Schlussbericht zur Verwaltungsvereinbarung Methodenforschung Verlaufsstatistik

Zuwendungsempfänger	Förderkennzeichen
Statistisches Bundesamt – Gruppe H2 Bildung, Forschung, Kultur, Rechtspflege	16P4273

Vorhabenbezeichnung

Methodenforschung Verlaufsstatistik - KI und Maschinelles Lernen in der Studienverlaufsdatenbank

Laufzeit des Vorhabens

01.10.2020 bis 30.12.2023

Berichtszeitraum

01.01.2020 bis 30.12.2023

I. Kurze Darstellung

1 Aufgabenstellung

Die auf Grundlage des novellierten Hochschulstatistikgesetzes von 2016 neu eingerichtete zentrale Studienverlaufsdatenbank mit Studienverlaufsinformationen auf Einzeldatenbasis bietet Analysemöglichkeiten, die mit den bisher in der amtlichen Hochschulstatistik eingesetzten Auswertungsverfahren allein schon wegen des zunehmend großen Datenbestands nicht ausgeschöpft werden (können). Vor diesem Hintergrund hatte das Bundesministerium für Bildung und Forschung (BMBF) ein Methodenprojekt ausgeschrieben, das darauf abzielte, den Einsatz von KI-Verfahren bzw. Verfahren des maschinellen Lernens in der Studienverlaufsdatenbank zu prüfen.

Die Leistungsbeschreibung für das Methodenprojekt umfasste zwei Teilaufgaben. Die erste Teilaufgabe zielte konkret auf eine „verbesserte Schätzung von Studienerfolgsquoten“, die für Hochschulforschung und Hochschulpolitik eine gleichermaßen hohe Bedeutung haben. Die zweite Teilaufgabe zielte allgemein darauf, die „Analysepotentiale der Studienverlaufsdatenbank besser auszuschöpfen“.

2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Auf Basis des Hochschulstatistikgesetzes von 1992 führen die Statistischen Ämter semesterweise Datenerhebungen zu Studierenden und Hochschulprüfungen als dezentrale Bundesstatistiken durch: Die im Hochschulstatistikgesetz geregelten Merkmale werden bei den Hochschulverwaltungen abgefragt und durch die regional jeweils zuständigen statistischen Ämter aufbereitet. Die Veröffentlichung der statistikspezifischen Ergebnisse erfolgt in aggregierter Form v.a. als tabellarische Darstellung. Das Statistische Bundesamt erstellt die Bundesergebnisse auf Basis der plausibilisierten Länderergebnisse.

Seit 2007 berechnet das Statistische Bundesamt auf Basis der Aggregatdaten aus Studierenden- und Prüfungsstatistik Studienerfolgsquoten. Hierbei wird zunächst für eine Anfängerkohorte der Gesamtanteil der Studierenden, die ihr Studium bis zum Datenrand bereits abgeschlossen haben, empirisch bestimmt. Im Rahmen der Prüfungsstatistik gemeldete Abschlussprüfungen werden über das Semester der Ersteinschreibung einer Anfängerkohorte aus der Studierendenstatistik zugeordnet. Für die Noch-Studierenden einer Anfängerkohorte zum Zeitpunkt der Erfolgsquotenberechnung wird auf Basis eines polynomialen Modells eine Abschlusswahrscheinlichkeit geschätzt.

Mit der Novelle des Hochschulstatistikgesetzes von 2016 wurde das Statistische Bundesamt beauftragt, eine zentrale Studienverlaufsdatenbank aufzubauen, um eine regelmäßige Studienverlaufsstatistik erstellen zu können. Datenbasis hierfür sind die Einzeldaten der Studierenden- und Prüfungsstatistik, außerdem die Einzeldaten der neuen Statistik über Promovierende, die ab dem Berichtsjahr 2017 von den Ländern an das Statistische Bundesamt geliefert werden müssen und in die Datenbank eingespeichert werden (dürfen). Die Verknüpfung der Einzeldatensätze im Zeitverlauf und über die Statistiken hinweg ermöglicht ein Pseudonym, das aus gemeinsamen und unveränderlichen Merkmalen der drei Statistiken gebildet und mit den berichtszeitraumbezogen erhobenen Statistik-Datensätzen in der Studienverlaufsdatenbank gespeichert wird.

3 Planung und Ablauf des Vorhabens

Die Projektplanung erfolgte gemeinsam durch das Fachreferat für Hochschulstatistik sowie das Methodenreferat für Künstliche Intelligenz/Big Data im Statistischen Bundesamt und wurde mit dem auftraggebenden Referat 123 im BMBF abgestimmt. Im Rahmen der Abstimmung wurden zu den beiden Teilprojekten („Schätzung von Studienerfolgsquoten verbessern“ sowie „Analysepotential der Studienverlaufsdatenbank ausschöpfen“, siehe I.1) jeweils 5 Einzelaufgaben konkretisiert. Die Bearbeitung der Aufgaben durch das vom Statistischen Bundesamt angeworbene Projektpersonal erfolgte sukzessive in enger Abstimmung sowohl mit dem Fach- als auch dem Methodenbereich im Statistischen Bundesamt sowie in regelmäßigen Austausch mit dem Auftraggeber sowie dem Ausschuss für die Hochschulstatistik. Die Berechnungen erfolgten auf Grundlage der Studienverlaufsdatenbank.

4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Die Arbeiten wurden auf Basis des aktuellen wissenschaftlichen und technischen Standes zu Verfahren und Methoden des maschinellen Lernens durchgeführt. Angeknüpft wurde zudem an den aktuellen wissenschaftlichen Stand der nationalen Studienverlaufsforschung.

5 Zusammenarbeit mit anderen Stellen

Bei den Arbeiten erfolgte eine enge Zusammenarbeit einerseits mit den Expertinnen und Experten für maschinelles Lernen im Statistischen Bundesamt sowie im Hochschulstatistikausschuss, andererseits mit diversen Stellen, die in Deutschland Studienverlaufsforschung betreiben.

II. Eingehende Darstellung

1 Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele

Die Verwendung der Zuwendung erfolgte für Personalausgaben beim Statistischen Bundesamt und für sonstige sächliche Verwaltungsausgaben. Personalausgaben entstanden für die projektbezogen eingeworbenen wissenschaftlichen Mitarbeitenden sowie für darüber hinaus zeitweise im Projekt eingesetztes Dauerpersonal.

2 Wichtigste Positionen des zahlenmäßigen Nachweises

Von den entstandenen Ausgaben entfallen auf Personalausgaben gemäß Position 0812 343.255,00 EUR und auf sonstige sachliche Verwaltungsaufgaben gemäß Position 0843 34.325,50 EUR (siehe auch Gesamtausgabennachweis für die Laufzeit 01.10.2020 bis 31.12.2023).

3 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die Arbeiten wurden vom Statistischen Bundesamt in Abstimmung mit dem BMBF geplant und gesteuert. Der Ausschuss für die Hochschulstatistik war eingebunden.

4 Voraussichtlicher Nutzen, insbesondere Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans

Die im Rahmen des Projekts entwickelten neuen Visualisierungsformen wurden in den Internetauftritt des Statistischen Bundesamts zur Hochschulstatistik aufgenommen und werden regelmäßig aktualisiert. Die im Rahmen des Projekts entwickelten Verbesserungen der Erfolgsquotenberechnung auf Aggregatdatenbasis wurden in die Aufbereitungsprogramme eingearbeitet und werden ab sofort bei der Veröffentlichung der auf dieser Basis berechneten Quoten berücksichtigt.

Voraussichtlich ab dem Jahr 2027 wird zudem die empirische Bestimmung der Noch-Studierenden und Absolvent/-innen bis zum aktuellen Datenrand auf die neue Einzeldatenbasis umgestellt. Dies verspricht eine deutlich verbesserte Qualität der Erfolgsprognose auch schon im Verfahren ohne KI-Einsatz.

Im Rahmen des Projekts konnte des Weiteren gezeigt werden, dass sich durch den Einsatz von KI-Verfahren in der Studienverlaufsdatenbank nicht nur die Qualität der Erfolgsquotenberechnung gegenüber dem bisherigen Verfahren verbessern, sondern zudem der zeitliche Abstand zwischen dem Start einer Anfängerkohorte und der Berechnung des Erfolgs für diese Kohorte erheblich reduzieren lässt. Das entwickelte Modell wurde für die Prognose von Studienerfolg und -abbruch in der Masteranfängerkohorte des WS 2018/2019 bereits erfolgreich verwendet und könnte, mit geringfügigen Anpassungen, auch für künftige Masteranfängerkohorten angewendet werden.

Wegen der noch begrenzten Verfügbarkeit historischer Daten in der Studienverlaufsdatenbank war es bisher nicht möglich, Erfolgsquoten für grundständige Studiengänge zu berechnen und zu evaluieren. Das bereits entwickelte KI-Modell für Masteranfängerkohorten bietet aber eine gute Basis für die Entwicklung eines KI-Modells zur Prognose von Studienerfolg und -abbruch in grundständigen Studiengängen.

Für die Anpassung des für Masterstudiengänge entwickelten KI-Modells, um damit Erfolg und Abbruch in grundständigen Studiengängen prognostizieren zu können, ist spezifische KI-Expertise erforderlich. Dies gilt auch für die Anwendung des bereits entwickelten KI-Modells auf künftige Masteranfängerkohorten. Da solche Expertise derzeit beim Stammpersonal der Hochschulstatistik wie auch in den meisten anderen Fachbereichen im Statistischen Bundesamt nicht vorhanden ist, kann ein weiterer KI-Einsatz in der Studienverlaufsstatistik nicht ohne Unterstützung zusätzlichen, entsprechend qualifizierten Personals erfolgen.

5 Während der Durchführung des Vorhabens dem Zuwendungsempfänger bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Während der Durchführung des Vorhabens wurde dem Zuwendungsempfänger kein Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen bekannt.

