Identifizierung von Klassen von Studierenden mit frühem Prüfungsmisserfolg

Bachelor-Thesis von Daniel Fath
Oktober 2013

TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich Informatik
Knowledge Engineering Group

Identifizierung von Klassen von Studierenden mit frühem Prüfungsmisserfolg

Vorgelegte Bachelor-Thesis von Daniel Fath

1. Gutachten: Prof. Johannes Fürnkranz
2. Gutachten: Dr. Frederik Janssen
3. Gutachten: Tim Neubacher

Tag der Einreichung:

Erklärung zur Bachelor-Thesis

Hiermit versichere ich, die vorliegende Bachelor-Thesis ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 28.10.2013

(Daniel Fath)

Zusammenfassung

Studierende, des Studiengangs *Bachelor of Science* Informatik an der TU Darmstadt, mit geringen Prüfungserfolg im ersten Semester sind dazu verpflichtet einen Fragebogen auszufüllen. Dieser dient dazu Gründe, für diesen Misserfolg, zu ermitteln. Diese Arbeit beschäftigt sich mit der Frage, ob sich anhand der gewonnen Daten aus den Fragebögen Klassen von Studierenden ermitteln lassen. Dazu wurden statistische Auswertungen, sowie Clusteranalysen, auf die vorhanden Daten aus den Fragebögen, vorgenommen. Die somit gefunden Ergebnisse werden in dieser Thesis beschrieben.

Inhaltsverzeichnis

1 Einleitung

Die Fachstudienberatung und das Mentorensystem des Fachbereichs Informatik der TU Darmstadt betreuen in jedem Semester Studierende des Studiengangs *Bachelor of Science* Informatik, die in ihrem ersten Hochschulsemester geringen Prüfungserfolg (keine oder eine bestandene Fachprüfung) vorweisen. Im Sommersemester 2013 betraf dies 123 Studierende. Um Gründe für den Misserfolg zu ermitteln sind die Betroffen dazu verpflichtet einen Fragebogen auszufüllen. Anhand der dort gegeben Antworten versucht die Fachstudienberatung Studierende mit besonderen Defiziten im Studienverhalten, den sogenannten *Problemfällen*, zu ermitteln.

Diese Arbeit soll durch die Identifizierung von Klassen von Studierenden mit frühem Prüfungsmisserfolg zukünftig dabei helfen die Auswahl dieser *Problemfälle* zu vereinfachen. Die bisher gesammelten Fragebögen dienen dabei als Datengrundlage. Mit Hilfe von Clusteranalysen sollen daraus Gruppen von Studierenden mit Prüfungsmisserfolg im ersten Hochschulsemester ermittelt werden.

Da an der TU Darmstadt ein Studienbeginn im Fach *Bachelor of Science* Informatik sowohl im Winter- als auch im Sommersemester möglich ist, soll zudem geprüft werden, ob sich die gefunden Klassen ähnlich oder voneinander verschieden sind.

In dieser Arbeit werden zunächst einige allgemeine Grundlagen (vgl. Kapitel 2), wie das Mentorensystem, die verwendeten Clusteranalyseverfahren, und vorhergehende Arbeiten näher erläutert. In einer Datenanalyse (Kapitel 3) werden die Antworten ausgewählter Fragen des Fragebogens über einen Beobachtungszeitraum (Sommersemester 2011 - Sommersemester 2013) mit statistischen Mitteln beschrieben. Eine, im Rahmen dieser Arbeit, entwickelte Software wird in Kapitel 4 betrachtet. Den Hauptteil bildet das Kapitel 5, die Clusteranalyse zur Identifizierung von Klassen von Studierenden mit geringem Prüfungserfolg. Dabei werden vier Leitfragen berücksichtigt:

- Gibt es, trotz unterschiedlicher Prüfungen und anderer Tätigkeiten vor Beginn des Studiums, ähnliche Gruppen im Winter- und Sommersemester?
- In welchen Clustern sind *Problemfälle* zu finden und was charakterisiert diese?
- Gibt es Unterschiede zwischen Studierenden mit einer bzw. keiner bestandenen Prüfung?
- Welche Attribute sind relevant, welche weniger?

2 Grundlagen

Dieses Kapitel beschreibt die Grundlagen, auf denen diese Arbeit aufbaut. Zunächst wird das *studentische Mentorensystem der Informatik* an der TU Darmstadt erläutert. Es folgt die Beschreibung, der in dieser Arbeit verwendeten Algorithmen aus dem Bereich des *Maschinellen Lernens*. Des Weiteren wird das zur Analyse verwendete Softwaretool kurz vorgestellt. Das Kapitel schließt mit dem Blick auf vorangegangene Arbeiten zu diesem Thema.

2.1 Mentorensystem

Das studentische Mentorensystem der Informatik dient der Unterstützung der Studierenden des Studiengangs *Bachelor of Science Informatik* im ersten Studienjahr. Dabei steht das Einfinden in den universitären Alltag im Vordergrund. Dieses umfasst hauptsächlich folgende Aspekte:

- Umgang mit den verschiedenen Veranstaltungsformen (Vorlesungen, Gruppenübungen, Praktika, Sprechstunden)
- Prüfungsvorbereitung (Schätzung des Lernaufwands, Erstellung von Lernplänen, Pseudoprüfung)
- Arbeiten in einer Lerngruppe, sowie Unterstützung beim Finden einer solchen
- Soziale Aspekte wie Wohnungssuche und private Schwierigkeiten

Im Wintersemester (WiSe) 2006/07 wurde das studentische Mentorensystem im ersten Fachsemester vom damaligen Dekan des Fachbereichs Informatik, Prof. Dr. Karsten Weihe, in Zusammenarbeit mit der Fachschaft als Pilotsystem für Erstsemester im *Bachelor of Science* Informatik erstmals durchgeführt. Im WiSe 2007/08 entstand daraus, im Rahmen der Initiative für gute Lehre der TU Darmstadt ein Kooperationsprojekt mit der Hochschuldidaktischen Arbeitsstelle (HDA). Seit dem Sommersemester (SoSe) 2010 gibt eine weitere Maßnahme. Im zweiten Fachsemester werden Studierende individuell beraten und geschult, die im ersten Fachsemester eine oder keine Prüfung bestanden haben[1]. Mit fachkundiger Begleitung durch die HDA wird das Programm für beide Fachsemester kontinuierlich verbessert und schrittweise auf andere Studiengänge der TU Darmstadt in fachspezifischer Ausprägung übertragen. [GBFT13]

2.1.1 Ziele des Mentorensystems

Das studentische Mentorensystem der Informatik beinhaltet eine enge Betreuung der Studierenden im ersten und zweiten Semester durch erfahrende Studierende des Studiengangs Informatik an der TU Darmstadt. Folgende Ziele [GBFT13] werden verfolgt:

- Anregung zur Reflexionen des Studierenden in Bezug auf die eigene Passung zum Studienfach Informatik an der TU Darmstadt
- Schaffung von Orientierung und Klarheit zum Studienanfang über die Anforderungen des Fachbereichs und des Studienfachs
- Einblick in das Berufsbild des Informatikers und die Struktur des Informatikstudiums
- Erzeugung einer Bindung des Studierenden zum Studienfach und Studienort
- Analyse von Gründen für einen eventuellen Studienabbruch bzw. -wechsel des Studierenden

[1] Probeklausuren oder Klausen zur Erlangung einer Studienleistung werden nicht mit einbezogen

2.1.2 Ablauf des Mentorensystems

Das studentische Mentorensystem der Informatik an der TU Darmstadt erstreckt sich über das erste Studienjahr. Es gliedert sich in zwei Teile auf:

1. Persönliche Betreuung aller Studierenden in ihrem ersten Hochschulsemester
2. Unterstützung aller Studierenden im zweiten Hochschulsemester mit geringen Prüfungserfolg (keine oder eine bestandene Prüfung im ersten Semester)

In der Vorlesungszeit finden wöchentlich Beratungsgespräche von ca. 15 Minuten zwischen einem Erstsemester-Studierenden und einem studentischen Mentor statt. Zum Erwerb des *Bachelor of Science* ist die Teilnahme am Mentorensystem nach Studienordnung des Fachbereichs im ersten Studienjahr erforderlich. Es handelt sich somit um ein verpflichtendes Programm für alle Studienanfänger im ersten Studienjahr. Das studentische Mentorensystem im zweiten Semester ist für alle Studierende verpflichtend, die in ihrem ersten Semester keine oder nur eine Prüfung bestanden haben. Die Studienordnung für den Studiengang *Bachelor of Science* Informatik an der TU Darmstadt [Dar] empfiehlt für das erste Semester und dem Beginn des Studiums zum Wintersemester folgende Fächerbelegung mit den dazugehörigen *Credit Points* (CP):

- Grundlagen der Informatik 1 (GdI 1) - 10 CP
- Technische Grundlagen der Informatik (TGdI) - 12 CP
- Mathematik 1 (Mathe 1) - 9 CP

Um vom Mentorensystem im zweiten Semester befreit zu werden, muss ein Studierender mindestens zwei Prüfungen bestanden haben. Nimmt ein Studierender sein Studium im Sommersemester auf, sieht der vom Fachbereich empfohlene Studienplan für das erste Semester wie folgt aus:

- Grundlagen der Informatik 1 (GdI 1) - 10 CP
- Formale Grundlagen der Informatik 1 (FGdI 1) - 5 CP
- Formale Grundlagen der Informatik 2 (FGdI 2) - 5 CP
- Einführung in Human Computer Systems (HCS) - 5 CP

Abbildung 2.1 zeigt den gesamten Ablauf des Mentorensystems der Informatik des ersten und zweiten Semesters für alle Studierenden des Studiengangs *Bachelor of Science* Informatik Studierende. Um am Mentorensystem teilnehmen zu können muss sich der Studierende über das *Lernportal Informatik* (Moodle) [moo] zu einem wöchentlichen Mentorengespräch anmelden. Freiwillige Angebote, wie ein Lerngruppentreffen, ergänzen das Angebot im ersten Semester. Eine finale Vorlesung, mit Lehrevaluation, bildet den Abschluss des Mentorensystems im ersten Semester.

Tritt der oben beschriebene Prüfungsmisserfolg (weniger als zwei Prüfungen bestanden) ein, ist eine Teilnahme am Mentorensystem im zweiten Semester verpflichtend. Die betroffenen Studierenden müssen zunächst einen Fragebogen ausfüllen, der Aufschluss über die Gründe des Prüfungsmisserfolgs geben soll. Die Auswertung dieser Fragebögen dient als Datengrundlage für diese Arbeit (vgl. Unterabschnitt 2.1.3).

Die Auftaktveranstaltung für das Mentorensystem des zweiten Semesters ist ein Vortrag der Fachstudienberatung. Dort werden die Studierenden unter anderem über den weiteren Ablauf des Mentorensystems informiert. Anhand der Fragebögen werden Studierende mit besonderem Betreuungsbedarf ermittelt. Diese müssen an einem Beratungsgespräch mit der Fachstudienberatung teilnehmen. Alle anderen an einem Gespräch mit einem studentischen Mentor. Nach diesen ersten Gesprächen folgen zwei Workshops zu den Themen Zeitmanagement und Prüfungsvorbereitung, sowie zwei weitere Mentorengespräche.

Der Mentor hat im Verlauf des zweiten Semesters jederzeit die Möglichkeit einen Studierenden aus dem Mentorensystem zu entlassen. Dieser Umstand tritt ein, falls der Mentor keinen weiteren Betreuungsbedarf des Studierenden erkennt. Nach regulärer Beendigung des Mentorensystems, oder nach vorzeitiger Entlassung aus diesem durch einen Mentor, wird dem Studierenden die Teilnahme durch Anerkennung in TUCaN[2], dem Campus-Management-System der TU Darmstadt, bestätigt. [GBFT13]

[2] http://www.tucan.tu-darmstadt.de

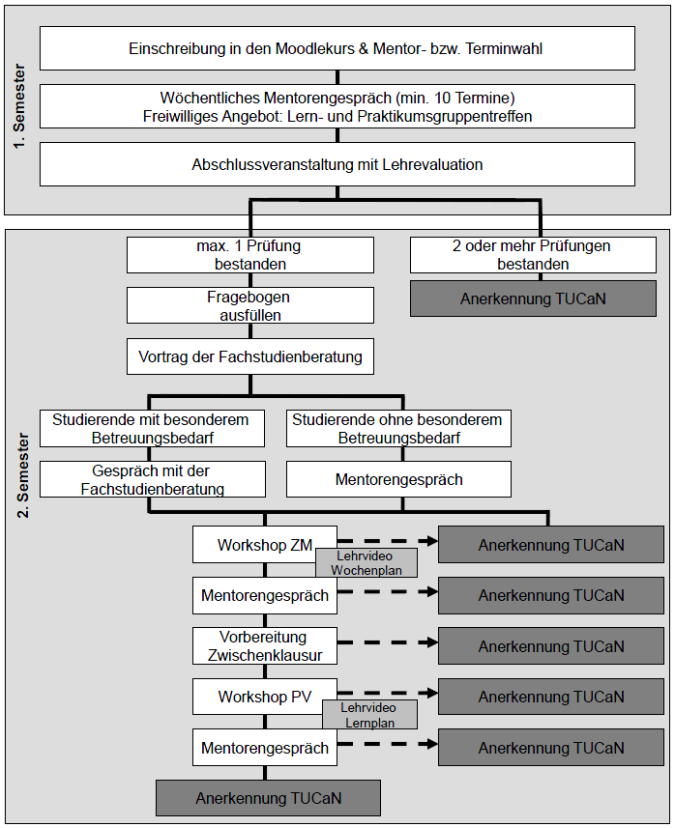


Abbildung 2.1: Ablauf des Mentorensystems des Fachbereichs Informatik im ersten Studienjahr. Quelle: [GBFT13]

2.1.3 Fragebogen zum Prüfungsmisserfolg

Wie im vorhergehenden Abschnitt erwähnt, sind alle Studierenden des Studiengangs Bachelor Informatik an der TU Darmstadt mit maximal einer bestandenen Prüfung im ersten Hochschulsemester dazu aufgefordert einen Fragebogen auszufüllen. Dieser wird im *Lernportal Informatik* [moo] für das Mentorensystem im ersten Semester zur Verfügung gestellt. Die erstmalige Befragung, der Studierenden mit geringem Prüfungserfolg, fand im Sommersemester 2010 statt. Den Fragebogen in seiner jetzigen Form gibt es seit dem Sommersemester 2011. Die für diese Arbeit zu Grunde liegenden Ergebnisse der Befragungen beginnen ab diesem Zeitpunkt.

Der Fragebogen erfasst Gründe für den Prüfungsmisserfolg und gibt Aufschluss über die Studiensituation der Betroffenen. Zudem können die Teilnehmer freiwillige Angaben zu ihrer privaten Situation machen, sofern diese aus ihrer Sicht zum Prüfungsmisserfolg beigetragen haben. Der Fragebogen erfasst u.a. Angaben zur Prüfungsanmeldung und -vorbereitung, Angaben zum Studienverhalten, zum Prüfungsmisserfolg und der privaten Situation.

Eine detaillierte Auflistung der Fragebogeninhalte ist in Abschnitt 3.2 dieser Arbeit zu finden. Die Ergebnisse des Fragebogens dienen dieser Thesis zum einen als Datengrundlage für eine statistische Datenanalyse (vgl. Kapitel 3), zum anderen als Basis für Clusteranalysen zur Identifizierung von Klassen von Studierenden (vgl. Abschnitt 2.2 und Kapitel 5).

2.2 Clusteranalyse

Die Clusteranalyse bezeichnet Verfahren, die zur Bildung von Gruppen eingesetzt werden. Ziel ist es „ähnliche Objekte zu Gruppen zusammenzufassen“[Gut]. Diese Gruppen werden als Cluster oder Klassen bezeichnet. Objekte können Personen (Teilnehmer einer Befragung), Aggregate (Organisationen, Nationen, Berufsgruppen usw.) oder Variablen (Merkmale) sein. Bei Personen und Aggregaten spricht man von einer *objektorientierten* Datenanalyse. Bei Variablen von einer *variablenorientierten* Datenanalyse [BPW10].

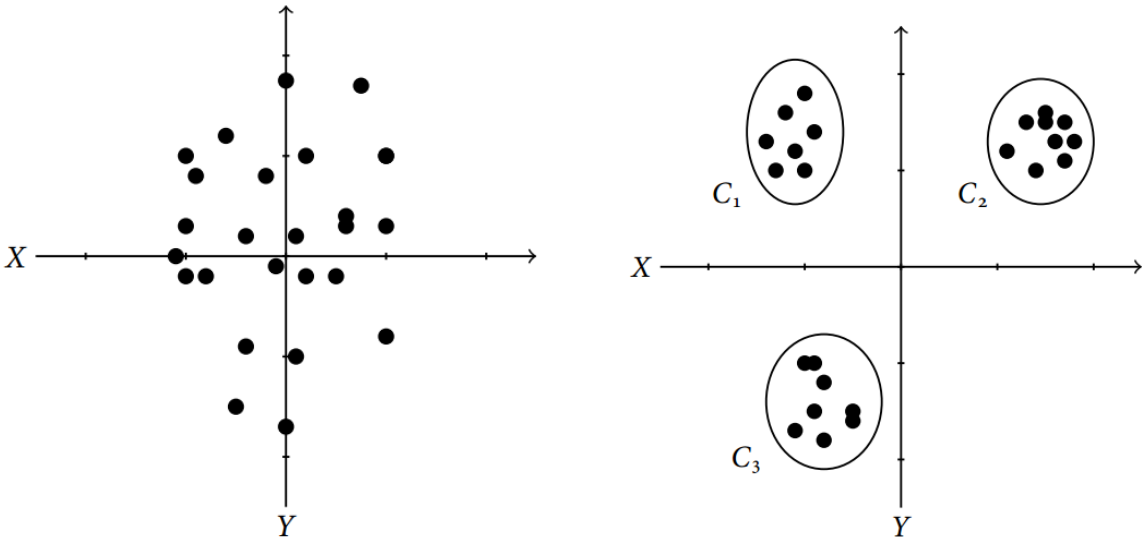
In Abbildung 2.2 werden die beiden Möglichkeiten der Clusteranalyse an einem Beispiel verdeutlicht. Die gezeigte Datenmatrix enthält die gegeben Antworten einer Befragung von Studierenden. Die erste Zeile enthält eine Kurzbe-

schreibung der Frage. Jede weitere Zeile steht für die gegeben Antworten eines Studierenden. Möchte man Gruppen von Studierenden finden, muss eine *objektorientierte* Clusteranalyse durchgeführt werden. Die Objekte sind demnach die Zeilen einer Datenmatrix. Will man hingegen Cluster anhand von Merkmalen finden, zum Beispiel soll jeder genannte Abiturnotenbereich genau einem bestimmten Cluster zugeordnet werden, sind die Objekte in der entsprechenden Spalte der Datenmatrix zu finden. In diesem Fall ist eine *variablenorientierte* Analyse notwendig.

	A	B	C	D
1	**Anfahrtsweg zur Uni:**	**Ich wohne:**	**Abiturnote**	**Mathematiknote**
2	bis zu 60 min	bei meinen Eltern	2,6-3,0	15-13
3	weniger als 15 min	im eigenen Haushalt (WG, allein, etc.)	2,6-3,0	15-13
4	bis zu 30 min	im eigenen Haushalt (WG, allein, etc.)	k.A.	3-1
5	weniger als 15 min	im eigenen Haushalt (WG, allein, etc.)	2,6-3,0	12-10
6	länger als 60 min	bei meinen Eltern	3,1-3,5	6-4
7	bis zu 60 min	bei meinen Eltern	2,1-2,5	9-7
8	länger als 60 min	bei meinen Eltern	1,6-2,0	15-13
9	bis zu 30 min	bei meinen Eltern	3,1-3,5	9-7

Abbildung 2.2: Ausschnitt aus einer Datenmatrix. Jede Zeile steht für einen Teilnehmer einer Umfrage (Objekt, rot markiert), die Spalten beinhalten die möglichen Antworten der einzelnen Fragen (Variablen, blau markiert).

Jede Clusteranalyse unterliegt der Vorstellung von *homogenen Gruppen* [Koz82]. Dies bedeutet einerseits, dass die zu klassifizierenden Objekte innerhalb eines Clusters ähnlich sind. Andererseits wird gefordert, dass die Objekte unterschiedlicher Cluster verschieden sind. Um die Ähnlichkeit zwischen Objekten zu beschreiben, muss ein geeignetes Distanzmaß gewählt werden [WZ01]. Dabei können sich die Ergebnisse der Analyse je nach gewähltem Distanzmaß unterscheiden. Für eine *objektorientierte* Clusteranalyse, wie sie in dieser Arbeit vorgenommen wird (vgl. Kapitel 5), eignet sich die quadrierte euklidische Distanz [BPW10]. Abbildung 2.3 zeigt zwei Beispiele für die grafische Darstellung der Ergebnisse einer Clusteranalyse, die von zwei Variablen (X und Y) abhängig ist. Das erste Beispiel (2.3(a)) zeigt eine nicht homogene Menge von Objekten. Hier hat das Clustering keine Cluster identifizieren können. Das zweite (2.3(b)) zeigt hingegen drei homogene und voneinander räumlich getrennte Cluster.



(a) Die Objekte bilden eine große Punktwolke.

(b) Objekte sind in drei homegene und räumlich voneinander getrennte Cluster unterteilt.

Abbildung 2.3: Objekte mit den Variablen X und Y im zweidimensionalen Raum R^2 dargestellt. Quelle: [BPW10]

2.2.1 Verfahrensgruppen zur Clusteranalyse

Für die Durchführung einer Clusteranalyse existieren eine Vielzahl von Verfahren. Je nach Anwendungsfall muss daraus ein geeignetes gewählt werden. *Bacher et al.* [BPW10] unterscheidet drei Verfahrensgruppen.

