Shabnam TAUBÖCK[1], Anna SCHÖFECKER, Karl LEDERMÜLLER, Maria KRAKOVSKY, Sukrit SHARMA, Markus REISMANN, Christian Gregor MARSCHNIGG, Gerhard MÜHLBACHER, Julia SPÖRK, Michael SCHADLER & Gabriel WURZER (Wien)

PASSt – Predictive Analytics Services für Studienerfolgsmanagement

Zusammenfassung

Hochschulen haben zunehmendes Interesse daran, den Studienerfolg ihrer Studierenden analysieren und quantifizieren zu können. In diesem Zusammenhang versucht das Projekt PASSt – Predictive Analytics Services für Studienerfolgsmanagement – einen Rahmen für die empirische Analyse und Vorhersage des Studienerfolges herzustellen: Studenten- und Studiendaten werden in eine generische Datenstruktur importiert, auf die Machine Learning und Simulationen angewendet werden. Die beiden wichtigsten Ergebnisse der Anwendung dieser Ansätze sind eine Vorhersage des Studienerfolgs und eine Strukturanalyse von Lehrplänen, die zur Verbesserung der Studienbedingungen für Studierende genutzt werden können. Das Framework verfügt darüber hinaus über eine zusammenfassende Visualisierung, die eine einfache Interpretation und Nutzung der Ergebnisse für die Curriculumsplanung ermöglicht.

Schlüsselwörter

Studierbarkeit, Studienerfolg, Machine Learning, Simulation, Visualisierung, Prognose, Datenschutz

1 E-Mail: shabnam.tauboeck@tuwien.ac.at

DOI: 10.3217/zfhe-SH-HL/13

PASSt – Predictive Analytics Services for Study success management

Abstract

Universities are continuously striving to understand and quantify their students' study success. In this context, the PASSt project – Predictive Analytics Services für Studienerfolgsmanagement – seeks to develop a framework for the empirical analysis and prediction of student success. To this end, student and study data is imported into a generic data structure, to which machine learning and simulation are then applied. This framework produces two key results – a forecast of study success and a structural analysis of curricula – which can be used to improve study conditions for students. In addition, the framework offers an intuitive summarising visualisation tool that allows for easy interpretation and use of the results for curricular planning.

Keywords

studyability, study success, machine learning, simulation, visualisation, prediction, information privacy

1 Einführung

Universitäten erheben im Rahmen ihrer Aufgaben eine Menge an Daten, die nicht nur für die Abwicklung der Studien bzw. den universitären Alltag von Relevanz sind, sondern auch zur Verbesserung der Studierbarkeit und des individuellen Studienerfolgs der Studierenden herangezogen werden können.

Eine entsprechende Auswertung der Daten erforderte bis jetzt Expert:innenwissen, nicht nur weil die Art der Abfrage über Data Warehouse sowie Datenbanken rein technisch herausfordernd ist, sondern auch weil jede Universität eine andere Organisation und Struktur – und damit verknüpft ein individuelles Datenmodell – aufweist. Das Projekt PASSt („Predictive Analytics Services für Studienerfolgsmanagement") vereint die individuellen Datenmodelle jeder Universität in einer gene-

rischen Datenstruktur (vgl. Abb. 1: oben; siehe Abschnitt 3). Auf dieser aufbauend werden Auswertungen hinsichtlich *Studierbarkeit* und *Studienerfolg* durchgeführt. Diese werden erstmals an verschiedenen Universitäten entlang eines einheitlichen, angeleiteten Vorgehens (vgl. Abb. 1: Mitte sowie Abschnitt 4, „Predictive Analytics Services für Studienerfolgsmanagement") – unter Rücksichtnahme auf die Unterschiede in den Daten – mit einem hochwertigen Standard etabliert. Die dabei gewonnenen Resultate können Studienverantwortlichen und Universitätsleitungen sowie potenziell auch Studierenden anschließend auf einfach zugängliche Weise zur Verfügung gestellt werden (z. B. interaktives Web-Portal, Reports, integriert mit individuellem Datensystem der Universität; vgl. Abb. 1: unten bzw. vgl. mit Abschnitt 5, „Nutzungsorientierte Präsentation").

Das Expert:innenwissen ist nach wie vor entscheidend für den erfolgreichen Einsatz der hier entwickelten Modelle – denn trotz generisch angelegter Datengrundlage und variablem Modelldesign gibt es keine „one-fits-all" Lösung, es bedarf immer noch einer finalen Kalibrierung. Die in PASSt entwickelten Tools liefern eine Fülle an Ergebnissen – wesentlich für die korrekte Interpretation derselbigen ist dabei ein grundlegendes Verständnis der jeweiligen Hochschule, bspw. auf Ebene des Studienangebots (wie sind Studien aufgebaut, was sind spezielle Spezifika), ein Verständnis der Datenstruktur in der Hochschule (wie ist die Datenstruktur der Hochschule aufgebaut) sowie ein Verständnis über das eingesetzte Instrument (z. B. welche Daten zur Prognose/Simulation herangezogen wurden, wie diese Daten in Modelle übersetzt wurden [und was dabei die Vereinfachungen waren], ob Resultate qualitativ oder quantitativ zu verstehen sind) erforderlich.

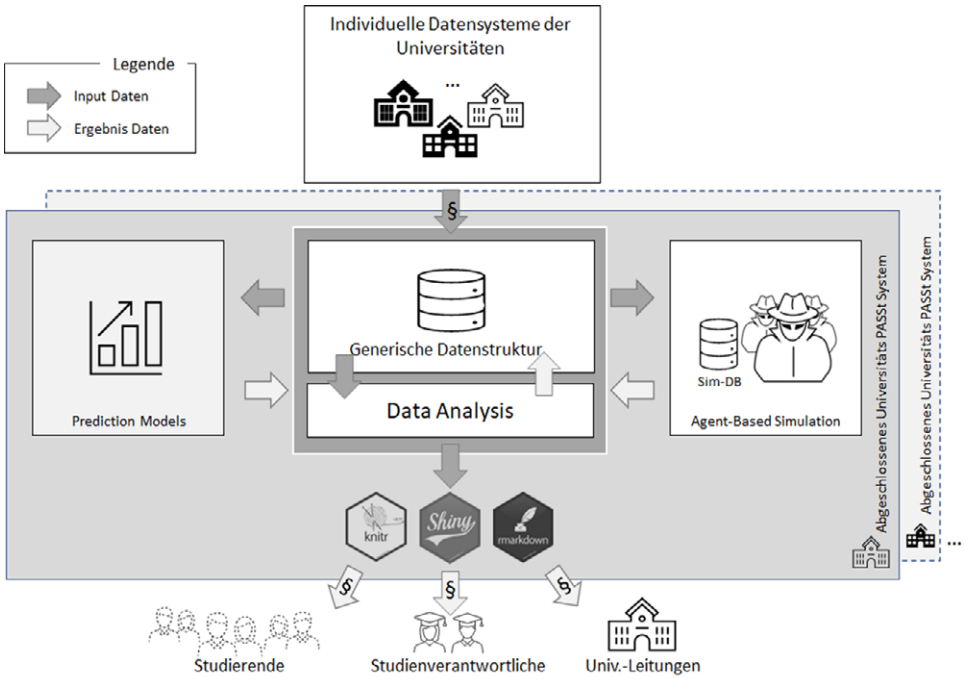

Abb. 1: Das PASSt Framework auf einen Blick

Der rechtliche Rahmen um PASSt wird, wie auch bei anderen Learning-Analytics-Projekten, primär durch das Datenschutzrecht gebildet. Soweit rechtliche Spielräume bestehen, soll ein solches Projekt dennoch ethischen Grundsätzen Genüge tun. Beiden Aspekten – sowohl dem rechtlichen als auch dem ethischen – wurde deshalb während der gesamten Projektlaufzeit, intensive Beachtung geschenkt. Die einzelnen Projektbestandteile (Daten, Abfragen sowie Präsentationsmethoden) wurden im Hinblick auf den aktuellen rechtlichen Rahmen in Österreich untersucht (siehe „Rechtlicher Rahmen“ in Abschnitt 6). Auf dieser Grundlage wird ein Katalog an Maßnahmen und zwingend zu beachtenden Kriterien ausgearbeitet, welcher auch über die Grenzen des Projekts hinaus für andere Ansätze mit demselben Ziel herangezogen werden kann (ebd.).