6 Erfolgte oder geplante Veröffentlichungen des Ergebnisses

Ergebnisse zu den einzelnen Arbeitspaketen des Projekts wurden dem Hochschulstatistikausschuss präsentiert, die Präsentationen finden sich als Anlagen zu den Protokollen der jeweiligen Sitzungen. Soweit im Rahmen des Projekts neue Visualisierungsformen oder Ergebnisformate entwickelt wurden bzw. werden, sind diese über das Internetangebot des Statistischen Bundesamts zur Hochschulstatistik bereits derzeit verfügbar bzw. werden verfügbar gemacht. Sobald die im Rahmen des Projekts entwickelten neuen Methoden bei der Abschätzung des Studienerfolgs der Noch-Studierenden sowie bei der Prognose von Studien(miss)erfolg für die regelmäßige Ergebniserstellung verwendet werden, erfolgt eine Beschreibung in den Qualitätsberichten.

III. Erfolgskontrollbericht

1 Beitrag zu den förderpolitischen Zielen

Trifft nicht zu.

2 Wissenschaftlich-technisches Ergebnis des Vorhabens, erreichte Nebenergebnisse und gesammelte wesentliche Erfahrungen

Siehe angehängter Ergebnisbericht.

3 Fortschreibung des Verwertungsplans

- Erfindungen/Schutzanmeldungen

  Trifft nicht zu.

- Wirtschaftliche Erfolgsaussichten nach Projektende

  Trifft nicht zu.

- Wissenschaftliche und/oder technische Erfolgsaussichten nach Projektende

  Siehe dazu II.4.

- Wissenschaftliche und wirtschaftliche Anschlussfähigkeit

  wissenschaftliche Anschlussfähigkeit:

  Die technische und konzeptionelle Ähnlichkeit der Studienverlaufsdatenbank mit dem geplanten Bildungsverlaufsregister verspricht, dass die entsprechenden Erkenntnisse zur Nutzbarkeit von KI zur Analyse von Studienverläufen auch für künftige Verlaufsanalysen zum gesamten Bildungsbereich genutzt werden können.

- wirtschaftliche Anschlussfähigkeit:

  Trifft nicht zu.

4 Arbeiten, die zu keiner Lösung geführt haben

Trifft nicht zu.

5 Präsentationsmöglichkeiten für mögliche Nutzer

Die im Rahmen des Projekts entwickelten neuen Visualisierungsformen sind frei verfügbar und können von möglichen weiteren Nutzerinnen und Nutzern verwendet werden. Die für den Hochschulstatistikausschuss erstellten Präsentationen zum entwickelten KI-Modell zur Prognose des Studienerfolgs können unter Verweis auf die Quelle verwendet werden.

6 Einhaltung des Ausgaben- und Zeitplans

Der zunächst bis zum 30.09.2023 befristete Projektzeitplan wurde aufgrund einer zwischenzeitlichen Vakanz der Projektmitarbeitenden im Jahr 2021 in Abstimmung mit dem Auftraggeber kostenneutral bis zum 30.12.2023 verlängert. Der verlängerte Projektplan wurde eingehalten.

Von den veranschlagten Mitteln wurden insgesamt 120.594 Euro nicht abgerufen, darunter 75.599 Euro, weil die veranschlagten Kosten für Dienstreisen, Veranstaltungen und Hardware/Lizenzgebühren nicht angefallen sind. Die restliche Differenz zwischen kalkulierten und abgerufenen Ausgaben entfällt auf Personal- und Gemeinkosten.

Für den veranschlagten Personalaufwand von 61,5 PM E13 wurde wissenschaftliches Zeitpersonal angeworben. Das angeworbene Zeitpersonal hat das Projekt teilweise vorzeitig verlassen. Da die Rekrutierung von Ersatzpersonal für die restliche Projektlaufzeit erfolglos blieb, hat der Auftraggeber im März 2023 zugestimmt, dass die Arbeitspakete 3 „Prüfung von Verfahren zur Verknüpfung externer Informationen

mit der Studienverlaufsdatenbank auf Basis statistischer Ähnlichkeiten“ und 5 „Konzeption eines Prognosemodells für studienverlaufsbezogene Ereignisse auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden“ der zweiten Teilaufgabe („Analysepotentiale der Studienverlaufsdatenbank besser ausschöpfen“) durch den Auftragnehmer nicht im Rahmen des Projekts bearbeitet werden. Um die anderen Arbeitspakete trotzdem termingerecht erledigen zu können, wurde zusätzlich zum projektbezogen eingeworbenen Zeitpersonal zeitweise Dauerpersonal eingesetzt.

Anlage zum Schlussbericht
Verwaltungsvereinbarung Methodenforschung Verlaufsstatistik

Wissenschaftlich-technisches Ergebnis des Vorhabens, erreichte Nebenergebnisse und gesammelte wesentliche Erfahrungen

I. Einleitung

Die auf Grundlage des novellierten Hochschulstatistikgesetzes von 2016 neu eingerichtete zentrale Studienverlaufsdatenbank mit Studienverlaufsinformationen auf Einzeldatenbasis bietet Analysemöglichkeiten, die mit den bisher in der amtlichen Hochschulstatistik eingesetzten Auswertungsverfahren allein schon wegen des zunehmend großen Datenbestands nicht ausgeschöpft werden (können). Vor diesem Hintergrund hatte das Bundesministerium für Bildung und Forschung ein Methodenprojekt ausgeschrieben, das darauf abzielte, den Einsatz von KI-Verfahren bzw. Verfahren des maschinellen Lernens in der Studienverlaufsdatenbank zu prüfen.

Die Leistungsbeschreibung für das Methodenprojekt umfasste zwei Teilaufgaben mit jeweils fünf Arbeitspaketen.

a. Verbesserte Schätzung von Studienerfolgsquoten
    1. Überprüfung der Ermittlung der Abschlusswahrscheinlichkeit der Noch-Studierenden bei der aktuellen Berechnung von Erfolgsquoten auf Basis der vorliegenden Summensätze der Studierenden- und Prüfungsstatistik
    2. Weiterentwicklung der Ermittlung der Abschlusswahrscheinlichkeit der Noch-Studierenden für die aktuelle Berechnung von Erfolgsquoten auf Basis der vorliegenden Summensätze der Studierenden- und Prüfungsstatistik und ggf. Einbau in Rechenmodell
    3. Modellierung der Abschlusswahrscheinlichkeit von Noch-Studierenden auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden
    4. Entwicklung eines Modells zur Berechnung von Erfolgsquoten auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden und Implementierung in Studienverlaufsdatenbank
    5. Entwicklung eines Modells zur Berechnung von Abbrecherquoten auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden und Implementierung in Studienverlaufsdatenbank

b. Analysepotentiale der Studienverlaufsdatenbank besser ausschöpfen
    1. Methodische und technische Vorarbeiten für geo-basierte Auswertungen und weitere geeignete graphische Analyseformen auf Basis der Studienverlaufsdatenbank
    2. Identifikation relevanter Gruppen und Faktoren mit KI-Verfahren zur Mustererkennung und Clustering
    3. Prüfung von Verfahren zur Verknüpfung externer Informationen mit der Studienverlaufsdatenbank auf Basis statistischer Ähnlichkeiten

4. Konzeption von Datenvisualisierungsverfahren, um die Informationen der Beobachtungen in der Studienverlaufsdatenbank schnell und intuitiv erschließen zu können
5. Konzeption eines Prognosemodells für studienverlaufsbezogene Ereignisse auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden

Nach vorzeitigen Personalabgängen im Projekt, die in der verbliebenen Projektlaufzeit nicht vollständig ausgeglichen werden konnten, hatte der Auftraggeber zugestimmt, dass – bei entsprechender Reduzierung der Vertragssumme – die Arbeitspakete 3 „Prüfung von Verfahren zur Verknüpfung externer Informationen mit der Studienverlaufsdatenbank auf Basis statistischer Ähnlichkeiten“ und 5 „Konzeption eines Prognosemodells für studienverlaufsbezogene Ereignisse auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden“ der zweiten Teilaufgabe („Analysepotentiale der Studienverlaufsdatenbank besser ausschöpfen“) nicht bearbeitet werden.

II. Detaillierte Ergebnisse zu den einzelnen Arbeitspaketen

1 Verbesserte Schätzung von Studienerfolgsquoten

1.1 Überprüfung der Ermittlung der Abschlusswahrscheinlichkeit der Noch-Studierenden bei der aktuellen Berechnung von Erfolgsquoten auf Basis der vorliegenden Summensätze der Studierenden- und Prüfungsstatistik

Seit 2007 berechnet und veröffentlicht das Statistische Bundesamt jährlich Studienerfolgsquoten, die auf Basis von Aggregatdaten aus den laufenden Statistiken über Studierende und Prüfungen ermittelt werden. Über das in beiden Statistiken erhobene Merkmal „Jahr der Ersteinschreibung“ werden Daten der Prüfungsstatistik zu Absolventen/Absolventinnen einer Studienanfängerkohorte zugeordnet. Ein Summenvergleich der Studienanfänger/Studienanfängerinnen und Absolventen/Absolventinnen ermöglicht dann Aussagen zum Anteil der erfolgreichen Studierenden in der entsprechenden Kohorte. Erfolg ist dabei als Erwerb eines ersten Hochschulabschlusses definiert.

Die geschätzte Erfolgsquote ergibt sich als Quotient aus

- der Summe der Absolventen/Absolventinnen bis zum aktuellen Datenrand und der mit der Abschlusswahrscheinlichkeit multiplizierten Noch-Studierenden am Datenrand und
- den Studierenden im 1. Hochschulsemester eines Anfängerjahrgangs.