- **Unvollständige Clusteranalyseverfahren**[3]: Diese Verfahren nehmen keine Zuordnung der zu klassifizierenden Objekte zu Clustern vor. Stattdessen wird eine räumliche Darstellung dieser Objekte berechnet; idealerweise im zweidimensionalen, bzw. im niedrigdimensionalen Raum. Mit dem Ergebnis kann eine visuelle Analyse vorgenommen werden.
- **Deterministische Clusteranalyseverfahren**: Bei dieser Art von Verfahren werden die Cluster berechnet und die Klassifikationsobjekte diesen deterministisch zugeordnet. Das bedeutet, dass jedes Objekt mit einer Wahrscheinlichkeit von 0 oder 1 einem (überlappungsfreie Verfahren) oder mehreren (überlappende Verfahren) Clustern zugeordnet ist.
- **Probabilistische Clusteranalyseverfahren**: Die Klassifikationsobjekte werden mit einer Wahrscheinlichkeit zwischen 0 und 1 einem bestimmten Cluster zugeordnet. Auch hier gibt es überlappungsfreie und überlappende Verfahren.

Ferner unterscheidet man zwei Vorgehensweisen bei der Clusteranalyse. Die *explorative Clusteranalyse* ist dadurch charakterisiert, dass die Zahl der Cluster unbekannt ist und ermittelt werden muss. Zudem sind die Merkmale der Cluster vorab nicht bekannt. Bei der *konfirmatorischen Clusteranalyse* ist die Zahl der Cluster im Vorfeld bekannt. Ebenfalls können auch Merkmale der einzelnen Cluster bekannt sein. Somit haben hier die Cluster bereits (teilweise) eine inhaltliche Bedeutung.

2.2.2 K-Means-Algorithmus

Die in dieser Arbeit durchgeführte Clusteranalyse verwendet den K-Means-Algorithmus [Wika]. Daher wird er in diesem Abschnitt kurz beschrieben. Der K-Means-Algorithmus ist ein überlappungsfreies, deterministisches Verfahren (vgl. Unterabschnitt 2.2.1) zur Clusteranalyse. Er eignet sich für die in Kapitel 5 durchgeführte Analyse, da nach *Bacher et al.* [BPW10] dieser gewählt werden kann wenn:

- eine Einteilung in Gruppen vorgenommen werden soll,
- eine Datenmatrix als Datengrundlage vorliegt und
- eine objektorientierte Clusteranalyse durchgeführt wird.

Die Grundidee ist es, den gegebenen Datensatz in K Gruppen (Cluster) aufzuteilen und jedes zu klassifizierende Objekt genau einem dieser Cluster zuzuordnen. Zu Beginn werden (zufällig) K Clusterzentren gewählt. Als Clusterzentrum bezeichnet man einen für dieses Cluster repräsentativen Wert der jeweiligen Variable (Attribut). Es ist der Wert, der am häufigsten in diesem Cluster für das jeweilige Attribut vorkommt. Diese Repräsentanten ermöglichen eine anschließende Benennung und Interpretation der Cluster.

Ziel des K-Means-Verfahrens ist es, die Clusterzentren so zu bestimmen, dass die Streuungsquadratsumme[4] in den Clustern $SQ_{in}(K)$ minimal ist. Formal lässt sich dies folgendermaßen beschreiben [BPW10]: K Clusterzentren $\overline{x}_{kj}$ ($k = 1, 2, ..., K$; $j = 1, 2, ..., m$; K = Anzahl der Cluster; m = Zahl der Variablen) werden so berechnet, dass die Streuungsquadratsumme in den Clustern

$$SQ_{in}(K) = \sum_{k} \sum_{g \in k} \sum_{j} (x_{gj} - \overline{x}_{kj})^2 \tag{2.1}$$

minimal wird. Die Streuung der Variablenwerte eines Clusters von den Clusterzentren

$$\sum_{k} (x_{gj} - \overline{x}_{kj})^2 = d^2_{g,k} \tag{2.2}$$

[3] Werden in der Literatur auch häufig als geometrische Verfahren bezeichnet.

[4] Summe der quadratischen Abweichungen von den Clusterzentren [Wika]

ist gleich der quadrierten euklidischen Distanz d^2 zwischen dem Objekt g und dem Clusterzentrum k. Daher kann die Optimierungsaufgabe folgendermaßen formuliert werden:

$$SQ_{in}(K) = \sum_j \sum_{g \in k} d^2_{g,k} \rightarrow \min \tag{2.3}$$

Für die Anwendung des K-Means-Algorithmus muss die Anzahl der Cluster (K) vorgegeben werden. Der Algorithmus gliedert sich in vier Schritte:
Schritt 1: Zufällige Zuteilung der Objekte zu den K Clustern.
Schritt 2: Neuberechnung der Clusterzentren über

$$\overline{x}_{kj} = \frac{\sum_{g \in k} x_{gj}}{n_{kj}} \tag{2.4}$$

mit n_{kj} als Anzahl der Objekte mit gültigen Angaben in der Variablen j.
Schritt 3: Die Objekte g werden demjenigen Clusterzentrum k zugeordnet, zu dem sie die minimale quadrierte euklidische Distanz aufweisen:

$$g \in k \Leftrightarrow k = \min_{k^*=1,2,\ldots,K} (d^2_{g,k^*}) \tag{2.5}$$

In jeder Iteration wir somit die Streuungsquadratsumme in den Clustern

$$SQ_{in}(K) = \sum_j \sum_{g \in k} d^2_{g,k} = \sum_g \min_{k^*=1,2,\ldots,K} (d^2_{g,k^*}) \tag{2.6}$$

minimiert.
Schritt 4: Hat sich im dritten Schritt die Zuordnung der Objekte zu den Clustern verändert, erfolgt eine erneute Durchführung der Schritte drei und vier. Andernfalls wird der Algorithmus beendet. Eine Beendigung kann ebenfalls erfolgen, wenn eine zuvor festgelegte Zahl an Iterationen erreicht ist.

Es gibt diverse Ausprägungen des K-Means-Algorithmus [Wika]. Beispielsweise kommen statt der euklidischen Distanz auch andere Distanzmaße, wie die Manhatten-Distanz [Wikc], zum Einsatz. Eine Software, die unter anderem eine Clusteranalyse mit Hilfe des K-Means-Verfahrens ermöglicht, ist WEKA (vgl. Abschnitt 2.3). Die im Rahmen dieser Arbeit vorgenommen Clusteranalysen wurden mit dieser Software durchgeführt.

2.3 WEKA

WEKA (Waikato Environment for Knowledge Analysis) [Wai] ist eine Software, die diverse Verfahren aus dem Bereich des *Maschinellen Lernens* zur Verfügung stellt. Sie ist in der Programmiersprache Java[5] geschrieben. Sie enthält neben einer grafischen Benutzerschnittstelle unter anderem auch Clusteranalyseverfahren, wie den K-Means-Algorithmus, und Visualisierungsmöglichkeiten. Als Eingabedaten werden Dateien vom Typ .arff verwendet (vgl. Seite 27). Ein Tool, welches die zur Analyse vorhandenen Datenmatrizen in dieses Format umwandelt, ist im Rahmen dieser Arbeit entstanden und in Kapitel 4 beschrieben. Vorhergehende Arbeiten (vgl. Abschnitt 2.4), die sich ebenfalls mit Anwendung von Algorithmen aus dem Gebiet des *Maschinellen Lernens* auf die Fragebögen des Mentorensystems im zweiten Semester beziehen, haben ebenfalls zur Durchführung WEKA verwendet.

2.4 Vorhergehende Arbeiten

Eine erste Arbeit, welche die Resultate der Fragebögen aus dem Mentorensystem (vgl. Unterabschnitt 2.1.3) und Verfahren des *Maschinellen Lernens* verbindet, ist ein Praktikumsbericht von Maxi Neubacher. Der Titel des Berichts ist: „Verbesserung und Weiterentwicklung der Abläufe im Mentorensystem des Fachbereichs Informatik der TU Darmstadt - Klassifizierung von Studierenden mit schlechtem Prüfungserfolg im ersten Semester“ [Neu12]. Die Aufgabenstellung bestand darin mit Verfahren aus dem *Maschinellen Lernen* Regeln zu lernen, mit denen sogenannte *Problemfälle*[6] anhand ihrer Angaben im Fragebogen identifiziert werden können. Mit der Software WEKA (vgl. Abschnitt 2.3) wurden diese

[5] http://www.java.com
[6] Studierende, mit besonderen Defiziten im Studienverhalten

Regeln, unter der Verwendung verschiedener Regel-Lern-Algorithmen, ermittelt. Als Datengrundlage dienten die Fragebögen des Wintersemesters 2010/11 und des Sommersemesters 2011.

Gefunden wurde unter anderem eine einfache Regelmenge, die besagt, dass Studierende, die im Fragebogen angeben keine Prüfung bestanden zu haben und auch keine mitgeschrieben haben, als *Problemfälle* zu klassifizieren sind. Mit dieser Regelmenge wurden 62,6%, der von der Fachstudienberatung als *Problemfälle* klassifizierten Studierenden, korrekt erkannt. Komplexere Regelmengen, mit einer höheren Regelanzahl, konnten nur unwesentlich mehr korrekt klassifizierte Studierende ermitteln (*Bsp.: 11 Regeln und 68,3% korrekt klassifizierte Studierende*).

Eine nachfolgende, auf den oben genannten Praktikumsbericht aufbauende, Arbeit von Maxi Neubacher hat den Titel „Analyse von Algorithmen des maschinellen Lernens für das Mentorensystem Informatik“ [Neu13]. Diese wurde im Rahmen einer Bachelor-Thesis an der Hochschule Darmstadt verfasst. Ziel der Arbeit war es auch hier geeignete Regeln zu finden, die Studierende als *Problemfälle* identifizieren. Dazu standen, im Vergleich zum Praktikumsbericht, zusätzlich die beantworteten Fragebögen des Wintersemesters 2011/12, sowie des Sommersemester 2012 zur Verfügung. Erneut wurden mit Hilfe diverser Regel-Lern-Algorithmen und dem Tool WEKA Regelmengen gebildet. Eine der gefundenen Regelmengen, lässt sich mit Hilfe eines Entscheidungsbaum (Abbildung 2.4) darstellen.

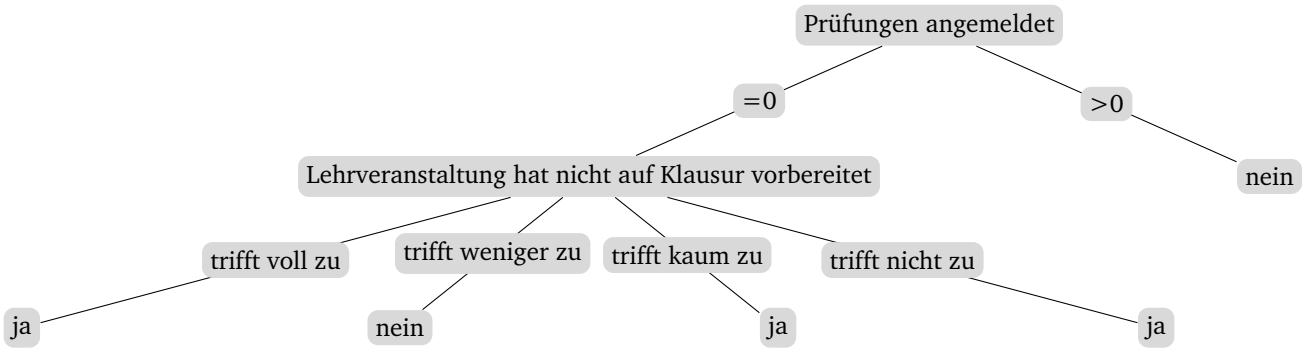


Abbildung 2.4: Entscheidungsbaum zur Identifizierung von *Problemfällen* nach [Neu13].

Mit den Regeln aus diesem Entscheidungsbaum wurden 71,8% der von der Fachstudienberatung ermittelten *Problemfälle* korrekt identifiziert. Weitere gefundene Regelmengen lieferten entweder ein schlechtere Genauigkeit oder beinhalteten zu viele Regeln. Eine finale Regelmenge, so das Resümee der Arbeit, sei damit noch nicht gefunden, da größere Datenmengen notwendig sein.

Eine weitere Arbeit, die sich mit der Klassifizierung von Studierenden beschäftigt, stammt von Jünger *et al.* und trägt den Titel „Subjektive Kompetenzeinschätzung von Studierenden und ihre Leistung im OSCE“ [JSN06]. Ziel dieser Arbeit war es Medizinstudierende der Universität Heidelberg anhand ihrer Prüfungsleistung im Verhältnis zur Selbsteinschätzung zu klassifizieren. Die Selbsteinschätzung wurde hierbei mit Hilfe eines Fragebogens ermittelt; die Prüfungsleistung anhand der sogenannten *klinisch-praktischen OSCE-Prüfungen*. Dabei wurden Studierende mit schlechten Prüfungsergebnissen und selbst hoch eingeschätzter Kompetenz als *Selbstüberschätzer* klassifiziert.

Die Studie „Klassifizierung europäischer Bildungssysteme anhand von OECD-Bildungsindikatoren“ von Nairz-Wirth *et al.* [NWE06] nutzt eine explorative Clusteranalyse zur Klassifizierung von nationalen Bildungssystemen. Ziel war es jene Länder zu gruppieren, die ein ähnliches Bildungssystem vorweisen. Als Datengrunde wurden sogenannte OECD[7]-Bildungsindikatoren (*Gesellschaftliche Wertigkeit, Bildungsinvestitionen pro BildungsteilnehmerIn, Allgemeinen Bildungsstand*) verwendet.

[7] Organisation für wirtschaftliche Zusammenarbeit und Entwicklung

3 Datenanalyse

Das Ziel dieser Arbeit ist es Klassen von Studierenden zu identifizieren. Dafür ist es notwendig, die aus dem Fragebogen gewonnen Daten darzulegen und einer ersten statistischen Auswertung zu unterziehen. In diesem Kapitel wird dazu zunächst der Fragebogen im Detail erläutert. Im weiteren Verlauf werden ausgewählte Ergebnisse der Befragung dargestellt und mit statistischen Verfahren ausgewertet. Dabei steht die Frage im Mittelpunkt, ob sich die Ergebnisse der Befragung in den einzelnen Semestern im Befragungszeitraum ähneln oder Unterschiede erkennbar sind. Zudem werden die Resultate des Winter- und Sommersemesters gegenübergestellt, da die vorgesehenen Fachprüfungen unterschiedlich sind und sich somit auch andere Gründe für den Prüfungsmisserfolg ergeben können.

3.1 Datengrundlage

Wie schon in Unterabschnitt 2.1.3 kurz eingeführt, gibt es einen Fragebogen, den alle Studierende des Studiengangs *Bachelor of Science* Informatik ausfüllen müssen, die in ihrem ersten Semester keine oder eine Prüfung bestanden haben. Die ausgefüllten Fragebögen dienen als Datengrundlage für diese Arbeit und die nachfolgend beschriebenen Analysen. Zur Verfügung stehen die Ergebnisse der Befragungen der Sommersemester[1] 2011, 2012 und 2013, sowie die der Wintersemester[2] 2011/12 und 2012/13. Im Sommersemester 2010 und dem Wintersemester 2011/12 lag der Fragebogen in einer anderen Form vor. Aus diesem Grund fließen die Ergebnisse aus den beiden Semestern nicht in diese Arbeit ein.

	WS 10/11	**WS 11/12**	**WS 12/13**	**SoSe 10**	**SoSe 11**	**SoSe 12**	**SoSe 13**
Anzahl	20	38	45	23	82	105	123

Tabelle 3.1: Anzahl ausgefüllter Fragebögen

Die Werte aus Tabelle 3.1 belegen eine kontinuierlich steigende Zahl von ausgefüllten Fragebögen. Grund dafür ist die Verpflichtung zum Ausfüllen seit dem Sommersemester 2011. Zudem erfolgt seit dem Wintersemester 2012/13 eine strengere Nachverfolgung der Studierenden, die den Fragebogen zunächst nicht ausgefüllt haben, obwohl das Kriterium (0 - 1 bestandene Prüfung) auf sie zutrifft. Diese Nachverfolgung wird durch die Fachstudienberatung und das Mentorensystem durchgeführt. Die Zahl der ausgefüllten Fragebögen entspricht daher seitdem auch der Anzahl der Studierenden mit keiner oder einer bestanden Prüfung.

3.2 Der Fragebogen im Detail

Im Folgenden sind die einzelnen Fragen des Fragebogens mit den möglichen Antwortmöglichkeiten tabellarisch gelistet. Der Fragebogen selbst teilt sich wiederum in sechs Kategorien auf:

1. Allgemeine Angaben
2. Angaben zur Prüfungsanmeldung
3. Angaben zur Prüfungsvorbereitung
4. Angaben zum Studienverhalten
5. Angaben zu Gründen des Prüfungsmisserfolgs
6. Private Situation

Fragen, die mit einen * gekennzeichnet sind, erlauben die Angabe von mehreren Antworten. Ist in einer Zeile im Antwortmöglichkeiten nichts angegeben, handelt es um ein Freitextfeld im Fragebogen.

[1] Die Befragung im Sommersemester bezieht sich immer auf die Studienanfänger im vorhergehenden Wintersemester
[2] Die Befragung im Wintersemester bezieht sich immer auf die Studienanfänger im vorhergehenden Sommersemester

1. Allgemeine Angaben: Tabelle 3.2 zeigt die Fragen dieser Kategorie. Darin nicht aufgeführt sind die im Fragebogen vorhandenen Fragen zu Name und Matrikelnummer, da diese für die Analyse in dieser Arbeit keine Rolle spielen. Die in der Tabelle angegebene Skala der Abiturnoten ist in dieser Form seit dem Fragebogen des Sommersemesters 2012 vorhanden. In den vorherigen Fragebögen teilt sich der *1er-Bereich* in die Werte *1,0 - 1,3; 1,4 - 1,7; 1,8 - 2,0* auf. Die restlichen Notenbereiche sind gleich geblieben. Die Frage nach der *letzten Informatiknote* ist erstmals im Fragebogen des Sommersemesters 2012 aufgeführt.

Frage	Antwortmöglichkeiten
Mein Anfahrtsweg zur Uni dauert einfach	weniger als 15 min; bis zu 30 min; bis zu 60 min; mehr als 60 min
Ich wohne während des Semesters	bei meinen Eltern; im eigenen Haushalt
In welchem Bereich liegt Ihre Abiturnote?	1,0 - 1,5; 1,5 - 2,0; 2,1 - 2,5; 2,6 - 3,0; 3,1 - 3,5; 3,6 - 4,0; ab 4,0; k.A.
In welchem Bereich liegt Ihre letzte Mathematiknote?	15-13; 12-10; 9-7; 6-4; 3-1; 0; k.A.
In welchem Bereich liegt Ihre letzte Informatiknote (falls vorhanden)?	15-13; 12-10; 9-7; 6-4; 3-1; 0; k.A.
Was haben Sie vor Beginn des Informatikstudiums gemacht?*	Bund; Zivildienst; freiwilliges soziales Jahr; Reisen; Ausbildung/Lehre; Jobben; anderes Studium; nichts Bestimmtes; Studienkolleg; Bundesfreiwilligendienst; k.A.

Tabelle 3.2: Allgemeine Angaben zu Wohnsituation, Schule und Tätigkeit vor Beginn des Studiums

2. Angaben zur Prüfungsanmeldung: In der zweiten Kategorie des Fragebogens wird nach allgemeinen Angaben zur Prüfungsanmeldung gefragt (vgl. Tabelle 3.3). Hierbei unterscheiden sich die Antwortmöglichkeit im Sommer- und Wintersemester, da unterschiedliche Prüfungen im Studienplan vorgesehen sind. In Tabelle 3.3 sind die Prüfungen für Studienanfänger des Wintersemesters gelistet. Im Sommersemester sind teilweise andere Prüfungen vorgesehen (vgl. Unterabschnitt 2.1.2).

Frage	Antwortmöglichkeiten
Zu welchen Prüfungen waren Sie zugelassen?*	GdI 1; TGdI; Mathe 1; sonstige; keine
Für welche Prüfungen waren Sie angemeldet?*	GdI 1; TGdI; Mathe 1; sonstige; keine
Welche Prüfungen haben Sie geschrieben?*	GdI 1; TGdI; Mathe 1; sonstige; keine
Von welchen Prüfungen haben Sie sich fristgerecht abgemeldet?*	GdI 1; TGdI; Mathe 1; sonstige; keine
Bei welchen Prüfungen waren Sie kurzfristig entschuldigt?*	GdI 1; TGdI; Mathe 1; sonstige; keine
Welche Prüfungen haben Sie bestanden?*	GdI 1; TGdI; Mathe 1; sonstige; keine
Zu welchen Klausuren sind Sie in die Klausureinsicht gegangen?*	GdI 1; TGdI; Mathe 1; sonstige; keine

Tabelle 3.3: Allgemeine Angaben zur Prüfungsanmeldung

3. Angaben zur Prüfungsvorbereitung: Die in Tabelle 3.4 aufgeführten Fragen sind pro vorgesehenem Prüfungsfach jeweils einmal im Fragebogen aufgeführt. So kann eine differenzierte Betrachtung für jedes Fach vorgenommen werden.

Frage	Antwortmöglichkeiten
Wie würden Sie Ihre Prüfungsvorbereitung selbst einschätzen?	sehr gut; gut; weniger gut; schlecht; k.A.
Haben Sie die Vorlesung regelmäßig besucht?	fast immer; häufig; unregelmäßig; selten; k.A.
Wie häufig haben Sie die Übungen besucht?	fast immer; häufig; unregelmäßig; selten; k.A.
Wie häufig haben Sie die Hausübungen gemacht?	fast immer; häufig; unregelmäßig; selten; k.A.

Tabelle 3.4: Allgemeine Angaben zur Prüfungsvorbereitung

4. Angaben zum Studienverhalten: Wie Tabelle 3.5 zeigt, wird hierbei nach der Zeiteinteilung und der Lernmethodik der Studierenden gefragt. Also wie viel Zeit, pro Tag oder Woche, im Durchschnitt für eine bestimmte Tätigkeit (z.B.: Zeitaufwand für Lehrveranstaltungsbesuche pro Woche) aufgebracht wird bzw. ob eine Lerngruppe vorhanden ist und eine Erwerbstätigkeit ausgeübt wird.