2 Hintergrund

Das Projekt PASSt wurde im Rahmen des Förderprogramms „Digitale und soziale Transformation“ des Österreichischen Bundesministeriums für Bildung, Wissenschaft und Forschung finanziert. Erklärtes Ziel ist, Studierende unter Berücksichtigung unterschiedlicher soziodemografischer und privater Rahmenbedingungen (soziale Herkunft, Erwerbstätigkeit, Betreuungspflichten etc.) bestmöglich bei der Absolvierung des Studiums – durch Einsatz von Prognosemodellen, Indikatoren und Planungswerkzeugen – zu unterstützen. Neben Methoden aus dem Bereich Predictive (Learning) Analytics sollen auch Methoden zur Maßnahmengenerierung sowie Simulation eingesetzt werden. Im Kern geht es bei diesen Maßnahmen um die Verbesserung der *Studierbarkeit* als Voraussetzung für die Förderung des *Studienerfolgs* (vgl. SPÖRK et al., 2021, S. 167f., sowie ZUCHA et al., 2020, S. 11). Das ist auch deshalb relevant, weil die Finanzierung von Universitäten mit einem Indikator verbunden ist, der mit Studienerfolg und Studierbarkeit im weiteren Sinne verbunden ist. Dieser Indikator wird als „Prüfungsaktivität“[2] bezeichnet.

Darüber hinaus ist PASSt auch mit einem zweiten Projekt (Learning Analytics – Studierende im Fokus) über einen inhaltlichen Cluster verbunden (siehe auch https://learning-analytics.at/learning-analytics-cluster). Im Rahmen dieses Clusters werden Fortschritte in den Projekten ausgetauscht und ein Wissenstransfer ermöglicht. Des Weiteren wird über eine Arbeitsgruppe zum Design von Vorhersagemodellen sichergestellt, dass es zu keinen Doppelentwicklungen kommt. Im Rahmen der Zusammenarbeit des Projektclusters werden Erfahrungen im Zusammenhang mit der Erstellung von Erklärungs- und Prognosemodellen für die Vorhersage des Studienerfolgs publiziert (BARTOK et al., 2023 [i.E.]).

2 Prüfungsaktive Studien werden in der österreichischen Wissensbilanz (BGBl. II Nr. 97/2016, §2.A.6) als Bachelor-, Diplom- und Masterstudien definiert, in denen im Studienjahr mindestens 16 ECTS-Punkte oder positiv beurteilte Studienleistungen im Umfang von acht Semesterstunden erbracht werden.

3 Generische Datenstruktur

Um eine einheitliche Auswertung der derzeit an den Universitäten existierenden Daten in puncto Studierende und deren Studien zu ermöglichen, wurde im Projekt eine gemeinsame, generische Datenstruktur definiert. Diese kann prinzipiell auch von nicht am Projekt teilnehmenden Universitäten befüllt werden. Die Datenstruktur wird an jeder Universität dezentral als Datenbank gehalten und auch nur dort befüllt, wodurch es zu keinem wechselseitigen Datenaustausch[3] kommt. Die Datenstruktur gliedert sich in die folgenden Tabellen (vgl. mit Abb. 2):

Daten Studierende. Enthält pseudonymisierte Daten zu Alter, Geschlecht, Herkunft, Bildungshintergrund sowie außeruniversitären Nebentätigkeiten und Betreuungspflichen von Studierenden.

Daten LV. Enthält die Beschreibung zu jeder Lehrveranstaltung (LV), welche in einem bestimmten Semester (z. B. 2022W) angeboten wurde; dazu zählen eine eindeutige ID[4], ein Titel sowie der Umfang in ECTS sowie Semesterwochenstunden.

3 D. h. die Quelldaten werden weder zwischen den (teilnehmenden) Universitäten noch nach oben hin mit dem Bundesministerium für Bildung, Wissenschaft und Forschung kommuniziert

4 Bei uns wird die ID – also die Einheit, die einer LV entspricht, „Inhaltsklammer“ genannt, weil manche Universitäten unter Umständen auch mehrere Kurse unter einer Klammer (bspw. wegen sehr hoher Nachfrage parallel organisierte Lehrveranstaltungen des gleichen Fachs) zusammenfassen und die inhaltlich einheitlichen Lehrveranstaltungen als Lehreinheit zusammenfassen. Die Inhaltsklammer kann somit als Klammer um die inhaltlich gleichen Lehrveranstaltungen definiert werden.

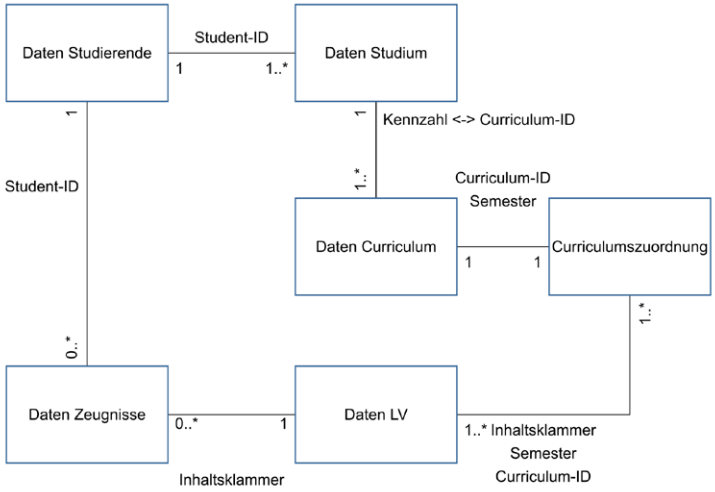

Abb. 2: Zusammenhang zwischen den Tabellen der generischen Datenstruktur

Daten Curriculum. Beschreibt einen Studienplan zum Stand eines bestimmten Semesters (z. B. 2022W), den Typ (Bachelor, Master, Diplom, Doktorat, Erweiterungsstudium) sowie seine Charakteristik (Ist Lehramt? Gibt es eine Studieneingangs- und Orientierungsphase (StEOP)? Sprache).

Curriculumszuordnung. Ordnet eine LV einem oder mehreren Curricula zu (1:n-Beziehung). An diese Zuordnung werden auch etliche Daten angehängt, die spezifisch für die Rolle des Fachs in dem benannten Curriculum gelten: Studienplan-Typ (Pflicht-, Wahl- oder Freifach), empfohlenes Stoffsemester, allfällige Voraussetzungen für das Fach (Prüfungsketten, StEOP).