Dabei erfolgt das bisherige Verfahren der Schätzung von Erfolgsquoten in den folgenden zwei Schritten:

a. Empirische Bestimmung der Noch-Studierenden und Absolventen/Absolventinnen bis zum Datenrand[1] für eine Studienanfängerkohorte 8 oder mehr Jahre nach deren Studienbeginn
b. Abschätzung der zukünftigen Absolventen/Absolventinnen am Datenrand für die Noch-Studierenden dieser Anfängerkohorte

Schritt b) unterteilt sich in folgende Unterschritte:

[1] Ausnahme sind Studierende, für die in den Prüfungsdatensätzen keine Angabe zum Jahr der Ersteinschreibung vorliegt. Diese werden anteilsmäßig den Anfängerkohorten zugeordnet.

b.1 Empirische Bestimmung der historischen Abnahme-Raten[2] und Abschluss-Rate (Anfänger/-innen mit erfolgreichem Abschluss), je Berichtsjahr und Anfänger-Kohorte, auf Basis der historischen Prüfungsdaten mindestens der letzten 20 Jahre

b.2 Berechnung der durchschnittlichen, historischen Abnahme-Rate sowie der Abschluss-Rate je n Jahre nach Studienbeginn

b.3 Jahrweise Vorhersage der zukünftigen Noch-Studierenden und Absolventen/Absolventinnen bis zu einem Horizont von 20 Jahren nach Start der Kohorte[3] auf Basis der Abnahme-Raten und Abschluss-Raten aus b.2 Ein so langer Vorhersagezeitraum ist erforderlich um zu erreichen, dass dann die gesamte Anfängerkohorte nahezu vollständig auf die Stadien „Erfolg“ bzw. „Abbruch/Misserfolg“ aufgeteilt werden kann

b.4 Bestimmung der noch zu erwartenden Absolventen/Absolventinnen bis zur kumulierten Absolventenzahl nach 20 Jahren (je Berichtsjahr und Anfänger-Kohorte). Daraus werden die folgenden Wertepaare ermittelt. Wert 1: Anteil der Noch-Studierenden in Prozent gemessen an der Anfänger-Zahl der Kohorte. Wert 2: Zugehörige zu erwartende Absolventen-Quote in Prozent gemessen an der kumulierten Absolventenzahl nach 20 Jahren

b.5 Ausschließen aller Wertepaare mit Noch-Studierenden-Anteilen von mehr als 20% und Ausschließen aller Wertepaare mit erwarteter Absolventenquote von 0%[4], jeweils in den für eine aktuelle Veröffentlichung relevanten fünf Anfängerjahren

b.6 Polynomiale Regression durch die verbleibenden Wertepaare. Das ergibt eine Funktion von Noch-Studierenden-Anteilen zu erwarteten Absolventenquoten

b.7 Anwendung der polynomialen Regression auf die Noch-Studierenden-Anteile aus Schritt a), um somit schließlich die zukünftigen Absolventen/Absolventinnen abzuschätzen

Die verwendete Schätzfunktion (b.6 und b.7) erfolgt derzeit durch ein Polynom dritten Grades.

Das erste Arbeitspaket des Projekts bestand darin, die derzeitige Berechnung von Studienerfolgsquoten auf Basis der Aggregatdaten aus Studierenden- und Prüfungsstatistik zu evaluieren. Insbesondere war zu prüfen, ob sich durch die Verwendung einer logarithmischen Schätzfunktion oder einer weiteren alternativen Schätzfunktion bessere Ergebnisse für die Abschlusswahrscheinlichkeit der Noch-Studierenden erzielen lassen als mit dem derzeit verwendeten Polynom 3. Grades (b.6 und b.7).

Es hat sich gezeigt, dass die derzeit verwendete Polynomfunktion vergleichbar gute Ergebnisse wie alternative Schätzverfahren erbringt. Eine grundlegende Umstellung des Berechnungsverfahrens ist somit nicht angezeigt, da die Erfolgsquotenberechnung nur noch für begrenzte Zeit auf Basis von Aggregatdaten erfolgen soll, bevor sie auf die neue Einzeldatenbasis der Studienverlaufsstatistik umgestellt werden kann (siehe Abschnitt 1.2).

Bei der Evaluation des derzeitigen Verfahrens auf Aggregatdatenbasis zur Schätzung der Abschlusswahrscheinlichkeit für die Noch-Studierenden wurden aber einige kleine Verbesserungspotentiale ins-

[2] Abnahme-Rate ist im aktuellen Ansatz der Erfolgsquotenberechnung mit Aggregatdaten definiert als Abnahme von Anfänger/-innen, die nicht mehr eingeschrieben sind. Bei der Berechnung auf Basis von Studienverlaufsdaten ist die Abnahmerate definiert als Rate der Anfänger/-innen, die nicht mehr eingeschrieben sind und keinen erfolgreichen Abschluss erworben haben.

[3] Die jahrweise Vorhersage erfolgt somit bis 12 Jahre über den Datenrand des aktuellen Prüfungsjahres hinaus, da, wie vorne beschrieben, die empirische Bestimmung der Noch-Studierenden einer Anfängerkohorte 8 Jahre nach deren Start erfolgt. Eine jahrweise Vorhersage der Noch-Studierenden- und Absolventenzahlen erfolgt gleichzeitig für die vier davor gestarteten Anfängerkohorten, da die Veröffentlichung des Statistischen Bundesamts zu den Erfolgsquoten jeweils einen Zeitraum von 5 Anfänger-Jahren umfasst.

[4] Erwartete Absolventenquoten von 0% sind ein Artefakt der Methode am Ende des Schätzhorizonts.

besondere bezüglich der Dokumentation des komplexen Ablaufs und der Nutzerunterstützung entdeckt, die bereits in die Aufbereitungsprogramme eingearbeitet wurden. Es ist zu erwarten, dass die angepassten Programme, die ab sofort zur Berechnung und Veröffentlichung der berechneten Studienerfolgsquoten verwendet werden können, zu einer besseren Ergebnisqualität führen.

1.2 Weiterentwicklung der Ermittlung der Abschlusswahrscheinlichkeit der Noch-Studierenden für die aktuelle Berechnung von Erfolgsquoten auf Basis der vorliegenden Summensätze der Studierenden- und Prüfungsstatistik und ggf. Einbau in Rechenmodell

Nachdem in Arbeitspaket 1.1 gezeigt werden konnte, dass die derzeitige Ermittlung der Abschlusswahrscheinlichkeit im Rahmen der Erfolgsquotenberechnung gute Ergebnisse erbringt, bestand kein grundlegender Weiterentwicklungsbedarf bezüglich der Berechnung auf Basis von Aggregatdaten. Ersatzweise wurden im Rahmen des Projekts methodisch untersucht, ob, wie und ab wann alleine durch eine Umstellung der Datenbasis (von Aggregatdaten auf Einzeldatensätze) für die Ermittlung der Abschlusswahrscheinlichkeit der Noch-Studierenden die Qualität der Schätzung auch ohne möglichen Einsatz von KI-Verfahren erhöht werden könnte.

In Schritt (a) des derzeitigen Verfahrens der Erfolgsquotenberechnung (siehe Abschnitt 1.1) werden Noch-Studierende sowie Absolventinnen/Absolventen einer Anfängerkohorte am Datenrand über einen Vergleich von Summendaten bestimmt. Dies ist ungenau und fehlerbehaftet. Dagegen könnte die Zahl der Noch-Studierenden sowie der Absolventinnen/Absolventen auf Grundlage der neuen Studienverlaufsstatistik exakt bestimmt werden, soweit der gesamte Studienverlauf mit der Studienverlaufsdatenbank abgebildet werden kann. Für alle Studienanfängerkohorten ab 2017 sind die Zahlen der Noch-Studierenden sowie der Absolventinnen/Absolventen am Datenrand genau quantifizierbar.

Sofern künftig die Berechnung der Erfolgsquoten vollständig auf Basis von Einzeldatensätzen erfolgen sollte, könnte dies erst ab dem Jahr 2038[5] erfolgen, denn – wie im derzeitigen Verfahren – wäre für die Schätzung der Abschlusswahrscheinlichkeit für die Noch-Studierenden gemäß Schritt (b) eine historische Datenreihe von 20 Jahren erforderlich.

Die konzeptionellen Arbeiten im Rahmen des (geänderten) Arbeitspakets 1.2 haben ergeben, dass eine Nutzung der Einzeldaten durch Einführung einer hybriden Verfahrensweise bei der Erfolgsquotenberechnung bereits deutlich früher erreicht werden kann. Dabei würde Schritt (a-neu) des hybriden Verfahrens, also die empirische Bestimmung der Noch-Studierenden sowie der Absolventinnen/Absolventen bis zum Datenrand für eine Datenanfängerkohorte zwar wie im bisherigen Verfahren 8 Jahre nach deren Studienbeginn, aber auf Basis von Einzeldaten erfolgen.

Schritt (b) wäre weitgehend unverändert. Die Schritte (b.1) bis (b.6) (siehe Abschnitt 1.1), mit denen das Regressions-Polynom bestimmt wird, würden wie bisher auf Basis der Aggregatdaten ausgeführt. Neu wäre in dem hybriden Verfahren lediglich der Schritt (b.7-neu), weil die Noch-Studierenden-Anteile der Anfängerkohorte, auf die das Regressions-Polynom angewendet wird, auf Basis von Einzeldaten ermittelt wurden. Das neue Verfahren wäre also insofern hybrid, als dass es zum Teil auf Basis von Aggregatdaten und zum Teil auf Basis von Einzeldaten durchgeführt wird.

[5] Der aktuelle Bestand an Prüfungsdaten in der Studienverlaufsdatenbank reicht Stand Ende 2023 vom Berichtszeitraum Sommersemester 2017 bis zum Berichtszeitraum Sommersemester 2022 bzw. bis zum Prüfungsjahr 2022, zu dem das Wintersemester 2021/2022 und das Sommersemester 2022 gehören. Vollständige Daten zu einem Prüfungsjahr liegen i.d.R. ab Oktober des folgenden Kalenderjahres in der Studienverlaufsdatenbank vor, zum Prüfungsjahr 2023 also im Oktober 2024.

