimmungen weckt und Rückschlüsse auf die inhaltliche Regelung zulässt.

7. Ergebnisse

Im Wintersemester 2020/2021 haben die ersten 15 Studierenden den Projektkurs Data Science & Law erfolgreich durchlaufen. Während der Projektphase teilten sich diese in drei Teams zu je fünf Personen auf. Alle drei Teams haben den Pitch-Aspekt bei Vortrag am 8. Februar 2021 sehr lebhaft umgesetzt und sind als Startup mitsamt eigener CI und Mitarbeiterorganigramm aufgetreten. Auch inhaltlich bekamen wir bereits im ersten Durchlauf interessante Ergebnisse zu sehen:

BV-Check: Automatisierte Beurteilung von Betriebsvereinbarungen bzgl. Attraktivität für AG/AN

Das Startup „BV-Check“ klassifizierte die zur Verfügung gestellten Betriebsvereinbarungen über die Kriterien Abfindung, Erstattung von Umzugskosten, Erstattung von Fahrtkosten sowie Ausstattung der Transfergesellschaft von Hand in die Kategorien „Above Average“, „Average“ und „Below Average“ und übersetzten dadurch die vielfältigen Aspekte einer BV in ein einfaches Ampelschema.

In einem zweiten Schritt wurde ein ML-Algorithmus trainiert, der diese Kategorisierung bzw. Einschätzung in Zukunft automatisiert durchführt und dazu nur die BV im Word- oder PDF-Format benötigt. Mit neuronalen Netzwerken wurde dabei eine Accuracy von 77% erzielt, was mit Blick auf den kleinen Datensatz und dessen Heterogenität ein beachtliches Ergebnis ist. Mithilfe eines besseren Datensatzes und einer professionell durchgeführten Vorklassifizierung könnte der Ansatz von BV-Check in ein Tool zur schnellen Einschätzung einer BV weiterentwickelt werden, welches sowohl von Arbeitgebern als auch von Arbeitnehmern und Gewerkschaften eingesetzt werden könnte.

24 Legal: Vorhersage der Abfindungshöhe aus der Situationsbeschreibung

Das Startup „24 Legal“ entwickelt ein Tool, welches basierend auf der Situationsbeschreibung eines Unternehmens die Art der in der BV festgesetzten Abfindungen vorhersagen kann. Aufgrund des kleinen Datensatzes erfolgt die Vorhersage im Rahmen eines groben Rasters: Keine Abfindung, Strukturierte Formel (nur Betriebszugehörigkeit), Strukturierte Formel (Betriebszugehörigkeit & Alter) oder pauschale Abfindung. Mit einem größeren Datensatz könnte der Ansatz von 24 Legal in ein Tool weiterentwickelt werden, welches genauere Vorhersagen machen kann, also z.B. auch die Höhe von Faktoren in der Formel vorhersagen kann. Eine Anwendungsmöglichkeit eines solchen Tools ist das Aufdecken von nicht offensichtlichen Zusammenhängen zwischen den Eigenschaften von Regionen und Betrieben und dem Verhandlungsergebnis bei der Verfassung einer BV mit Sozialplan. Ggf. lässt sich ein solches Tool auch zur Aufdeckung von Diskriminierung oder zur Erlangung von Vorteilen bei Verhandlungen einsetzen.

Das Startup ABEEM verfolgt einen ähnlichen Ansatz, wobei die Kategorisierung der Abfindung hier bewusst noch gröber gewählt wird: keine Abfindung oder eine Abfindung in nicht weiter spezifizierter Höhe. Die Idee dahinter ist sich noch stärker auf die Aufdeckung von nicht offensichtlichen Zusammenhängen zu fokussieren. Eines der Ergebnisse: wenn die BV den Wortteil Rahmen enthält, also auf eine Rahmenbetriebsvereinbarung verweist, dann ist die Wahrscheinlichkeit einer Abfindungszahlung signifikant höher. Bei der Interpretation eines solchen Ergebnisses ist allerdings Vorsicht geboten. Oft steht ein doch allgemein bekannter Zusammenhang indirekt dahinter; hier z.B.: Größere Betriebe haben sowohl eine höhere Wahrscheinlichkeit eine Rahmen-BV aufzusetzen als auch gleichzeitig mehr politischen Druck im Falle von Entlassungen.