Frage	Antwortmöglichkeiten
Zeitaufwand für Lehrveranstaltungen in Stunden pro Woche	< 5h; 6-10h; 11-15h; 16-20h; 21-25h; > 25h
Zeitaufwand für Selbststudium in Stunden pro Woche	< 5h; 6-10h; 11-15h; 16-20h; 21-25h; > 25h
Haben Sie eine Lerngruppe?	ja; nein
Lernen Sie bevorzugt in Gruppen oder alleine?	eher in Lerngruppen; eher alleine
Verwenden Sie folgende Hilfsmittel zur Prüfungsvorbereitung?*	Karteikarten; Lerntagebuch; Lernplan; Lesemethoden; empfohlene Lehrbücher; selbstrecherchierte Lehrbücher
Gehen Sie neben dem Studium einer Erwerbstätigkeit nach?	ja; nein
Wenn ja, wie viele Stunden pro Woche?	< 5h; 5-10h; 11-20h; 21-30h; 31-40h; > 40h
Wie viel Zeit verbringen Sie mit Ihren Hobbies (pro Woche)?	< 5h; 5-10h; 11-15h; 16-20h; 21-25h; 26-30h
Wie viel Zeit verbringen Sie mit Ihren Freunden und Ihrer Familie (pro Woche)?	< 5h; 5-10h; 11-15h; 16-20h; 21-25h; 26-30h
Wie viele Stunden verbringen Sie durchschnittlich pro Tag mit Computer spielen?	< 1/2h; 1/2-1h; 1-1 1/2h; 1 1/2 -2h; > 2h
Wie viele Stunden verbringen Sie durchschnittlich pro Tag mit der Pflege von sozialen Onlinenetzwerken?	< 1/2h; 1/2-1h; 1-1 1/2h; 1 1/2 -2h; > 2h
Wie viele Stunden sehen Sie durchschnittlich pro Tag fern (Serien, Filme,...)?	< 1/2h; 1/2-1h; 1-1 1/2h; 1 1/2 -2h; > 2h

Tabelle 3.5: Angaben zum Studienverhalten

5. Angaben zu Gründen des Prüfungsmisserfolgs: Der fünfte Fragenblock (vgl. Tabelle 3.6) zielt auf die subjektive Einschätzung von Gründen des Prüfungsmisserfolgs ab. Hier soll der Studierende aus eigener Sicht angeben, welche Faktoren zu dem Ergebnis geführt haben, dass maximal eine oder keine Prüfung im vergangenen Semester bestanden wurde. Die Frage nach *speziellen Gründen für Prüfungsmisserfolg* ist für jede vorgesehene Prüfung im betreffenden Semester genau einmal im Fragebogen vorhanden.

Frage	Antwortmöglichkeiten
Der Lernstoff war zu schwer	trifft voll zu; trifft weniger zu; trifft kaum zu; trifft nicht zu
Der Lernstoff war zu viel	trifft voll zu; trifft weniger zu; trifft kaum zu; trifft nicht zu
Der Lernstoff war anders als erwartet	trifft voll zu; trifft weniger zu; trifft kaum zu; trifft nicht zu
Ich habe zu wenig Zeit investiert	trifft voll zu; trifft weniger zu; trifft kaum zu; trifft nicht zu
Ich habe mir zu viel vorgenommen	trifft voll zu; trifft weniger zu; trifft kaum zu; trifft nicht zu
Ich hatte keinen Lernplan	trifft voll zu; trifft weniger zu; trifft kaum zu; trifft nicht zu
Klausurinhalte waren anders als erwartet	trifft voll zu; trifft weniger zu; trifft kaum zu; trifft nicht zu
Ich war zu nervös in der Klausur	trifft voll zu; trifft weniger zu; trifft kaum zu; trifft nicht zu
Die Lehrveranstaltung hat nicht auf die Klausur vorbereitet	trifft voll zu; trifft weniger zu; trifft kaum zu; trifft nicht zu
Spezielle Gründe für Prüfungsmisserfolg	
Haben Sie vor weiterhin Informatik an der TU Darmstadt zu studieren?	ja; nein
Wenn nein, warum nicht? (optional)	
Welche Prüfungen haben Sie vor im nächsten Semester zu schreiben?	

Tabelle 3.6: Angaben zu den Gründen des Prüfungsmisserfolgs

6. Private Situation: Die Übergangsphase von Schule zu Hochschule bedeutet häufig auch eine deutliche private Veränderung. Wenn dies bei einem Studierenden der Fall ist, kann dieser in einem Freitextfeld optional *private Gründe* angegeben.

Die gesammelten Daten aus den Fragebögen dienen der Fachstudienberatung des Fachbereichs Informatik, bzw. den Mentoren aus dem Mentorensystem zur Vorbereitung auf die im Ablauf des zweiten Semesters beschriebenen Beratungsgespräche (vgl. Abbildung 2.1). Im nachfolgenden Abschnitt werden einige Kernfragen aus dem Fragebogen über die verschiedenen Semester verglichen.

3.3 Auswertung der Befragungen

In diesem Abschnitt werden ausgewählte Antworten auf Fragen aus den Fragebögen (vgl. Abschnitt 3.1) beginnend ab dem Sommersemester 2011 bis zum Sommersemester 2013 ausgewertet. Dabei wird auch auf Unterschiede zwischen dem Studienbeginn im Sommer und Winter eingegangen.

Zu allen Ergebnissen werden Konfidenzintervalle C [Wikb][PHRB09] gebildet, um *Ausreißer* in den Befragungen zu identifizieren. Diese, vom Erwartungswert $\overline{x}$ signifikant (dies entspricht einem Signifikanzniveau von $\alpha = 0,05$ [Wikf]) abweichenden Ergebnisse, sind in den Tabellen rot markiert. Zudem zeigt das Konfidenzintervall, dass der wahre Wert der Grundgesamtheit zu 95% $(1-\alpha)$ in diesem liegt. So lässt sich beispielsweise die Aussage formulieren, dass 25,78% bis 28,17% der Studierenden mit einer Wahrscheinlichkeit von 95% einen einfachen Anfahrtsweg zwischen 30 und 60 Minuten zur TU Darmstadt haben (vgl. Unterabschnitt 3.3.1).

Ein Konfidenzintervall mit hoher Breite deutet auf eine starke Variabilität der Grundgesamtheit hin. Die durchschnittliche Breite der Konfidenzintervalle liegt bei der folgenden Datenanalyse bei 11,67%. Die Standardabweichung [Wike] für die Breite beträgt 9,21%. Aus diesem Grund wird im Folgenden von einer *auffälligen* Breite des Konfidenzintervalls gesprochen, wenn diese größer als 20,88% ist. Da für den Erwartungswert $\overline{x}$ eine unbekannte Varianz vorliegt, wird eine *t-Verteilung* [Wikc] angenommen.

Alle Befragten haben in ihrem ersten Semester keine oder eine Prüfung im Studiengang *Bachelor of Science* Informatik bestanden. Die nachfolgenden Betrachtungen und Aussagen beziehen stets auf diese Gruppe von Studierenden.

3.3.1 Anfahrtsweg

Dieser Fragenkomplex behandelt die Frage nach der Dauer des einfachen Anfahrtswegs zur TU Darmstadt. Wie Tabelle 3.7 und Abbildung 3.1 zeigen gibt es im Wintersemester 2011/12 eine zu niedrige Zahl von Studierenden, die einen Anfahrtsweg von mehr als 60 Minuten angeben, im Vergleich zu den anderen aufgeführten Semestern. In allen anderen Semestern stellt der Teil der Studierenden mit einem Anfahrtsweg von über 60 Minuten die größte Gruppe dar. Im Sommersemester 2011 die Anzahl der Studierenden mit einem Anfahrtsweg von weniger als 15 Minuten deutlich niedriger, als in den übrigen Semestern.

Auch wenn das Ergebnis im Sommersemester 2012 leicht außerhalb des Konfidenzintervalls liegt, liefert die Dauer des Anfahrtswegs von $30 < x \leq 60$ Minuten den konstantesten Wert über den gesamten Befragungszeitraum. Hier beträgt die Breite des Konfidenzintervalls lediglich 2,39%. Anhand der Mittelwerte lässt sich erkennen, dass ca. 56% der Studierenden einen Anfahrtsweg von > 30 Minuten haben. Bei den Studienanfängern im Wintersemester ist Anteil der Studierenden, die eine Anfahrtszeit von ≤ 15 Minuten angeben, stets höher (25% und 24,44%), als bei Studienanfängern im Wintersemester (14,63%, 20,66% und 22,76%).

	SoSe 11	**WS 11/12**	**SoSe 12**	**WS 12/13**	**SoSe 13**	$\bar{x}$	C
$\leq$ 15 min	14,63%	25,00%	20,66%	24,44%	22,76%	21,50%	$16,29\% - 26,71\%$
$\leq$ 30 min	28,05%	28,13%	16,53%	17,78%	22,76%	22,65%	$15,84\% - 29,46\%$
$\leq$ 60 min	26,83%	28,13%	25,62%	26,67%	27,64%	26,98%	$25,78\% - 28,17\%$
$>$ 60 min	30,49%	18,75%	37,19%	31,11%	26,83%	28,87%	$20,47\% - 37,28\%$

Tabelle 3.7: Übersicht über die Angaben zum Anfahrtsweg

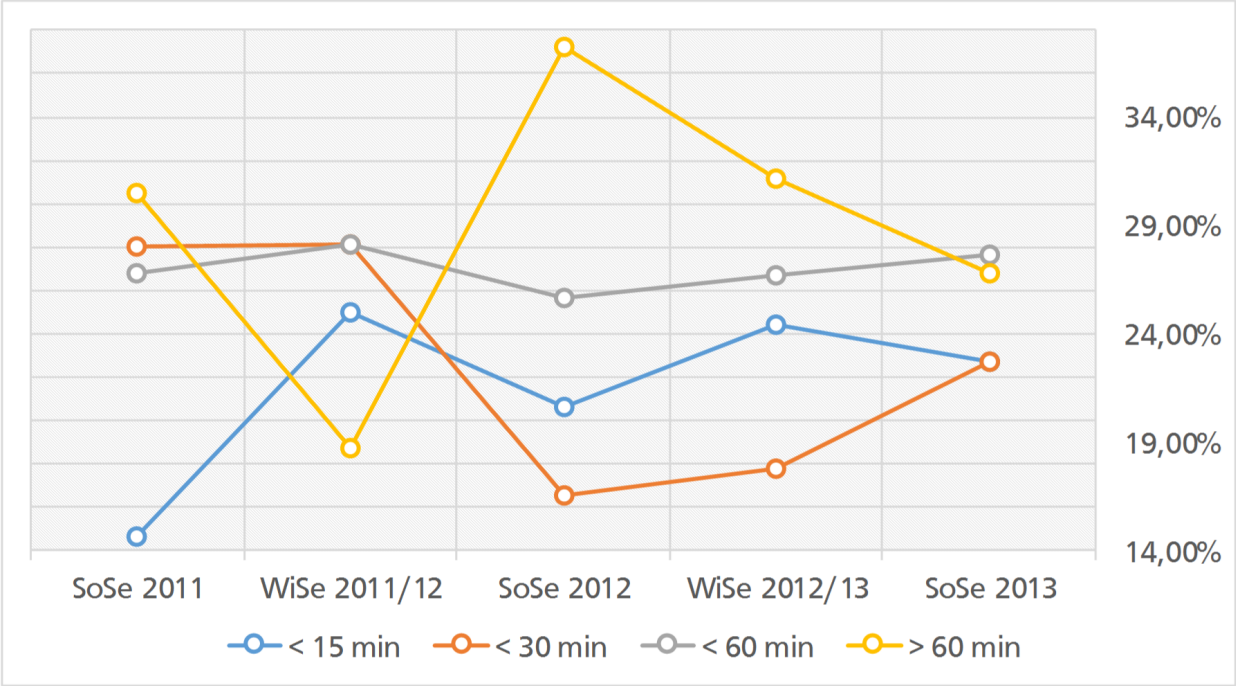


Abbildung 3.1: Angaben zum Anfahrtsweg

3.3.2 Wohnort

Die Beobachtungen bei der Betrachtung des Wohnorts zeigen (vgl. Tabelle 3.8) signifikante Abweichungen beim Wintersemester 2011/12. Im Vergleich zu den weiteren Semestern sind die Werte hier gedreht und fallen daher auch aus den entsprechenden Konfidenzintervallen. Zudem ist die Breite der Intervalle mit jeweils 27,35% auffällig hoch im Vergleich zu den übrigen Intervallbreiten in der Fragebogenauswertung dieser Arbeit. Berücksichtigt man aufgrund dessen die Ergebnisse des Wintersemesters 2011/12 nicht, ergeben sich engere Intervalle, als in der ursprünglichen Berechnung (vgl. blaue Werte in Tabelle 3.8). Die Breite der Konfidenzintervalle verringert sich durch diese Maßnahme auf 6,06%.

	SoSe 11	**WS 11/12**	**SoSe 12**	**WS 12/13**	**SoSe 13**	$\bar{x}$	C
bei den Eltern	59,76%	34,38%	61,16%	55,56%	57,72%	53,71%	$40,04\% - 67,39\%$
						58,55%	$55,52\% - 61,58\%$
eigener Haushalt	40,24%	65,63%	38,84%	44,44%	42,28%	46,29%	$32,61\% - 59,96\%$
						41,45%	$38,43\% - 44,49\%$

Tabelle 3.8: Übersicht über die Angaben zum Wohnort (blaue Werte: Berechnung ohne WS 11/12)

Insgesamt gesehen lässt die Betrachtung den Schluss zu, dass mehr Studierende dieser Befragung bei ihren Eltern wohnen, als im eigenen Haushalt. Die Anzahl der Studierenden, die zum Studienbeginn im Sommersemester angeben im eignen Haushalt zu wohnen ist im Beobachtungszeitraum stets höher (65,63% und 44,44%), als bei denen, die zum Wintersemester ihr Studium beginnen (40,25%, 37,84% und 42,28%).

3.3.3 Abiturnote

Die Verteilung der Abiturdurchschnittsnote ist, bis auf wenige Ausnahmen, über den Beobachtungszeitraum sehr ähnlich (vgl. Abbildung 3.2). Anhand dieser Abbildung lässt sich eine *gauß-artige Verteilung* [Wikd] erkennen. Wie sich anhand Tabelle 3.9 ablesen lässt, ist mit knapp einem Drittel der Notenbereich *2,6 - 3,0* der am häufigsten angegebene. Lediglich im Wintersemester 2011/12 zeigt sich eine signifikante Abweichung nach unten. Weiterhin lässt sich beobachten, dass der Notenbereich von *1,0 - 2,0*, über den gesamten Beobachtungszeitraum gesehen, im Mittel mit 10,6 Prozent repräsentiert wird.

	SoSe 11*	WS 11/12*	SoSe 12	WS 12/13	SoSe 13	$\bar{x}$	C
1,0 - 1,5	4,88%	3,13%	1,65%	0%	1,63%	2,26%	$0\% - 4,54\%$
1,6 - 2,0	4,88%	9,38%	7,44%	11,11%	8,94%	8,35%	$5,44\% - 11,25\%$
2,1 - 2,5	20,73%	28,13%	22,31%	17,78%	28,46%	23,48%	$17,66\% - 29,30\%$
2,6 - 3,0	31,71%	25,00%	35,54%	33,33%	34,15%	31,95%	$26,83\% - 37,07\%$
3,1 - 3,5	28,05%	21,88%	28,93%	24,44%	22,76%	25,21%	$21,30\% - 29,12\%$
3,6 - 4,0	1,22%	0%	2,48%	2,22%	2,44%	1,67%	$0,35\% - 3,0\%$
> 4,0	0%	0%	0%	2,22%	0%	0,44%	$0\% - 1,68\%$
k.A.	8,54%	12,50%	1,65%	8,89%	1,63%	6,64%	$0,65\% - 12,63\%$

Tabelle 3.9: Übersicht über die Angaben zur Abiturnote

Die mit einem * gekennzeichneten Semester in Tabelle 3.9 hatten in ihrem Fragebogen eine andere Aufteilung der Notenbereiche. Statt der Intervalle *1,0 - 1,5* und *1,6 - 2,0* existiert dort die Aufteilung in die Bereiche *1,0 - 1,3*, *1,4 - 1,7* und *1,8 - 2,0*. Für diese Auswertung wurde folgende Ersetzung gewählt:

- $1,0 - 1,3 \rightarrow 1,0 - 1,5$
- $1,4 - 1,7 \rightarrow k.A.$
- $1,8 - 2,0 \rightarrow 1,6 - 2,0$

Durch diese Anpassung ist es möglich die Ergebnisse, der Befragung aus den Semestern SoSe 2011 und WiSe 2011/12, den restlichen der Befragungen gegenüberzustellen. Da der Notenbereich *1,4 - 1,6* nicht eindeutig zuzuordnen ist, werden diese der Antwort *k.A.* (keine Angabe) zugeteilt.

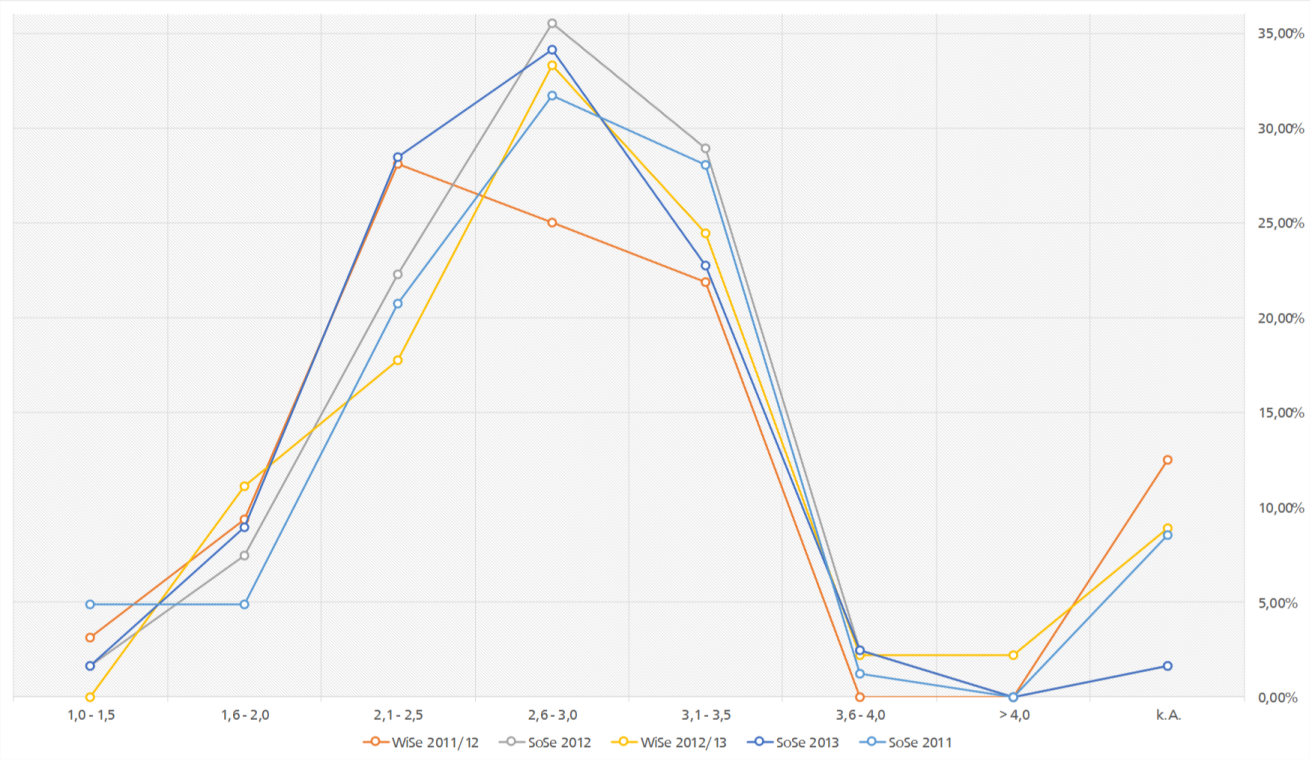


Abbildung 3.2: Angaben zur Abiturnote

3.3.4 Mathematiknote

Naben der Abiturdurchschnittsnote wird im Fragebogen auch nach der letzten Note im Fach Mathematik gefragt. Als Notenskala dient die in gymnasialen Oberstufe verwendete Punkteskala [BD13]. Diese lässt sich jedoch leicht auf die *6-Noten-Skala* übertragen:

$$\{15, 14, 13\} \rightarrow \text{Note 1}, \{12, 11, 10\} \rightarrow \text{Note 2}, \{9, 8, 7\} \rightarrow \text{Note 3}, \{6, 5, 4\} \rightarrow \text{Note 4}, \{3, 2, 1\} \rightarrow \text{Note 5}, \{0\} \rightarrow \text{Note 6}$$

Wie Tabelle 3.10 zeigt, gibt es insbesondere bei den Befragungen der Sommersemester 2011 und 2012 signifikante Abweichungen zu den übrigen Semestern. So sind im erstgenannten Semester auffällig viele Studierende mit einer Mathematiknote im Bereich der Note 1. Im Sommersemester 2012 fällt eine hohe Zahl an Studierenden mit einer Mathematiknote im Bereich der Note 3 auf. Wie auch Abbildung 3.3 verdeutlicht liegt, über alle Semester gesehen, die *letzte Mathematiknote* im Schnitt bei über 75 Prozent der Teilnehmer der Befragung im Bereich der Note 3 oder besser. Im Schnitt haben nur rund zwölf Prozent eine Note im Bereich *ausreichend*, *mangelhaft* oder *ungenügend*. Mit im Schnitt 30 bzw. 33 Prozent sind die Bereiche *12 - 10* und *9 - 7* die meistgenannten.