Eine erhebliche Qualitätsverbesserung der Schätzung kann trotz des weitgehend unveränderten Schritts (b) erreicht werden, weil die empirische Bestimmung der Noch-Studierenden sowie der Absolventinnen/Absolventen aus Schritt (a) des derzeitigen Verfahrens die Qualität der geschätzten Erfolgsquoten deutlich stärker beeinflusst als die eigentliche Schätzung aus Schritt (b). Denn in Schritt (b) muss die Abschlusswahrscheinlichkeit nur für die Noch-Studierenden solcher Kohorten geschätzt werden, deren Studienbeginn bereits 8 Jahre oder mehr zurückliegt. Der Großteil der Anfängerinnen/Anfänger hat nach 8 Jahren aber bereits ein Studium erfolgreich abgeschlossen oder das Hochschulsystem endgültig verlassen, so dass der Anteil der Noch-Studierenden 8 Jahre nach Studienbeginn vergleichsweise klein ist (üblicherweise zwischen 5% und 20%)

Das beschriebene geänderte Verfahren könnte erstmals Ende 2026 angewendet werden und würde sich auf die Studienanfängerkohorte 2017 und die Absolventendaten aus dem Prüfungsjahr 2025 beziehen. Wird der übliche Zeitraum für Datenprüfung und Vorbereitung einer Veröffentlichung berücksichtigt, könnten die auf Basis von Einzeldatensätzen berechneten Erfolgsquoten für die Studienanfängerkohorte 2017 im Frühjahr 2027 veröffentlicht werden.[6]

1.3 Modellierung der Abschlusswahrscheinlichkeit von Noch-Studierenden auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden

1.4 Entwicklung eines Modells zur Berechnung von Erfolgsquoten auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden und Implementierung in Studienverlaufsdatenbank

1.5 Entwicklung eines Modells zur Berechnung von Abbrecherquoten auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden und Implementierung in Studienverlaufsdatenbank

Die Arbeitspakete 1.3 bis 1.5, die wegen ihrer engen inhaltlichen Verbindung gemeinsam bearbeitet wurden und deren Lösung aufeinander aufbaut, werden im folgenden Abschnitt des Ergebnisberichts auch textlich gemeinsam abgehandelt. Bei diesen Arbeitspaketen sollte untersucht werden, ob und wie das Verfahren der Erfolgsquotenberechnung durch den Einsatz von KI-Verfahren verbessert werden kann. Denn die Einzeldatenbasis, die durch die Einführung der Studienverlaufsstatistik ermöglicht wurde und beginnend mit dem Studienanfängerjahr 2017 sukzessive erweitert werden kann, verspricht eine gute Einsatzmöglichkeit für KI-Verfahren.

Durch den Einsatz von KI-Verfahren, so die Erwartung, könnte nicht nur die Qualität der Schätzung verbessert, sondern die Schätzung für eine Studienanfängerkohorte auch deutlich schneller vorgelegt werden als bisher. Wie unter 1.1 und 1.2 beschrieben, besteht im derzeitigen Verfahren jeweils ein Zeitverzug von 8 Jahren, bis die Berechnung einer Erfolgsquote für eine neue Anfängerkohorte vorgelegt werden kann. Die höhere Aktualität der Schätzung soll aber nicht durch eine Verkürzung des Zeitraums, für den der zukünftige Abschluss der Noch-Studierenden einer Anfängerkohorte in die Zukunft (derzeit 20 Jahre) geschätzt wird, erkauft werden.

Das Ziel des KI-Einsatzes bei der Erfolgsquotenberechnung besteht somit darin, für möglichst junge Kohorten eine Schätzung des Studienerfolgs für einen Zeitpunkt durchzuführen, der gleichzeitig möglichst weit in der Zukunft liegt. Möglichst jung sollen die Kohorten sein, damit die Ergebnisse für die Nutzer/-innen der Statistik möglichst relevant sind. Der Zeitpunkt, bis zu dem geschätzt wird, soll gleichzeitig

[6] Zu berücksichtigen ist dabei, dass die derzeitige Veröffentlichung des Statistischen Bundesamts zu Erfolgsquoten immer Ergebnisse für fünf Anfängerkohorten umfasst. Eine Veröffentlichung in 2027 würde nur für die dann aktuellste Studienanfängerkohorte 2017 Erfolgsquoten enthalten, bei denen Noch-Studierenden- und Absolventenanteile auf Basis von Einzeldaten bestimmt wären. Die entsprechenden Anteile für die anderen, in der Veröffentlichung nachgewiesenen Anfängerkohorten wären noch auf Basis von Aggregatdaten berechnet.

möglichst weit in der Zukunft liegen, damit bis dahin, wie im bisherigen Verfahren, möglichst alle Studierenden der Kohorte entweder dem Stadium „Erfolg“ oder „Abbruch/Misserfolg“ zugeschrieben werden können, mit anderen Worten, dass bis dahin der Anteil der Noch-Studierenden der Kohorte einen möglichst geringen Wert erreicht hat.

Auch für ein KI-Modell gilt, dass sich die beabsichtigte Länge des Schätzzeitraums in der empirischen Datenbasis wiederspiegeln muss. Soll wie im bisherigen Verfahren der Erfolg/Misserfolg einer Anfängerkohorte bis zu einem Zeitpunkt von t=20 Jahren (bzw. 40 Semestern) nach deren Studienbeginn erfolgen, werden dafür Daten benötigt, die Informationen über zurückliegende Verläufe einer entsprechenden Zeitspanne enthalten.

Für die Umsetzung eines KI-Modells zur Schätzung des Studienerfolgs sind unterschiedliche Realisierungsoptionen denkbar, deren Vor- und Nachteile sowie insbesondere ihren frühestmöglichen Einsatzzeitpunkt im Folgenden abgewogen werden.

Vorüberlegungen zur Modellentwicklung I: Grundsätzliche Abwägung: „Schätzung bis“ oder „Schätzung in“

- **Ansatz (a) „Schätzung bis“**

Eine erste Option besteht darin, auf Basis von vorliegenden historischen Einzeldaten direkt den „Endzustand“ einer Kohorte (Erfolg/Misserfolg) abzuschätzen. Der große Vorteil dieses einfachen Ansatzes besteht darin, dass nur ein Rechenschritt gemacht werden muss, da für die dazwischenliegenden Semester keine Berechnung erfolgt.

Der große Nachteil besteht darin, dass das Modell erstmals im Jahr 2038 zur Anwendung kommen könnte, wenn in der Studienverlaufsdatenbank für die erste Anfängerkohorte 2017 auf Einzeldatenbasis ein Zeitraum von 20 Jahren nach Studienstart abgebildet werden kann.[7]

Ein weiterer Nachteil dieses einfachen Ansatzes liegt darin, dass der Erfolg/Misserfolg einer aktuellen Anfängerkohorte stets auf Basis der ältesten verfügbaren Kohorten abgeschätzt werden, sodass potentielle, neuere Trends im Hochschulsystem nicht abgebildet werden.

- **Ansatz (b) „Schätzung in“ (Mittel über gesamte Historie)**

Eine naheliegende Erweiterung von Ansatz (a) ist es, nicht nur die Daten zu denjenigen Anfängerkohorten einfließen zu lassen, für die 20 Jahre bzw. 40 Berichtssemester vorliegen, sondern auch die zu den 39 Anfängerkohorten mit weniger Historie. Es ergibt sich somit eine stufenförmige Datengrundlage: Zur ältesten Anfängerkohorte liegen 40 Berichtssemester Historie vor, zur zweitältesten 39 und so weiter.

Mit dem stufenförmigen Ansatz (b) würde der Vorteil aufgegeben, eine direkte Abschätzung des Endzustandes ableiten zu können. Stattdessen würde die Gesamt-Schätzung aus einzelnen jährlichen bzw. semesterweisen Schätzungen zusammengebaut, also iterativ vorgegangen. Für die Schätzung von Erfolg, Abbruch bzw. Schwund einer Kohorte nach x Studienjahren würde in Ansatz (b) die Erfahrung von 20-x Studienjahren bzw. 2*(20-x)+1 Semestern genutzt. Aus diesen Erfahrungswerten ließe sich beispielsweise der Mittelwert je Abstand in Studienjahren bilden, um den Übergang von x-1 auf x zu schätzen.

[7] Um für diese Verläufe außerdem den Abbruch zu ermitteln, bräuchte man bei einer Karenzzeit von 3 Semestern noch weitere 1,5 Jahre Daten. Insgesamt müsste man somit mindestens bis zum Jahr 2040 warten, um erste Anfängerkohorten mitsamt Abbruchquoten abzuschätzen. Sinnvollerweise würde man das Verfahren zudem nicht auf Basis einer einzigen Kohorte aufbauen, sondern warten, bis mehrere Kohorten mit dieser Eigenschaft vorliegen und den Mittelwert dieser Kohorten als Grundlage für das Verfahren nehmen.

Das iterative Vorgehen in Ansatz (b) ähnelt dem unter 1.1 skizzierten Vorgehen aus der aktuellen Schätzung auf Basis von Aggregatdaten. Auch im derzeitigen Verfahren, bei dem die Schätzung der Noch-Studierenden über eine Polynomfunktion erfolgt, wird dieses Polynom aus den Abschluss- und Schwundquoten aller einzelnen Studienjahre gebildet, ist also das Resultat einer polynomiellen Regression über viele einzelne iterative Abschätzungen.

Bei Ansatz (b) fließen – anders als bei (a) – somit auch Trends der aktuellsten Kohorten ein. Allerdings wird dieser Einfluss stark gedämpft, weil in allen Mittelwerten (je Abstand in Studienjahren) stets auch die ältesten Kohorten der gesamten Historie enthalten sind. Zudem ist der Zeitraum an benötigten historischen Daten weiterhin genauso lang, wie der Zeitraum, für den der Erfolg abgeschätzt werden soll, nämlich 20 Jahre.

- **Ansatz c) „Schätzung in“ (Mittel über jüngste Historie)**

Bei genauerer Analyse des iterativen Ansatzes (b) hat sich gezeigt, dass eine durchgehende Stufenform an Daten für ein KI-Modell nicht benötigt wird. Um einen Einblick in Verläufe im 40. Fachsemester zu erhalten, ist es ausreichend, von der ältesten Anfängerkohorte nur das letzte Semester in der Studienverlaufsdatenbank zu betrachten. Das letzte vorliegende Semester enthält zudem Informationen zu Studienanfänger/-innen, bei denen zwischen Studienbeginn und aktuellem Datenrand 39 oder 38 etc. Semester lagen. Entscheidend für den aus dieser Erkenntnis entwickelten Ansatz (c) ist, dass für alle Datensätze der Abstand zwischen Studienbeginn und Datenrand und damit die Kohortenzugehörigkeit bestimmt werden kann[8]. Für Studienanfängerkohorten ab 2017, also dem Start der Studienverlaufsstatistik, ist dies über die Zahl der Fachsemester im Studiengang leicht möglich. Für ältere Studienanfängerkohorten wurde hierfür eine Übergangslösung entwickelt, die unten im Abschnitt Linkszensierung und fehlende Fachsemesterangaben für Anfängerkohorten vor 2017 beschrieben ist.

Da unter den genannten Voraussetzungen das letzte vorliegende Berichtssemester die aktuellste Information zu allen relevanten Anfängerkohorten enthält, wäre es im Prinzip möglich, ein KI-Verfahren nach Ansatz (c) auf der Informationsgrundlage eines einzigen Semesters aufzubauen. Das letztendlich entwickelte KI-Verfahren nutzt aber, um Ausreißer zu vermeiden, die gemittelten Erwartungswerte der jeweils letzten 10 Semester.

