



Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Statistische Prognosen zukünftigen Studienerfolges, Visualisierung und Ausblick

Oliver Steenbuck
Projekt 2

Inhaltsverzeichnis

Tabellenverzeichnis	2
Abbildungsverzeichnis	3
Listings	3
1. Einleitung	3
1.1. Daten	3
1.2. Knowledge Discovery in Databases	4
1.3. Architektur	4
2. Grafische Aufbereitung	5
2.1. Navigation	5
2.2. Kohortenvergleich	5
2.3. Studentenvergleich	6
2.4. Studentenansicht	7
2.5. Analyse	8
3. RapidMiner	9
3.1. Datentransformation	9
3.2. Entscheidungsbaum	10
4. Ausblick	11
4.1. Validierung der Ergebnisse für prospektive Aussagen	11
4.2. Anbindung weiterer Datenquellen	11
4.3. Toolsupport	12
4.4. Ethische Fragen	13
4.5. Vergleichende Betrachtungen	13
4.6. Weitere Auswertungen	14
Literatur	15
A. Listings	16
A.1. Datentransformation	16

Tabellenverzeichnis

1. Zieldatenstruktur für Entscheidungsbaum	10
--	----

Abbildungsverzeichnis

1. Technische Architektur/Datenfluss	4
2. Vergleich der betrachteten Kohorten nach CP pro Student	6
3. Verlauf der Studenten eines Semesters nach CPs	7
4. Screenflow	8
5. Entscheidungsbaum, ausgewählte Klausuren des ersten Studienjahres	10

Listings

1. Datentransformation SQL	16
--------------------------------------	----

1. Einleitung

In diesem Projektbericht werden aufbauend auf ([Steenbuck, 2015](#))¹ die ersten Schritte zur Auswertung der dort beschriebenen Daten und der Visualisierung dieser Daten gezeigt. Weiterhin wird eine prototypische Implementierung eines Entscheidungsbaumes in Rapid-Miner durchgeführt und die Ergebnisse gezeigt. Abschließend wird ein Ausblick auf weitere Möglichkeiten zur Datenanalyse gegeben, die in der folgenden Masterarbeit betrachtet werden (können). Im folgenden Teil der Einleitung wird zuerst ein kurzer Überblick über die Datengrundlage dieses Projektes und über die hier behandelten Schritte des KDD² Prozesses gegeben. Sowie abschließend darauf eingegangen, wo die hier behandelten Funktionalitäten sich in der in ([Steenbuck, 2015](#)) beschriebenen Architektur befinden.

1.1. Daten

Grundlage auch dieser Projektarbeit ist sind ca. 32600 Prüfungsereignisse, die von 1038 Studierenden des Studienganges Technische Informatik von 2004 bis 2012 erbracht wurden. Die Daten stammen aus dem Studierenden Informations System³, das an der HAW Hamburg verwendet wird.⁴

¹dort im Kapitel Ausblick aufgeführt

²Knowledge Discovery in Databases

³im Folgenden StISys

⁴Daten sind in ([Steenbuck, 2015](#)) ausführlich beschrieben

1.2. Knowledge Discovery in Databases

Betrachtet man den unten dargestellten KDD Prozess nach (vgl. [Beierle und Kern-Isberner, 2008](#), S. 143 ff.) wurden in ([Steenbuck, 2015](#)) primär die ersten 5 Schritte (hier rot hervorgehoben), betrachtet. Diese Projektarbeit betrachtet die darauffolgenden letzten beiden Schritte des KDD (hier blau hervorgehoben) Data Mining und Interpretation.

1. Hintergrundwissen und Zielsetzung
2. Datenauswahl
3. Datenbereinigung
4. Modellfunktionalität
5. Verfahrensauswahl
6. Data Mining
7. Interpretation

Hier werden prototypisch dynamische Visualisierungen gezeigt sowie erste Analysen getroffen.

1.3. Architektur

Die hier gezeigte grafische Funktionalität befindet sich im in mit Ruby on Rails implementierten Teil des Systems. Abbildung 1 zeigt blau hervorgehoben die Position dieser Komponente in der technischen Architektur des Gesamtsystems.