Daten Studium. Entspricht der Belegung eines Curriculums durch Studierende (1:n-Beziehung); für jedes solches Studium werden hier individuelle Metadaten pro Student:in (z. B. Beginnsemester, Aufnahmeverfahren StEOP-Absolvierung, Studienabschluss oder -abbruch) geführt.

Die generische Datenstruktur ist für eine gemeinsame Vorgehensweise bei der Auswertung notwendig; einzelne Applikationen können diese Datenstruktur entweder

direkt verwenden (z. B. nutzungsorientierte Präsentation [Abschnitt 5]), oder sie leiten aus der generischen Datenstruktur eigene Daten ab (z. B. eigene Datenschnittstelle für Prediction [Abschnitt 4.1], Simulationsdatenbank für die agentenbasierte Simulation [mittig in Abb. 1 bzw. Abschnitt 4.2]). In letzterem Fall kann auch eine Rückführung von den applikationsspezifischen Resultaten zurück in die generische PASSt Datenstruktur erfolgen.

4 Predictive Analytics Services für Studienerfolgsmanagement

Der Bereich Evaluierung an Hochschulen bezieht sich längst nicht mehr nur auf einzelne Lehrveranstaltungen, Prüfungen, Programme oder den Lifecycle der Studierenden. Seit einigen Jahren werden die Studierenden in den Fokus gerückt und Hochschulen sind zunehmend daran interessiert, den Studienerfolg ihrer Studierenden besser zu verstehen und empirisch zu quantifizieren. Mit verbesserter und differenzierterer Wissenslage fällt es leichter, interne Maßnahmen und Handlungen zu setzen, um vielfältige Ziele zu begründen, zu unterstützen und zu erreichen.

Basierend auf Prognosemodellen können neben der Vorhersage des individuellen Studienabbruchrisikos (BEAULAC & ROSENTHAL, 2018) oder der Darstellung der individuellen Studierbarkeit und der Diversitätsgerechtigkeit eines Studienprogramms (BUSS, 2019) auch externe Faktoren – wie die Koppelung der Universitätsfinanzierung an das Aktivitätsniveau der Studierenden bei den Zielen – mitberücksichtigt werden.

Black-Box-Modelle, wie die im folgenden Abschnitt präsentierten Machine-Learning-Ansätze, liefern akkurate Ergebnisse in der Vorhersage abhängiger Variablen, sind jedoch gleichzeitig in Bezug auf die Erklärbarkeit der Wichtigkeit und inneren Zusammenhänge zwischen Modellvariablen eingeschränkt (vgl. MOLNAR, 2022). Ergänzend können auch White-Box-Ansätze, wie die im weiteren Verlauf präsentierte Agent-Based-Simulation (ABS; siehe z. B. WILENSKY & RAND, 2015) eingesetzt werden. Diese bieten zwar wenig akkurate Ergebnisse als Black-Box-Modelle, sind aber von ihren inneren Regeln her komplett transparent gehalten. Fundamental ist hierbei, dass es kein Entweder-oder gibt; beide Ansätze haben ihre

Anwendungsbereiche innerhalb der zugrundeliegenden Fragestellung. Werden sie aufeinander aufbauend angewendet, ergibt sich ein ausbalanciertes Bild hinsichtlich Nachvollziehbarkeit und Genauigkeit. Im gegenständlichen Fall liefert die ABS eine Sicht auf strukturelle Defizite innerhalb eines Curriculums (z. B. zu wenig Durchsatz durch zu wenig Ressourcen oder zu geringe Erfolgsraten), ohne dass die dabei simulierten Studierenden 1:1 auf die realen Studierenden pro LV übertragbar wären. Machine Learning wird im Gegenzug dazu herangezogen, sehr präzise Leistungsprognosen zu erstellen, die auf einer Vielzahl individueller Faktoren und deren Verknüpfung untereinander basieren.

4.1 Prediction Models

Fragestellungen zum Aktivitätsniveau, zur Studierbarkeit, zur Studienaktivität oder zu diversen Abbruchrisiken können mithilfe von Prognose- und Beschreibungsmodellen, wie sie in der multivariaten Statistik häufig angewandt werden, zielgerichtet analysiert und beantwortet werden. Durch die Analyse der Wichtigkeit einzelner Einflussfaktoren, den Fokus auf die Relevanz der Variablen und die Komplexitätsreduktion, wo dies ohne nennenswerte Informationsverluste geschehen darf, wird versucht, Zusammenhänge zu erklären und zu beschreiben, Entwicklungen werden prognostiziert und die Anwendungen in den Modellen versuchen, Entscheidungsgrundlagen zu liefern (STOETZER, 2020, S. 151f.). Regressions- bzw. Prognosemodelle werden verwendet, um unterschiedliche Fragestellungen zu beantworten, wie beispielsweise:

- Welche Studierenden haben erhöhtes Risiko hinsichtlich eines Studienabbruchs oder Prüfungsinaktivität?
- Welche Faktoren beeinflussen diese Risiken?
- Welche Faktoren bedingen bzw. hemmen einen zügigen Studienfortschritt?
- Wie identifizieren wir relevante Zielgruppen für Serviceleistungen?
- Welche Zielgruppen haben spezifischen Unterstützungsbedarf?

4.1.1 Die statistische Modellierung von Studienerfolg in PASSt

Um Studienerfolg (in ECTS-Punkten oder Prüfungsaktivität) zu modellieren, werden unterschiedliche Regressionsmodelle und andere Ansätze aus dem Bereich des Machine Learning verwendet, die dabei helfen, komplexe Sachverhalte als Modell formal beschreibbar und/oder prognostizierbar zu machen. Klassische Ansätze wie OLS-Regressionen, generalisierte additive Modelle oder logistische Regressionen werden dabei ebenso angewandt (siehe z. B. KREMPKOW, 2020). Diese sind etwa die in Kapitel 4.1.3 vorgestellten Ansätze Random Forests, Gradient Boosting Machine-Modelle oder Support Vector Machine. Die Modelle werden pro Studiengang und Studienjahr berechnet, das Modell mit den besten Vorhersagewerten wird für die Modellierung automatisiert ausgewählt und die Ergebnisse werden dargestellt. Dabei wird folgendermaßen vorgegangen.

4.1.2 Teilung des Datensatzes

Basierend auf den Daten aus einem vergangenen Studienjahr wird der Gesamtdatensatz randomisiert und in einen Trainings- und Validierungsdatensatz (Verhältnis 70 % : 30 %) geteilt. Im Trainingsdatensatz wird das Modell trainiert. Mittels Validierungsdatensatz bewertet das Modell automatisiert seine eigene Leistung. Daten des vorherzusagenden Studienjahres werden eingegeben, die Prüfungsaktivität oder der Studienerfolg wird prognostiziert.