Zusammenfassung und weitere Ansätze Im Rahmen der Projektphase gelang es den Kursteilnehmern vielversprechende Erkenntnisse für die juristische Praxis zu erlangen. Durch die Anwendung von Maschine-Lear-

ning-Methoden konnten die Studenten Muster und Zusammenhänge in den Betriebsvereinbarungen erkennen und dadurch Betriebsvereinbarungen für den Arbeitgeber, unter anderem vor dem Hintergrund der konkret geregelten Abfindungshöhe, als besonders günstig oder teuer einordnen. Darüber hinaus entwickelten die Kursteilnehmer ein Tool, welches die Vorhersage von Abfindungsformeln innerhalb einer Betriebsvereinbarung ermöglichte. Hiermit gelang es, die Abfindungsformel in Abhängigkeit von Eigenschaften innerhalb einer Betriebsvereinbarung, die auf den ersten Blick nicht offensichtlich mit der Abfindung im Zusammenhang stehen, zu antizipieren. Ferner verfolgte eine Gruppe die Idee, zu untersuchen, ob die Existenz einer Abfindungsregel innerhalb eines Dokuments von einzelnen Wörtern oder Wortteilen abhängt.

Einschränkungen Die Ergebnisse der ersten Gruppen sind sehr vielversprechend, aber natürlich noch nicht so wirklich greifbar. Um aus den vorhandenen Ansätzen kommerziell nutzbare Softwareprodukte zu gewinnen sind nicht nur größere Datensätze und Parametertuning notwendig - es bedarf insbesondere einer konkreten Umsetzung mit einer benutzerfreundlichen UI, da die momentan in der Entwicklungsumgebung R-Studio laufenden Skripte für kommerzielle Anwender kaum verwendbar sind.

8. Ausblick

Nachdem der Projektkurs bereits zwei Semester nacheinander dem Lehrprogramm der Universität Ulm angehört, ist die Pilotphase abgeschlossen. Da die Studierenden mithilfe von Maschine-Learning-Methoden vielversprechende Ansätze und Erkenntnisse bei der Analyse von Rechtstexten gewinnen konnten, ist der Projektkurs Data Science & Law auch in den kommenden Semestern als fester Bestandteil des Lehrangebots der Universität Ulm eingeplant.

Aufgrund der begrenzten Anzahl an „Betriebsvereinbarungen“ und der Heterogenität der vorliegenden Vertragstexte sollen die Studierenden künftig mit deutlich homogeneren Arbeitsverträgen arbeiten. Sowohl aufgrund der Vielzahl als auch vor dem Hintergrund der Standardisierung der Arbeitsverträge und der daraus resultierenden größeren sowie homogeneren Datenmenge, insbesondere im Vergleich zu den bis dato verfügbaren Interessensausgleichen und Sozialplänen, könnten Algorithmen anhand von Arbeitsverträgen effektiver und schneller lernen und somit genauere Ergebnisse im Rahmen der Muster- und Zusammenhangserkennung liefern. Bei der Ausarbeitung der einzelnen Projekte wurde deutlich, dass sowohl die Genauigkeit als auch die Leistungsfähigkeit der Algorithmen mit größeren Datensätzen beträchtlich verbessert werden könnte.

Die Arbeitsverträge sollen von den Mitgliedsunternehmen des Arbeitgeberverbands Südwestmetall beschafft werden. Da eine manuelle Pseudonymisierung der Arbeitsvertragstexte, wie sie etwa im Rahmen der Anonymisierung der „Betriebsvereinbarungen“ geschah, aufgrund der Vielzahl an Verträgen mit enormem Aufwand verbunden wäre, könnte der Einsatz einer Pseudonymisierungssoftware Zeit und Aufwand sparen. Eine solche Software wäre zwar mit gewissen Kosten verbunden. Der Nutzen würde jedoch die Kosten deutlich überwiegen.

Insgesamt würde die Verwendung von Arbeitsverträgen im Rahmen der Projektphase die Genauigkeit und Leistungsfähigkeit der Muster- und Zusammenhangserkennung unter Anwendung Künstlicher Intelligenz und Maschine-Learning-Methoden wesentlich verbessern.

9. Literatur

Anzinger, Heribert M., 10 Jahre Modria: KMS und Online-Mediation auf dem Weg zur Digitalisierung der Justiz – Teil 1, Zeitschrift für Konfliktmanagement (ZKM), 2021, Heft 2, S. 53–57.

Goodenough, Oliver, Legal Technology 3.0, Huffpost, https://www.huffpost.com/entry/legal-technology-30_b_6603658 (aufgerufen am 19. Dezember 2022), Erscheinungsjahr 2015.