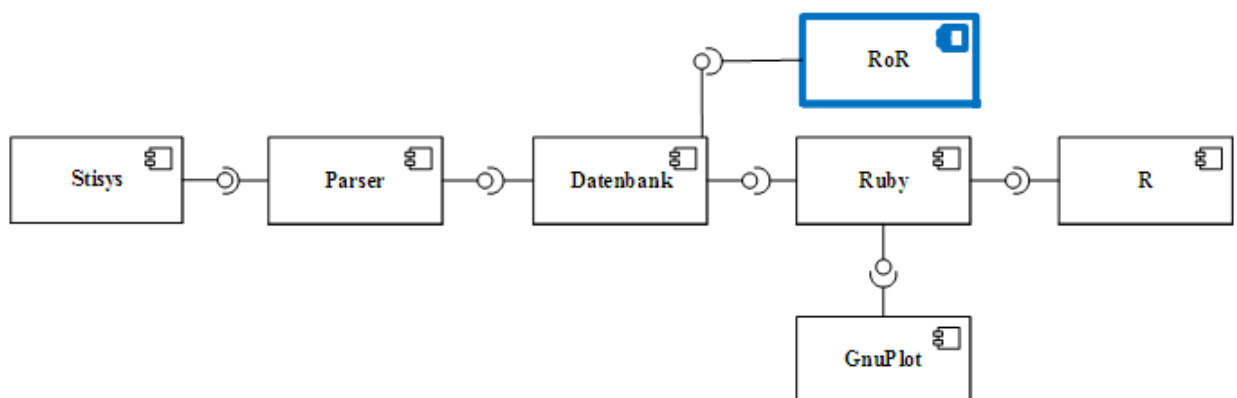


Abbildung 1: Technische Architektur/Datenfluss

2. Grafische Aufbereitung

Hier werden die ersten prototypischen Visualisierungen gezeigt. Diese dienen der explorativen Datenvisualisierung und sind damit im Schema des KDD in Schritt 7 Interpretation angesiedelt. Dies sind zum einen der Vergleich zwischen Semestern in Abschnitt 2.2 und der Vergleich zwischen den Studenten eines Semesters in Abschnitt 2.3 sowie die Anzeige der Daten eines/einer Studierenden. Die Implementierung ist hier in Ruby on Rails erfolgt, um die bereits bestehenden Ruby-Komponenten aus dem ersten Teil des Projektes (([Steenbuck, 2015](#))) nutzen zu können. Als Visualisierungsbibliothek wird Fusioncharts⁵ verwendet. Als Statistikbibliothek wird Artificial Intelligence for Ruby ⁶ genutzt. Um im späteren Verlauf individuellere Grafiken zu ermöglichen, ist die Nutzung von D3.js⁷ zur Erweiterung der Visualisierung geplant.

2.1. Navigation

Die Navigation in der Anwendung ist zum einen durch das Menü (in Abbildung 2.4 abgebildet) als auch durch Links in den einzelnen Visualisierungen möglich. Die inhaltsbasierte Navigation durch die Daten wird für die betreffenden Statistiken bei diesen beschrieben⁸.

2.2. Kohortenvergleich

Abbildung 2, zeigt wie die Kohorten sich im Vergleich zueinander verhalten. Studienerfolg wird hier gemessen an den durchschnittlichen CP (y-Achse), die ein Student aus der Kohorte im jeweiligen Fachsemester (x-Achse) erreicht hat. Das deutliche Abflachen einiger Semester ist dem Ende der Datenerfassung geschuldet, da dann keine CPs mehr erfasst wurden.