4.1.3 Automatisierte Wahl des besten Regressionsverfahrens

Startend mit **klassischen** linearen bzw. logistischen **Regressionen** werden in den Methoden zwei weitere Verallgemeinerungen von klassischen Regressionsmodellen verwendet. Generalisierte additive Modelle (gam) identifizieren automatisch geeignete Transformationen der Prädiktoren. Diese Modelle können nichtlineare und nicht-monotone Beziehungen zwischen abhängigen und unabhängigen Variablen berücksichtigen. Auch die **(boosted) logistische Regression** wird durch eine Verallgemeinerung des linearen Modells erreicht. Sie dient zur Vorhersage einer nominalskalierten (dichotomen) abhängigen Variable durch mehrere unabhängige Variablen. Die vorherzusagende Variable in diesen Modellen ist die Prüfungsaktivität, die bei unseren Daten die Ausprägungen 0 oder 1 hat. Die Schätzung der Regressionsparameter findet üblicherweise über sogenannte Maximum-Likelihood-Schätzungen

unter Zuhilfenahme eines iterativen Algorithmus statt. Bei diesem Supervised-Machine-Learning-Verfahren werden innerhalb des Trainingsdatensatzes Studien mit der Ausprägung prüfungsaktiv bzw. prüfungsinaktiv versehen. Diese Klassifikation passiert im Lernalgorithmus, der ein Modell generiert, welches die Kennzeichnung für weitere Punkte (aus dem Evaluierungsdatenset) vornimmt. Im davon getrennten Evaluierungsdatenset werden falsch klassifizierte Datenpunkte identifiziert und unterschiedlich stark gewichtet (boosting), um eine gute Anpassung für neue Daten zu erzielen (SCHAPIRE, 1990; BURKOV, 2019). Für das Boosting wird der **LogitBoost-**Algorithmus von FRIEDMAN et al. (2000) verwendet.

Die Methode **Random Forest** (ebenso eine Supervised-Machine-Learning-Methode) (BREIMAN, 2001) kombiniert eine Vielzahl an Entscheidungsbäumen[5], um die neuen Daten vorherzusagen. Dabei gibt es einige sogenannte Bäume, die das Ergebnis korrekt vorhersagen, andere, die nicht korrekt sind. Durch die Kombination und das Training, welches parallel über einen Algorithmus bei der Erstellung geschieht, sagen alle Bäume zusammen sehr gut geschätzte Ergebnisse vorher. Wichtig ist, dass die Bäume untereinander eine sehr niedrige Korrelation aufweisen. Dadurch sollen sich Schwächen der einzelnen Entscheidungsbäume durch unterschiedliche Zufallsstichproben und unterschiedliche Zufallsteilmengen an Prädiktoren bei jeder Verzweigung ausgleichen. Gleichzeitig wird auch das Risiko des Overfittings (Überanpassung an die Trainingsdaten) minimiert.

Gradient-Boosting-Machine-Modelle (gbm) basieren ebenfalls auf Entscheidungsbäumen, welche jedoch sequenziell unter Verwendung von Informationen aus den vorherigen Bäumen aufeinander aufgebaut werden. Während bei Random Forests viele unterschiedliche Bäume für die finale Prognose herangezogen werden, die gegenseitig ihre Schwächen ausgleichen, wird im Gradient-Boosting-Machine-Modell jeder neue Baum so gebaut, dass der Prognosefehler aus dem vorherigen Baum im neuen Baum geringer ist. Diese Methode erzielt bei unserem Beispiel gemeinsam mit dem „LogitBoost“ die beste Treffsicherheit.

Die **Support Vector Machine** (svmLinear) versucht, Objekte mithilfe von Trennungslinien oder -ebenen zu teilen und die Daten so zu separieren. Diese Ebenen

5 Entscheidungsbäume stellen Entscheidungsregeln dar, sie bestehen aus einem Wurzelknoten, auf den weitere Knoten mit jeweils mindestens zwei Blättern folgen. Die Knoten stellen jeweils eine Entscheidungsregel, die Blätter die Ausprägungen (ja oder nein) dar.

werden so gewählt, dass der Abstand zwischen den verschiedenen Klassen maximiert wird. Im 2-dimensionalen Bereich wird eine Trennlinie gezogen, im 3-dimensionalen Bereich eine Trennfläche eingezogen und bei 4 oder mehr Dimensionen eine sogenannte Hyperplane. Durch die Anwendung des Kernel-Tricks lässt sich die Methode auch bei nicht-linearen Entscheidungsgrenzen einsetzen: Hierfür werden die Trennungsvektoren in eine zusätzliche Dimension transformiert. Verglichen mit den anderen Methoden ist die Treffsicherheit in diesem Beispiel mit der Methode etwas niedriger.

4.1.4 Ergebnisse der Regressionsverfahren

Zum besseren Verständnis der Regressionsverfahren wurden zwei Plots angefertigt (Abb. 3): Im Importance Plot (Abb. 3 oben) ist erkennbar , welche Variablen für die Prognosen von Prüfungsaktivität unter Verwendung des GBM entscheidend sind. Die Darstellung entlang des standardisierten Regressionskoeffizienten bei der Modellierung von Studienerfolg [ECTS] (Abb. 3 unten) zeigt für die lineare Regression darüber hinaus, welche Variablen signifikanten Einfluss auf den Studienerfolg haben.

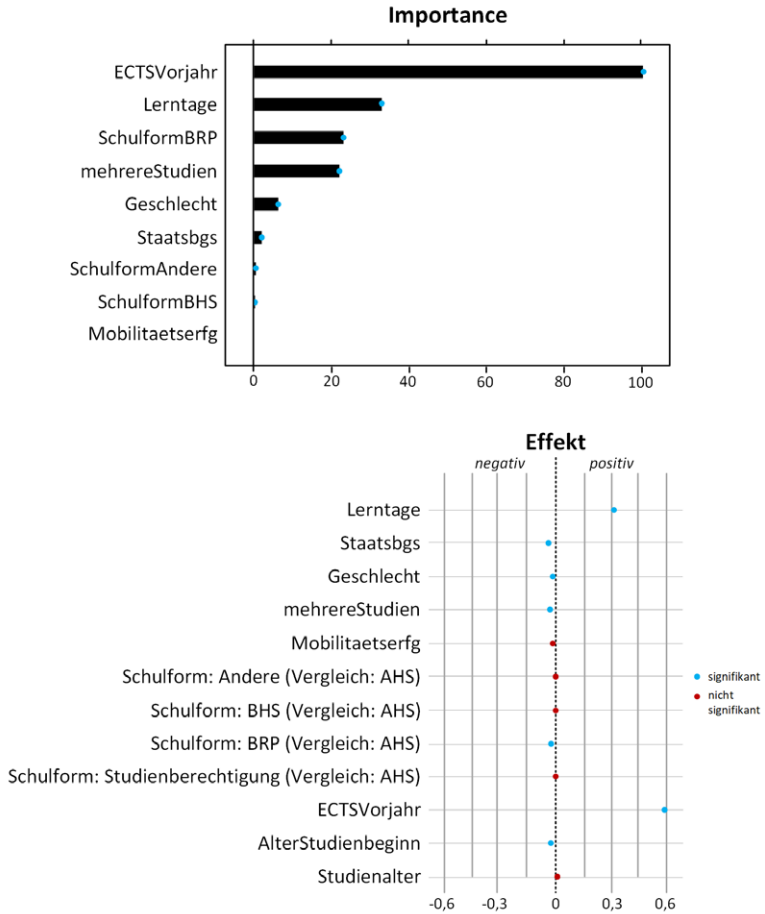

Abb. 3: Plots: (oben) Wichtigkeit von Variablen (unten) Einfluss auf Studienerfolg; vgl. BARTOK et al., 2021

Die größte Rolle für die gute Prognose spielen in diesem Beispiel die ECTS-Punkte aus dem vorgehenden Studienjahr, doch auch die Intensität der Verwendung der Lernplattform (hier als Lerntage bezeichnet) hat starken Einfluss. Jene Studierende, welche die Berufsreifeprüfung ablegten, und die Information, ob mehrere Studien gleichzeitig besucht werden, beeinflussen die Prognoseergebnisse. Hingegen scheinen die Staatsbürgerschaft, das Geschlecht oder ob vor dem Studium eine allgemeine oder berufsbildende höhere Schule besucht wurde, weniger relevant zu sein. In der Grafik ist ersichtlich, dass die Variable Mobilitätserfahrung, die die Absolvierung eines Auslandssemesters darstellt, als nicht wichtig einzustufen ist, was, ebenso wie die anderen Variablen zwischen den Studienprogrammen und Personengruppen, die in der Stichprobe sind, variiert.

