Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods[†]

Johannes Berens
WIB, University of Wuppertal
berens@wiwi.uni-wuppertal.de

Kerstin Schneider
WIB, University of Wuppertal and CESifo
schneider@wiwi.uni-wuppertal.de

Simon Görtz
University of Wuppertal
sgoertz@uni-wuppertal.de

Simon Oster
WIB, University of Wuppertal
oster@wiwi.uni-wuppertal.de

Julian Burghoff
WIB, University of Düsseldorf
burghoff@wiwi.uni-wuppertal.de

To successfully reduce student attrition, it is imperative to understand what the underlying determinants of attrition are and which students are at risk of dropping out. We develop an early detection system (EDS) using administrative student data from a state and private university to predict student dropout as a basis for a targeted intervention. To create an EDS that can be used in any German university, we use the AdaBoost Algorithm to combine regression analysis, neural networks, and decision trees—instead of relying on only one specific method. Prediction accuracy at the end of the first semester is 79% for the state university and 85% for the private university of applied sciences. After the fourth semester, the accuracy improves to 90% for the state university and 95% for the private university of applied sciences.

Keywords: student dropout, early detection, administrative data, higher education, AdaBoost

[†] We gratefully acknowledge funding for the project “FragSte” by the Federal Ministry of Education and Research

Die Zahl der Kröten und Kraniche ist in Deutschland besser erfasst als die Zahl der Studienabbrecher.

The number of toads and cranes in Germany is better understood than the number of university dropouts
Dr. Angela Merkel - 09.12.2016

1. Introduction

Student attrition at universities has a negative impact on all parties involved: students, institutions, and the general public (Bowen et al., 2009; Bound et al., 2010). Notwithstanding the educational gain of a student prior to dropping out, university attrition represents a misuse of public and private resources. In addition to monetary losses, dropping out can cause feelings of inadequacy and lead to one being socially stigmatized (Larsen et al., 2013).

Facing high attrition rates and an ever-greater demand for a workforce qualified in STEM-related subjects, education policymakers are increasing their efforts to reduce the number of student dropouts (Gaebel et al., 2012). There are a large number of support programs that aim at reducing student attrition at German universities. However, those programs are not explicitly targeted at the group of students at risk of dropping out but are offered to the general student body. Students, thus, have to self-select into a program. Hence, due to a matching problem, individual support networks and assistance programs may go underutilized. An appropriate initiative to combat student attrition must be cost-efficient and should target students in danger of dropping out. First, students at risk need to be identified using available administrative data; second, at-risk students need to come into contact with a relevant outreach program, and finally, the intervention needs to be evaluated.

Therefore, it is essential to predict dropouts early, which is a central task of educational data mining (EDM), according to Baker (2010), as well as Baker and Yacef (2009). In the present paper, we present an accurate, cost-efficient, self-adjusting early detection system (EDS) that can be implemented at any point in time within a student's career at any German university. Unlike in other countries, students in Germany are not closely monitored in most universities and study programs. In German universities, it is not uncommon to allow students to stay enrolled for extended periods without progressing towards the completion of their degree program. More than that, German students even have financial incentives to enroll. For instance, there is a monthly child allowance for students and a very generous public transportation ticket valid within the city or federal state of their university. Given the very liberal system and the absence of student fees in state universities, an EDS in German universities constitutes an innovative feature, which might counterbalance the liberal but possibly inefficient system currently in place. To make the system applicable to all universities, the proposed EDS uses the limited student administrative data that is regularly updated and maintained by legal mandate. The data is restricted to the demographic data that is collected when students enroll and the achievement data that is collected at the end of each semester. After some cleaning, this data can be used for all other EDM tasks, like clustering of students or analyzing study programs. Thus, the EDS provides a good starting point for research on student attrition using administrative data, it offers important insights for university administration, and it can be useful in the efficient allocation of support and intervention measures to reach at-risk students.

To the best of our knowledge, there is no comparable system being implemented at a German university. Recently some early detection systems based on static thresholds have been implemented. However, due to data protection restrictions, the system is based on achievement data only (Schulze-Stocker et al., 2017; Westerholt et al., 2018). But clearly, static thresholds chosen for one university might not be appropriate for other universities, and they might also differ between fields of study. In addition, Germany has two types of institutions for tertiary education: universities and universities of applied science, both of which can be either public or private institutions. While universities are predominantly public institutions with no student fees, universities of applied science are often private institutions and charge tuition. These two types of universities differ regarding admission requirements and research focus. Universities are generally more research-oriented, whereas universities of applied science focus more on professional training. Universities of applied science have lower admission standards and tend to work with smaller classes and offer more guidance to students. Having a strong focus on research, only universities can offer Ph.D. programs.