⁵<http://www.fusioncharts.com>

⁶<http://www.ai4r.org>

⁷<http://d3js.org/>

⁸siehe 2.2 und 2.3

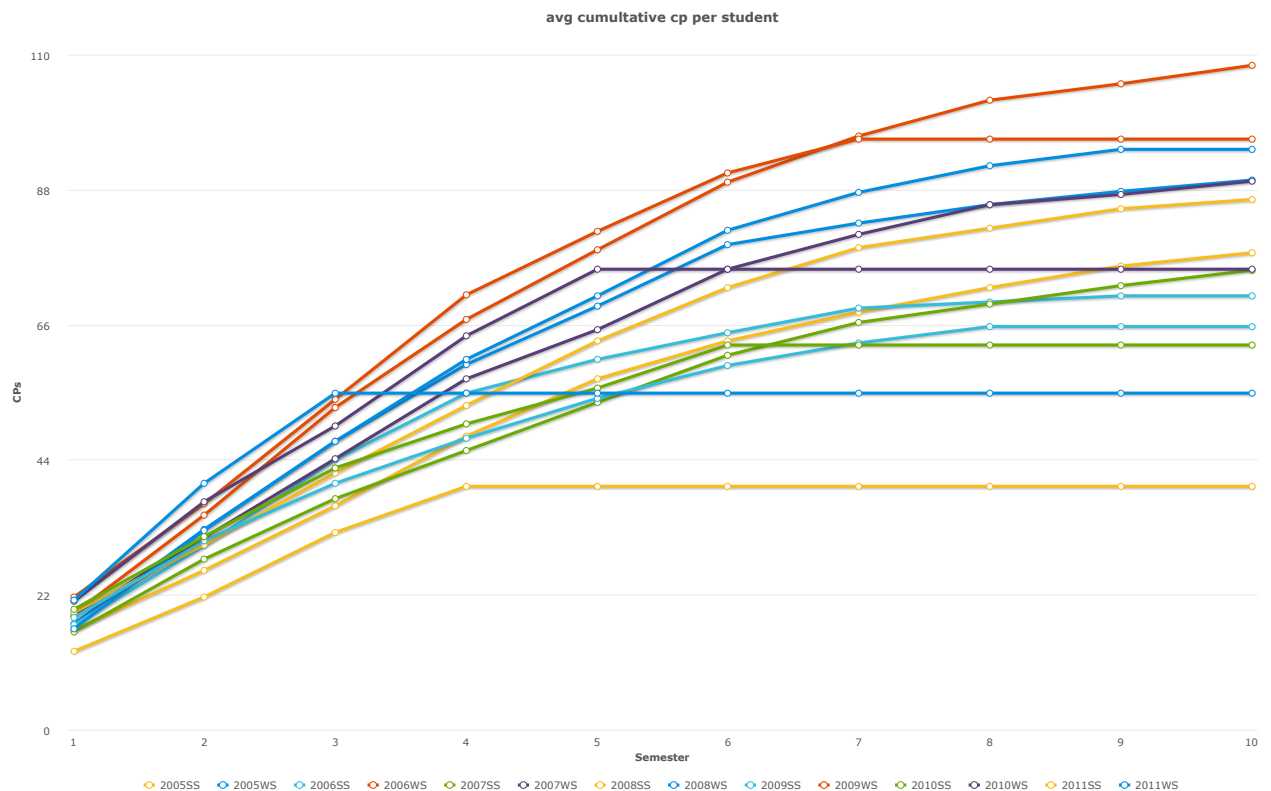


Abbildung 2: Vergleich der betrachteten Kohorten nach CP pro Student

Aus dieser Statistik ist die Navigation zur vergleichenden Ansicht (2.3) zwischen den Studenten einer Kohorte durch die Auswahl der betreffenden Kohorte in der Grafik möglich.

2.3. Studentenvergleich

Abbildung 3 zeigt den Verlauf einer Kohorte aufgeschlüsselt nach den Credit Points, die Studenten nach jedem Semester erreicht haben. Die Tabelle ist sortiert nach den CP die die Studenten im ersten Semester erreicht haben.