	SoSe 11	**WS 11/12**	**SoSe 12**	**WS 12/13**	**SoSe 13**	$\bar{x}$	C
15 - 13	23,17%	6,25%	8,26%	15,56%	15,45%	13,74%	$5,38\% - 22,10\%$
12 - 10	23,17%	37,50%	30,58%	24,44%	35,77%	30,29%	$22,27\% - 38,31\%$
9 - 7	29,27%	28,13%	41,32%	35,56%	31,71%	33,20%	$26,54\% - 39,85\%$
6 - 4	8,54%	3,13%	9,09%	4,44%	8,13%	6,67%	$3,32\% - 10,01\%$
3 - 1	1,22%	9,38%	4,96%	6,67%	0,81%	4,61%	$0,09\% - 9,13\%$
0	1,22%	0%	0%	0%	0%	0,24%	$0\% - 0,92\%$
k.A.	13,41%	15,63%	5,79%	13,33%	8,13%	11,26%	$6,15\% - 16,37\%$

Tabelle 3.10: Übersicht über die Angaben zur Mathematiknote

Im Vergleich zur Abiturnote (Abbildung 3.2) lässt sich in Abbildung 3.3 eine Verschiebung der Kurven nach links, hin zu einer besseren Mathematiknote, erkennen. Während der größte Anteil, der Teilnehmer dieser Befragung, eine befriedigende Abiturnote hat, hat die Mehrzahl bei der Mathematiknote eine gute (12-10 Punkte) bis sehr gute (15-13 Punkte) Note.

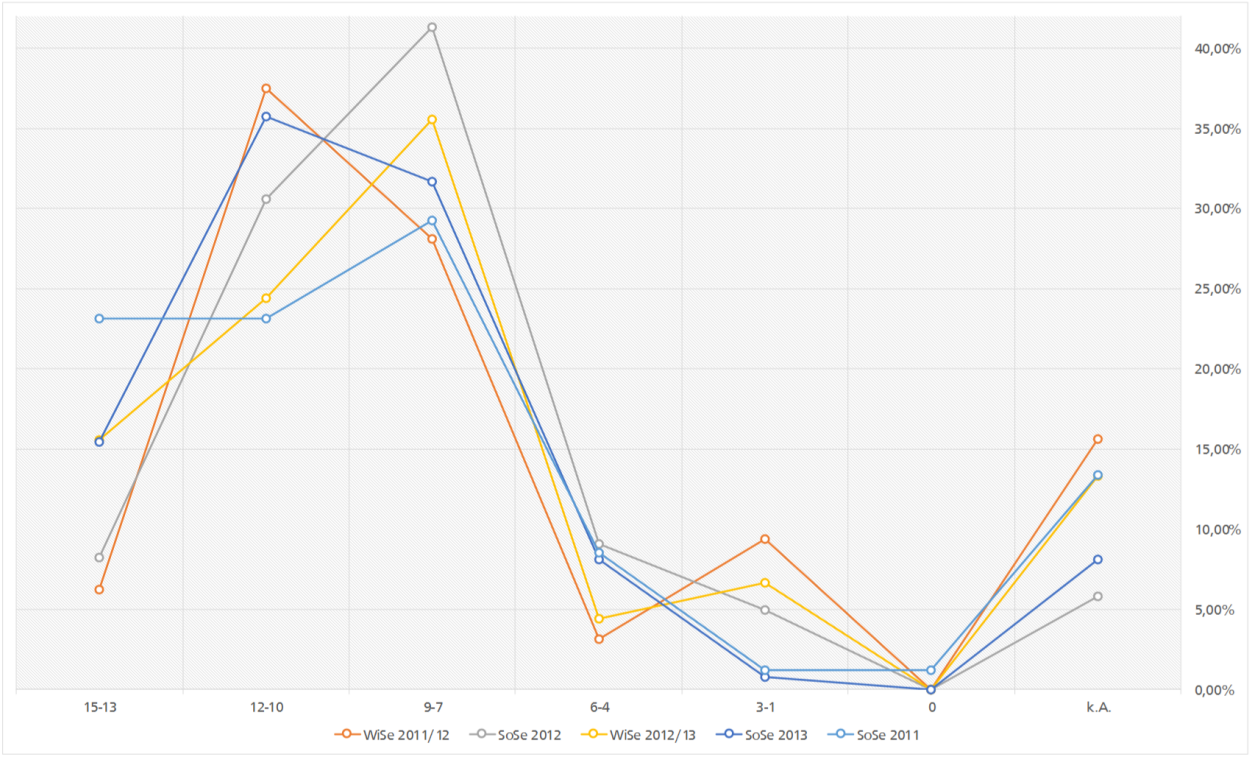


Abbildung 3.3: Angaben zur Mathematiknote

3.3.5 Tätigkeit vor dem Studium

Bei der Frage nach der Tätigkeit vor Beginn des Informatikstudiums lassen sich klare Unterschiede in der Beantwortung der Fragen zwischen Sommer- und Wintersemester feststellen (vgl. Abbildung 3.4). In Tabelle 3.11 zeigen sich zudem bei den Antwortmöglichkeiten *Zivildienst*, *anderes Studium*, *nicht bestimmtes* und *k.A.* (keine Angabe) auffällig breite Konfidenzintervalle, was auch mit den beiden Startzeitpunkten in das Studium zu erklären ist.

Die Antwortmöglichkeit *Jobben* wird von den Studienanfängern zum Sommersemester deutlich häufiger (36,84%, 41,67%) genannt als bei denen zum Wintersemester (27,06%, 28,10%, 20,16%). Noch deutlicher wird die Differenz bei den Studierenden, die ein anderes Studium begonnen haben. Im Schnitt waren 10,87% der Starter im Wintersemester vorher bereits in einem anderen Studium eingeschrieben. Zum Sommersemester sind dies im Schnitt 30,38%. Das Studienkolleg für ausländische Studierende besuchen ebenfalls vor dem Sommersemesterstart mehr als vor dem Wintersemesterstart (im Schnitt 10,48% zu 3,98%). Umgekehrt ist die Tendenz bei der Antwort *nichts Bestimmtes*. Diese Antwort wird von den Studienbeginnern im Wintersemester häufiger genannt als von denen im Sommersemester.

	SoSe 11	WS 11/12	SoSe 12	WS 12/13	SoSe 13	$\bar{x}$	C
Bund	9,41%	10,53%	3,31%	12,50%	2,33%	7,62%	$1,99\% - 13,24\%$
Zivildienst	11,76%	28,95%	7,44%	10,42%	2,33%	12,18%	$0\% - 24,66\%$
Freiw. soziales Jahr	2,35%	2,63%	4,13%	8,33%	1,55%	3,80%	$0,45\% - 7,15\%$
Reisen	7,06%	7,89%	8,26%	4,17%	1,55%	5,79%	$2,23\% - 9,34\%$
Ausbildung	10,59%	10,53%	10,74%	10,42%	9,30%	10,32%	$9,60\% - 11,04\%$
Jobben	27,06%	36,84%	28,10%	41,67%	20,16%	30,77%	$20,21\% - 41,32\%$
anderes Studium	9,41%	31,58%	11,57%	29,17%	11,63%	18,67%	$5,32\% - 32,03\%$
nichts Bestimmtes	23,53%	2,63%	23,97%	12,50%	36,43%	19,81%	$3,91\% - 35,71\%$
Studienkolleg	4,71%	10,53%	4,13%	10,42%	3,10%	6,58%	$2,10\% - 11,05\%$
Bundesfreiwilligendienst	-	-	-	2,08%	1,55%	1,82%	$1,35\% - 2,28\%$
k.A.	14,12%	2,63%	23,14%	10,42%	26,36%	15,33%	$3,40\% - 27,27\%$

Tabelle 3.11: Angaben zur Tätigkeit vor Beginn des Informatikstudiums (Mehrfachnennungen möglich)

Die Antwort Bundesfreiwilligendienst wurde erst im Fragebogen des Wintersemesters 2012/13 aufgenommen. Die Wehrpflicht (Antwort: Bund) und der Zivildienst wurden zum 1. Juli 2011 von der Bundesregierung ausgesetzt [ste13].

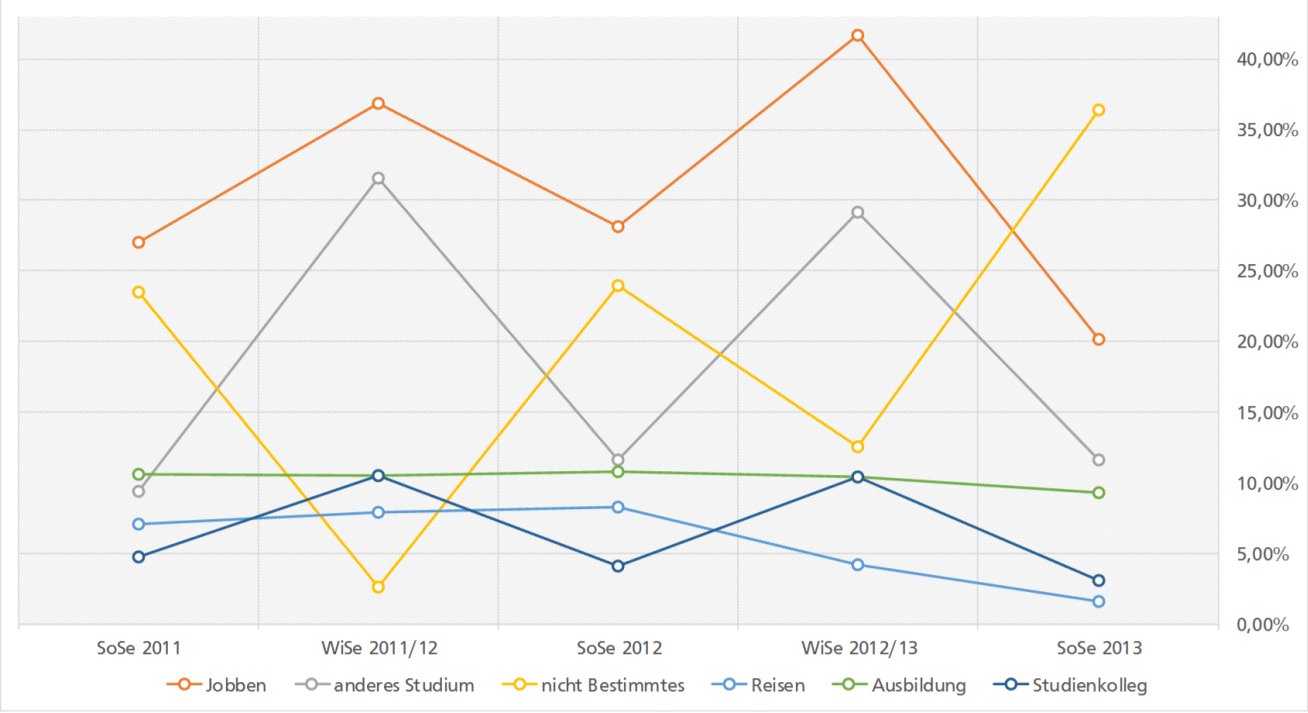


Abbildung 3.4: Tätigkeit vor Beginn des Studiums (ausgewählte Antworten)

3.3.6 Bestandene Prüfungen

Die Antwortmöglichkeiten zur Frage nach den bestandenen Prüfungen unterscheiden sich von Winter- zu Sommersemester. Für Studienbeginner zum Wintersemester sieht der empfohlene Studienplan drei Fachprüfungen vor: *Grundlagen der Informatik 1* (GdI 1), *Technische Grundlagen der Informatik* (TGdI) und *Mathematik 1* (Mathe 1). Für Studieneinsteiger zum Sommersemester sind vier vorgesehen: *Grundlagen der Informatik 1*, *Formale Grundlagen der Informatik 1 und 2* (FGdI 1+2) und *Einführung in Human Comuter Systems* (HCS).

Tabelle 3.12 und Abbildung 3.5 zeigen die Ergebnisse der Befragung für Studienanfänger zum Wintersemester. Sehr konstant zeigt sich hierbei die Antwort *keine bestandene Prüfungen*, da die Breite des Konfidenzintervalls bei lediglich 2,52% liegt. Im Mittel liegt die Zahl der Studierenden, die keine Prüfung in ihrem ersten Semester bestanden haben, in den Wintersemestern, bei 70,98 Prozent. Insgesamt deutet dies auf eine geringe Variabilität der Grundgesamtheit hin. Wenn eine Prüfung bestanden wird, ist es in den meisten Fällen die zu *GdI 1* mit im Mittel 17,56%.

	SoSe 11	**SoSe 12**	**SoSe 13**	$\bar{x}$	C
GdI 1	15,85%	15,70%	21,14%	17,56%	$13,72\% - 21,41\%$
TGdI	8,54%	11,57%	2,44%	7,52%	$1,74\% - 13,29\%$
Mathe 1	3,66%	1,65%	1,63%	2,31%	$0,87\% - 3,76\%$
sonstige	0%	0%	4,88%	1,63%	$0\% - 5,13\%$
keine	71,95%	71,07%	69,92%	70,98%	$69,72\% - 72,24\%$

Tabelle 3.12: Bestandene Prüfungen im Wintersemester (Befragung im Sommersemester)

Die Häufigkeit von bestanden Prüfungen zu *Mathe 1* oder in *sonstigen* Fächern ist bei Studienanfängern im Wintersemester sehr gering (2,31% und 1,63% im Durchschnitt). Bei TGdI ist bei der Befragung im Sommersemester 2013, im Vergleich zu den beiden vorhergehenden Befragungen, mit 2,44% ein vergleichsweise niedriger Wert festzustellen.

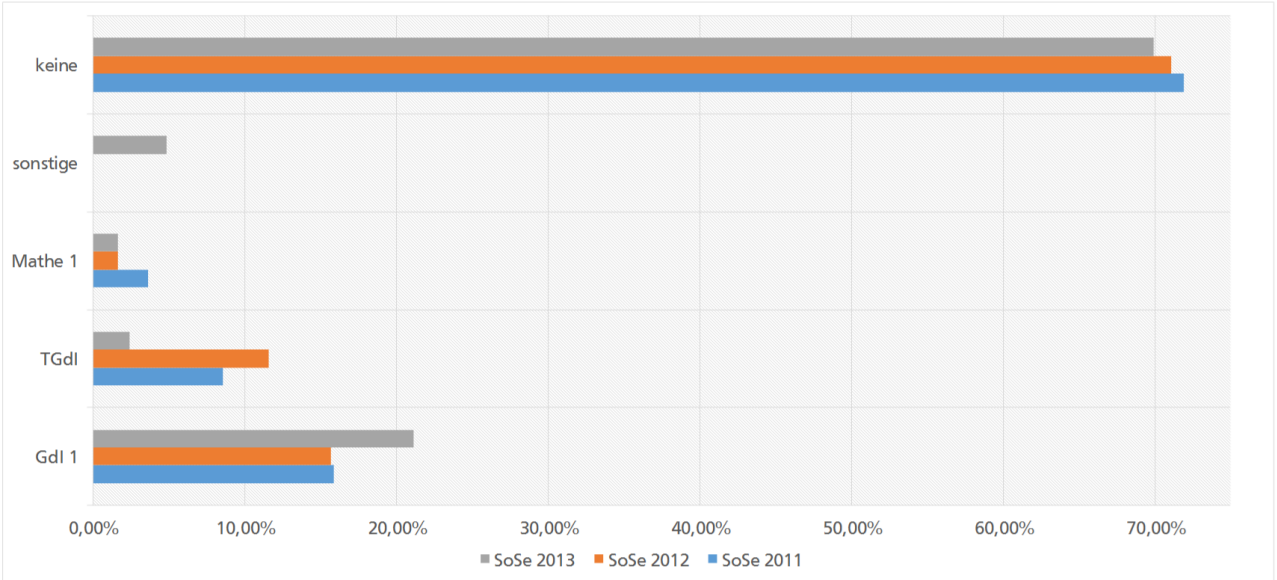


Abbildung 3.5: Bestandene Prüfungen im Wintersemester (Befragung im Sommersemester)

Wie bereits oben beschrieben sind bei den Studienanfängern zum Sommersemester andere Prüfungen im ersten Semester vorgesehen. Tabelle 3.13 und Tabelle 3.3.6 zeigen die Ergebnisse der Befragungen. Auffällig ist hier ein hoher Wert bei der Befragung im Wintersemester 2011/12 bei den *sonstigen* Prüfungen. Dieser hohe Wert erklärt sich durch die dort angebotene Veranstaltung *Einführung in wissenschaftliches Arbeiten*[3], die auf Empfehlung der Mentoren von vielen Erstsemestern belegt wurde. Die Prüfung zu *FGdI 2* wird von keinem der Befragten bestanden. Nur jeder Zehnte hat an dieser Prüfung teilgenommen.

[3] `http://www.ra.informatik.tu-darmstadt.de/lehre/eiwa/`

	WS 11/12	WS 12/13	$\bar{x}$	C
GdI 1	15,63%	13,33%	14,48%	$12,47\% - 16,49\%$
FGdI 1	3,13%	6,67%	4,9%	$1,79\% - 8,01\%$
FGdI 2	0%	0%	0%	-
HCS	9,38%	8,89%	9,13%	$8,71\% - 9,56\%$
sonstige	18,75%	0%	9,38%	$0\% - 25,84\%$
keine	53,13%	71,11%	62,12%	$46,33\% - 77,91\%$

Tabelle 3.13: Bestandene Prüfungen im Sommersemester (Befragung im Wintersemester)

Die Fachprüfung zu GdI 1 bestehen Studienanfänger im Sommersemester mit 14,48% im Schnitt leicht weniger Studierende, als im Wintersemester (17,56%). Die Fachprüfung zu FGdI 1 bestehen im Schnitt 4,9% der Befragten. Die Bestehensquote in HCS liegt im Mittel bei 9,13%.

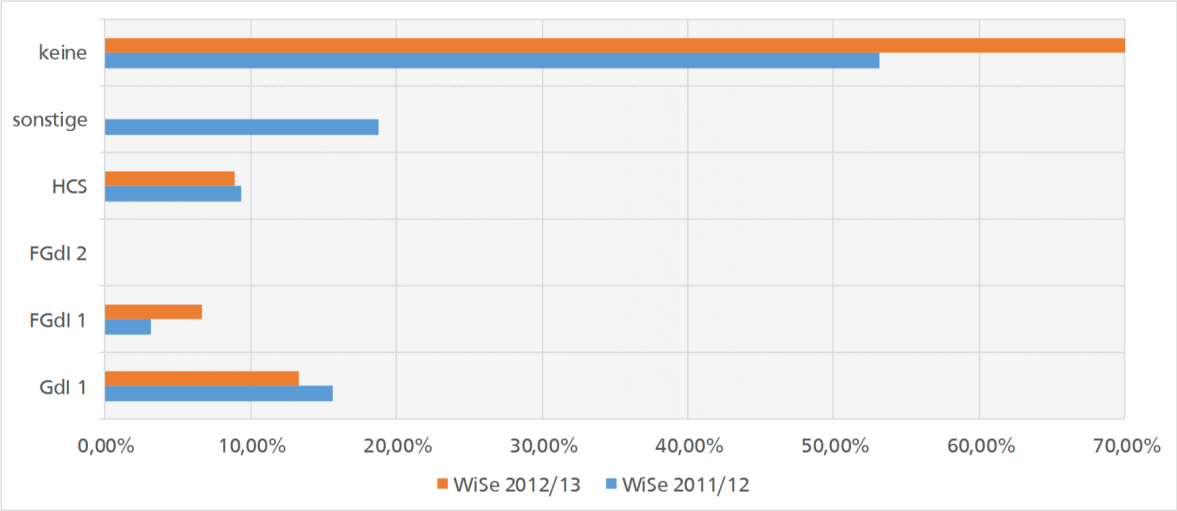


Abbildung 3.6: Bestandene Prüfungen im Sommersemester (Befragung im Wintersemester)

3.4 Erkenntnisse der Datenanalyse

Betrachtet man die Auswertung der Befragung zeigt sich, dass es sowohl Veränderungen in den Antworten über die Semester, als auch sehr konstante Verläufe zu beobachten sind. Zudem sind stellenweise auch Unterschiede zwischen den Befragungen im Sommer- und Wintersemester zu erkennen. Sehr konstante Verläufe lassen sich unter anderem bei den Abiturnoten der Teilnehmer erkennen. Der größte Anteil der Studierenden gibt einen Notenbereich von 2,6 bis 3,0 an. Sehr konstant ist ebenfalls die Angabe der Studierenden, die einen einfachen Anfahrtsweg 30 bis 60 Minuten angeben. Auch bei den bestanden Prüfungen sind Gemeinsamkeiten über den gesamten Befragungszeitraum erkennbar. So ist stets die Prüfung zum Fach GdI 1, die meist bestandene. Die Zahl der Befragten, die eine Ausbildung vor ihrem Studienbeginn absolviert haben, ist ebenfalls konstant.

Stellt man die Ergebnisse der Befragung der Studienanfänger von Winter- und Sommersemester gegenüber, lassen sich Unterschiede in der Häufigkeit einzelner Antworten erkennen. Studierende, die im Sommersemester ihr Studium aufnehmen geben häufiger an einen Anfahrtsweg von weniger als 15 Minuten zu haben, als ihre Kommilitonen, die im Winter beginnen. Ebenfalls ist die Zahl der Studierenden, die im eigenen Haushalt leben, bei denen im Sommersemester höher. Insgesamt wohnen jedoch mehr Befragte bei ihren Eltern, als im eigenen Haushalt. Große Differenzen gibt es bei der Angabe der Tätigkeiten vor dem Beginn des Studiums. Studienanfänger im Sommersemester haben zu durchschnittlich 30% zuvor ein anderes Studium begonnen. Bei denen zum Wintersemester sind es im Schnitt lediglich 11%. Ebenfalls höher zum Start des Studiums im Sommer ist stets die Angabe zuvor einen Job ausgeübt zu haben oder am Studienkolleg teilgenommen zu haben. Die Studienanfänger zum Wintersemester geben hingegen häufiger an *nichts Bestimmtes* getan zu haben.