Die zwei wesentlichen Vorteile von Ansatz (c) in der gewählten Ausgestaltung sind:

1. Bereits mit ca. 10 Semestern Datenbestand kann das Verfahren aufgebaut werden und es muss somit nicht bis zum Jahr 2038 gewartet werden wie in Ansatz a) oder Ansatz b) sondern nur bis zum Jahr 2024.
2. Aktuelle Trends werden stärker berücksichtigt als in Ansatz (b) (und ohnehin stärker als in Ansatz (a)).

Vorüberlegungen zur Modellentwicklung II: Linkszensierung der Studienverlaufsdatenbank und fehlende Fachsemesterangaben für Anfängerkohorten vor 2017

Der Datenbestand der Studienverlaufsdatenbank beginnt mit der Anfängerkohorte 2017. Nur für diese und folgende Kohorten kann der Studienverlauf auf Basis von Einzeldaten eindeutig beschrieben werden.

[8] Studierende mit Zeiten ohne Einschreibung im Fachstudium werden in diesem Verfahren nicht der korrekten Startkohorte zugerechnet. Mangels Alternativen wurde dies aber hingenommen, zumal Fälle mit zwischenzeitlichen Exmatrikulationsphasen quantitativ unbedeutend sind.

Für alle Studienanfänger/-innen vor 2017 ist nicht nur unbekannt, welches Fach an welcher Hochschule vor dem Start der Studienverlaufsstatistik studiert wurde, sondern auch das Fachsemester des aktuellen Studiengangs, sofern hierfür das erste Fachsemester vor 2017 lag. Dieser Umstand führt dazu, dass für alle Verläufe, die vor 2017 ein Studium begonnen haben, die Kohorten-Zugehörigkeit nicht einfach bestimmt werden kann. Wie oben geschildert, ist für den gewählten Ansatz (c) aber unbedingt erforderlich, zu allen Verläufen den Abstand zwischen Studienbeginn und Datenrand in Fachsemestern bestimmen zu können. Ersatzweise wurde daher der Abstand (und somit die Kohorten-Zugehörigkeit) auf Basis des Hochschulsemesters approximiert.[9]

Modellausgestaltung

- **Gewählte KI-Methode**

Zur Abschätzung der Noch-Studierenden wurde ein baumbasiertes Gradient-Boosting-Verfahren gewählt (R-Paket „xgboost: Extreme Gradient Boosting“ als Teil der Machine Learning Bibliothek „caret“). Dabei wurde die Schätzung als Multiclass-classification-Problem formuliert.

Baumbasierte Gradient-Boosting-Verfahren bieten folgende Vorteile:

1) Baumbasierte Verfahren können sehr flexibel mit kategoriellen Merkmalen umgehen.
2) Gradient Boosting-Verfahren sind in der Lage, sehr feine Unterscheidungen der Klassen vorzunehmen (hohe Auflösung der „decision boundary“).

- **Auswahl der Merkmale**

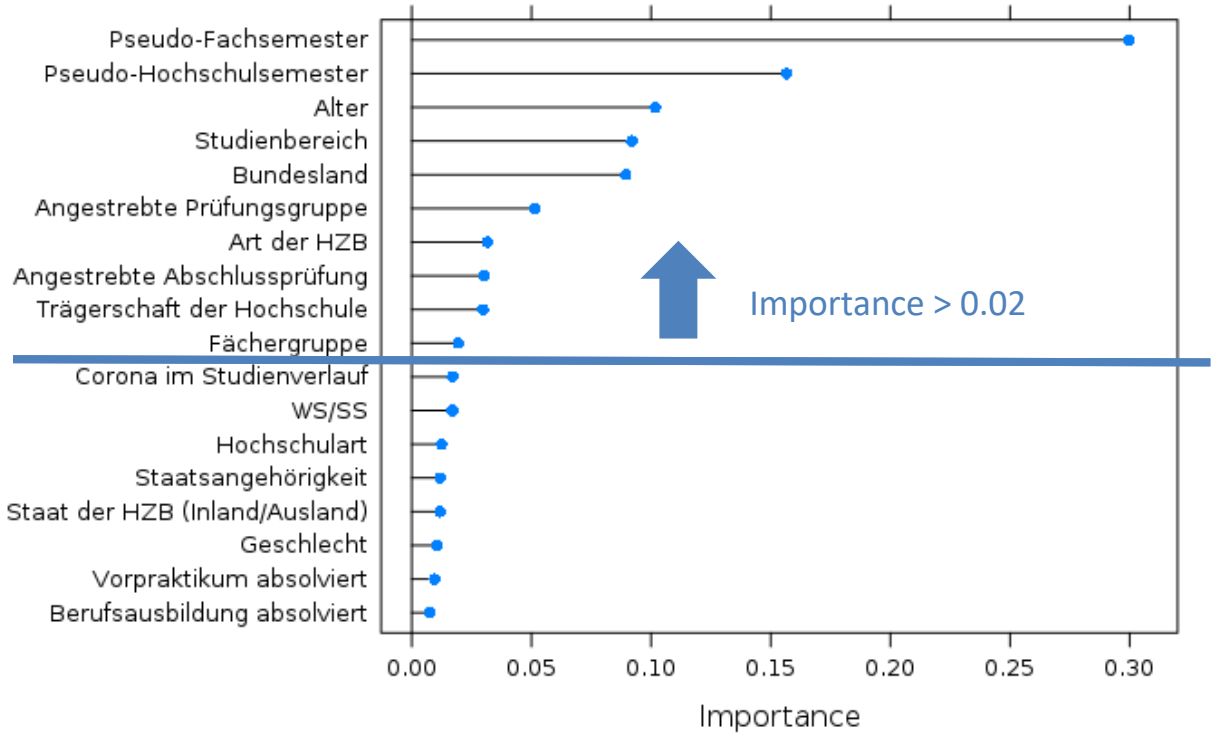
Die Auswahl der Merkmale, die in das KI-Verfahren zur Prognose von Studienerfolg bzw. -misserfolg einbezogen werden können, ist auf das Datenangebot der Studienverlaufsstatistik beschränkt. Merkmal der Studienverlaufsstatistik ist etwa die Art der Hochschulzugangsberechtigung, die aus der Forschung zu Studienerfolg und Studienabbruch als relevanter Einflussfaktor für Studienerfolg bzw. -misserfolg bekannt ist. Der Großteil der ansonsten von der Wissenschaft benannten (individuellen) Einflussfaktoren gehört dagegen nicht zum Merkmalskatalog der Studienverlaufsstatistik. Die Auswahl der Merkmale, die in das aktuelle KI-Modell einfließen, wurde daher ergänzt durch weitere Merkmale der Studienverlaufsstatistik, für die in einem datengetriebenen Ansatz zumindest eine indirekte Aussagekraft in Hinblick auf Studienerfolg festgestellt wurde.

Der statistische Einfluss der verfügbaren Studienverlaufs-Merkmale auf den Studienerfolg wurde über die so genannte Feature Importance gemessen (siehe Abbildung 1). Auf dieser Grundlage wurden insgesamt zehn in der Studienverlaufsdatenbank gespeicherte Merkmale in das Modell einbezogen. Drei Merkmale davon enthalten reine Zeitangaben (Hochschulsemester, Fachsemester, Alter der Person), zwei Merkmale sind zeitunabhängig (Art der HZB sowie Land der HZB) und fünf Merkmale sind im Studienverlauf veränderlich (Studienbereich, Angestrebte Abschlussprüfung, Angestrebte Prüfungsgruppe, Hochschulträger, Berichtsland).

[9] Das impliziert die Annahme, dass das Studium im selben Studiengang ohne Unterbrechung fortgesetzt wird.

Abbildung 1:

Feature Importance der in das KI-Modell zur Prognose von Studienerfolg einbezogenen Merkmale der Studienverlaufsstatistik[10]



XGBoost-Parameter:

* nrounds = 650, * max_depth = 10, * colsample_bytree = 0.9, * eta = 0.03, * gamma=0, * min_child_weight = 1, * subsample = 0.8

Von einer Einbeziehung der weiteren, in der Studienverlaufsdatenbank gespeicherten Merkmale in das Modell wurde abgesehen, da dies die Prognose des Studienerfolgs nicht verbessert hätte. Das gilt auch für das gebildete Merkmal „Corona im Studienverlauf“, das Studierende kennzeichnet, die im Sommersemester 2020 sowie im Wintersemester 2020/2021 eingeschrieben waren, als der Studienbetrieb aufgrund der Corona-Pandemie eingeschränkt war. Das wurde dahingehend interpretiert, dass die Corona-Pandemie bei den betroffenen Studierenden ggf. zu verlängerten Studienverläufen führt, jedoch deren endgültiges Erfolgs- oder Abbruch-Verhalten nicht wesentlich verändert.

Die Einbeziehung der im Studienverlauf veränderlichen Merkmale (Studienbereich, Angestrebte Abschlussprüfung, Angestrebte Prüfungsgruppe, Hochschulträger, Berichtsland) in das Modell ist nicht unproblematisch. Eigentlich müsste sich die Information der Merkmale in der Trainings-Phase einheitlich auf das 1. Fachsemester beziehen. Denn die Merkmale werden in der Trainings-Phase des gewählten Verfahrens am linken Rand ausgewertet und mit dem jeweiligen ersten Wert festgehalten. In der Schätz-Phase des Verfahrens wird das zugehörige Modell auf dem Stand der Studierenden-Einträge am

[10] Die textliche Ergänzung „Pseudo“ vor Fachsemester und Hochschulsemester zeigt an, dass die Angaben teilweise errechnet wurden.

linken Rand ausgeführt. Auch hier wird der jeweilige Wert der im Studienverlauf veränderlichen Merkmale festgehalten und bis zum Horizont von 40 Semestern als Grundlage der Schätzung (siehe vorne) verwendet.