Wie bereits erläutert, wird über die beste Methode zur Prognose die Prüfungsaktivität des neuen Studienjahres vorhergesagt. Dies kann in anonymer Weise passieren, wenn sich die Fragestellung zum Beispiel auf Bereiche bezieht, in denen die Hochschule nach Verbesserungspotenzial sucht. Bei manchen Fragestellungen liefert bereits die deskriptive Analyse des Studienerfolgs interessante Einblicke, andere Fragestellungen können besser mittels Entscheidungsbäumen oder Clusteranalysen erörtert werden. Im Rahmen der rechtlichen Voraussetzungen können aus technischer Sicht auch personalisierte Prognosen über den Studienerfolg erstellt werden.

4.2 Agent-Based Simulation

In der agentenbasierten Simulation eines Studiums (WURZER et al., 2022) werden „virtuelle Studierende“ entsprechend den im Vorfeld analysierten (Winter/Sommer-) Beginnzahlen generiert[6], anschließend wird deren Fortschreiten durch das Studium simuliert. Genauer gesagt werden „virtuelle Studierende“ in Form von *Agenten* (aktiver Teil der Simulation) repräsentiert, welche eine bestimmte Anzahl an ECTS in Pflicht-, Wahl- und Freifächern absolvieren müssen, um ein Studium abzuschließen. Die *tatsächliche Studienleistung pro Semester* ergibt sich dabei (a.) aus einer im

6 Die Generierung erfolgt komplett anonym; aus dem historischen (pseudonymisierten) Studierendenvolumen wird im Zuge der Kalibrierung die Verteilung von für die Simulation wichtigen Merkmalen ermittelt, anschließend wird in der Simulation ein komplett neues Studierendenvolumen mit derselben Merkmalsverteilung in der gewünschten Zahl erzeugt.

Vorfeld kalibrierten durchschnittlichen Anzahl an ECTS, welche pro Semester geleistet werden können (siehe links in Abb. 4), (b.) einer prognostizierten Anzahl an ECTS, welche sich aus den soziodemografischen Parametern der Agenten ergeben (vgl. mit Prognose in Abschnitt 4.1.1), andererseits (c.) steht jeder Agent in Konkurrenz zu anderen Agenten, wenn es um die Belegung der Fächer (passiver Teil der Simulation) geht. Aus dem Zusammenspiel zwischen den Agenten und den zu belegenden Fächern ergibt sich ein dynamisches System, in dem mögliche Handlungsfelder (Engpässe/Flaschenhälse, limitierendes Studienverhalten) aufgezeigt und mittels Veränderung der Parameter einer Lösung zugeführt werden können (z. B. Änderung des Studienplans, siehe rechts in Abb. 4).

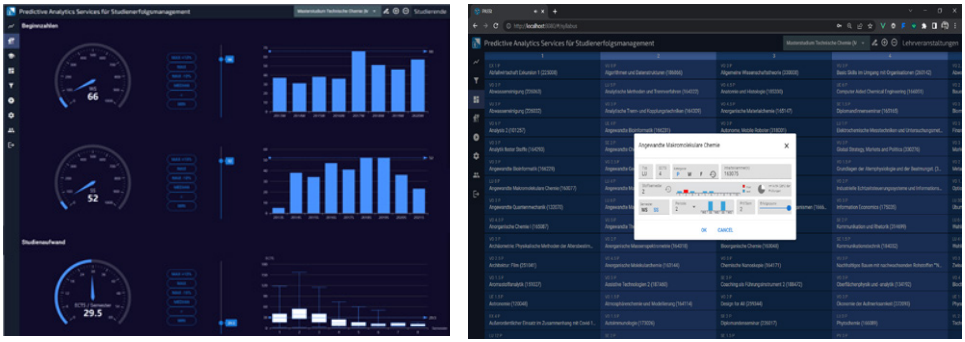

Abb. 4: Simulation: (links) Einstellung von Parametern [Anzahl Studienbeginn-ende WS/SS, durchschnittliche Leistung in ECTS/Semester] (rechts) Studienplan mit Detailansicht zu Fach (Semester/Periode/Kapazität/…)

Die zu belegenden Fächer bilden die *Ressourcen* der Simulation: Aus der Prüfungshistorie je Studium wird – egal ob es sich beim betreffenden Fach um ein Pflicht-, Wahl- oder Freifach handelt – eine Kapazität an Prüfungsplätzen je Semester errechnet. Agenten können ein Fach genau dann belegen, wenn es noch freie Prüfungsplätze gibt und dieses Fach im betreffenden Stoffsemester angeboten wird. Das Stoffsemester wird wahlweise (a.) statisch hinterlegt, oder (b.) ebenfalls durch eine im Vorfeld vollzogene Kalibrierung bestimmt (Verteilung von Erstprüfungen auf Stoffsemester je Fach). Nachprüfungen sind vom Stoffsemester ausgenommen, sprich: Studierende können prinzipiell jederzeit misslungene Prüfungen wiederho-

len, es sei denn, es gibt keine freien Prüfungsplätze. In diesem Fall werden die Agenten zur Warteschlange des Fachs hinzugefügt und dürfen je nach gewählter Prüfungsstrategie (Parameter) vor etwaigen Erstbelegenden die Prüfung nachmachen. Tritt eine durchgängige Überbelegung (durch zu wenig Kapazität oder hohe Anzahl an Wiederholenden) in einem oder mehreren Fächern auf, so werden diese als Flaschenhälse markiert.

Das Thema Studierbarkeit kann nun durch (1.) Beeinflussung der Prüfungskapazitäten bei den jeweiligen Fächern, (2.) Verschiebung in ein anderes Semester sowie (3.) Einflussnahme auf die Anzahl der Beginnenden geschehen. Des Weiteren wird die Sensitivität (bzw. vice versa die Resilienz) eines gegebenen Studienplans gegenüber Veränderungen der Anzahl an Beginnenden aus dem Unterschied zwischen minimaler und maximaler Auslastung jedes Fachs errechnet (Details dazu in 4.2.1). Studienverzögerungen können zudem mittels der vonseiten der Agenten verzeichneten Agentenhistorie (ähnlich einem Logbuch aller Antritte) genauer eingegrenzt werden; es ergibt sich damit auch eine potenzielle Kausalitätskette aus verzögerten Antritten, welche longitudinal über alle Agenten verglichen werden können.