4 Fragebogenkonverter

Der Fragebogen (vgl. Abschnitt 3.2), den die Studierenden mit geringen Prüfungserfolg über das Lernportal *moodle* des Fachbereichs Informatik der TU Darmstadt ausfüllen müssen, lässt sich im xls.-Format exportieren. Um, die aus den Fragebögen gewonnen Daten, mit Hilfe der Clusteranalyse auszuwerten (vgl. Kapitel 5) wird das Softwaretool WEKA (vgl. Abschnitt 2.3) benötigt. Dieses unterstützt als Eingabeformat nur Dateien vom Typ .arff. Neben der Clusteranalyse sind auch weitere Verfahren des *Maschinellen Lernens* auf diese Daten anwendbar. Aus diesem Grund ist das Tool *qTac* (questionnaire to arff converter) entwickelt worden. Es ermöglicht das Erstellen einer arff-Datei aus xls-Dateien. In diesem Kapitel wird kurz auf den Aufbau der Dateitypen eingegangen, sowie die Funktionalität des Programms *qTac* erläutert.

4.1 Aufbau der Fragebogentabelle

Für jedes Semester kann, mit Hilfe der Exportfunktion im Lernportal Informatik, die Befragung als Microsoft-Excel-Datei (.xls) [Ren08] exportiert werden. In der ersten Zeile der Tabelle ist in jeder Spalte eine Frage (Attribut) des Fragebogens aufgeführt. In jeder weiteren Zeile ist jeweils eine Abgabe des Fragebogens (Instanz) gelistet. Abbildung 4.1 zeigt einen Ausschnitt aus einer solchen Excel-Datei. Alle xls-Dateien mit diesem Aufbau können von dem hier vorstellten Fragebogenkonverter (*qTac*) eingelesen werden. Ein abweichender Aufbau wird nicht unterstützt.

	A	B	C
1	**Mein Anfahrtsweg zur Uni dauert einfach:**	**Ich wohne während des Semesters:**	**In welchem Bereich liegt Ihre Abiturnote?**
2	bis zu 60 min	bei meinen Eltern	2,6-3,0
3	weniger als 15 min	im eigenen Haushalt (WG, allein, etc.)	2,6-3,0
4	bis zu 30 min	im eigenen Haushalt (WG, allein, etc.)	k.A.
5	weniger als 15 min	im eigenen Haushalt (WG, allein, etc.)	2,6-3,0
6	länger als 60 min	bei meinen Eltern	3,1-3,5
7	bis zu 60 min	bei meinen Eltern	2,1-2,5
8	länger als 60 min	bei meinen Eltern	1,6-2,0
9	bis zu 30 min	bei meinen Eltern	3,1-3,5
10	länger als 60 min	bei meinen Eltern	2,1-2,5
11	länger als 60 min	bei meinen Eltern	k.A.
12	weniger als 15 min	im eigenen Haushalt (WG, allein, etc.)	2,6-3,0
13	bis zu 30 min	im eigenen Haushalt (WG, allein, etc.)	1,6-2,0
14	weniger als 15 min	im eigenen Haushalt (WG, allein, etc.)	2,1-2,5
15	bis zu 30 min	im eigenen Haushalt (WG, allein, etc.)	2,1-2,5
16	bis zu 60 min	im eigenen Haushalt (WG, allein, etc.)	2,1-2,5
17	länger als 60 min	bei meinen Eltern	3,6-4,0
18	bis zu 60 min	bei meinen Eltern	2,6-3,0

Abbildung 4.1: Aufbau des Fragebogenexports in Excel

4.2 Aufbau der ARFF-Datei

Das Softwaretool WEKA benötigt als Eingabe Dateien im ARFF [WEK] Datenformat. Diese Dateien bestehen aus einem Dateikopf (Header), gefolgt von einem Datenfeld. Im Header sind der Name der Relation, eine Liste von Attributen und deren Datentypen aufgeführt. Jedes Attribut wird mit der Angabe *@ATTRIBUTE* eingeleitet. Danach folgt der Name des Attributs und dessen Datentyp. Erlaubte Datentypen sind:

- Numeric → reelle oder ganzzahlige Werte
- <nominal specification> → eine Liste möglicher Werte z.B.: {*häufig, selten, nie*}

- String → Text
- Date → Datumswert im ISO-8601-Format[1]: yyyy-MM-dd'T'HH:mm:ss

Ein Beispiel für einen Header in einer ARFF-Datei zeigt Abbildung 4.2. Darauf zu sehen sind Attribute mit den Datentypen *Nominal* und *String*.

```
1  @RELATION fragebogen
2
3  @ATTRIBUTE Ich_wohne_während_des_Semesters {imeigenenHaushalt(WG.allein.etc.),beimeinenEltern}
4  @ATTRIBUTE In_welchem_Bereich_liegt_Ihre_Abiturnote {2.6-3.0,1.0-1.5,1.6-2.0,k.A.,3.6-4.0,2.1-2.5,3.1-3.5}
5  @ATTRIBUTE In_welchem_Bereich_liegt_Ihre_letzte_Mathematiknote {12-10,k.A.,15-13,9-7,3-1,6-4}
6  @ATTRIBUTE In_welchem_Bereich_liegt_ihre_letzte_Informatiknote_falls_vorhanden {12-10,k.A.,15-13,9-7,6-4}
7  @ATTRIBUTE Was_haben_Sie_vor_Beginn_des_Informatikstudiums_gemacht_Mehrfachnennungen_möglich String
```

Abbildung 4.2: Ausschnitt aus einem Header einer ARFF-Datei

Das Datenfeld besteht aus den auszuwertenden Daten, den Instanzen. Es wird einmalig mit der Angabe *@DATA* eingeleitet. Jede Zeile stellt genau Instanz dar. Jeder darin aufgeführte Wert muss zuvor im Header deklariert werden und zum dort angegebenen Datentyp passen. Abbildung 4.3 zeigt exemplarisch einen Ausschnitt aus dem Datenteil einer ARFF-Datei.

```
56 @DATA
57 imeigenenHaushalt(WG.allein.etc.),2.6-3.0,9-7,k.A.,anderesStudium,Mathe1,Mathe1,keine,Mathe1,keine,
58 beimeinenEltern,2.6-3.0,15-13,k.A.,k.A.,TGdIMathe1,GdI1TGdIMathe1,keine,sonstige,Mathe1,keine,keine
59 imeigenenHaushalt(WG.allein.etc.),2.1-2.5,12-10,k.A.,k.A.,GdI1TGdIMathe1,GdI1TGdIMathe1,GdI1TGdI,Ma
60 imeigenenHaushalt(WG.allein.etc.),2.6-3.0,6-4,12-10,Jobben,TGdIMathe1,TGdI,TGdI,Mathe1,keine,keine,
61 beimeinenEltern,3.6-4.0,15-13,k.A.,k.A.,GdI1TGdIMathe1,GdI1,TGdIMathe1,keine,keine,keine,keine,gut,
62 imeigenenHaushalt(WG.allein.etc.),3.1-3.5,9-7,6-4,nichtsBestimmtes,keine,GdI1TGdI,keine,GdI1TGdI,ke
```

Abbildung 4.3: Ausschnitt aus dem Datenteil einer ARFF-Datei

4.3 Programmfunktionalität

Das im Rahmen dieser Arbeit erstellte Programm *qTac* dient zur Konvertierung der Fragebogentabelle (vgl. Abschnitt 4.1) im xls-Format in das Dateiformat ARFF (vgl. Abschnitt 4.2). Es ist in der Programmiersprache Java[2] geschrieben. Die Konvertierung ist notwendig, um spätere Clusteranalysen mit dem *Data-Mining-Tool* WEKA (vgl. Abschnitt 2.3) vorzunehmen. Abbildung 4.4 zeigt schematisch den Programmablauf des Konverters.

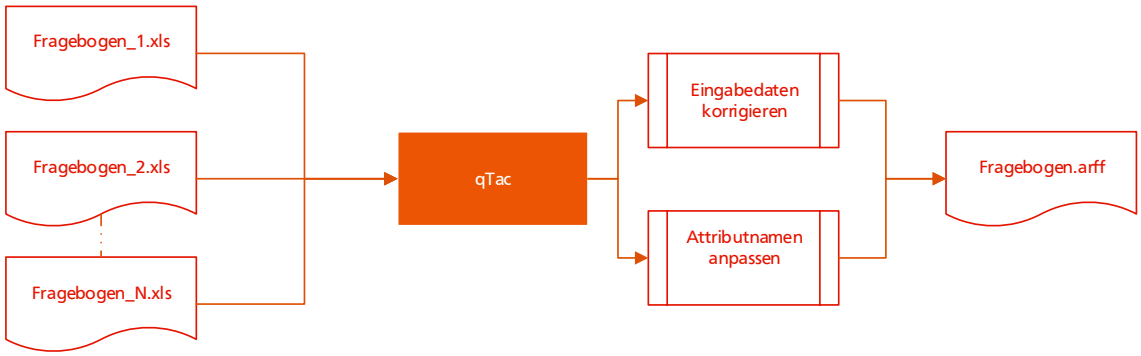


Abbildung 4.4: Ablaufdiagramm: Von der Eingabe zur Ausgabe in qTac

[1] http://www.iso.org/iso/home/standards/iso8601.htm
[2] http://www.java.com

Der Konverter *qTac* ermöglicht es mehrere Fragebogentabellen einzulesen und daraus eine ARFF-Datei zu erstellen. Das Einlesen erfolgt mit Hilfe der *Java Excel API* [JL]. Des Weiteren gibt das Programm Warnungen aus, falls sich die möglichen Attributswerte beim Einlesen mehrerer Tabellen nicht decken. Wie auf Abbildung 4.5 zu sehen, hat der Anwender die Möglichkeit Korrekturvorschläge einzugeben. Die Werte werden dann automatisch korrigiert. Zum Beispiel die Notenskala bei der Frage nach der Abiturnote. In den Fragebögen vor dem Sommersemester 2012 ist diese anders aufgeteilt (vgl. Abschnitt 3.2). Diese Skala kann nun nachträglich auf die *neue* Skala übertragen werden.

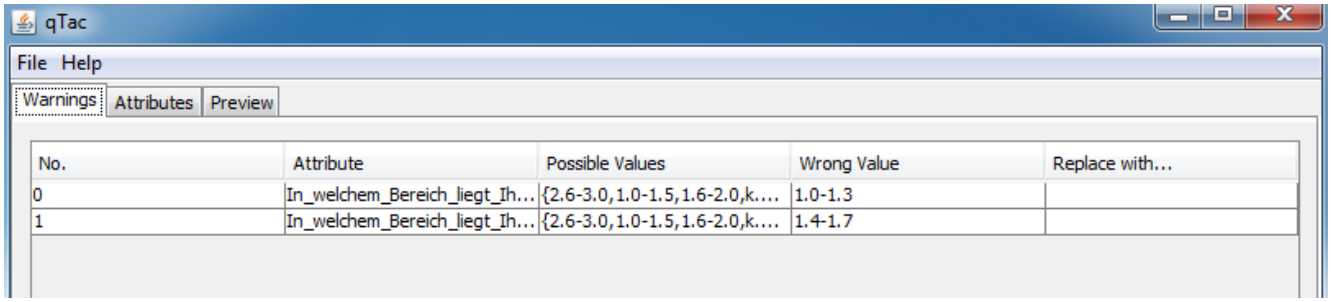


Abbildung 4.5: Automatisch erkannte nicht zuordenbare Attributeswerte

Der Konverter legt zudem automatisch die Attributsnamen und zugehörigen Datentypen fest. Die Attributsnamen kann der Nutzer nachträglich manuell anpassen. Abbildung 4.6 zeigt die Darstellung der Attribute und Datentypen im Konvertierungstool.

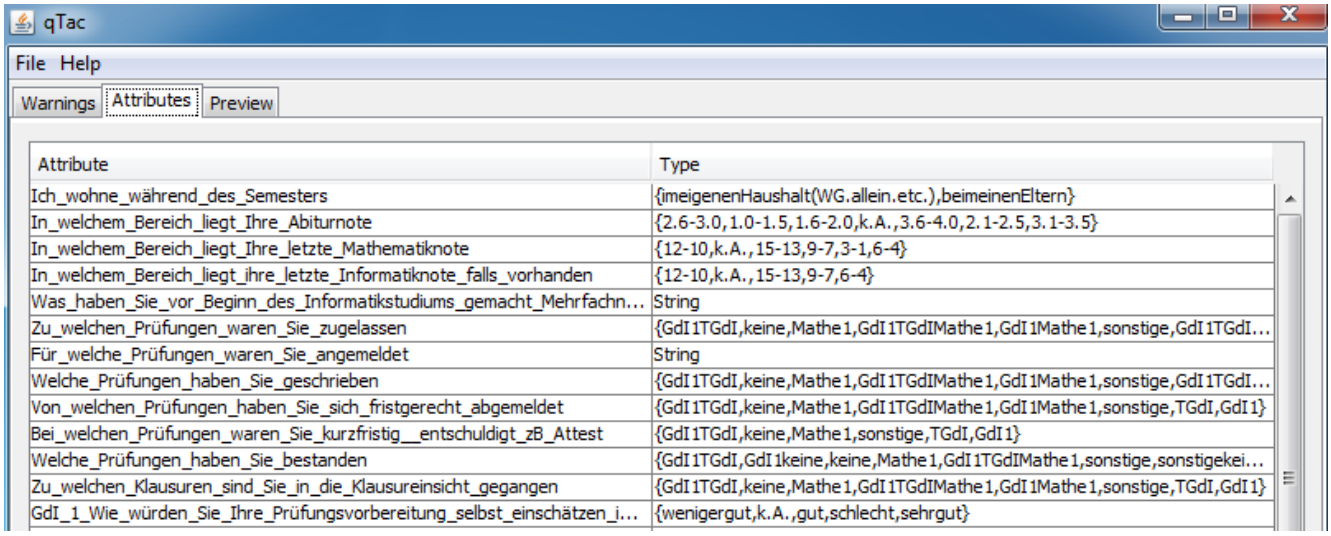


Abbildung 4.6: Anzeige der gefundenen Attribute zum zugehörigen Datentypen

5 Clusteranalyse

Im Kapitel Datenanalyse (vgl. Kapitel 3) wurde gezeigt, dass über den Zeitraum, in dem die Befragung von Studierenden mit geringen Prüfungserfolg im ersten Semester durchgeführt wurde, sind sowohl konstante Ergebnisse, als auch Unterschiede zu erkennen. Insbesondere fallen diese Unterschiede beim Vergleich zwischen Studienanfängern zum Winter- und Sommersemester (vgl. Abschnitt 3.3) auf. In diesem Kapitel wird, mit Hilfe der aus Clusteranalysen (vgl. Abschnitt 2.2) gewonnenen Daten, eine Identifizierung von Klassen von Studierenden mit geringem Prüfungserfolg erläutert. Für die Analysen wurde der K-Means-Algorithmus verwendet, da sich dieser, wie in Unterabschnitt 2.2.2 beschrieben, für diese Fragestellung eignet und insbesondere, durch die Ermittlung von Repräsentanten, Beschreibungen und Interpretationen der Cluster ermöglicht.

5.1 Vorgehensweise

Die nachfolgend beschriebenen Ergebnisse der Clusteranalyse wurden durch exploratives Vorgehen (vgl. Unterabschnitt 2.2.1) ermittelt. Die Anzahl der Cluster war im Vorfeld der Analyse nicht bekannt. Auch die Zahl der Attribute variierte. Nach diversen Versuchen mit unterschiedlicher Anzahl von Clustern wurde diese auf vier festgelegt. Ab dieser Aufteilung ist es in diesem Fall möglich die Ergebnisse zu interpretieren. Diese Zahl ist jedoch nicht als fixiert anzusehen, sondern dient als Startwert für eine erste Analyse. Bei weiteren Analysen kann die Zahl auf eine größere Clusterzahl erweitert werden, um auf ausgewählte Fragestellungen gezielt eingehen zu können. Bei weniger als vier Clustern entsteht das Problem, dass der Algorithmus bei vielen Attributen in den Clustern die gleichen Repräsentanten auswählt.

Die Beschreibung und Interpretation der Ergebnisse, die Anhand der gefunden Repräsentanten in den Clustern vorgenommen wird, unterliegt folgenden Leitfragen:

- Gibt es, trotz unterschiedlicher Prüfungen und anderer Tätigkeiten vor Beginn des Studiums, ähnliche Gruppen im Winter- und Sommersemester?
- In welchen Clustern sind *Problemfälle* zu finden und was charakterisiert diese?
- Gibt es Unterschiede zwischen Studierenden mit einer bzw. keiner bestandenen Prüfung?
- Welche Attribute sind relevant, welche weniger?

In den nachfolgenden Abschnitten werden zunächst die Ergebnisse der Clusteranalyse, unter Berücksichtigung aller Attribute, für die Befragungen im Sommer- (Unterabschnitt 5.2.1) und Wintersemester (Unterabschnitt 5.2.2) dargelegt. Gemeinsamkeiten und Unterschiede werden in Unterabschnitt 5.2.3 beschrieben. Es wird hierbei insbesondere auf die Frage nach den Problemfällen eingegangen.

In Abschnitt 5.3 werden Ergebnisse einer Clusteranalyse im Sommer- und Wintersemester gezeigt. Jedoch werden nur noch ausgewählte Attribute berücksichtigt, um die Gruppierungen eindeutiger beschreiben zu können. Abschließend werden in Abschnitt 5.4 die Resultate einer Clusteranalyse, basierend auf einer Klasseneinteilung der Fachstudienberatung, erläutert.

5.2 Analyse mit allen Attributen

Um einen ersten Überblick zu erhalten wird in diesem Abschnitt das Ergebnis der Clusteranalyse mit allen Attributen vorgestellt. Im Fragebogen, der im Sommersemester gestellt wird sind dies 45. Aus Gründen der besseren Lesbarkeit sind die Ergebnisse in mehreren Tabellen dargestellt. Zudem sind die Attribute (Fragen) abgekürzt angegeben. Die vollständige Fragestellung ist in Abschnitt 3.2 nachzulesen.

5.2.1 Sommersemesterbefragung

Im ersten Schritt der Clusteranalyse dieser Arbeit werden die ausgefüllten Fragebögen des Sommersemesters[1] analysiert. Hierfür liegen 246 Instanzen (auswertbare Fragebögen) vor. 45 Attribute (Fragen) werden für diese erste Auswertung

[1] Die Daten beziehen sich auf die Resultate der Studienanfänger zum Wintersemester.

berücksichtigt. Wie oben bereits beschrieben werden vier Cluster gewählt. *Cluster #0* werden 82 (33%), *Cluster #1* 54 (22%), *Cluster #2* 53 (22%) und *Cluster #3* 57 (23%) der Instanzen durch den K-Means-Algorithmus zugeordnet. Zur besseren Lesbarkeit werden die Ergebnisse dieser Analyse in mehreren Tabellen dargestellt.

Tabelle 5.1 zeigt die Repräsentanten der einzelnen Cluster der Fragebogenkategorie *allgemeine Angaben*. Hier zeigt sich, dass in drei Clustern (#0,#1 und #3) die Zentren der Attribute *Anfahrtsweg* und *Ich wohne* jeweils gleich gewählt wurden. Dem Anfahrtsweg wird in diesen Clustern das Zentrum *länger als 60min* zugeordnet. Dem Wohnort bei diesen Clustern wird der Wert *bei meinen Eltern* zugeteilt. *Cluster #2* erhält als einziges den Wert *im eigenen Haushalt* und *bis zu 60 Minuten* als Zentrum des jeweiligen Attributs zugeteilt. Zudem hat dieses Cluster den besten Abiturschnitt im Vergleich mit den anderen.

Attribute	**Cluster #0**	**Cluster #1**	**Cluster #2**	**Cluster #3**
Anfahrtsweg	länger als 60min	länger als 60min	bis zu 60min	länger als 60min
Ich wohne	bei meinen Eltern	bei meinen Eltern	im eigenen Haushalt	bei meinen Elterm
Abiturnote	2,6-3,0	3,1-3,5	2,1-2,5	2,6-3,0
Mathematiknote	12-10	9-7	9-7	9-7
Informatiknote	k.A.	k.A.	15-13	k.A

Tabelle 5.1: Ergebnis der Clusteranalyse mit K-Means (4 Cluster); Kategorie *allgemeine Angaben*; Fragebögen der Sommersemester

In Tabelle 5.2 sind Ergebnisse der Clusteranalyse des Fragebogenkomplexes *Angaben zur Prüfungssituation* dargestellt. Bei den Attributen *Anmeldung*, *Entschuldigt*, *Bestanden* und *Klausureinsicht* ergab die Durchführung der Clusteranalyse, mit dem Wert *keine*, jeweils den gleichen Repräsentanten. In *Cluster #3* wurde als Zentrum des Attributs *Zulassung TGdI, Mathe1* ermittelt. In diesem Cluster findet sich demnach ein hoher Anteil von Studierenden, die keine Zulassung in GdI 1 erhalten haben. Auffällig ist auch, dass in den Clustern *#1* und *#3* im Attribut *Geschrieben* der Algorithmus den Repräsentanten *keine* gewählt hat. In *Cluster #3* beträgt die Quote der Studierenden ohne Prüfungserfahrung 56%. In *Cluster #1* sind dies noch 40%. In den Clustern *#0* und *#2* sind es hingegen nur 12% und 9%, die an keiner Prüfung teilgenommen haben.

Attribute	**Cluster #0**	**Cluster #1**	**Cluster #2**	**Cluster #3**
Zulassung	GdI1, TGdI, Mathe1	GdI1, TGdI, Mathe1	GdI1, TGdI, Mathe1	TGdI, Mathe1
Anmeldung	GdI1, TGdI, Mathe1	GdI1, TGdI, Mathe1	GdI1, TGdI, Mathe1	GdI1, TGdI, Mathe1
Geschrieben	GdI1, TGdI	keine	TGdI	keine
Abgemeldet	Mathe 1	keine	keine	keine
Entschuldigt	keine	keine	keine	keine
Bestanden	keine	keine	keine	keine
Klausureinsicht	keine	keine	keine	keine

Tabelle 5.2: Ergebnis der Clusteranalyse mit K-Means (4 Cluster); Kategorie *Angaben zur Prüfungssituation*; Fragebögen der Sommersemester

Um detaillierter auf die Fragestellung einzugehen, in welcher Gruppierung sich Studierende mit zumindest einer bestanden Prüfung befinden, kann Abbildung 5.1 herangezogen werden. Diese Abbildung zeigt die Verteilung der Studierenden auf die vier Cluster und ihre bestanden Prüfungen. Es ist zu erkennen, dass in jedem Cluster die Mehrzahl der Teilnehmer keine Prüfung bestanden hat. In *Cluster #1* ist die Quote derer ohne Prüfungserfolg mit 91% am höchsten, gefolgt von *Cluster #3* mit 84%. Studierende mit einer bestanden Prüfungen beobachtet man hauptsächlich in *Cluster #2*. Hier haben 43% eine Prüfung bestanden. In *Cluster #0* sind es noch 34%.