Tatsächlich bezieht sich aber die erste, in der Studienverlaufsdatenbank gespeicherte Information derzeit nur für einen Teil der Studierenden-Pseudonyme auf das 1. Fachsemester. Für den derzeit noch erheblichen Teil der Studierenden-Pseudonyme, die beim Start der Studienverlaufsstatistik bereits im zweiten oder höheren Fachsemester waren, beziehen sich die Informationen aus den Merkmalen am linken Rand der Studienverlaufsdatenbank somit nicht auf das 1. Fachsemester, sondern auf den Start der Studienverlaufsstatistik.

Die Option, die veränderlichen Merkmale bis zu dem Zeitpunkt, ab dem für Fachstudienanfänger/-innen ab 2017 Verläufe von 40 Semestern in der Studienverlaufsdatenbank abgebildet werden können, nicht in das Modell einzubeziehen, wurde verworfen, weil dies bis auf Weiteres keine fachlich und/oder regional differenziertere Schätzung des Studienerfolgs von Anfängerkohorten ermöglicht hätte.

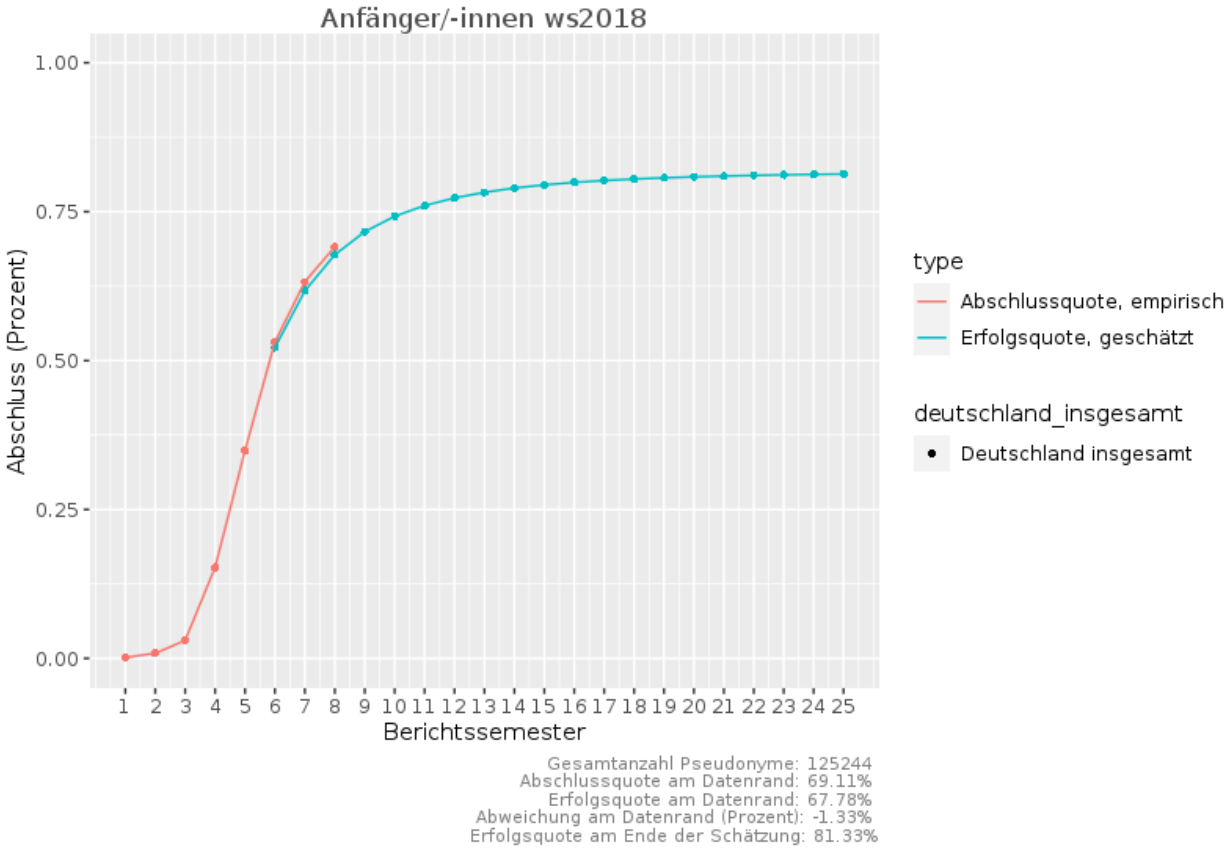
Mit dem entwickelten KI-Modell erzielte Ergebnisse zu Erfolg bzw. Misserfolg von Anfängerkohorten in Masterstudiengängen

Die mit dem entwickelten KI-Modell durchgeführten Berechnungen bezogen sich bisher im Wesentlichen auf Masterstudierende, da für die Betrachtung grundständiger Studiengänge noch nicht ausreichend Berichtszeiträume in der Studienverlaufsdatenbank vorlagen. Als erfolgreich werden dabei die Pseudonyme einer Masteranfängerkohorte gewertet, die ein Masterstudium erfolgreich abgeschlossen haben bzw. abschließen werden.

Es hat sich gezeigt, dass sich durch das entwickelte KI-Modell bereits auf Basis des aktuell verfügbaren Datenbestands in der Studienverlaufsdatenbank Abschlussquoten für Master-Studierende in guter Ergebnisqualität schätzen lassen. Die Abbildung 2 zeigt beispielhaft die geschätzte Abschlussquote der Anfängerkohorte des Wintersemesters 2018/2019 bis zu einem Horizont von 25 Semestern nach Studienbeginn. Basis für die Schätzung sind die Noch-Studierenden im 5. Semester, der erste Schätzpunkt ist das 6. Semester dieser Anfängerkohorte. Die Abweichung der auf dieser Basis geschätzten von der empirischen Abschlussquote am rechten Datenrand (hier das 8. Semester der Anfängerkohorte) beträgt -1,3 Prozentpunkte.

Abbildung 2:

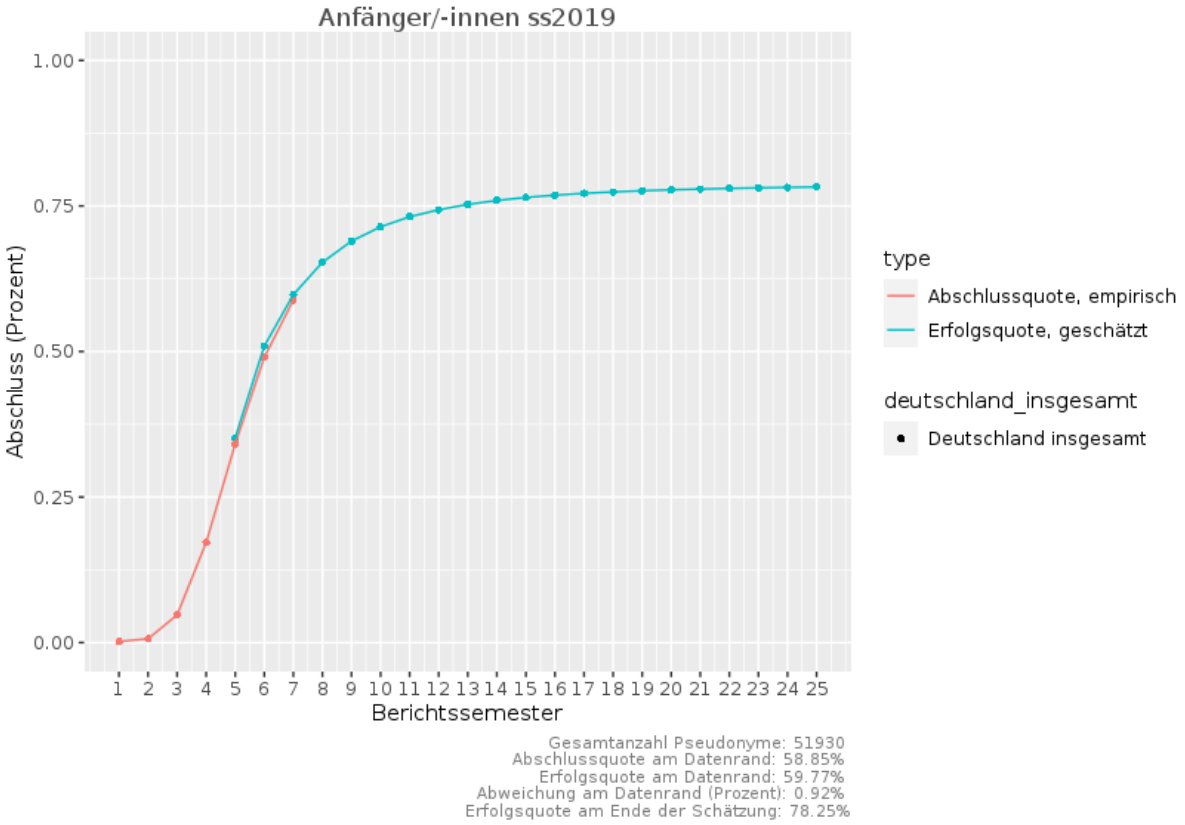
Geschätzte Abschlussquote der Anfängerkohorte Wintersemester 2018/2019 (ws2018) bis zu einem Horizont von 25 Semestern



Eine absolut noch geringe Abweichung von 0,9 Prozentpunkten der geschätzten von der empirischen Abschlussquote am rechten Datenrand findet sich bei der Berechnung für die Anfängerkohorte des Sommersemesters 2019 (Abbildung 3). Schätzbasis sind hier die Noch-Studierenden im 4. Semester, der erste Schätzpunkt das 5. Semester dieser Anfängerkohorte.