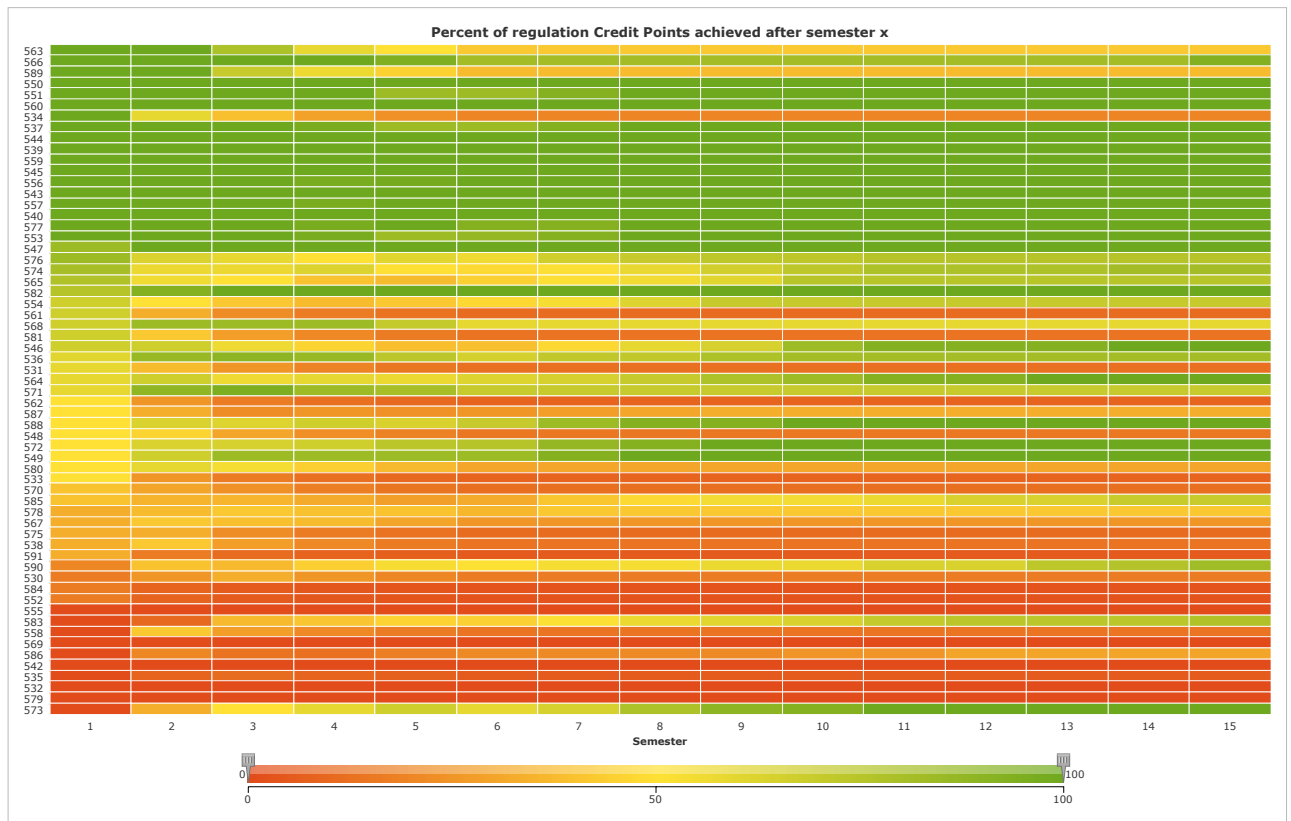


Abbildung 3: Verlauf der Studenten eines Semesters nach CPs

Die Auswahl eines Studenten in der Grafik führt zur in 2.1 gezeigten Einzelansicht eines Studenten.

2.4. Studentenansicht

Nachfolgend wird die derzeitige (Beta-) Implementierung der in Ruby on Rails entwickelten Applikation gezeigt, über die dynamische Auswertungen generiert und dargestellt werden. Die hier ausgewählte Anzeige 'Student' (hier der Student mit der Nummer 1027) zeigt die Details eines Studenten im Vergleich mit dem Durchschnitt seiner Semestergruppe.

verlauf	student	cluster	semestercompare			
---------	---------	---------	-----------------	--	--	--

Student#1027						
---------------------	--	--	--	--	--	--

Id	Kohorte	Studiengang	IsGraduated	Studiendauer	GPA	
1027	2010WS	B-TI	M	4.0	8.199999999999999	

Courses:

Titel	avg	Note	Note	Note	Note
PR1	6.96	10			
PRP1	5.67	6			
GE1	6.51	2	8		
MG	5.67	9			
GT	7.65	12			
GS	9.45	8			
AF	6.86	7			
PR2	7.2	3	9		
GE2	8.02	8			
GSP	6.7	14			
BS	7.78	6			
SE1	9.07	7			
AD	8.24	7			
DB	8.13	6			
RN	7.23	6			

Abbildung 4: Screenflow

Die Navigation ist von von diesem Bildschirm möglich auf die Kohorten-, Studenten- und Kursvergleiche⁹.

2.5. Analyse

Im Folgenden werden einige Auffälligkeiten, die sich in den Graphen ergeben, beschrieben. Diese sind in einer zukünftigen Arbeit statistisch zu untersuchen.

⁹noch nicht implementiert

Aus dem grafischen Kohortenvergleich ergibt sich als erste Fragestellung der auffällige Unterschied im durchschnittlichen Studienerfolg zwischen unterschiedlichen Semestern. So erreicht z.B. das Wintersemester 2005 nur knapp die Hälfte der CP pro Student und Semester, die das Wintersemester 2008 erreicht.

Aus dem Studentenvergleich lässt sich ablesen, dass in erster Näherung pro Semester ähnlich viele Credit Points erreicht werden. Es drängt sich die Vermutung auf, dass eine grobe Gruppeneinteilung der Studenten schon durch eine einfache Betrachtung der Credit Points erreicht werden kann.

3. RapidMiner

RapidMiner ist eine Umgebung für maschinelles Lernen und Data Mining. Programme werden in einer grafischen Entwicklungsumgebung erstellt und als XML gespeichert. Es existieren eine Serverkomponente, um entwickelte Prozesse auf Enterprise-Hardware auszuführen¹⁰, sowie eine Version, um Hadoop-Infrastrukturen zu nutzen¹¹. Im Rahmen dieser Projektarbeit wurde die aktuelle Version 6.4 des RapidMiner-Studios mit einer akademischen Lizenz verwendet.

Es wurde prototypisch eine erste Analyse in RapidMiner durchgeführt, um die Anwendbarkeit des Tools auf die bestehende Datenbasis sicherzustellen. Hier wird ein Teil der dazu durchgeführten Transformation der in (Steenbuck, 2015) gezeigten Datenstrukturen sowie der erzeugte Entscheidungsbaum als Ergebnis des Testlaufes gezeigt. Es wurde der auf dem C4.5 (vgl. Beierle und Kern-Isberner, 2008, S. 116 ff.) basierenden Algorithmus für Entscheidungsbäume genutzt, RapidMiner bietet auch einen auf ID3 (vgl. Beierle und Kern-Isberner, 2008, S. 116 ff.) basierenden Algorithmus.