Die Clustermittelpunkte der Kategorie *Angaben zur Prüfungsvorbereitung* in GdI 1 finden sich in Tabelle 5.3. In den Clustern *#0* und *#2* findet sich beim Attribut *Prüfungsvorbereitung* auf die Prüfung in GdI 1 der Wert *gut* in den Clusterzentren. Diese Beobachtung passt wiederum zu den Resultaten, die Abbildung 5.1 zeigt. In der Abbildung ist zu erkennen, dass in diesen Clustern (*#0* und *#2*) deutlich mehr Studierende GdI 1 bestanden haben, als in den übrigen beiden.

Ein vergleichbares Bild zu den Angaben der Prüfungsvorbereitung zeigt im Prüfungsfach TGdI. Auch hier finden sich in den Clustern *#1* und *#3* nur jeweils ein Studierender, der diese Prüfung bestanden hat (vgl. Abbildung 5.1). Zwar

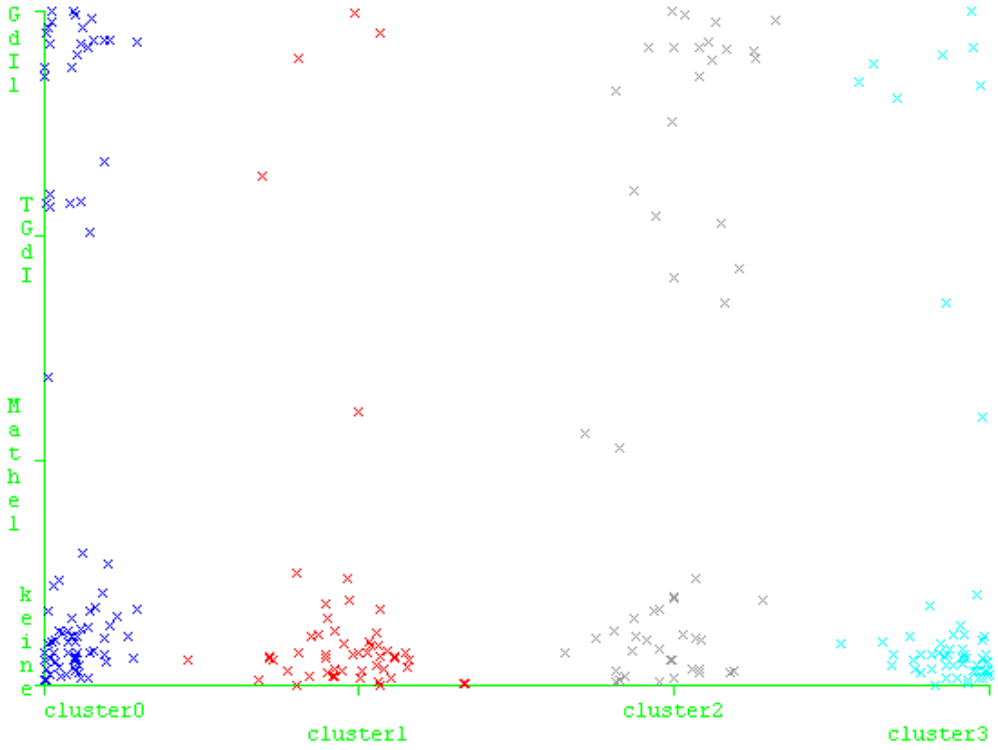


Abbildung 5.1: Visualisierung der bestandenen Prüfungen bei 4 Clustern und allen Attributen im Fragebogen des Sommersemesters

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Prüfungsvorbereitung	gut	weniger gut	gut	schlecht
Vorlesung	fast immer	häufig	fast immer	unregelmäßig
Gruppenübungen	unregelmäßig	unregelmäßig	fast immer	selten
Hausübungen	fast immer	fast immer	fast immer	fast immer

Tabelle 5.3: Ergebnis der Clusteranalyse mit K-Means (4 Cluster); Kategorie *Angaben zur Prüfungsvorbereitung in GdI 1*; Fragebögen der Sommersemester

liegt die Gesamtzahl an bestanden TGdI-Prüfung bei dieser Betrachtung sehr niedrig (vgl. Unterabschnitt 3.3.6), dennoch zeigt sich auch hier, dass eher Studierende der Cluster *#0* und *#2* diese Prüfung bestehen. Diese Beobachtung wird auch durch die gefunden Zentren dieser Fragekategorie gedeckt (vgl. Tabelle 5.4).

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Prüfungsvorbereitung	weniger gut	weniger gut	gut	schlecht
Vorlesung	fast immer	häufig	fast immer	häufig
Gruppenübungen	unregelmäßig	unregelmäßig	fast immer	selten
Hausübungen	unregelmäßig	unregelmäßig	häufig	unregelmäßig

Tabelle 5.4: Ergebnis der Clusteranalyse mit K-Means (4 Cluster); Kategorie *Angaben zur Prüfungsvorbereitung in TGdI*; Fragebögen der Sommersemester

Die Prüfung zur Vorlesung Mathe 1 wird in der Regel nur in seltenen Fällen von Studierenden mit keiner oder einer bestanden Prüfung im ersten Semester erfolgreich absolviert. Im Schnitt liegt hier die Quote bei lediglich 2,31%. Wie in Tabelle 5.5 zu sehen ist, liefert der Algorithmus, ausgenommen bei *Cluster #2*, *unregelmäßige* oder *seltene* Teilnahme an Vorlesungen, Gruppenübungen oder Hausübungen als Clusterrepräsentanten.

Tabelle 5.6 zeigt die Clusterzentren der Fragekategorie *Angaben zum Studienverhalten* der Sommersemesterfragebögen. *Cluster #2* liefert bei den Attributen *Lehrveranstaltungsbesuche* und *Selbststudium* Repräsentanten, die eine sinnvolle

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Prüfungsvorbereitung	k.A.	weniger gut	weniger gut	schlecht
Vorlesung	unregelmäßig	unregelmäßig	fast immer	selten
Gruppenübungen	unregelmäßig	unregelmäßig	fast immer	selten
Hausübungen	unregelmäßig	unregelmäßig	fast immer	selten

Tabelle 5.5: Ergebnis der Clusteranalyse mit K-Means (4 Cluster); Kategorie *Angaben zur Prüfungsvorbereitung in Mathe 1*; Fragebögen der Sommersemester

Aufwandseinschätzung für das erste Semester darstellen. Insbesondere *Cluster #3* zeigt bei den selben Attributen, den geringsten Zeitaufwand, als Zentrum. Ferner fällt bei diesem Cluster der Zentrumswert für das Attribut *Computerspiele* auf. Hier ist der höchstmögliche Zeitaufwand von im Schnitt ≥2 Stunden pro Tag angegeben.

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Lehrveranstaltungsbesuche	11-15h	5-10h	11-15h	< 5h
Selbststudium	5-10h	5-10h	11-15h	< 5h
Lerngruppe	ja	ja	ja	nein
Jobben	nein	ja	nein	nein
Stunden Job pro Woche	11-20h	11-20h	11-20h	11-20h
Hobbies	5-10h	5-10h	5-10h	5-10h
Freunde und Familie	5-10h	5-10h	11-15h	5-10h
Computerspiele	< 1/2h	< 1/2h	<1/2h	> 2h
Soziale Netzwerke	< 1/2h	< 1/2h	< 1/2h	< 1/2h
Fernsehen	< 1/2h	< 1/2h	< 1/2h	< 1/2h

Tabelle 5.6: Ergebnis der Clusteranalyse mit K-Means (4 Cluster); Kategorie *Angaben zum Studienverhalten*; Fragebögen der Sommersemester

Tabelle 5.7 zeigt die Clustermittelpunkte der Kategorie *Angaben zu Gründen des Prüfungsmisserfolgs*. Alle Cluster haben beim Attribut *Zu wenig Zeit investiert* den Wert *trifft voll zu* als Zentrum. Ebenfalls gleiche Werte liefert das Attribut *Weiterhin Informatik an der TU*. Die Auswertung der Fragebögen ergab, dass 232 von 246 Studierenden hier den Wert *ja* angegeben haben. In Tabelle 5.8 zeigt abschließend die Häufigkeit der Angabe von privaten Gründen im jeweiligen Cluster. *Cluster #2* hat hier mit 45,28% den größten Anteil.

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Lernstoff zu schwer	trifft weniger zu	trifft weniger zu	trifft weniger zu	trifft kaum zu
Lernstoff zu viel	trifft weniger zu	trifft voll zu	trifft weniger zu	trifft weniger zu
Lernstoff anders als erwartet	trifft weniger zu	trifft voll zu	trifft nicht zu	trifft nicht zu
Zu wenig Zeit investiert	trifft voll zu	trifft voll zu	trifft voll zu	trifft voll zu
Zu viel vorgenommen	trifft kaum zu	trifft weniger zu	trifft nicht zu	trifft kaum zu
Keinen Lernplan	trifft voll zu	trifft weniger zu	trifft nicht zu	trifft voll zu
Andere Klausurinhalte	trifft kaum zu	trifft weniger zu	trifft nicht zu	trifft nicht zu
Nervös in der Klausur	trifft voll zu	trifft weniger zu	trifft nicht zu	trifft nicht zu
LV nicht auf Klausur vorbereitet	trifft kaum zu	trifft weniger zu	trifft nicht zu	trifft nicht zu
Weiterhin Informatik an der TU	ja	ja	ja	ja

Tabelle 5.7: Ergebnis der Clusteranalyse mit K-Means (4 Cluster); Kategorie *Angaben zu Gründen des Prüfungsmisserfolgs*; Fragebögen der Sommersemester

Betrachtet man diese Ergebnisse vor dem Hintergrund *Problemfälle* zu identifizieren, sind diese am ehesten in *Cluster #1* oder *#3* zu finden. In beiden Gruppen finden sich viele Studierende, die keine Prüfungserfahrung vorweisen können. So erhält in diesen Clustern das Attribut *geschriebene Prüfungen* den Repräsentanten *keine*. Auch die Wahrscheinlichkeit wenigstens eine Prüfung zu bestehen ist in diesen Clustern am geringsten. Der zeitliche Aufwand für das Studium (investierte Zeit für Lehrveranstaltungsbesuche und Selbststudium pro Woche) ist in diesen Clustern mit niedrigen Werten

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
private Gründe	29,27%	33,33%	45,28%	26,32%

Tabelle 5.8: Häufigkeit im angegeben Cluster mit K-Means (4 Cluster); Kategorie *private Gründe*; Fragebögen der Sommersemester

belegt. Insbesondere in *Cluster #3*. Dort ist als Zeitaufwand, für diese beiden Attribute, jeweils der Wert $< 5h$ als Repräsentant gewählt. Eine Lerngruppe ist in *Cluster #3* in der Regel auch nicht vorhanden. Zudem wartet dieses Cluster noch mit dem höchstmöglichen Wert beim Attribut *Computerspielen* (mehr als zwei Stunden pro Tag) auf. Aufgrund dieser Ergebnisse liegt die Vermutung nahe, dass Studierende mit diesen Angaben im Fragebogen womöglich als *Problemfälle* einzustufen sind.

Die Cluster *#0* und *#2* zeigen bei einer Vielzahl von Attributen gleiche Clusterzentren auf. Die dort zugeteilten Studierenden konnten in der Regel zumindest Prüfungserfahrung sammeln, teilweise sogar Prüfungen bestehen. Daher sind diese am eher nicht als Problemfälle zu klassifizieren. Im nächsten Abschnitt sind die Ergebnisse der Clusteranalyse unter Berücksichtigung aller Attribute mit den Ergebnissen der Befragungen im Wintersemester dargelegt.

5.2.2 Wintersemesterbefragung

Die Befragungen im Wintersemester[2] ergaben 76 Instanzen. Für die Clusteranalyse stehen hierbei 38 Attribute zur Verfügung. Wie bei der Analyse der Befragung im Sommersemester (vgl. Unterabschnitt 5.2.1) ist eine Aufteilung in vier Cluster vorgenommen worden. Den Clustern #0 bis #3 wurden 18 (24%), 31 (41%), 10 (13%) und 17 (22%) der Studierenden zugeordnet.

Die Repräsentanten der vier Cluster bei den *allgemeinen Angaben*, die der Algorithmus ermittelt hat, sind in Tabelle 5.9 aufgeführt. Bei den Attributen *Anfahrtsweg* und *Abiturnote* sind in jedem Cluster voneinander verschiedene Werte als Mittelpunkte zu sehen. Beim Attribut *Ich wohne* ist nur *Cluster #2* mit dem Zentrumswert *bei meinen Eltern* versehen. Außerdem ist diesem Cluster der schlechteste Abiturschnitt (3,1 - 3,5) als Mittelwert zugewiesen.

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Anfahrtsweg	weniger als 15min	bis zu 60min	länger als 60min	bis zu 30min
Ich wohne	im eigenen Haushalt	im eigenen Haushalt	bei meinem Eltern	im eigenen Haushalt
Abiturnote	1,6-2,0	2,1-2,5	3,1-3,5	2,6-3,0
Mathematiknote	k.A.	12-10	9-7	9-7

Tabelle 5.9: Ergebnis der Clusteranalyse mit K-Means (4 Cluster); Kategorie *Allgemeine Angaben*; Fragebögen der Wintersemester

Bei der Clusteranalyse der Wintersemesterbefragungen steht in der Kategorie *Angaben zur Prüfungszulassung* nur das Attribut *Bestandene Prüfungen* zur Verfügung. Die weiteren Fragen dieser Kategorie sind für die Analyse nicht berücksichtigt worden. Grund dafür ist die Vielzahl an aufgetreten Antwortmöglichkeiten. Die so gefundenen Repräsentanten lassen daher keine Interpretationen zu. Das Ergebnis der Analyse zeigt, dass bei der Frage nach den bestandenen Prüfungen für alle Clusterzentren den Wert *keine* ermittelt wurde. Abbildung 5.2 stellt die Verteilung der Studierenden mit ihren bestandenen Prüfungen auf die vier Cluster dar. Wie auf dieser Abbildung zu sehen ist werden Studierende mit zumindest einer bestanden Prüfung, in ihrem ersten Semester, überwiegend *Cluster #1* zugeordnet. 61% beträgt die Quote von Teilnehmern mit einer bestanden Prüfung in diesem Cluster. In den übrigen drei Clustern sind mehrheitlich Studierende mit keiner bestanden Prüfung zu finden. Dies sind 88% in *Cluster #3*, 80% in *Cluster #2* und 72% in *Cluster #0*.

Die Ergebnisse der Clusteranalyse der Kategorie *Angaben zur Prüfungsvorbereitung in GdI 1* sind in Tabelle 5.10 dargestellt. Hier zeigt sich anhand der durch den Algorithmus ermittelten Clusterzentren, dass die Hausübungen in GdI 1 *häufig* oder *fast immer* gemacht werden. Das Attribut Prüfungsvorbereitung liefert in den Clustern *#0* und *#1* den Mittelwert *weniger gut*. In den anderen beiden den Wert *schlecht*.

[2] Beginn des Studiums zum Sommersemester

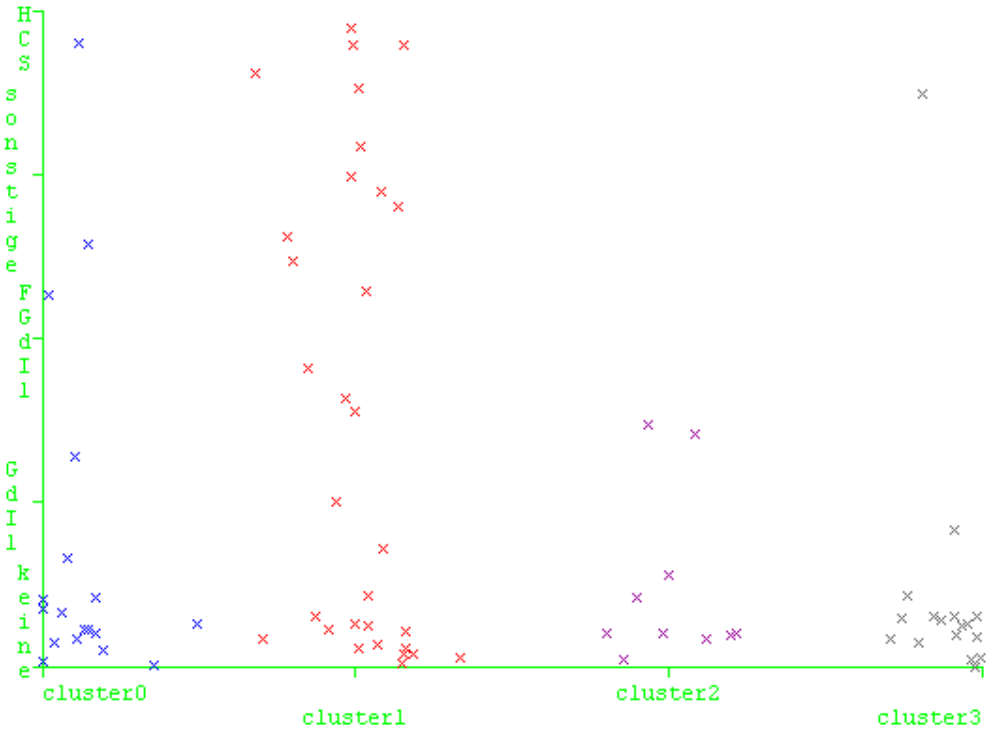


Abbildung 5.2: Visualisierung der bestandenen Prüfungen bei 4 Clustern und allen Attributen im Fragebogen des Wintersemesters

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Prüfungsvorbereitung	weniger gut	weniger gut	schlecht	schlecht
Vorlesung	unregelmäßig	fast immer	fast immer	fast immer
Gruppenübungen	selten	fast immer	unregelmäßig	häufig
Hausübungen	fast immer	fast immer	fast immer	häufig

Tabelle 5.10: Ergebnis der Clusteranalyse mit k-Means (4 Cluster); Kategorie *Angaben zur Prüfungsvorbereitung in GdI 1*; Fragebögen der Wintersemester

Die Clusteranalyse für die Kategorie *Angaben zur Prüfungsvorbereitung in FGdI 1* ergibt häufig negative Bewertungen als Clusterzentren (vgl. Tabelle 5.11). Beispielsweise liefert die Analyse beim Attribut *Prüfungsvorbereitung* bei den Clustern #0, #2 und *#3* den Wert *schlecht*.

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Prüfungsvorbereitung	schlecht	gut	schlecht	schlecht
Vorlesung	selten	fast immer	fast immer	häufig
Gruppenübungen	selten	fast immer	unregelmäßig	selten
Hausübungen	selten	fast immer	unregelmäßig	selten

Tabelle 5.11: Ergebnis der Clusteranalyse mit k-Means (4 Cluster); Kategorie *Angaben zur Prüfungsvorbereitung in FGdI 1*; Fragebögen der Wintersemester

Tabelle 5.12 zeigt die gewählten Zentren der Kategorie *Angaben zur Prüfungsvorbereitung in HCS*. Hier zeigt sich, dass die Hausübungen in diesem Fach *häufig* oder *fast immer* bearbeitet werden. Die *Prüfungsvorbereitung* wird mit den Clusterzentren *weniger gut* (*Cluster #0* bis *#2*) oder *gut* (*Cluster #3*) versehen. Die meisten Studierenden, die diese Prüfung bestehen finden sich, bis auf zwei Ausnahmen, in *Cluster #1*.

Bei den Angaben zum Studienverhalten (vgl. Tabelle 5.13) fällt auf, dass in allen Clustern das Zentrum für das Attribut *Lehrveranstaltungsbesuche* mit dem gleichen Wert belegt ist. Beim *Selbststudium* hingegen unterscheiden sich die Zentren.

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Prüfungsvorbereitung	weniger gut	weniger gut	weniger gut	gut
Vorlesung	unregelmäßig	fast immer	fast immer	fast immer
Gruppenübungen	selten	fast immer	k.A.	häufig
Hausübungen	fast immer	fast immer	fast immer	häufig

Tabelle 5.12: Ergebnis der Clusteranalyse mit k-Means (4 Cluster); Kategorie *Angaben zur Prüfungsvorbereitung in HCS*; Fragebögen der Wintersemester

Cluster #1 bekommt hierbei mit *11-15h* den größten Wert aller Cluster zugewiesen. Im Gegensatz dazu hat *Cluster #2* den niedrigsten Wert (*< 5h*). *Cluster #3* weist die höchsten Werte für den Zweitaufwand für *Hobbies* und *Freunde und Familie* auf.

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Lehrveranstaltungsbesuche	11-15h	11-15h	11-15h	11-15h
Selbststudium	5-10h	11-15h	< 5h	5-10h
Lerngruppe	nein	ja	nein	ja
Jobben	nein	nein	ja	ja
Stunden Job pro Woche	11-20h	11-20h	11-20h	11-20h
Hobbies	5-10h	5-10h	< 5h	11-15h
Freunde und Familie	5-10h	5-10h	< 5h	11-15h
Computerspiele	< 1/2h	< 1/2h	<1/2h	1 1/2-2h
Soziale Netzwerke	< 1/2h	< 1/2h	1/2-1h	< 1/2h
Fernsehen	< 1/2h	1/2-1h	1/2-1h	< 1/2h

Tabelle 5.13: Ergebnis der Clusteranalyse mit k-Means (4 Cluster); Kategorie *Angaben zum Studienverhalten*; Fragebögen der Wintersemester

Betrachtet man bei den, aus Sicht der Studierenden, *Angaben zu den Gründen des Prüfungsmisserfolgs* (vgl. Tabelle 5.14) das Attribut *Zu wenig Zeit investiert*, fällt auf, dass alle Cluster den Wert *triff voll zu* zugewiesen bekommen. Weniger entscheidend für den Misserfolg scheinen die Attribute *Lernstoff zu schwer*, *Zu viel vorgenommen*, *Lehrveranstaltung hat nicht auf die Prüfung vorbereitet* und *Lernstoff zu viel* zu sein. Hier finden sich lediglich die Werte *trifft weniger zu* oder *trifft nicht zu* in den Zentren der Cluster.