Since we propose an EDS that can be implemented at all universities, we have to ensure that the method for prediction is optimally chosen and is flexible enough to account for differences between universities regarding the structure of the university as well as available data. We therefore use and discuss various common classification methods starting with regression models, followed by different machine learning methods, and finally combining all of the approaches in a boosting algorithm.

We set up and test the EDS using two medium-sized universities in the federal state of North Rhine-Westphalia: a state university (SU) with about 23,000 students and 90 different bachelor programs and a private university of applied science (PUAS) with about 6,700 students and 26 bachelor programs. The state university is tuition-free, while tuition at the private university is about 400 Euros per month.

Our results indicate that 79% of SU and 85% of PUAS student outcomes are correctly identified at the end of the first semester; furthermore, the accuracy of the EDS increases as new student performance data becomes available at the end of each consecutive semester: after the fourth semester, the EDS correctly predicts 90% (95%) of the SU (PUAS) student outcomes. Confirming earlier studies, early stage performance data is particularly important for predicting student attrition, while demographic data has limited predictive value once performance data is made available.

This paper is organized as follows. Section 2 reviews related literature. Section 3 offers a description of the data. Section 4 explains the empirical strategy. In Section 5, we present the results. Section 6 concludes.

2. Related literature on student attrition

The quality of empirical research on student attrition depends on the availability of good data. There are two types of data that have been exploited in the literature: administrative data and survey data. Due to the lack of available administrative data, information on student attrition in German universities is largely gathered from surveys (Larsen et al., 2013). However, student surveys have significant limitations when investigating the causes of attrition. In ex-ante interviews, the dependent variable, student attrition, must be replaced with the intention of dropping out. Using the intention to drop out as a predictor for actually dropping out is, however, controversial in the literature as it assumes that the intention is not exaggerated or otherwise subjected to self-adjustment (Brandstätter et al., 2006). But clearly, one advantage of survey data as compared to administrative data is that survey data allows for learning more about determinants of the dropout decision. Tinto's (1975) "student integration model" established the central importance of the social and academic integration of the student. Pascarella and Terenzini (1979) adopt the idea of integration and extend the model by distinguishing between forced and voluntary attrition. Bean (1983), on the other hand, presents the importance of integration as a main predictor of attrition and adds student satisfaction as a central variable. The importance of academic performance and informational frictions for explaining attrition has been stressed in recent literature (Stinebrickner & Stinebrickner, 2008; 2012; 2013; 2014; Arcidiacono et al., 2016). But despite the importance of the topic, there is still much that is unknown about the underlying determinants of attrition and the effective means for reducing it. But all of the aforementioned determinants, namely, integration, identification, and satisfaction of students are not components of student administrative data. Analyzing administrative data implies knowingly using incomplete data and making the best of it.

Aside from the acknowledged shortcomings of administrative data, they are much better suited for studying the extent of dropout occurrences and, more importantly, the use of administrative data allows predicting student dropout and analyzing student study behavior.

Using that information, support programs can be developed and explicitly targeted at students in need. In addition to the availability of administrative data, the application of machine learning methods in the field of educational data mining can be an improvement over using traditional regression models to predict dropouts (Baker & Yacef, 2009). One of the main goals of educational mining is to predict the class or label of educational outcomes (Baker, 2010; Baker & Yacef, 2009). Research in this field can be distinguished according to the granularity of the predicted outcome: at the tutoring level, at the course level, or at the degree level (Asif et al., 2017). For example, Feng et al. (2006) predict test scores with intelligent tutoring systems by integrating the amount of assistance a student needs to solve problems, while Strecht et al. (2015) and Barber and Sharkey (2012) predict the success or failure in a course or program. In one of the first papers to identify successful degree-level students, Kotsiantis et al. (2003) analyze demographic and performance data using machine learning methods. They correctly predicted more than 70% of successful students using various methods such as decision trees, neural networks, a naive Bayes method, logistic regression analysis, support vector machines, and instant learning algorithms. Subsequent studies have largely followed a similar structure and methodology. Examples are Xenos (2004), Minaei-Bidgoli et al. (2004), Nghe et al. (2007), Dekker et al. (2009), Zhang et al. (2010), Bayer et al. (2012), Yukselturk et al. (2014), Sara et al. (2015), Santana et al. (2015), and Kemper et al. (2018). While the studies are not easily comparable due to differences in sample size, variable settings, research methods, and research questions, the different methods employed within a given study resulted in only marginal differences in predictive accuracy. Significant differences in between-study results primarily reflect the predictive quality of the data, i.e., the power of the data to predict study outcomes. Although there is a tendency for random forests to provide good and robust results, the most accurate method for a data set cannot be determined generally. Reviews of advanced methodologies within educational data mining can be found in Romero and Ventura (2010) and Pena-Ayala (2014).