3.1. Datentransformation

Die in 3NF¹² vorliegenden Daten müssen dahingehend transformiert werden, dass eine Zeile pro Student mit einer Spalte pro Prüfung vorliegt. Die Zieldatenstruktur wird in Tabelle 1 beispielhaft gezeigt¹³. Hier bildet `Abschluss` die Zielspalte, nach deren Ausprägungen 1 (hat einen Abschluss) und 2 (hat keinen Abschluss) der Entscheidungsbaum klassifizieren soll.

¹⁰<https://rapidminer.com/products/server/>

¹¹<https://rapidminer.com/products/radoop/>

¹²Normalform

¹³mit verkürzter Attributsauswahl

StudentId	Alter	PR1	MG	DB	Abschluss
7	34	10	NULL	6	0
385	25	9	5	10	1

Tabelle 1: Zieldatenstruktur für Entscheidungsbaum

Da eine ähnliche Struktur auch in anderen Algorithmen in RapidMiner notwendig sein wird, wurde entschieden, die Transformation aus Tabelle 1 direkt in der Datenbasis in SQL umzusetzen. Listing A.1 im Anhang zeigt das SQL, durch das die benötigte Tabelle als View erzeugt wird.

3.2. Entscheidungsbaum

Für die prototypische Implementierung wurde als Ziel ein Entscheidungsbaum gewählt, da die Ausgabe (gezeigt in Abbildung 5) intuitiv verständlich und dadurch leicht verifizierbar ist. Die Knoten des Baumes repräsentieren hier Prüfungsleistungen, die Kanten Bedingungen auf den Noten dieser Prüfungsleistungen¹⁴ und die Blätter, ob die Studenten einen Abschluss haben (1) oder nicht (0). Die Farbaufteilung des Balkens in den Blättern repräsentiert die genaue Zusammensetzung der Population des Blattes.

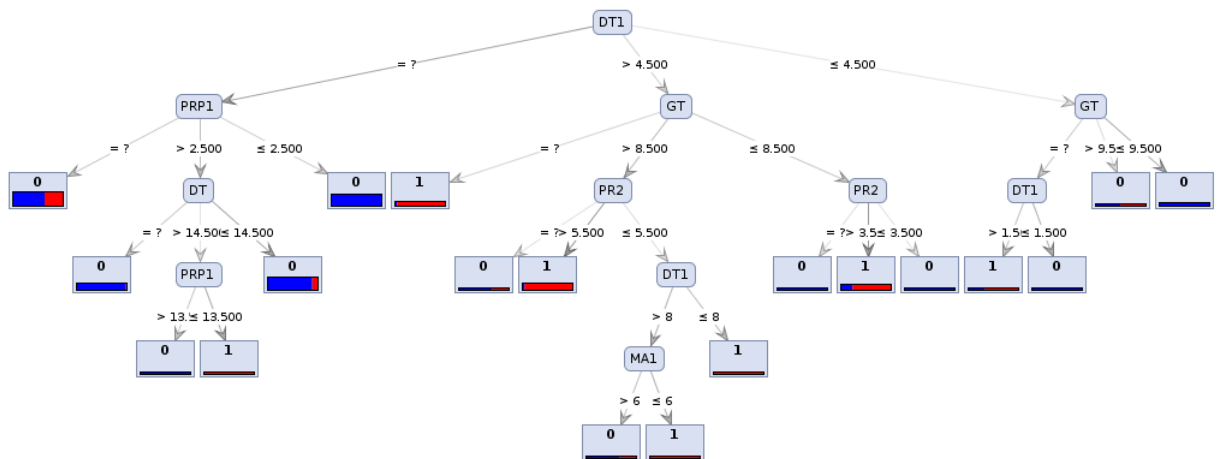


Abbildung 5: Entscheidungsbaum, ausgewählte Klausuren des ersten Studienjahres

¹⁴'=?' repräsentiert eine nicht abgelegte Prüfung

4. Ausblick

Im weiteren Verlauf der Analyse ergeben sich diverse Optionen und Aufgaben. Im Folgenden werden mögliche weitere Datenquellen beschrieben, deren Anbindung tiefergehende Analysen ermöglichen würde, sowie Techniken dargelegt, die mit der systematischen Nutzung von Werkzeugen ermöglicht werden können. Weiterhin werden ethische Fragestellungen aufgezeigt, die durch die Analyse personenbezogener Daten entstehen und weitere Veröffentlichungen genannt, die zur Kalibrierung und zum Vergleich der Ergebnisse genutzt werden können, sowie abschließend weitere mögliche Auswertungen gezeigt.