Abbildung 3:

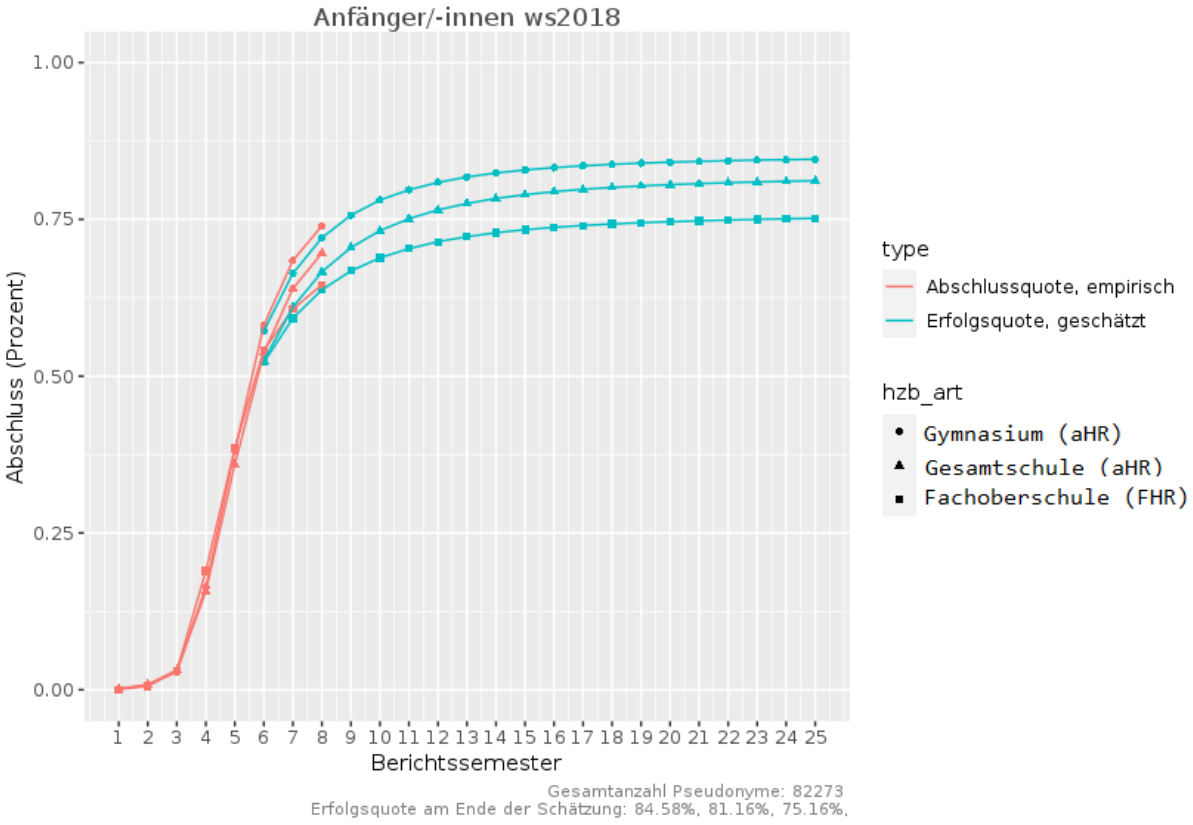
Geschätzte Abschlussquote der Anfängerkohorte des Sommersemesters 2019 (ss2019) bis zu einem Horizont von 25 Semestern



Während Abbildungen 2 und 3 Ergebnisse für Deutschland insgesamt zeigen, veranschaulicht Abbildung 4 die geschätzten Erfolgsquoten für die Anfängerkohorte des Wintersemesters 2018/2019 nach ausgewählten Arten der Hochschulzugangsberechtigung. Schätzbasis sind wie bei der in Abbildung 2 veranschaulichten Schätzung die Noch-Studierenden im 5. Semester, der erste Schätzpunkt das 6. Semester dieser Anfängerkohorte. Auch bei der Schätzung in dieser stärkeren fachlichen Differenzierung werden, gemessen an der Abweichung der geschätzten von der empirischen Abschlussquote am rechten Datenrand, qualitativ gute Ergebnisse erzielt.

Abbildung 4:

Geschätzte Abschlussquote der Anfängerkohorte Wintersemester 2018/2019 (ws2018) bis zu einem Horizont von 25 Semestern für ausgewählte Arten der Hochschulzugangsberechtigung



Zusammenfassendes Ergebnis der Arbeitspakete 1.3 – 1.5:

Durch den Einsatz eines KI-Modells bei der Erfolgsquotenberechnung auf Basis von Einzeldaten der Studienverlaufsstatistik lässt sich nicht nur die Prognosequalität gegenüber dem bisherigen Verfahren auf Aggregatdatenbasis verbessern, sondern zudem der Zeitverzug, bis die Berechnung einer Erfolgsquote für eine neue Anfängerkohorte vorgelegt werden kann, deutlich verkürzen.

Mit dem im Rahmen der Projektarbeit entwickelten KI-Modell konnten bereits Studien(miss)erfolg für die ersten Masteranfängerkohorten seit 2017 in guter Ergebnisqualität geschätzt werden. Grundsätzlich dürfte sich das Konzept, ggf. mit geringfügigen Anpassungen, auch für grundständige Studiengänge einsetzen lassen. Allerdings reicht derzeit der Datenbestand der Studienverlaufsdatenbank noch nicht aus, um eine verlässliche Evaluierung vorzunehmen.

Das entwickelte KI-Modell konnte und kann allerdings nicht auf eine Weise in die Studienverlaufsdatenbank integriert werden, dass die Erfolgsquoten für neue Masteranfängerkohorten künftig automatisiert berechnet werden könnten. Bereits die Berechnung von Studien(miss)erfolgsquoten zu neuen Jahreskohorten von Masteranfängern mit dem entwickelten KI-Modell setzt spezifische KI-Kenntnisse voraus. Erst recht sind solche spezifischen Kenntnisse für die (wenn auch geringfügigen) Anpassungen erforderlich, damit das KI-Modell künftig für die Schätzung des Studien(miss)erfolgs erster Anfängerkohorten auch in grundständigen Studiengängen eingesetzt werden kann. Derzeit ist eine entsprechende

KI-Expertise beim Stammpersonal der Hochschulstatistik wie auch in den meisten anderen Fachbereichen im Statistischen Bundesamt nicht vorhanden. Sofern die im Rahmen des Projekts durchgeführten Erfolgsprognosen aktualisiert werden sollen oder künftig regelmäßig KI in der Studienverlaufsstatistik eingesetzt werden soll, erfordert dies eine Unterstützung des Stammpersonals der Hochschulstatistik im Statistischen Bundesamt durch geeignete Expert/-innen.

2 Analysepotential der Studienverlaufsdatenbank ausschöpfen

2.1 Methodische und technische Vorarbeiten für geo-basierte Auswertungen und weitere geeignete graphische Analyseformen auf Basis der Studienverlaufsdatenbank

- Geo-basierte Auswertungen

Im Rahmen des Projekts wurde eine Kartendarstellung entwickelt, mit der auf Basis der Studienverlaufsdaten Regionalwanderungen veranschaulicht werden können. Im März 2022 wurde eine interaktive Kartendarstellung zu Regionalwanderungen von Studienanfängerinnen und -anfängern im Internetauftritt des Statistischen Bundesamts veröffentlicht, die zeigt, woher die Erstsemester an die Hochschulen eines Bundeslandes kamen und wohin die Studienberechtigten eines Bundeslands zur Studienaufnahme gewandert sind
(siehe https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/studierende-wanderungen.html).
Die interaktive Kartendarstellung zu Regionalwanderungen von Studienanfängerinnen und -anfängern wird regelmäßig aktualisiert. Grundsätzlich sind auf Basis der Studienverlaufsdaten weitere Kartendarstellungen möglich, die dieselbe Technik nutzen.

- Weitere graphische Analyseformen

Im Rahmen des Projekts wurde zudem der Beitrag der Hochschulstatistik zum neuen Veröffentlichungsformat der Bildungsindikatoren entwickelt, das im August 2023 erstmals veröffentlicht wurde. Mit den Bildungsindikatoren wird die quantitative Entwicklung in den Bereichen Schule, Berufsbildung, Hochschule und Bildungsstand der Bevölkerung anhand von jeweils fünf Statistikindikatoren veranschaulicht. Die Entwicklung je Indikator, der jeweils kurz definiert erläutert ist, wird jeweils mit tabellarischen Zeitreihen für den Zeitraum der letzten 10 Jahre, sowie mit Grafiken veranschaulicht. Per Mouse-Over können in den Grafiken die dahinterstehenden Werte angezeigt werden. Das Kontextmenü erlaubt zudem den Download sowohl der Basisdaten als auch den der Grafik in verschiedenen Formaten
(siehe https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsindikatoren/_inhalt.html).

Auch das neue Veröffentlichungsformat zu Bildungsindikatoren im Internetangebot des Statistischen Bundesamts soll regelmäßig aktualisiert werden.

2.2 Identifikation relevanter Gruppen und Faktoren mit KI-Verfahren zur Mustererkennung und Clustering

Ein weiterer Untersuchungsgegenstand im KI-Projekt zur Studienverlaufsstatistik war, ob und wie Entwicklungen und Zusammenhänge, die mit den herkömmlichen Methoden über Fallzahlen und Kennzahlen nur begrenzt veranschaulicht werden können, durch die Identifikation von Mustern und Gruppen aufgedeckt werden können. Operationalisiert wurde die Untersuchung anhand der Fragestellung, ob

und wie sich mit den in der Studienverlaufsstatistik verfügbaren Daten typische Verläufe bzw. Ereignisse identifizieren lassen, die einen messbaren Einfluss insbesondere auf Studien(miss)erfolg und Studiendauer haben.

Als Verfahren wurde die Sequenzmusteranalyse gewählt, die ursprünglich für die Lebensverlaufsforschung entwickelt wurde. Die Sequenzmusteranalyse betrachtet Paneldaten, die eine Abfolge von verschiedenen Zuständen beschreiben. Mit der Sequenzmusteranalyse wird untersucht, wie ähnlich bzw. unterschiedlich einzelne Sequenzen in der zeitlichen Abfolge dieser Zustände sind und ob sich daraus Sequenzcluster ergeben.