In principle, however, all studies show that the forecasts become more accurate in later semesters. But since, in particular, the first year in higher education is highly correlated with student success (Arnold & Pistilli, 2012; Barefoot et al., 2005), accurate forecasts are especially beneficial for first-year students.

3. Administrative student data used in the EDS

The EDS developed in this paper uses student administrative and performance data to predict whether a student will drop out of his/her program. Using historical student data from dropouts and graduates, our system identifies the demographic and performance characteristics of students who are at risk of dropping out. The current analysis restricts itself to bachelor level degree programs; however, the method can be easily applied to master level programs, as well. Overall, the EDS is designed in such a way that it can be introduced and operationally maintained at low cost in German state and private universities as well as in universities of applied sciences. Provided that the administrative data requirement is met, the implementation is, of course, not limited to Germany. For ease of implementation, however, it is necessary that only standardized data—data which is necessarily collected by law at all universities—be required for implementing the system. The standardized and nationally available student data used in this paper is collected and stored by mandate of the Higher Education Statistics Act (HStatG). The HStatG established a nationwide standard for collecting specific student data. Furthermore, §3 HStatG, which is relevant to the present analysis, was last modified in 1997 (BGBI I, 1997, p. 3158). According to §3 HStatG, both public and state-recognized private universities have to collect, store, and regularly report the student data outlined in Table 1.

The EDS can be expanded to accommodate additional relevant variables available at universities. For example, the university entrance qualification grade is collected at the time of enrollment but not necessarily stored at all universities. The information is, however, according to prevailing opinion, a well-suited predictor of study outcomes (Danilowicz-Gösele et al., 2017; Trapmann et al., 2007; Brandstätter & Farthofer, 2002).

Using the standardized student data referenced in Table 1 has advantages, but it certainly limits the dimensions of the EDS in explaining and predicting dropouts. Some of the reasons cited in the literature for dropping out are not captured by the student data collected at universities. In the literature reviewed above, it is agreed that the determinants of attrition are not mono but rather multi-dimensional in character and include the student's self-concept (Burrus et al., 2013; Larsen et al., 2013, p. 47). With regard to German universities, Heublein, et al. (2014) identified seven causes for attrition: performance requirements, finances, exam failure, lack of motivation, study conditions, professional reorientation, and illness. Wiers-Jenssen et al. (2002), on the other hand, state that student satisfaction is a key factor for student success, although the direction of causality or whether they are mutually dependent is unclear.

While possibly important for explaining student dropout rates, the following data are not available for use in an EDS: information on student satisfaction, financial circumstances, family situation, personal motivation, individual fit of the institutional framework, diligence while choosing the course of study, professional interest in the subject of study, professional inclination, academic or social student integration, and the student's state of health. Thus, the EDS is based on student demographics and academic achievement data that are collected according to §3 HStatG. Moreover, the central importance of academic achievement as a predictor for dropping out is emphasized again and again in the literature (Larsen et al., 2013).