4.1. Validierung der Ergebnisse für prospektive Aussagen

Die derzeit vorhandenen Aussagen auf den Daten sind retrospektiv geprägt, eine zukunftsgerichtete Betrachtung der Daten erfordert Aussagen unter anderem zur Stetigkeit der Datenbasis unter möglicherweise wechselnden Rahmenbedingungen. So wird beispielsweise zu untersuchen sein, inwiefern sich die während der Datenerhebung wechselnden Prüfungsordnungen¹⁵ auf mögliche Aussagen aus den Daten auswirken.

4.2. Anbindung weiterer Datenquellen

Die HAW führt umfangreiche Statistiken über neu immatrikulierte Studenten, die sich aus den Bewerbungsdaten speisen. Insbesondere interessant erscheint hier die Verbindung der vorliegenden Daten mit den Statistiken zu Studienanfängerzahlen (C1) und deren Qualifizierten Schulabschlüssen (C3). Mithilfe dieser Daten könnte unter anderem untersucht werden, ob es einen Zusammenhang zwischen der Anzahl der Bewerber bzw. deren Qualifikation und dem späteren Studienerfolg gibt¹⁶.

Sofern diese Daten verfügbar sind, würde ein Abgleich mit den entsprechenden Statistiken der anderen Hamburger Hochschulen erste Hinweise auf mögliche Unterschiede im Bewerbungsverhalten geben.

Der Abgleich des Studien-/Anmeldeverhaltens mit den gesamtwirtschaftlichen Rahmenbedingungen würde es erlauben, Rückschlüsse darauf zu ziehen, ob es einen anekdotisch berichteten Zusammenhang zwischen hohen Bewerberzahlen (schlechte wirtschaftliche Randbedingungen) und kurzer Studiendauer (gute wirtschaftliche Randbedingungen) gibt.

¹⁵PO2004 zu PO2008

¹⁶Eine Hypothese die das Bewerbungsverfahren mit Zentrierung auf die Abiturnote implizit aufstellt

Die HAW wird im CHE Hochschulranking geführt. Durch die Integration dieser Daten kann möglicherweise gezeigt werden, ob das CHE Ranking einen Einfluss auf die Qualität¹⁷ und die Quantität der Bewerbungen an der HAW hat.

Die Erweiterung der Datenbasis um eine Zuordnung einer Veranstaltung zu einem Lehrenden zu ermöglichen, drängt sich auf, um betrachten zu können, ob/welchen Einfluss ein Lehrender auf den konkreten Erfolg in einer Veranstaltung bzw. auf den gesamten Studienverlauf hat. Eine solche Zuordnung wäre händisch über die historischen Stundenplandaten der entsprechenden Semester möglich. Es ist vorher zu klären, wie sich eine solche Zuordnung auf die Akzeptanz der Anwendung bei den Lehrenden auswirkt¹⁸.

In der zentralen Verwaltung (genauer im zentralen Studierendenbüro) liegen detailliertere Daten aus den Bewerbungsverfahren der Studierenden vor, als diese in den C1- und C3-Statistiken gezeigt werden. Eine Einbindung dieser Datenbestände würde den Blick auf sozioökonomische Aspekte¹⁹ und den Einfluss der Note der Hochschulzugangsberechtigung auf den Studienerfolg eröffnen. Wenn diese Daten zur Verfügung gestellt werden, ist ein weiterer Verarbeitungsschritt notwendig, um diese personenbezogenen Daten den vorliegenden anonymisierten Daten zuzuordnen.

Während der Implementierung des Prototypen hat sich gezeigt, dass eine Anreicherung der Daten um die Informationen, in welchem Fachsemester eine Prüfung liegt, notwendig ist. Die entsprechenden Daten liegen in den Prüfungsordnungen 2004/2008 vor und sind händisch in die Datenbank einzupflegen²⁰.

4.3. Toolsupport

Prototypisch wurden schon erste Versuche mit RapidMiner durchgeführt, um die Nutzbarkeit des Tools mit den Daten zu zeigen. Aufgrund der gezeigten Machbarkeit ist eine weitere Analyse mit publizierten Techniken und darauffolgendem Abgleich der Ergebnisse mit folgenden Publikationen geplant:

- Backpropagation [Fausett und Elwasif \(1994\)](#)
- Bayesian Networks [Itoh u. a. \(2012\)](#)
- Decision Trees [Bunkar u. a. \(2012\)](#)

¹⁷gemessen an der Studiendauer

¹⁸siehe auch die Regelungen zur Veranstaltungsevaluation

¹⁹indirekt über den Wohnort

²⁰es handelt sich um ca. 40 Datensätze

4.4. Ethische Fragen

Sollten die Ergebnisse der auf dieser Projektarbeit aufbauenden weiteren Arbeiten in der Realität angewendet werden, ist wahrscheinlich eine Nutzung der entwickelten Funktionalitäten ohne Anonymisierung der Daten notwendig. Bei einer solchen algorithmischen Auswertung personenbezogener Daten stellen sich unter anderem folgende Fragen:

Wer darf die Daten und Auswertungen einsehen? Mögliche Zugriffsgruppen wären unter anderem alle Lehrenden, ein ausgewählter Kreis der Lehrenden²¹ sowie die Studierenden selber. Hier gilt es, den Datenschutz gegen den Nutzen der Datenverarbeitung abzuwägen.