4.2.1 Technische Details

Die Agentensimulation hält ihre Daten in einer eigenen Datenbank, welche im ersten Schritt (Kalibrierung) aus der generischen PASSt Datenstruktur befüllt wird. Nach diesem Schritt wird eine Validierung vorgenommen, um die Plausibilität der kalibrierten Daten zu gewährleisten (z. B. Beginnzahlen). Die Daten stehen dann für die Parametrisierung zur Verfügung. Es können sowohl Beginnzahlen als auch Studienleistung in ECTS für ein spezifisches Studium eingestellt werden; diese Durchschnittszahlen können weiters durch ECTS-Berechnung mittels Prognose über die soziodemografischen Charakteristika der virtuellen Agenten (= generierten Studierenden) errechnet werden. In der Simulation werden anschließend mittels einer Variation der eingestellten Parameter (= „Parameter Sweep“) mehrere Durchläufe gemacht, um auf eine Aussage hinsichtlich Durchsatz (Abschlüsse je Semester) und Länge (Anzahl an Semestern pro Abschluss) des Studiums zu kommen; durch Vergleich zwischen errechneten Minima und Maxima in den Resultaten kann auf die Sensibilität geschlossen werden. Die Ergebnisse werden weiters im Zuge einer Validierung den historischen Daten gegenübergestellt. Die aufbereitete Ausgabe erfolgt

mittels Datentransformation, welche auf die nutzungsorientierte Visualisierung (siehe nächster Abschnitt) zurechtgeschnitten ist.

5 Nutzungsorientierte Präsentation

Wie in Abb. 1 (unten) dargestellt, gibt es drei Hauptzielgruppen für die Anwendung von PASSt: Studierende[7], Studienverantwortliche sowie die Universitätsleitung. Dazu kommen etliche Sichten, die je nach Anwendungsfall, variieren. Um ein einheitliches Aussehen des Info-Cockpits zu gewährleisten, wurde ein Portal geschaffen, in dem die einzelnen Datenbestände aus dem PASSt-Projekt zusammengefasst und nach Studienrichtung sowie Zeitraum gefiltert werden können:

Überblick: Die in der generischen Datenstruktur hinterlegten Tabellen (vgl. Abschnitt 3) werden direkt auf der Hauptseite des Portals angezeigt (vgl. Abb. 5). Hierzu zählen die Anzahl an Beginnenden/Abschlüssen/Abbrüchen und Fortmeldung pro Jahr, Studienleistung (0–3; 4–7; 8–15; 16–39; 40+ ECTS); weitere, mögliche Sichten sind: die durchschnittliche Studienlänge in Semestern, Anzahl (positiv/negativer) Prüfungen.

7 Aufgrund Reduktion des Projektumfangs wird auf die Zielgruppe der Studierenden im Projekt nur teilweise eingegangen.

Abb. 5: Überblicksdarstellung (Konzept): Visualisierung der generischen Datenstruktur nach Studierenden [Beginner:innen, Abbrecher:innen, Fortgemeldet, Absolvent:innen] sowie Prüfungsleistung je Semester [ECTS]

Prognose/Regression (vgl. Abschnitt 4.1): Eine Darstellung der prognostizierten Studienleistung pro Studium [ECTS] kann als Dichteverteilung (Abb. 6: unten) präsentiert werden. Weiters können Konfidenz und Fehlerwerte der Prognose angezeigt werden (Abb. 6: oben), um eine Einschätzung ihrer Güte zu erlauben.

Simulation (vgl. Abschnitt 4.2): Eine Zusammenfassung der Ergebnisse aus der Simulation kann interaktiv und pro Studium gezeigt werden (Abb. 7); weiters kann die Simulation als Planungstool geöffnet werden, um Veränderungen am Studienplan vorzunehmen und deren Einflüsse zu visualisieren (Abb. 4: rechts).

Weitere Implementierungen und Anwendungen im Cockpit umfassen Features wie bspw. den Export von Reports, um Ergebnisse auch über die direkte Anwendung hinaus (bspw. in Meetings) nutzbar machen zu können.

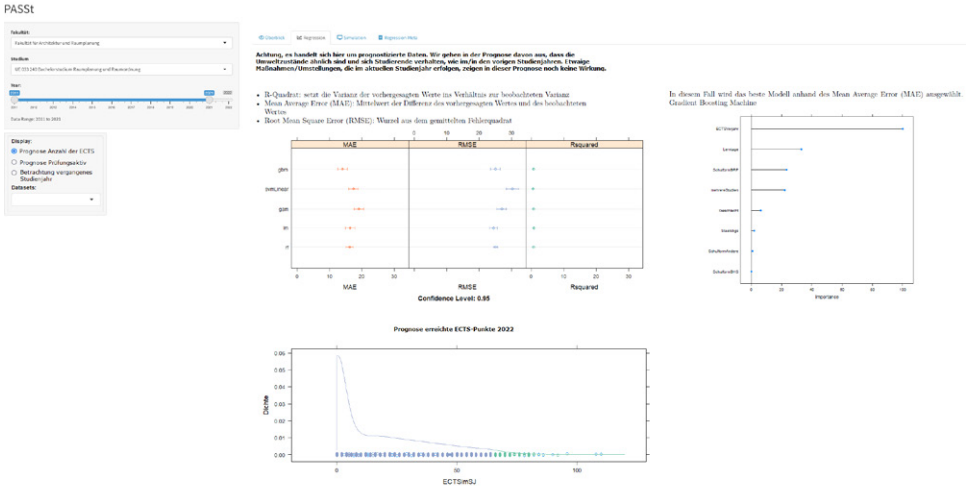

Abb. 6: Visualisierung der Regression (Konzept): Prognose ECTS, Fehlerwerte, verwendete Machine-Learning-Methode

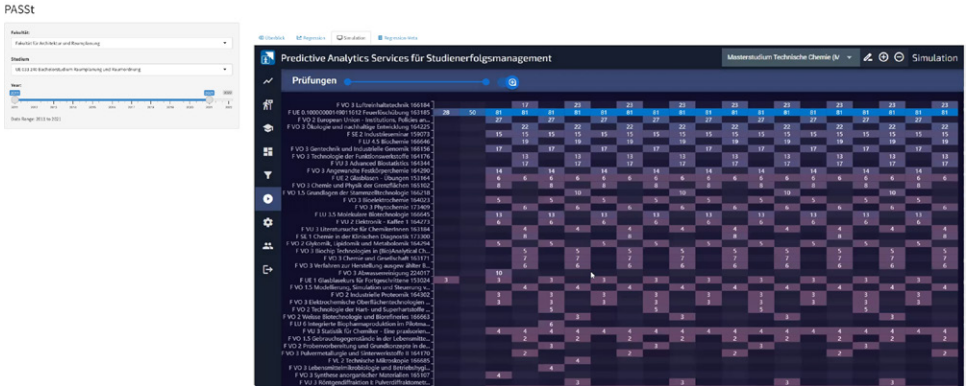

Abb. 7: Visualisierung der Simulation (Konzept): Simulierter Studienplan mit errechneten Auslastungen der Lehrveranstaltungen (absolut; farbcodiert)

6 Rechtlicher Rahmen und Ethik

6.1 Einleitung

Wie bereits eingangs erwähnt, unterliegt das Projekt einer laufenden rechtlichen und ethischen Begleitung, damit von Beginn an mögliche Risiken eingeschätzt und proaktiv Maßnahmen zu deren Eindämmung ergriffen werden können. Ziel der Begleitung ist die Sicherstellung der Einhaltung des rechtlichen und ethischen Rahmens sowie die Ableitung transferierbarer Erkenntnisse.