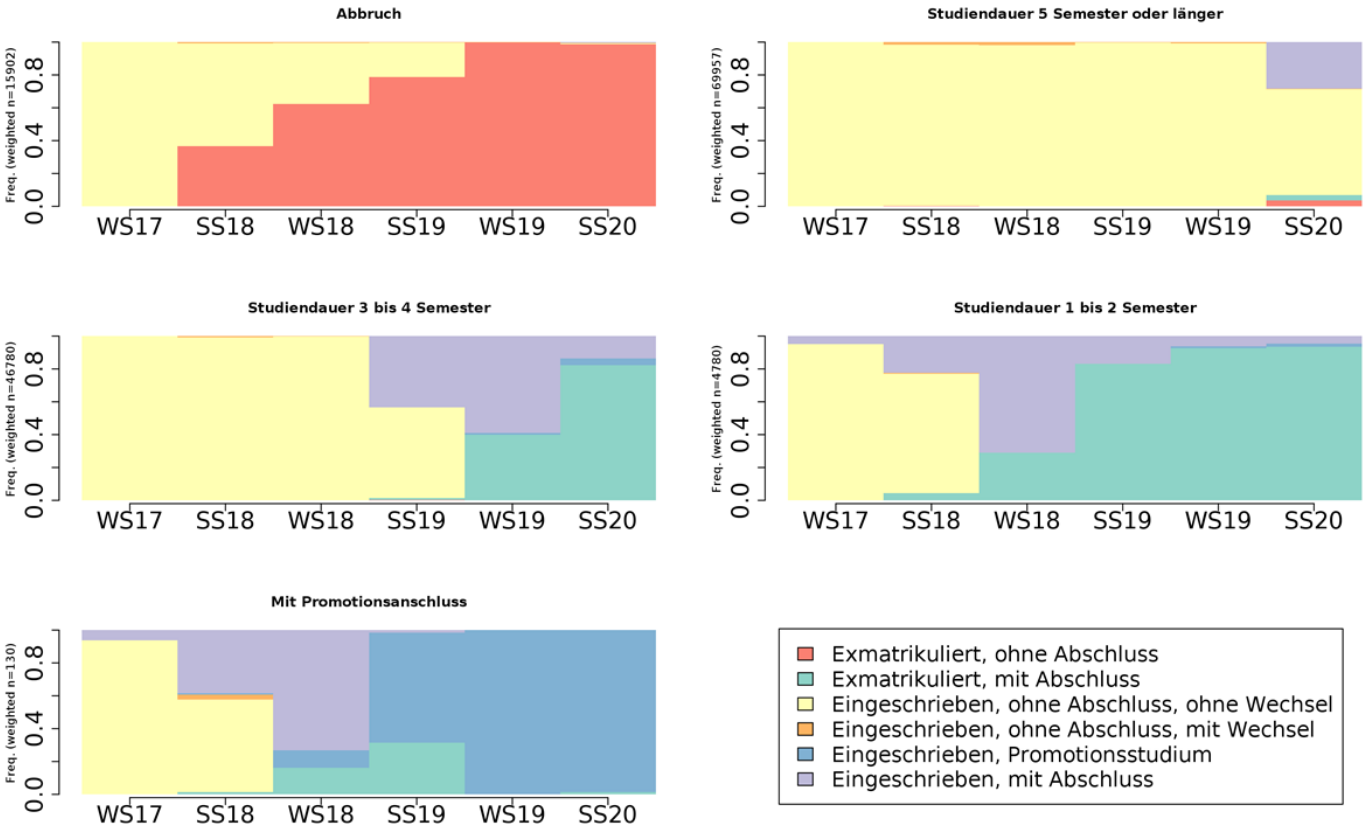
Bezogen auf den Studienverlauf lassen sich sechs disjunkte Zustände identifizieren, die mit den Studienverlaufsdaten quantifiziert werden können, nämlich
(1) eingeschrieben, ohne Abschluss im begonnenen Studiengang, ohne Hochschulwechsel,
(2) eingeschrieben, ohne Abschluss im begonnenen Studiengang, mit Hochschulwechsel,
(3) eingeschrieben, mit Abschluss im begonnenen Studiengang,
(4) eingeschrieben, mit Abschluss im begonnenen Studiengang und in weiterführendem Studium,
(5) exmatrikuliert, mit Abschluss im begonnenen Studiengang,
(6) exmatrikuliert, ohne Abschluss im begonnenen Studiengang.

Da zum Zeitpunkt der methodischen Untersuchungen (Ende 2021) erst sechs Berichtssemester in der Studienverlaufsdatenbank verfügbar waren (WS 2017/18 bis Sommersemester 2020), konzentrierten sich diese notwendigerweise auf die Masterstudierenden, weil anders als bei grundständigen Studiengängen erwartet werden kann, dass ein maßgeblicher Teil der Masteranfängerinnen und -anfänger das Studium nach Semestern (also in dem Zeitraum, der mit der Studienverlaufsdatenbank bereits abgebildet werden konnte) bereits beendet hat.

Im Rahmen der Untersuchungen für die Masterstudienanfängerinnen und -anfänger des WS 2017/2018 ergaben sich 5 typische Cluster, die plakativ wie folgt beschrieben werden können: (a) Studiendauer 1-2 Semester, (b) Studiendauer 3-4 Semester, (c) Studiendauer 5 Semester oder länger, (d) Abbruch und (e) mit Promotionsanschluss (siehe nachfolgendes Schaubild).

Abbildung 5:

5 typische Cluster von Masterstudierenden



Von den 5 typischen Verläufen bei den Masteranfängerinnen und Masteranfängern des WS 2017/2018 führen somit drei zum Abschluss des Masterstudiums, wobei die Dauer bis zum Abschluss variiert, ein Cluster führt zum Studienabbruch und für ein Cluster bleibt der Abschluss im verfügbaren Zeitraum unklar.

Wurden im Rahmen der Clusteranalyse weitere demographische und fachliche Merkmale kontrolliert, zeigte sich, dass das Geschlecht keinen Einfluss auf die Größe der einzelnen Cluster hat. Dagegen führt die Kontrolle von Fächergruppe, Hochschulart und der Staat des HZB-Erwerbs (Inland/Ausland) zu veränderten Cluster-Anteilen am Insgesamt der Studienanfängerkohorte. Bezogen auf das Cluster „Abbruch“ ließ sich zeigen, dass hier Studierende mit HZB-Erwerb im Ausland deutlich, Studierende an Universitäten etwas überrepräsentiert, Studierende der Fächergruppe Mathematik/Naturwissenschaften dagegen deutlich unterrepräsentiert sind.

Im zahlenmäßig kleinsten Cluster „Promotionsanschlüsse“ finden sich hauptsächlich Studierende an Universitäten, Personen mit HZB-Erwerb im Ausland sind leicht, Personen in den Fächergruppen Mathematik, Naturwissenschaften sowie Medizin deutlich überrepräsentiert.

Da die methodischen und technischen Voraussetzungen zum Erstellen einer Sequenzmusteranalyse in der Studienverlaufsdatenbank implementiert wurden, könnte das Verfahren bei Bedarf für spätere Studienanfängerkohorten oder für andere Studiengänge durchgeführt werden.

2.3 Identifikation relevanter Gruppen und Faktoren mit KI-Verfahren zur Mustererkennung und Clustering

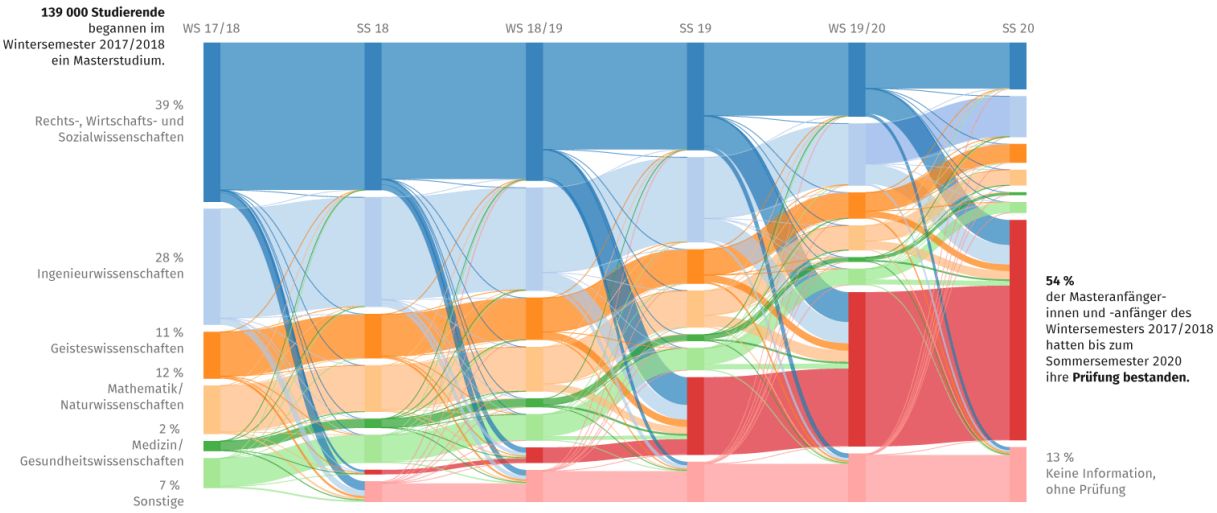
In Abstimmung mit dem Auftraggeber keine Bearbeitung im Rahmen des Projekts.

2.4 Konzeption von Datenvisualisierungsverfahren, um die Informationen der Beobachtungen in der Studienverlaufsdatenbank schnell und intuitiv erschließen zu können

Im Rahmen des Projekts wurde mit einem sogenannten Sankey zum Studienfachwechsel im Zeitverlauf eine Visualisierungsform entwickelt, mit dem Studienverlaufsdaten schnell und intuitiv erschlossen werden können. Das Sankey, das erstmals im März 2022 veröffentlicht wurde, zeigt für die Masteranfängerinnen und -anfänger des Wintersemesters 2017/2018, in welchen Fächergruppen sie ihr Masterstudium begonnen haben und zudem, wie sich die fachliche Ausrichtung im Zeitverlauf bis zum Sommersemester 2020 entwickelt hat. Mit der entwickelten Sankeytechnik, die künftig regelmäßig für das Internetangebot des Statistischen Bundesamts zur Studienverlaufsstatistik genutzt werden soll, können längere Studienverläufe derselben Anfängerkohorte, Studienverläufe aktuellerer Masteranfängerinnen und -anfänger, zudem Studienverläufe beispielsweise auch von Bachelorstudiengängen intuitiv erschlossen werden.

Abbildung 7:

Studienverlauf nach Fächergruppen für die Masteranfängerinnen und -anfänger (1. Fachsemester)
Wintersemester 2017/2018 bis Sommersemester 2020



© Statistisches Bundesamt (Destatis), 2022 Rundungsbedingte Abweichung möglich.

2.5 Konzeption eines Prognosemodells für studienverlaufsbezogene Ereignisse auf Basis der Studienverlaufsdatenbank unter Verwendung von KI-Methoden

In Abstimmung mit dem Auftraggeber keine Bearbeitung im Rahmen des Projekts.