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Lernstoff zu schwer	trifft weniger zu	trifft weniger zu	trifft weniger zu	trifft nicht zu
Lernstoff zu viel	trifft weniger zu	trifft weniger zu	trifft weniger zu	trifft nicht zu
Lernstoff anders als erwartet	trifft voll zu	trifft voll zu	trifft voll zu	trifft nicht zu
Zu wenig Zeit investiert	trifft voll zu	trifft voll zu	trifft voll zu	trifft voll zu
Zu viel vorgenommen	trifft kaum zu	trifft nicht zu	trifft weniger zu	trifft nicht zu
Keinen Lernplan	trifft voll zu	trifft nicht zu	trifft voll zu	trifft nicht zu
Andere Klausurinhalte	trifft weniger zu	trifft nicht zu	trifft voll zu	trifft nicht zu
Nervös in der Klausur	trifft nicht zu	trifft nicht zu	trifft voll zu	trifft nicht zu
LV nicht auf Klausur vorbereitet	trifft nicht zu	trifft weniger zu	trifft nicht zu	trifft nicht zu

Tabelle 5.14: Ergebnis der Clusteranalyse mit k-Means (4 Cluster); Kategorie *Angaben zu Gründen des Prüfungsmisserfolgs*; Fragebögen der Wintersemester

Bei der letzten Kategorie des Fragebogens, der Frage nach privaten Gründen, werden in Tabelle 5.15 die Häufigkeit der Vorkommen von privaten Gründen in den einzelnen Clustern dargestellt. Hier zeigt sich, dass jeder zweite Studierende aus *Cluster #0* private Gründe für seine Schwierigkeiten im ersten Semester angibt.

Möchte man *Problemfälle* aufgrund dieser Analyse identifizieren, kann man Studierende aus *Cluster #1* am ehesten ausschließen. Das Zeitmanagement scheint hier zu stimmen (vgl. Tabelle 5.13). Die dort angegeben Zentren für die

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
private Gründe	50,0%	32,26%	30,0%	41,18%

Tabelle 5.15: Häufigkeit im angegeben Cluster mit k-Means (4 Cluster); Kategorie *private Gründe*; Fragebögen der Wintersemester

Attribute Selbststudium, Zeit für Hobbies und Zeit für Freunde und Familie liegen im durchschnittlichen Bereich. Das deckt sich auch mit den Zentren der Attribute für die Prüfungsvorbereitung. Bei den Attributen zu Vorlesungsbesuchen, Gruppenübungsbesuchen und Hausübungen wird stets der Wert *fast immer* als Clusterzentrum gewählt. Dies unterstützt die These, dass das Studienverhalten der Studierenden aus *Cluster #1* angemessen ist. *Problemfälle* verteilen sich daher eher auf die übrigen Cluster. Außerdem lässt die Betrachtung aller Cluster die Vermutung zu, dass für den geringen Prüfungserfolg im ersten Semester die zu wenig investierte Zeit ein entscheidender Faktor ist.

5.2.3 Vergleich Sommer- und Wintersemester

Vergleicht man die Resultate der vorangegangen Clusteranalysen von Sommer- und Wintersemester, lässt sich ein sehr ähnliches Cluster in beiden Analysen finden. *Cluster #2* der Sommersemesterbefragung und *Cluster #1* aus dem Wintersemester zeigen ein hohes Maß an übereinstimmenden Clusterzentren. Es handelt sich zudem um die Cluster, in denen bereits vermutet wurde eher wenige *Problemfälle* zu finden. Der Zeitaufwand für Lehrveranstaltungsbesuche und Selbststudium liegt mit jeweils 11-15 Stunden pro Woche im, für das erste Semester, gewöhnlichen Bereich. Auffällig ist jedoch, dass die Anzahl an zugeteilten Studierenden zu den Clustern unterschiedlich ist. 41% sind es bei der Analyse des Wintersemesters, nur 22% sind es im Sommersemester.

Aufgrund dieser Erkenntnisse lässt sich, unabhängig vom Semester, ein Profil eines Studierenden beschreiben, der nicht als *Problemfall* zu bezeichnen ist. Ein solcher Studierender wohnt im eigenen Haushalt und besucht fast immer die Vorlesungen und Übungsgruppen im Fach GdI 1. Ebenfalls werden fast immer die Hausübungen in diesem Fach erledigt. Die Abiturdurchschnittsnote liegt im Bereich von 2,1 bis 2,5. Der Zeitaufwand für Lehrveranstaltungsbesuche und Selbststudium pro Woche beträgt 11-15 Stunden. Eine Lerngruppe ist vorhanden. Die investierte Zeit für die Prüfungsvorbereitung wird als zu gering eingeschätzt. Der Stoff hingegen wird als nicht zu schwer und auch nicht zu viel bewertet. Finden sich diese Angaben im Fragebogen, handelt es sich eher nicht um einen *Problemfall*. Spezielle private Gründe, sofern angegeben, können allerdings dennoch auf einen *Härtefall* hindeuten. Dies muss im Einzelfall entschieden werden.

Ein *Problemfall* lässt sich anhand der obigen Ergebnisse ebenfalls skizzieren. Seltene Teilnahme an den Lehrveranstaltungen (Vorlesungen, Übungen und Sprechstunden) und geringer Aufwand im Selbststudium (jeweils weniger als fünf Stunden im Durchschnitt pro Woche) scheinen Merkmale eines solchen zu sein. Auch haben solche Studierende in der Regel keine Prüfung bestanden oder erst gar nicht an diesen teilgenommen. Auch großer Zeitaufwand für Freunde und Familie und Hobbies, insbesondere das Computerspielen, kann *Problemfälle* charakterisieren.

Sowohl in der Analyse der Befragung im Sommer als auch im Winter zeigt sich, dass die Studierenden in den meisten Fällen zu wenig Zeit in ihr Studium investiert haben. Der Repräsentant des Attributs *Zu wenig Zeit investiert* ist in allen Clustern, unabhängig ob Winter- oder Sommersemester, stets *trifft voll zu*.

5.3 Analyse unter Ausschluss von Attributen

In diesem Abschnitt werden die Resultate der Clusteranalyse unter Ausschluss verschiedener Attribute erläutert. Dieses Nichtberücksichtigen vereinfacht die Interpretation der gefunden Cluster. Wie schon im vorhergehenden Abschnitt wurde diese Analyse sowohl für die Befragung im Sommer- als auch im Wintersemester durchgeführt. Hauptaugenmerk liegt in diesem Abschnitt auf der Fragestellung, ob sich Studierende mit einer bestanden Prüfung von denen unterscheiden, die keine bestanden haben.

5.3.1 Sommersemesterbefragung

Die Ergebnisse der Clusteranalyse, unter Ausschluss von Attributen, der Befragungen im Sommersemester, sind nachfolgend beschrieben. Das Augenmerk dieser Analyse liegt auf der Leitfrage, ob es Unterschiede zwischen Studierende mit einer bestanden Prüfung und keiner bestanden Prüfung im ersten Semester gibt. Für eine erste Analyse wurden 14

Attribute ausgewählt. Diese Zahl wurde explorativ ermittelt. Eine genauere Erläuterung folgt im Verlauf dieses Abschnitts.

Die Anzahl der Cluster liegt bei vier, da erst ab dieser Zahl sinnvolle Beschreibungen der Cluster möglich sind. In einer zweiten Analyse folgt die Kürzung auf acht ausgewählte Attribute. Dabei wurde die Zahl der Cluster schrittweise auf acht angehoben, um gezielt Gruppen zu identifizieren, in denen sich möglichst viele Studierende wiederfinden, die zumindest eine Prüfung bestanden haben. Bei acht Clustern ist erstmalig die bestandene Klausur in GdI 1 als Clusterzentrum ermittelt worden. Die Quote der Studierenden mit einer bestanden Prüfung liegt beim vorhanden Datensatz bei 26,42%.

Tabelle 5.16 zeigt die Ergebnisse der Analyse mit 14 ausgewählten Attributen und vier Clustern. Den Clustern #0 bis #3 wurden 62 (25%), 82 (33%), 74 (30%) und 28 (11%) Studierende zugewiesen. Verzichtet wurde auf folgende Attribute bzw. Fragenkomplexe aus dem Fragebogen:

- Informatiknote - 40% der Teilnehmer beantworteten diesen Punkt *ohne Angabe*. Einen geeigneten Repräsentanten zu finden, ist aufgrund dieser hohen Zahl nicht möglich.
- Zulassung, Anmeldung, geschriebene Klausuren, von Prüfungen abgemeldet, entschuldigte Abmeldungen, Klausureinsicht - Diese Angaben werden vernachlässigt, da sie nur in der Analyse der Befragung im Sommersemester vorkommen. So können die Ergebnisse von Winter- und Sommersemster im Verlauf dieses Abschnitts verglichen werden.
- Angaben zur Prüfungsvorbereitung in TGdI und Mathe 1 - Nur wenige Studierende bestehen in diesen beiden Fächern die Abschlussprüfungen. Daher wird nur die Prüfungsvorbereitung zu GdI 1 berücksichtigt. Zudem ist GdI 1 das einzige Fach, was im Studienplan für Erstsemester im Sommer- und Wintersemester vorgesehen ist.
- Stunden pro Woche im Job - Die Frage, ob einer Erwerbstätigkeit nachgegangen wird, reicht für diese Betrachtung aus.
- Zeit für Hobbies, Freunde und Familie, Computerspiele, soziale Onlinenetzwerke, Fernsehen - Das Augenmerk wird nur auf den Zeitaufwand von universitären Tätigkeiten gelegt, um erkennen zu können, ob genügend Zeit in Lehrveranstaltungsbesuche und Selbststudium investiert wurde. Zudem wurde im vorhergehenden Abschnitt gezeigt, dass insbesondere die zu wenig investierte Zeit in das eigene Studium ein wichtiges Anzeichen für den frühen Prüfungsmisserfolg ist.
- Angaben zu den Gründen des Prüfungsmisserfolgs - Da in diesem Abschnitt gezielt nach Gründen für das Bestehen von einer Prüfung gesucht wird, sind die Angaben zum Prüfungsmisserfolg nicht relevant.

Attribute	**Cluster #0**	**Cluster #1**	**Cluster #2**	**Cluster #3**
Anfahrtsweg	bis zu 30min	bis zu 60min	weniger als 15min	länger als 60min
Wohnort	im eigenen Haushalt	bei meinen Eltern	im eigenen Haushalt	bei meinen Eltern
Abiturnote	2,1-2,5	2,6-3,0	3,1-3,5	2,6-3,0
Mathenote	12-10	9-7	9-7	9-7
Prüfung bestanden	keine	keine	keine	keine
Prüfungsvorbereitung GdI 1	gut	weniger gut	gut	schlecht
Vorlesungsbesuche GdI 1	häufig	unregelmäßig	fast immer	selten
Übungsbesuche GdI 1	unregelmäßig	unregelmäßig	fast immer	selten
Hausübungen GdI 1	fast immer	fast immer	fast immer	selten
Lehrveranstaltungsbesuche	11-15h	5-10h	11-15h	< 5h
Selbststudium	5-10h	5-10h	11-15h	< 5h
Lerngruppe	nein	ja	ja	nein
Jobben	nein	nein	nein	nein
private Gründe	24,19%	26,83%	45,95%	35,71%

Tabelle 5.16: Ergebnis Clusteranalyse mit K-Means (4 Cluster) und 14 ausgewählten Attributen; Fragebögen der Sommersemester

Die Ergebnisse zeigen, dass kein Cluster eine bestandene Prüfung als Repräsentanten aufweist. Dennoch ist die Quote der Studierenden mit einer bestanden Prüfung in den Clustern unterschiedlich. *Cluster #2* weist mit 37,38% die höchste auf. Gefolgt von *Cluster #0* mit 32,25%. In beiden ist der Anteil an Studierenden mit einer bestanden Prüfung höher

als der Durchschnitt. Im Gegensatz dazu liegt die Rate in den Clustern *#1* und *#3* mit 17,07% und 10,71% unter dem Durchschnittswert von 26,42%.

Durch einen Vergleich der Clusterzentren lassen sich mögliche Gründe für das Bestehen einer Prüfung ableiten. Bei den Clustern mit einer überdurchschnittlichen Quote bei den Studierenden mit einer bestanden Prüfung ist der Wohnort mit dem Wert *im eigenen Haushalt* repräsentiert. In den anderen beiden Clustern ist hier der Repräsentant *bei meinen Eltern*. Die Clusterzentren beim Attribut *Anfahrtsweg* passen zu dieser Beobachtung. Die Lehrveranstaltungsbesuche unterscheiden sich ebenfalls. Während in den Clustern *#0* und *#2* der Wert *11-15h* gewählt wurde liegt er in den anderen beiden jeweils niedriger. Die Angaben zur subjektiven Einschätzung der eigenen Prüfungsvorbereitung in GdI 1 zeigen, dass auch hier nur den Clustern *#0* und *#2* den Wert *gut* zugewiesen wurde. Anhand dieser Ergebnisse lassen sich mögliche Attribute für den Erfolg in zumindest einer Prüfung identifizieren. Anfahrtsweg und Wohnort, sowie die investierte Zeit für das eigene Studium scheinen als Indikatoren zu dienen.

Um gezielt ein Cluster zu finden, in dem besonders viele Studierende mit einer bestanden Prüfung zugeordnet wurden, wurde die Zahl der Cluster schrittweise auf acht angehoben. Ab dieser Zahl ist erstmals einem Cluster die bestandene Prüfung zu GdI 1 als Zentrum zugewiesen. Zudem ist die Zahl der Attribute auf acht reduziert worden. Die Attribute *Prüfungsvorbereitung GdI 1*, *Vorlesungsbesuche GdI 1*, *Übungsbesuche GdI 1* und *Hausübungen GdI 1*, werden dabei nicht berücksichtigt, da diese durch Zeitangabe von Lehrveranstaltungsbesuchen und Selbststudium abgedeckt sind. Zudem werden die Attribute *Job* und *Lerngruppe* außer Acht gelassen, um den Fokus nur auf die investierte Zeit für das Studium zu legen. Die Ergebnisse der Analyse sind in Tabelle 5.17 dargestellt.

Den Clustern *#0* bis *#7* wurden jeweils 33 (13%), 51 (21%), 25 (10%), 19 (8%), 14 (6%), 28 (11%), 52 (21%) und 24 (10%) Studierende zugewiesen. *Cluster #5* wird unter anderem durch die bestandene Prüfung in GdI 1 repräsentiert. Die Quote an Studierenden in diesem Cluster mit einer bestanden Prüfung liegt bei 71,43%. Dieser Wert liegt deutlich über dem Gesamtschnitt der Studierenden mit einer bestanden Prüfung von 26,42%. Aus diesem Grund können die Zentren aus *Cluster #5* als möglicher Repräsentanten für Studierende mit einer bestanden Prüfung gesehen werden. Die übrigen Cluster weisen keinen solch hohen Anteil auf (*Cluster #0* hat den zweithöchsten mit 27,27%).

Attribute	**Cluster #0**	**Cluster #1**	**Cluster #2**	**Cluster #3**
Anfahrtsweg	bis zu 30min	bis zu 60min	weniger als 15min	länger als 60min
Wohnort	im eigenen Haushalt	bei meinen Eltern	im eigenen Haushalt	bei meinen Eltern
Abiturnote	2,1-2,5	2,6-3,0	3,1-3,5	2,1-2,5
Mathenote	12-10	9-7	6-4	9-7
Prüfung bestanden	keine	keine	keine	keine
Lehrveranstaltungsbesuche	11-15h	5-10h	16-20h	< 5h
Selbststudium	5-10h	5-10h	16-20h	< 5h
private Gründe	57,58%	21,57%	52,0%	36,84%
	Cluster #4	**Cluster #5**	**Cluster #6**	**Cluster #7**
Anfahrtsweg	bis zu 30min	weniger als 15min	länger als 60min	länger als 60min
Wohnort	im eigenen Haushalt	im eigenen Haushalt	bei meinen Eltern	bei meinen Eltern
Abiturnote	2,6-3,0	2,6-3,0	2,6-3,0	3,1-3,5
Mathenote	9-7	12-10	12-10	12-10
Prüfung bestanden	keine	GdI 1	keine	keine
Lehrveranstaltungsbesuche	< 5h	11-15h	11-15h	11-15h
Selbststudium	5-10h	11-15h	5-10h	11-15h
private Gründe	35,71%	17,86%	9,62%	66,67%

Tabelle 5.17: Ergebnis der Clusteranalyse mit K-Means (8 Cluster) und acht ausgewählten Attributen; Fragebögen der Sommersemester

Um festzustellen, was *Cluster #*5 von den anderen unterscheidet, werden diese miteinander verglichen. So lassen sich mögliche Gründe aufzeigen, was Studierende mit einer oder keine bestandenen Prüfung unterscheidet. Es wird insbesondere auf Clusterzentren geachtet, bei denen sich die Werte um zwei Stufen (z.B.: Mathenote 6-4 zu 12-10) von denen aus *Cluster #5* unterscheiden.

Cluster #0

In diesem Cluster liegt die Anteil Studierender mit einer bestanden Prüfung bei 27,27%. Mit *Cluster #5* stimmen drei Zentren (Wohnort, Mathenote und Lehrveranstaltungsbesuche) überein. Ein auffälliger Unterschied ist zwischen den beiden Gruppierungen in der Häufigkeit der Angabe von privaten Gründen zu erkennen. In *Cluster #0* sind dies 57,58% (17,86% in Cluster #5). Auch die Zeitangabe für das Selbststudium liegt mit 5-10 Stunden pro Woche niedriger.

Cluster #1

Der Anteil Studierender mit einer bestanden Prüfung liegt hier bei 21,57%. Es gibt ein übereinstimmendes Clusterzentrum mit *Cluster #5* (Abiturnote). Die Hauptunterschiede zeigen sich beim Anfahrtsweg und Wohnort, sowie der investierten Zeit in das Studium.

Cluster #2

24% der Studierenden in diesem Cluster haben eine Prüfung bestanden. Zwei Clusterzentren stimmen mit denen von *Cluster #5* überein (Anfahrtsweg und Wohnort). Auffällig sind die Unterschiede beim Zeitaufwand für das Studium. Sowohl bei den Lehrveranstaltungsbesuchen, als auch beim Selbststudium, werden mit jeweils 16-20 Stunden pro Woche höhere Angaben als Clusterrepräsentaten gefunden, als bei *Cluster #5*. Der Zeitaufwand scheint demnach nicht der Grund für die niedrigere Quote derer zu sein, die eine Prüfung bestehen. Vielmehr liegt die Erklärung in der hohen Zahl an Studierenden mit der Angabe von privaten Gründen (52%) Zudem fällt die schlechtere Abitur- (3,1-3,5) und Mathenote (6-4) auf.

Cluster #3

In Cluster #3 sind 15,79% Studierende mit einer bestanden Prüfung enthalten. Es bestehen keine gemeinsamen Clusterzentren mit Cluster #5. Die größten Abweichungen zu Cluster #5 gibt es bei den Attributen Anfahrtsweg ($>$ 60 Minuten), Wohnort (bei den Eltern) und dem Zeitaufwand für Lehrveranstaltungsbesuche und Selbststudium (jeweils $<$ 5h).

Cluster #4

Diese Gruppierung enthält 7,14% Studierende mit einer bestanden Prüfung. Die Attribute Wohnort und Abiturnote haben die gleichen Clustermittelpunkte wie *Cluster #5*. Den größten Unterschied gibt es beim Zeitaufwand für Lehrveranstaltungsbesuche ($<$ 5h).

Cluster #6

Cluster #6 weist einen Anteil von 23,08% Studierender auf, die eine Prüfung bestanden haben. Es bestehen drei übereinstimmende Clusterzentren mit *Cluster #5* (Abitur-, Mathenote und Lehrveranstaltungsbesuche). Die größten Unterschiede liegen im Anfahrtsweg ($>$ 60 Minuten) und Wohnort (bei den Eltern).

Cluster #7

Die Quote Studierender mit einer bestanden Prüfung liegt in *Cluster #7* bei 12,5%. Übereinstimmungen bei Clusterzentren mit *Cluster #5* gibt es bei den Attributen Mathematiknote, Lehrveranstaltungsbesuche und Selbststudium. Auffällig ist in diesem Cluster die hohe Zahl der Studierenden mit der Angabe von privaten Gründen (66,67%). Große Unterschiede sind beim Anfahrtsweg ($>$60 Minuten), Wohnort (bei meinen Eltern), sowie der Abiturnote (3,1 - 3,5) zu beobachten.

Die Betrachtung der gefunden Cluster zeigt, dass die Angabe von privaten Gründen ein wichtiges Merkmal bei der Unterscheidung von Studierenden mit einer und keiner bestanden Prüfung im ersten Semester ist. *Cluster #5* weist mit 71,43% die höchste Quote von Studierenden mit einer bestanden Prüfung aus. Die Zahl der Studierenden, die private Gründe im Fragebogen angeben liegt in diesem Cluster lediglich bei 17,86%. Mit den Clustern *#0*, *#2* und *#7* wurden gleich drei ermittelt, die sowohl eine hohe Zahl an Studierenden mit der Angabe von privaten Gründen, als auch einer niedrigen Zahl von Studierenden mit einer bestanden Prüfung aufweisen. Ein weiteres Merkmal ist in der investierten Zeit für Lehrveranstaltungsbesuche und Selbststudium zu sehen. Studierende, die hierfür jeweils $\leq$ 10 Stunden pro Woche investieren bestehen häufiger keine Prüfung, als ihre Kommilitonen mit mehr als 10 Stunden. Dies ist in den Clustern *#1*, *#3* und *#4* zu beobachten.