6.2 Rechtliche Rahmenbedingungen und Problemfelder

Die Identifikation und Analyse der datenschutzrechtlichen Rahmenbedingungen für die Verarbeitung von Studierendendaten sind eine zentrale Forschungsaufgabe im Rahmen des Projekts.

Universitäten verfügen naturgemäß über Studierendendaten. Sie dürfen sie aber nur für festgelegte, eindeutige und legitime Zwecke erheben und nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeiten (Art. 5 Abs. 1 lit. b DSGVO). Die Grundlage für die Erhebung und (Weiter-)Verarbeitung dieser Daten bilden einerseits die Datenschutzgrundverordnung (DSGVO) und das österreichische Datenschutzgesetz (DSG), andererseits die hochschulrechtlichen spezifischen Regelungen des Universitätsgesetzes (UG) und des Bildungsdokumentationsgesetzes (BilDokG 2020).

Je nach Zweck und den damit einhergehenden Use Cases ist zu prüfen, ob eine mit den ursprünglichen Zwecken vereinbare Weiterverarbeitung der Studierendendaten vorliegt oder ob zur Gewährleistung der Rechtmäßigkeit der Verarbeitung ein anderer Rechtfertigungsgrund für diese Verarbeitung einschlägig sein muss, wobei insbesondere an die Gründe der Einwilligung oder einer einschlägigen (unionalen oder staatlichen) Rechtsvorschrift zu denken ist (Art. 6 Abs. 4 DSGVO).

6.3 Datenschutz durch Pseudonymisierung

Aus Performancegründen und zur erhöhten Sicherheit werden die Daten durch jede Universität in eine jeweils separate Datenbank importiert. Im Zuge dessen werden die Daten zunächst pseudonymisiert.

Zu Beginn des Projekts wurde auch eine Anonymisierung angedacht, weil eine agentenbasierte Simulation nicht unbedingt eine Rückführbarkeit auf eine bestimmte Person benötigt, sondern ihre Berechnung mit anonymen bzw. anonymisierten Daten durchführt. Im Gegensatz dazu ist jedoch für einige andere Use Cases oder Methoden, wie eine angedachte Regressionsanalyse, die Rückführbarkeit gerade auch für Überprüfung der Vorhersagequalität sinnvoll.

Zur Verdeutlichung und zum besseren Verständnis dessen kann folgendes Beispiel betrachtet werden: Es findet eine Betrachtung der Kohorte „RE“ zum Zeitpunkt X und eine Vorhersage der Prüfungsaktivität für X1 (p) statt. Um eine Verifizierung für X1 (p) zu erlangen, ist es notwendig, auf die reale Kohorte für X1 zugreifen zu können. Dies dient einerseits zur Feststellung der Prüfungsaktivität dieser Kohorte, andererseits bildet dies eine essenzielle Kontrollfunktion für das PASSt-Tool. Ein fortlaufendes Monitoring und gegebenenfalls Anpassungen der Modelle und Methoden sind wesentliche Kernelemente jeder Anwendung und daher sinnvollerweise vorzusehen.

Die Anonymisierung unterscheidet sich von der Pseudonymisierung insofern, als die (Studierenden-)Daten bei Ersterem ausnahmslos anonym sein sollen, womit diese anonymisierten Daten im weiteren Prozess nicht mehr einer Person zugeordnet werden können. Damit wird der Personenbezug dauerhaft ausgeschlossen (ROSSNAGEL & SCHOLZ, 2000; LEITNER, MAYRHOFER & STADLBAUER, 2022). Pseudonymisierung hingegen wird definiert als „die Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden“ (Art. 4 Z. 5 DSGVO).

Aus datenschutzrechtlicher Sicht gilt die Pseudonymisierung unter anderem als eine geeignete technische und organisatorische Maßnahme im Sinne des Art. 32 DSGVO.

Das bedeutet, dass die Pseudonymisierung als eine zentrale Standardmaßnahme zur Gewährleistung der Sicherheit der Verarbeitung angesehen werden kann. Dieser erste Verarbeitungsvorgang der Pseudonymisierung dient dem sogenannten Grundsatz der Datenminimierung nach Art. 5 Abs. 1 lit. c DSGVO, wonach Daten dem Zweck angemessen und erheblich sowie auf das notwendigste Maß beschränkt sein müssen. Erwägungsgrund 28 DSGVO zufolge soll die Anwendung der Pseudonymisierung die Risiken der betroffenen Person senken und die Verantwortlichen oder Auftragsverarbeiterinnen und Auftragsverarbeiter bei der Einhaltung der Datenschutzpflichten unterstützen. Weiters ist zu bemerken, dass es keine Verpflichtung der DSGVO gibt,[8] eine Pseudonymisierung von personenbezogenen Daten durchzuführen. Diese Maßnahme wurde von den beteiligten Universitäten als Verantwortliche gesetzt, da die Verarbeitungssicherheit von hoher Relevanz ist. Im weiteren Projektverlauf wird ein allgemeiner rechtlicher Rahmen erstellt werden, der als Handlungsempfehlung für einen Roll-out dient.

6.4 Ethik: Code of Practice

Zusätzlich zur Einhaltung des rechtlichen Rahmens soll (und darf) auch die ethische Dimension eines Vorhabens im Bereich Learning Analytics nicht außer Acht gelassen werden. Daher ist Ziel des begleitenden ethischen Arbeitspakets ein daraus resultierender „Code of Practice", der einen Anhaltspunkt für den Umgang mit Studierendendaten einerseits für ein Roll-Out, andererseits aber für eine Erweiterung der Funktionen bieten soll.

Um ein großes Spektrum an Wissen dafür generieren zu können, wurde zusätzlich zur Berücksichtigung der einschlägigen Literatur eine interdisziplinäre Herangehensweise in der Art gewählt, dass nicht nur potenzielle Anwender:innen der Universitätsverwaltung und Curriculakommission bereits Feedback einbringen konnten, sondern auch Expert:innen aus den Bereichen des Antidiskriminierungsrechts, der Bildungsforschung und der Didaktik sowie dem Bereich des „Secure Systems". Die Erkenntnisse und Ergebnisse dessen werden entsprechend aufberei-

8 Art 32 Abs 1 DSGVO sieht die Pseudonymisierung nur als eine der technischen und organisatorischen Maßnahmen an, die ein dem Risiko angemessenes Schutzniveau gewährleisten sollen.

tet und als „Code of Practice“ den Projektpartnern zur Verfügung gestellt. Darin werden die wesentlichen Gesichtspunkte aus ethischer Sicht dargestellt, sodass bei entsprechenden Anwendungen eine fundierte Abwägungsentscheidung durch die betroffenen Stellen gewährleistet werden kann.

7 Schlussbetrachtung und Ausblick

Mit dem Projekt PASSt (TU Wien, WU Wien, JKU Linz) wird derzeit auf Basis einer generischen Datenstruktur eine Prognose von Studienerfolg mittels Machine Learning und eine Auswertung von Studierbarkeit mittels agentenbasierter Simulation erprobt. Parallel dazu wird ein begleitender Maßnahmenkatalog angefertigt, der die rechtlichen und ethischen Rahmenbedingungen für den Einsatz des Projekts – auch für einen künftigen Einsatz an weiteren Universitäten – möglich machen soll.