Es muss betrachtet werden, ob und welche möglicherweise negativen Auswirkungen eine Schaffung zusätzlicher personalisierter Statistiken und insbesondere deren institutionalisierte Nutzung hat. Intuitiv wäre zum Beispiel die Schaffung einer sich selbsterfüllenden Prophezeiung zu befürchten, wenn Studenten nach relativ kurzer Studiendauer auf der Basis prognostizierter Daten in eine Gruppe von erfolgreichen und eine Gruppe von nicht erfolgreichen Studenten aufgeteilt würden.

Sollten auf Basis von Prognosen tatsächlich Entscheidungen getroffen werden, scheint eine Diskussion der tolerierbaren positiven und negativen Fehler unumgebar, um die Nutzer der Prognosen/Statistiken in die Lage zu versetzen diese korrekt zu interpretieren.

Und nicht zuletzt ist zu diskutieren, was die möglichen Zielsetzungen sind. Zielkonflikte zwischen verschiedenen Beteiligten scheinen unter dem steigenden Kostendruck, dem die Hochschulen ausgesetzt sind, vorprogrammiert zu sein.

4.5. Vergleichende Betrachtungen

Bisher wurden die vorhandenen Daten mit Teilergebnissen aus zwei Veröffentlichungen (([Golding und Donaldson, 2006](#)) und ([Borrego u. a., 2005](#))) verglichen. Angestrebt wird für weitere Arbeiten ein kompletter Vergleich mit den Ergebnissen aus diesen beiden Arbeiten sowie die Betrachtung weiterer veröffentlichter Vergleichsdaten wie etwa:

In ([Sanjeev und Zytkow, 1995](#)) werden unter anderem Zusammenhänge zwischen dem GPA²² an der High School und dem GPA an der Universität und den erreichten Credit Points betrachtet. Sollten die entsprechenden Daten aus den Studienplatzbewerbungen durch die Hochschule²³ bereitgestellt werden, kann eine ähnliche Betrachtung für die vorliegende Datenbasis durchgeführt werden.

²¹ Prüfungsausschuss, Studienfachberater

²² Grade Point Average

²³ siehe 4.2

4.6. Weitere Auswertungen

Aus dem derzeitigen Stand der Analyse ergeben sich unter anderem folgende Ansätze für weitergehende Betrachtungen der Daten.

Gibt es signifikante Unterschiede für die gleiche Prüfungsleistung zwischen verschiedenen Semestern? Und können diese darauf aufbauend auf Sommer/Wintersemester oder Prüfende²⁴ zurückgeführt werden?

Aus der Betrachtung, welche Veranstaltungen mehr als andere in spätere Semester geschoben werden²⁵, können Aufschlüsse darüber gewonnen werden welche Veranstaltungen die Studierenden als Problemfächer empfinden.

Aus einer ähnlichen Betrachtung sollten sich die Showstopper²⁶ ergeben, also die Veranstaltungen, mit denen Studenten endgültig durchfallen²⁷.

Werden in unterschiedlichen Fachsemestern/Hochschulsemestern im Mittel unterschiedlich viele CP erbracht oder können die CP, die ein Student pro Semester erbringt, als Konstante betrachtet werden? Sollte es Unterschiede nach Fachsemester geben, können diese auf konkrete Prüfungen zurückgeführt werden?

Die Prüfungsordnung 2008 hat die Gewichtung aller Prüfungen für die Endnote vereinheitlicht²⁸. Für komplett rational agierende Studenten sollten sich also Unterschiede im Prüfungserfolg in dem Fall ergeben, dass Prüfungen aus unterschiedlichen Fachsemestern gleichzeitig abgelegt werden, je nachdem ob der Student in der PO2008 oder in der PO2004 studiert. Kann ein solcher Effekt in der Datenbasis gezeigt werden?

Aus den C1- und C3-Statistiken der Hochschule können im Aggregat²⁹ Aussagen über einen Zusammenhang zwischen Art der Hochschulzugangsberechtigung und Studienerfolg sowie über einen eventuellen Zusammenhang zwischen Quantität der Bewerbungen und späterem Studienerfolg getroffen werden³⁰.