5.3.2 Wintersemesterbefragung

Die Darstellung der Ergebnisse der Clusteranalyse unter Ausschluss von Attributen der Befragungen im Wintersemester ist analog zum Vorgehen im Sommersemester (vgl. Unterabschnitt 5.3.1). Auch in diesem Abschnitt wird die Frage verfolgt, ob sich Studierende mit einer bestanden Prüfung von denen mit keiner unterscheiden lassen. Dabei werden die

gleichen Attribute wie oben berücksichtigt, um die Ergebnisse abschließend vergleichen zu können.

Die Clusteranalyse liefert, bei der Verwendung des K-Means-Algorithmus und der Aufteilung in vier Cluster, die in Tabelle 5.18 gezeigten Clusterzentren. Dabei sind den Clustern *#0* bis *#3* 11 (14%), 30 (39%), 18 (20%) und 20 (26%) Studierende zugewiesen. Der Gesamtanteil von Studierenden mit einer bestanden Prüfung liegt bei 36,84%. Alle vier Cluster enthalten für das Attribut *Prüfung bestanden* als Repräsentanten den Wert *keine*.

Attribute	**Cluster #0**	**Cluster #1**	**Cluster #2**	**Cluster #3**
Anfahrtsweg	bis zu 60min	weniger als 15min	länger als 60min	bis zu 30min
Wohnort	im eigenen Haushalt	im eigenen Haushalt	bei meinen Eltern	im eigenen Haushalt
Abiturnote	3,1-3,5	2,6-3,0	2,6-3,0	3,1-3,5
Mathenote	12-10	12-10	15-13	9-7
Prüfung bestanden	keine	keine	keine	keine
Prüfungsvorbereitung GdI 1	schlecht	weniger gut	schlecht	weniger gut
Vorlesungsbesuche GdI 1	unregelmäßig	fast immer	fast immer	häufig
Übungsbesuche GdI 1	selten	fast immer	häufig	unregelmäßig
Hausübungen GdI 1	selten	fast immer	fast immer	häufig
Lehrveranstaltungsbesuche	< 5h	11-15h	16-20h	11-15h
Selbststudium	5-10h	11-15h	5-10h	16-20h
Lerngruppe	nein	ja	nein	ja
Jobben	ja	nein	ja	ja
private Gründe	54,55%	43,33%	26,67%	30%

Tabelle 5.18: Ergebnis der Clusteranalyse mit K-Means (8 Cluster) und acht ausgewählten Attributen; Fragebögen der Wintersemester

In *Cluster #1* haben 53,33% der Studierenden eine Prüfung bestanden. Damit weist es von allen höchsten Wert aus (*Cluster #0*: 18,18%, *Cluster #2*: 22,22%, *Cluster #3*: 30%). Betrachtet man die Clustermittelpunkte aus *Cluster #1* fallen eine kurzer Anfahrtsweg (< 15 Minuten), *fast immer* besuchte Vorlesungen und Übungen in GdI 1, *fast immer* abgegebene Hausübungen in GdI 1 und das Vorhandensein einer Lerngruppe auf. Zudem liegen die Repräsentanten für die Attribute Lehrveranstaltungsbesuche und Selbststudium bei 11-15 Stunden pro Woche.

Cluster #1 weist die niedrigste Quote von Studierenden mit einer bestanden Prüfung aus (18,18%). In dieser Gruppe geben 54,55% der zugeteilten Studierenden private Gründe an. In keinen anderen Cluster ist dieser Wert so hoch. Vorlesungen ist von den meisten *unregelmäßig* besucht. Hausübungen werden *selten* abgegeben. Der Zeitaufwand für den Besuch von Lehrveranstaltungen wird in diesem Cluster mit dem Wert *weniger als 5 Stunden pro Woche* repräsentiert. Eine Lerngruppe ist nicht vorhanden.

Tabelle 5.19 zeigt die Ergebnisse der Analyse mit acht Attributen und acht Clustern. Den Clustern *#0* bis *#7* sind 9 (12%), 16 (21%), 8 (11%), 7 (9%), 14 (18%), 5 (7%), 6 (8%) und 11 (14%) Studierende zugeordnet. Im Vergleich zu den Analyseergebnissen der Sommersemesterbefragungen wurde, im Attribut *bestandene Prüfung*, auch bei er Erweiterung auf acht Cluster, stets der Repräsentant keine ermittelt. Daher ist auch kein Cluster zu finden bei dem mehr als 50% der zugeteilten Studierenden eine Prüfung bestanden haben. Die Cluster *#1* und *#4* weisen allerdings exakt diesen Wert aus. Beide Cluster kennzeichnet eine vergleichsweise hohe Angabe für den Zeitaufwand (Lehrveranstaltungsbesuche und Selbststudium). Besonders stark unterscheiden sich hingegen die Repräsentanten für Anfahrtsweg und die Angabe von privaten Gründen (78,57% in Cluster #4). Betrachtet man die übrigen Cluster lassen sich keine klaren Aussagen darüber treffen, warum Studierende keine Prüfung bestanden haben. Insgesamt sind die gefunden Repräsentanten zu verschieden, um ein Muster zu erkennen.

5.3.3 Vergleich Sommer- und Wintersemester

In diesem Abschnitt werden die Ergebnisse aus den vorherigen beiden verglichen. Im Vordergrund steht die Frage, ob sich Studierende, die eine Prüfung bestanden haben, von denen unterscheiden, die keine bestanden haben. Betrachtet man sich die Ergebnisse aus Unterabschnitt 5.3.1 (Analyse der Befragung im Sommersemester), lässt sich diese Frage mit *ja* beantworten. Ein überdurchschnittlicher Zeitaufwand von 11-15 Stunden pro Woche, jeweils bei Lehrveranstaltungsbesuchen und Selbststudium ist hierbei ein entscheidendes Merkmal für das Bestehen einer Prüfung. Studierende,

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3
Anfahrtsweg	weniger als 15min	weniger als 15	länger als 60min	bis zu 30min
Wohnort	bei meinen Eltern	im eigenen Haushalt	bei meinen Eltern	im eigenen Haushalt
Abiturnote	k.A.	2,6-3,0	3,1-3,5	2,6-3,0
Mathenote	k.A.	9-7	12-10	9-7
Prüfung bestanden	keine	keine	keine	keine
Lehrveranstaltungsbesuche	16-20h	11-15h	16-20h	5-10h
Selbststudium	11-15h	16-20h	5-10h	5-10h
private Gründe	77,78%	12,50%	25,0%	42,86%
	Cluster #4	**Cluster #5**	**Cluster #6**	**Cluster #7**
Anfahrtsweg	länger als 60min	bis zu 60min	länger als 60min	bis zu 60min
Wohnort	im eigenen Haushalt	im eigenen Haushalt	bei meinen Eltern	bei meinen Eltern
Abiturnote	2,1-2,5	2,1-2,5	2,6-3,0	3,1-3,5
Mathenote	12-10	12-10	15-13	9-7
Prüfung bestanden	keine	keine	keine	keine
Lehrveranstaltungsbesuche	11-15h	< 5h	11-15h	11-15h
Selbststudium	11-15h	< 5h	5-10h	11-15h
private Gründe	78,57%	40,0%	0%	18,18%

Tabelle 5.19: Ergebnis der Clusteranalyse mit K-Means (8 Cluster) und acht ausgewählten Attributen; Fragebögen der Wintersemester

die keine Prüfung bestehen haben entweder private Gründe angegeben oder zu wenig Zeit in ihr Studium investiert.

Durch die Clusteranalyse der Befragungen im Wintersemester lässt sich nicht klar feststellen, warum Studierende keine Prüfung bestehen. Die gefunden Clustermittelpunkte sind zu verschieden, um ein Muster zu erkennen. Es lässt sich lediglich feststellen, dass Studierende mit einer bestandenen Prüfung einen Zeitaufwand für Lehrveranstaltungsbesuche und Selbststudium von ≥ 11 Stunden pro Woche angeben.

5.4 Supervised Clustering

Die Fachstudienberatung hat mit Hilfe der Fragebögen bisher händisch *Problemfälle* identifiziert. Die Studierenden wurden nach den Beratungsgesprächen von der Fachstudienberatung in drei bzw. vier Klassen unterteilt, um die Qualität der vorhergehenden Auswahl zu überprüfen. Folgende Klassen wurden festgelegt:

- *True Positives* (TP) - Richtig erkannte *Problemfälle*.
- *False Positives* (FP) - Fälschlicherweise als *Problemfall* identifiziert.
- *semi False Positives* (sFP) - Fälschlicherweise als *Problemfall* identifiziert, dennoch mit gesondertem Beratungsbedarf. (Einstufung erst ab dem WiSe 2012/13)
- *unclassified* (N)- Nicht als *Problemfall* eingestufte Studierende.

Für die Befragungen im Sommersemester 2012, sowie das Wintersemester 2012/13, liegen die ausgewerteten Fragebögen mit dieser Klassifizierung vor. Dies ermöglicht ein sogenanntes *Supervised Clustering* (konfirmatorisches Vorgehen, vgl. Unterabschnitt 2.2.1). Hierbei werden die oben genannten Klassen, jeweils einem Cluster zugewiesen. Als Algorithmus diente auch für diese Analyse der K-Means. Es kann somit überprüft werden, ob die Klassenzuteilung durch den Algorithmus die gleiche ist, wie die der Fachstudienberatung. Zudem liefern die Repräsentanten der Cluster eine Beschreibung der obigen Klassen.

Im Sommersemester 2012 stehen 45 ausgefüllte Fragebögen zur Verfügung. Die Zahl der Cluster ist auf drei festgelegt, da drei Klassen (TP, FP und N) gegeben sind. Der K-Means-Algorithmus ermittelt für die Cluster *#0* bis *#2* die Zuteilung von 20 (44%), 15 (33%) und 10 (22%) Studierenden. 51,11% der Teilnehmer wurden vom Algorithmus anderen Klassen zugeteilt, als von der Fachstudienberatung. Folgende Aufteilung liefert der Algorithmus:

- Cluster #0 (TP) → 7 x N, 10 x TP, 3 x FP

- Cluster #1 (N) → 9 x N, 3 x TP, 3 x FP
- Cluster #2 (FP) → 4 x N, 3 x TP, 3 x FP

Wie diese Aufteilung zeigt, wird *Cluster #0* mit den *True Positives*, den *Problemfällen*, assoziiert. Anhand der Repräsentanten des Clusters wären demnach Problemfälle solche Studierende, die:

- bei ihren Eltern wohnen und einen Anfahrtsweg von weniger als 60 Minuten haben;
- einen Abiturschnitt von 2,6 - 3,0 und eine Mathenote von 9-7 Punkten haben;
- keine Prüfung geschrieben und bestanden haben;
- ihre Prüfungsvorbereitung eher als schlecht einstufen und keine Lerngruppe haben;
- regelmäßig Vorlesungen und Übungen besuchen, jedoch eher weniger Zeit mit Selbststudium verbringen;
- mehr als zwei Stunden pro Tag mit dem Computer spielen;
- insgesamt der Ansicht sind zu wenig Zeit in ihr Studium zu investieren.

Im Wintersemester 2012/13 haben 123 Studierende einen Fragebogen ausgefüllt. Die Zahl der Cluster entspricht den vier Klassen (TP, FP, sFP und N). Den Clustern *#0* bis *#3* wurden jeweils 27 (22%), 35 (28%), 31 (25%) und 30 (24%) Teilnehmer zugeordnet. 71,55% der Studierenden wurden vom Algorithmus anderen Klassen zugeteilt, als von der Fachstudienberatung. Folgende Aufteilung liefert der Algorithmus:

- Cluster #0 (sFP) → 24 x N, 0 x FP, 1 x sFP, 2 x TP
- Cluster #1 (FP) → 25 x N, 4 x FP, 3 x sFP, 3 x TP
- Cluster #2 (TP) → 23 x N, 2 x FP, 2 x sFP, 4 x TP
- Cluster #3 (N) → 26 x N, 0 x FP, 1 x sFP, 3 x TP

Die Problemfälle (True Positives) werden vom Algorithmus *Cluster #2* zugeordnet. Dieses beschreibt die Problemfälle als Studierende, die:

- einen Anfahrtsweg von mehr als 60 Minuten haben und bei ihren Eltern wohnen;
- einen Abiturschnitt von 2,1 - 2,5 und eine Mathenote von 12-10 Punkten haben;
- keine Prüfung geschrieben und bestanden haben;
- in der Regel selten Vorlesungen und Übungen besuchen und keine Lerngruppe haben;
- jeweils weniger als fünf Stunden pro Woche für Lehrveranstaltungsbesuche und Selbststudium aufwenden;
- mehr als zwei Stunden pro Tag mit dem Computer spielen;
- insgesamt der Ansicht sind zu wenig Zeit in ihr Studium zu investieren.

Beim Vergleich der jeweils gefunden Cluster, denen die *Problemfälle* zugeordnet werden, fallen einige Gemeinsamkeiten auf. Gleiches gilt, wenn man die Ergebnisse aus Unterabschnitt 5.2.1 hinzuzieht. Es zeigt sich, dass es sich bei Problemfällen um Studierende handeln könnte, die neben zu geringem Zeitaufwand für ihr Studium, keine Prüfung mitgeschrieben haben, keiner Lerngruppe angehören und viel Zeit mit dem Computerspielen verbringen. Dieses Muster ist bei allen, in dieser Arbeit aufgeführten Analysen, zu erkennen.

6 Fazit

Die Analyse der zugrunde liegenden Daten, der Studierenden mit frühem Prüfungsmisserfolg, zeigt, dass über den gesamten Befragungszeitraum Gemeinsamkeiten über die Semester hinweg gibt. So lassen sich gauß-artige Kurvenverläufe bei der Abitur- und Mathematiknote der Studierenden erkennen. Während die Studierenden zumeist durchschnittliche Abiturnoten haben, ist die Kurve bei der Mathematiknote hin zu einer im Schnitt guten Note verschoben. Unterschiede lassen sich beim Studienbeginn erkennen. Die Zahl der Studierenden, die vor Beginn des Informatiksstudiums an der TU Darmstadt, bereits ein anderes Studium begonnen haben, ist beim Studienstart zum Sommersemester deutlich höher, als zum Wintersemester.

Die Ergebnisse der Clusteranalyse zeigen, dass sich Studierende in *Problemfälle* und Nichtproblemfälle klassifizieren lassen. Seltene Teilnahme an den Lehrveranstaltungen (Vorlesungen, Übungen und Sprechstunden) und geringer Aufwand im Selbststudium (jeweils weniger als fünf Stunden im Durchschnitt pro Woche) scheinen Merkmale eines *Problemfalls* zu sein. Auch haben solche Studierende in der Regel keine Prüfung bestanden oder erst gar nicht an diesen teilgenommen. Auch großer Zeitaufwand für Freunde und Familie und Hobbies, insbesondere das Computerspielen, kann *Problemfälle* charakterisieren. Zu diesem Ergebnis kommen unabhängig voneinander die explorative (vgl. Unterabschnitt 5.2.1), als auch die konfirmatorische (vgl. Abschnitt 5.4) Analyse.

Ein Student, der aufgrund seines Fragebogens eher kein *Problemfall* darstellt zeichnet sich durch einen Zeitaufwand für Lehrveranstaltungsbesuche und Selbststudium von jeweils 11-15 Stunden aus. Zudem ist eine Lerngruppe in der Regel vorhanden. Vorlesungen und Übungen im Fach GdI 1 werden fast immer besucht, die dazugehörigen Hausübungen erledigt.

Generell zeigt die Clusteranalyse, dass die zu wenig in das eigene Studium investierte Zeit, ein Merkmal für den frühen Prüfungsmisserfolg sein kann. Für das Bestehen von zumindest einer Prüfung ist ein überdurchschnittlicher Zeitaufwand von 11-15 Stunden pro Woche, jeweils bei Lehrveranstaltungsbesuchen und Selbststudium ein entscheidendes Kriterium.

Mit Hilfe der Ergebnisse der Clusteranalyse lassen sich in Zukunft Studierende anhand von bestimmten Attributen als mögliche *Problemfälle* klassifizieren und können somit gezielt durch die Fachstudienberatung beraten werden. Ferner besteht die Möglichkeit eine Software zu entwickeln, die Studierende automatisch klassifiziert oder eine Empfehlung abgibt, ob es sich bei einem Studierenden um einen potenziellen *Problemfall* handelt.

Zudem können die Ergebnisse dieser Arbeit im Rahmen des Mentorensystems im ersten Semester eingesetzt werden. Eine Früherkennung von *Problemfällen*, anhand der beschriebenen Eigenschaften eines solchen, in den wöchentlichen Mentorengesprächen wäre denkbar. Eine somit mögliche und rechtzeitig vor den Prüfungen beginnende gezielte Beratung könnte im Einzelfall dem frühen Prüfungsmisserfolg vorbeugen.

Literaturverzeichnis

[BD13] Bundesrepublik Deutschland, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in d.: *Vereinbarung zur Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II*. http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1972/1972_07_07-Vereinbarung-Gestaltung-Sek2.pdf. Version: Juni 2013. – (abgerufen: 04.09.2013)

[BPW10] Bacher, Johann ; Pöge, Andreas ; Wenzig, Knut: *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*. 3. Auflage. Oldenbourg Wissenschaftsverlag, 2010

[Dar] Darmstadt, Fachbereich Informatik T.: *Studienordnung des Bachelor- und Master-Studiengangs Informatik*. http://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Dekanat/Pruefungssekretariat/Allgemeines/Informatik-BA-MA-StudOrd2009.pdf. – (abgerufen: 01.10.2013)

[GBFT13] General, Sabine ; Bolte, Katharina ; Fath, Daniel ; Tenenbaum, Kateryna: *Mentorensystem der Informatik - Bericht für den Studiendekan*. Juli 2013

[Gut] Gutfleisch, Dr. R.: *Was ist eine Clusteranalyse, wann und wie wird sie angewendet?* http://www.staedtestatistik.de/fileadmin/vdst/ag-methodik/Leitfaeden/2008_AGMethodik_LeitfadenClusteranalyse_Teil2.pdf. – (abgerufen: 08.10.2013)

[JL] Jung, Eric H. ; Lambert, David L.: *JExcelApi*. http://jexcelapi.sourceforge.net/. – (abgerufen: 01.09.2013)

[JSN06] Jünger, Jana ; Schellberg, Dieter ; Nikendei, Christoph: Subjektive Kompetenzeinschätzung von Studierenden und ihre Leistung im OSCE. In: *GMS Zeitschrift für Medizinische Ausbildung* 23 (2006), Nr. 3

[Koz82] Kozelka, Robert M.: How to Work Trough a Clustering Problem. In: *Classifying Social Data. New Applications of Analytic Methods for Social Science Research.* (1982), S. 1–12

[moo] moodle: *Lernportal Informatik*. https://moodle.informatik.tu-darmstadt.de. – (abgerufen: 01.10.2013)

[Neu12] Neubacher, Maxi: Verbesserung und Weiterentwicklung der Abläufe im Mentorensystem des Fachbereichs Informatik der TU Darmstadt - Klassifiezierung von Studierenden mit schlechtem Prüfungserfolg im ersten Semester / Hochschule Darmstadt. 2012. – Praktikumsbericht

[Neu13] Neubacher, Maxi: *Analyse von Algorithmen des maschinellen Lernens für das Mentorensystem Informatik*, Hochschule Darmstadt, Bachelor-Thesis, April 2013

[NWE06] Nairz-Wirth, Erna ; Eeckert, Daniel: Klassifizierung europäischer Bildungssysteme anhand von OECD-Bildungsindikatoren. In: *Zeitschrift für Hochschulentwicklung* 4 (2006), Dezember, Nr. 4, S. 31–42

[PHRB09] Prel, Jean-Baptist du ; Hommel, Gerhard ; Röhrig, Bernd ; Blettner, Maria: Konfidenzintervall oder p-Wert? In: *Deutsches Ärzteblatt* 106(19) (2009), 8. Mai, S. 335–339

[Ren08] Rentz, Daniel: *Microsoft Excel File Format*. http://www.openoffice.org/sc/excelfileformat.pdf. Version: April 2008. – (abgerufen: 01.09.2013)

[ste13] stern.de: *Ende der Wehrpflicht - Wir. Dienen. Deutschland ab jetzt freiwillig*. http://www.stern.de/politik/deutschland/ende-der-wehrpflicht-wir-dienen-deutschland-ab-jetzt-freiwillig-1701057.html. Version: Juni 2013. – (abgerufen: 04.09.2013)

[Wai] Waikato, Machine Learning G. o.: *WEKA*. http://www.cs.waikato.ac.nz/ml/weka/. – (abgerufen: 27.08.2013)

[WEK] WEKA: *ARFF Data Format*. http://weka.wikispaces.com/ARFF. – (abgerufen: 27.08.2013)

[Wika] Wikipedia.org: *k-Means-Clustering*. http://de.wikipedia.org/wiki/K-Means-Algorithmus. – (abgerufen: 29.08.2013)

[Wikb] Wikipedia.org: *Konfidenzintervall*. http://de.wikipedia.org/wiki/Konfidenzintervall. – (abgerufen: 05.09.2013)

[Wikc] Wikipedia.org: *Manhattan-Metrik*. http://de.wikipedia.org/wiki/Manhattan-Metrik. – (abgerufen: 06.10.2013)

[Wikd] Wikipedia.org: *Normalverteilung*. http://de.wikipedia.org/wiki/Normalverteilung. – (abgerufen: 10.10.2013)

[Wike] Wikipedia.org: *Standardabweichung*. http://de.wikipedia.org/wiki/Standardabweichung. – (abgerufen: 10.10.2013)

[Wikf] Wikipedia.org: *Statistische Signifikanz*. http://de.wikipedia.org/wiki/Statistische_Signifikanz. – (abgerufen: 05.09.2013)

[WZ01] Wiedenbeck, Michael ; Züll, Cornelia: Klassifikation mit Clusteranalyse: Grundlegende Techniken hierarchischer und K-means-Verfahren / Zentrum für Umfragen, Methoden und Analysen, Mannheim. 2001. – Forschungsbericht

Abbildungsverzeichnis

Tabellenverzeichnis