Die Übertragbarkeit der Projektergebnisse – nicht aber der gegenseitigen Austausch von Daten zwischen den Universitäten – wird durch die beschriebenen Applikationen zur Visualisierung und Prognose sowie der Curriculumsplanung ermöglicht: Die Übertragbarkeit ist auf mehreren Ebenen gegeben. Inhaltlich: Es könnten durch die einheitliche Datenstruktur unterschiedliche Studien an verschiedenen Standorten über eine gemeinsame Metrik verglichen werden bzw. ergeben sich durch die Verwendung desselben Datenmodells auch vergleichbare Prognosen. Besonders sinnvoll ist so ein Vergleich im Fall gemeinsamer Studien mehrerer Universitäten (z. B. Lehramt), da somit ein übergreifendes Gesamtbild geschaffen und das Lehrangebot jeder einzelnen Universität innerhalb solcher Verbundstudien geschärft werden kann. Technisch: die zur Entwicklung verwendeten Open-Source-Lösungen helfen anderen Universitäten dabei, die Projektergebnisse (im Sinne der technischen Infrastruktur) zu implementieren, auf die eigenen Bedürfnisse, über das gemeinsame Projektziel hinaus, maßzuschneidern und weiterzuentwickeln. Prozessual: Es gibt mehrere Diskussionen, im Rahmen des Projekts, aber auch im Rahmen des gemeinsamen projektübergreifenden Clusters (bspw. im Rahmen von Workshops), mit Kolleg:innen anderer Hochschulen in Austausch zu treten. Auch hier könnte eine Verwertbarkeit der Projektergebnisse implementiert werden – dieser Verwertungsstrang (also prozessual) befindet sich aber derzeit noch in der frühen Konzeptionsphase.

8 Literaturverzeichnis

Bartok, L., Gleeson, R. & Kriegler-Kastelic, G. (2021). The impact of individual factors on definitions of academic success at an Austrian University. *Studierbarkeit und Studienerfolg: Zwischen Konzepten, Analysen und Steuerungspraxis, 4*, 119.

Bartok, L., Spörk, J., Gleeson, R., Krakovsky, M. & Ledermüller, K. (2023 [i.E.]). *Handreichung zu Erfahrungen bei der Anwendung unterschiedlicher Machine Learning Ansätze in Prognosemodellen zum Studienerfolg.*

Beaulac, C. & Rosenthal, J. S. (2018). Predicting university students' academic success and choice of major using random forests. *ArXiv e-prints*. https://arxiv.org/abs/1802.03418v1

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Burkov, A. (2019). *Machine Learning kompakt: Alles, was Sie wissen müssen*. Luxemburg: MITP-Verlags GmbH & Co. KG.

Buß, I. (2019). The relevance of study programme structures for flexible learning: an empirical analysis. *Zeitschrift für Hochschulentwicklung*, *14*(3), 303–321.

Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337–407.

Krempkow, R. (2020). Determinanten der Studiendauer – individuelle oder institutionelle Faktoren? Sekundärdatenanalyse einer bundesweiten Absolvent(inn)enbefragung. *Zeitschrift für Evaluation*, *19*(1), 37–63.

Leitner, P., Mayrhofer, M. & Stadlbauer, M. (2022). Datenschutzrechtliche Aspekte der Nutzung von Krankenhausinformationssystemen für Forschungszwecke. *Zeitschrift für Technikrecht (ZTR), 4*(4), 207–222.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd Edition)*. ISBN 9798411463330 (Taschenbuch). https://christophm.github.io/interpretable-ml-book

Roßnagel, A. & Scholz, P. (2000). Datenschutz durch Anonymität und Pseudonymität – Rechtsfolgen der Verwendung anonymer und pseudonymer Daten. *Multimedia und Recht (MMR) 3*(12), 721–729.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, *5*(2), 197–227.

Spörk, J., Ledermüller, K., Krikawa, R., Wurzer, G. & Tauböck, S. (2021). Analyse von Studierbarkeit mittels Prognose und Simulationsmodellen. *Zeitschrift für Hochschulentwicklung, 16*(4), 163–182. https://doi.org/10.3217/zfhe-16-04/09

Stoetzer, M. W. (2020). *Regressionsanalyse in der empirischen Wirtschafts-und Sozialforschung. Bd. 2*. Berlin, Heidelberg: Springer.

Wilensky, U. & Rand, W. (2015). *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. Cambridge, Ma.: MIT Press.

Wurzer, G., Reismann, M., Marschnigg, C., Dorfmeister, A., Tauböck, S., Ledermüller, K. & Spörk, J. (2022). PASSt-A: Agent-based student analytics aimed at improved feasibility and study success. *IFAC-PapersOnLine, 55*(20), 361–366.

Zucha, V., Zaussinger, S. & Unger, M. (2020). Studierbarkeit und Studienzufriedenheit. Zusatzbericht der Studierenden-Sozialerhebung 2019. Institut für Höhere Studien – Institute for Advanced Studies (IHS).

Autor:innen

Dipl.-Ing. Dr. Shabnam TAUBÖCK || TU Wien, Zentrum für strategische Lehrentwicklung || Karlsplatz 13, A-1040 Wien

www.tuwien.ac.at

shabnam.tauboeck@tuwien.ac.at

Mag. Anna SCHÖFECKER, LL.B. || JKU Linz, LIT Law Lab und Institut für Verwaltungsrecht und Verwaltungslehre || Altenberger Straße 69, A-4040 Linz

www.jku.at

anna.schoefecker@jku.at

Dr. Karl LEDERMÜLLER || WU Wien, Evaluierung & Qualitätsentwicklung || Welthandelsplatz 1, A-1220 Wien

www.wu.ac.at

karl.ledermueller@wu.ac.at

Mag. Maria KRAKOVSKY || WU Wien, Evaluierung & Qualitätsentwicklung || Welthandelsplatz 1, A-1220 Wien

www.wu.ac.at

maria.krakovsky@wu.ac.at

Sukrit SHARMA, BSc || TU Wien, Information Technology Solutions || Operngasse 11, A-1040 Wien

www.tuwien.ac.at

sukrit.sharma@tuwien.ac.at

Markus REISMANN || TU Wien, Studienbezogenes Daten- und Projektmanagement || Karlsplatz 13, A-1040 Wien

www.tuwien.ac.at

markus.reismann@tuwien.ac.at

Christian Gregor MARSCHNIGG, BSc || TU Wien, Studienbezogenes Daten- und Projektmanagement || Karlsplatz 13, A-1040 Wien

www.tuwien.ac.at

christian.marschnigg@tuwien.ac.at

Mag. Gerhard MÜHLBACHER || JKU Linz, Qualitätsmanagement & Berichtswesen || Altenberger Straße 69, A-4040 Linz

www.jku.at

gerhard.muehlbacher@jku.at

Julia SPÖRK, MA || WU Wien, Evaluierung & Qualitätsentwicklung || Welthandelsplatz 1, A-1220 Wien

www.wu.ac.at

julia.spoerk@wu.ac.at

Ing. Michael SCHADLER || TU Wien, Information Technology Solutions || Operngasse 11, A-1040 Wien

www.tuwien.ac.at

michael.schadler@tuwien.ac.at

Priv.-Doz. Dipl.-Ing. Dr. Gabriel WURZER || TU Wien, Zentrum für strategische Lehrentwicklung || Karlsplatz 13, A-1040 Wien

www.tuwien.ac.at

gabriel.wurzer@tuwien.ac.at