²⁴momentan liegen keine Daten vor, die eine Zuordnung zwischen einer einzelnen Veranstaltung und einem Lehrenden zulassen, siehe 4.2

²⁵gemessen daran, ob die Veranstaltung mehr als die Varianz aller Veranstaltungen geschoben wird

²⁶Veranstaltungen, mit denen Studenten herausgeprüft werden

²⁷je nach Prüfungsordnung und Fachsemester der Veranstaltung nach 3 Versuchen oder nach 3 Versuchen + 1 mündlichen Überprüfung

²⁸vorher Multiplikator abhängig vom Fachsemester der Prüfung

²⁹auf Kohortenebene

³⁰siehe auch 4.2

Literatur

- [Beierle und Kern-Isberner 2008] BEIERLE, Christoph ; KERN-ISBERNER, Gabriele: *Methoden wissensbasierter Systeme*. 4. Edition. Wiesbaden : Friedrich Vieweg & Sohn Verlagsgesellschaft mbH, 2008. – ISBN 978-3-8348-0504-1
- [Borrego u. a. 2005] BORREGO, M.J. ; PADILLA, M.A. ; OHLAND, M.W. ; ANDERSON, T.J.: Graduation Rates, Grade-Point Average, and Changes of Major of Female and Minority Students Entering Engineering. In: *Proceedings Frontiers in Education 35th Annual Conference*, Ieee, 2005, S. T3D–1–T3D–6. – URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1611931>. – ISBN 0-7803-9077-6
- [Bunkar u. a. 2012] BUNKAR, Kamal ; SINGH, Umesh K. ; PANDYA, Bhupendra ; BUNKAR, Rajesh: Data mining: Prediction for performance improvement of graduate students using classification. In: *2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN) (2012)*, September, S. 1–5. – URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6335530>. ISBN 978-1-4673-1989-8
- [Fausett und Elwasif 1994] FAUSETT, L.V. ; ELWASIF, W.: Predicting performance from test scores using backpropagation and counterpropagation. In: *Neural Networks, 1994. IEEE World Congress on Computational Intelligence.*, Florida Institute of Technology, 1994, S. 3398–3402. – URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=374782. – ISBN 078031901X
- [Golding und Donaldson 2006] GOLDING, Paul ; DONALDSON, Opal: Predicting Academic Performance. In: *Proceedings. Frontiers in Education. 36th Annual Conference*, Ieee, 2006, S. 21–26. – URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4117161>. – ISBN 1-4244-0256-5
- [Itoh u. a. 2012] ITOH, K. ; ITOH, Hirotaka ; FUNAHASHI, Kenji: Forecasting Students' Grades Using a Bayesian Network Model and an Evaluation of Its Usefulness. In: *13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Ieee, August 2012, S. 331–336. – URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6299301>. – ISBN 978-1-4673-2120-4
- [Kim 2008] KIM, Y.: Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. In: *Expert Systems with Applications* 34 (2008), Februar, Nr. 2, S. 1227–1234. – URL <http://linkinghub.elsevier.com/retrieve/pii/S0957417406004118>. – ISSN 09574174

[Sanjeev und Zytkow 1995] SANJEEV, P. A. ; ZYTKOW, M. J.: Discovering Enrollment Knowledge in University Databases Regularities. In: *Proc. Int. Conf. Knowledge Discovery and Data Mining*, 1995, S. 246–251

[Steenbuck 2015] STEENBUCK, Oliver: Statistische Prognosen zukünftigen Studienerfolges / University of Applied Science Hamburg. Hamburg, 2015. – Forschungsbericht. – 1–17 S

A. Listings

A.1. Datentransformation

Listing 1: Datentransformation SQL

```
SET @sql = NULL;
SET @@group_concat_max_len = 64000;

SELECT
  GROUP_CONCAT(DISTINCT
    CONCAT(
      'MAX(IF(courses.fachkrz = ''',
      courses.fachkrz,
      ''', attendings.note, NULL)) AS ''',
      courses.fachkrz,
      ''''
    )
  ) INTO @sql
FROM attendings
  JOIN courses ON attendings.course_id = courses.id
  WHERE attendings.klausur = 1;

SET @sql = CONCAT('
  CREATE OR REPLACE VIEW studentsKlausuren
  AS SELECT
    students.id
    , students.ageAtImma
    , students.isGraduated
    , students.kohorte, ', @sql, '
  FROM attendings JOIN students
```

```
        ON attendings.student_id = students.id  
        JOIN courses  
        ON attendings.course_id = courses.id  
    GROUP BY students.id');  

```

```
PREPARE stmt FROM @sql;  
EXECUTE stmt;  
DEALLOCATE PREPARE stmt;
```