Table 1: Data collected according to the Higher Education Statistics Act

		Data collected according to the Higher Education Statistics Act	Variables	Values
Demographic data	Personal	Year of birth	Age at enrolment	Age in years
		Gender	Gender	1=male; 0=female
		Place of birth	Federal state of birth	16 German federal states
		Nationality	Nationality	1=foreign; 0=German
			Region and country of origin (either by birth or imputed immigration background)	11 regions and 5 countries
		First and last name		
			Immigration background of students	Probability in percent
		Health insurance company	Health insurance (private/state)	1=private; 0=public
	Previous education	Type of university entrance qualification	Type of entrance degree (AHR, FHR, fgHR, foreign)	1 to 4
		City where university entrance qualification was earned	City where university entrance degree was earned	1 = City of university; 0 = else
		Grade of university entrance qualification	Grade of entrance degree	1.00 to 4.00
		No. of semesters in previously enrolled study programs	Lateral entrants	1=yes; 0=no
		Number of study programs previously enrolled in at this university	Number of previous semesters	0 to max
			Number of previous courses of study at this university	0 to max
	Study	Course of study	Course of study or number of simultaneous programs enrolled in	1 to max
		Type of study program	Study form (full time/part time/dual)	1 to 3
Academic Performance data		Name of exam	Number of important successfully completed exams	1 to 9
			Number of other successfully completed exams	0 to max
		Exam grade	Average grade per semester	1.00 to 4.00
		Date of exam		
		Result of enrolled exams (pass/fail/withdrawn/no-show)	Number of failed exams per semester	0 to max
			Number of exams per semesters not participated in	0 to max
			Number of no-show exams per semester	0 to max
Outcome		Ex-matriculation date	Graduate or drop out	1=drop out, 0=graduate
		Reason for ex-matriculation		

Notes:

Nationality: Citizenship and place of birth distinguish foreign from non-foreign students and students without an immigration background from students that are first-generation immigrants.

Immigration background: Name-based imputation of immigration background distinguishes between students that are second-generation immigrants and those that are not.

Type of entrance degree: AHR = university entrance degree, FHR = university of applied science entrance degree, fgHR = restricted subject-specific entrance degree, foreign = foreign entrance degree.

Average grade: Failed exams have to be rewritten; thus, they don't lower the GPA.

Number of important successfully completed exams: The nine exams each semester that are most closely correlated with graduation in a study program.

Number of exams per semester not participated in: When available, some universities register when a student has withdrawn from an exam, others don't. Furthermore, some universities register non-participation—when a student neither withdraws nor presents a medical excuse—as a "no-show," others as a "not-pass." The latter can't be distinguished from failed exams.

The extent to which data limitations impede the efficacy of the EDS depends on whether and how quickly the above-mentioned factors influence academic performance before leading to student attrition.

Table 1 shows how the §3 HStatG student data are transformed into the variables used in the EDS. In summary, the demographic variables consist of the following information:

- Personal: age, gender, address, place of birth, immigration background
- Previous education: type and place of university entrance qualification, previous academic experience
- Study: course of study, type of enrollment (i.e., full- or part-time)

Additional information for students with an immigration background includes nationality, domestic or foreign university entrance qualification, and whether the student is a first or second-generation immigrant (cf. Section 3.1. below).

In addition to the demographic data, student performance data are also made available at the end of each subsequent semester. The student performance data collected at the end of each completed semester include the average semester grade, average semester credit points earned, the number of registered but unattended exams, and the number of attempted but failed exams. In addition, it is determined how many of the most important exams were passed in a given semester. An exam is determined to be most important when its successful completion is highly correlated with the successful completion of the degree. We restrict the number of exams to nine, which is about 150% of the regular number of exams per semester. This accounts for the possibility of shifting exams.

Finally, in order to fit our model, former students are classified as dropouts or graduates.

As already mentioned, the EDS was developed and tested at two medium-sized universities in the federal state of North Rhine-Westphalia: SU with about 23,000 students and 90 different bachelor programs and a PUAS with about 6,700 students and 26 undergraduate programs. The machine learning process was performed using administrative data from former bachelor students, who either dropped out or graduated between 2007 and 2017. The forecasting system was then tested using student data that had not been included in the training data. Data from students who matriculated in 2012 and 2010 were chosen for the test data of PUAS and the SU, respectively.[1] The training data of former SU students who matriculated in 2007-2009 and 2011-2017 comprised a total of 12,730 observations; the 2010 data used for testing consisted of 1,766 bachelor students. PUAS training data from former students who matriculated in 2007-2011 and

[1]We choose the year 2010 as our test cohort at the SU because, the actual observed duration of studies is longer at the SU as compared to the PUAS, even though a bachelor degree is scheduled as a 6 semester program at both universities. Hence choosing a later cohort, would result in a test cohort with many students still enrolled. The cohort of 2010 at the SU and the cohort of 2012 at the PUAS are the latest cohorts with (almost) no remaining students. See also chapter 5.2.4

2013-2017 included a total of 6,297 observations; the test data from 2012 comprised 1,303 bachelor students.

3.1. Expanding the data by imputing information on migration status and additional information

In addition to the student data collected pursuant to §3 HStatG, additional variables can be derived from the available data. A limitation of the available administrative data is the missing information on socioeconomic status, which is known to predict educational achievement, in particular, for Germany (OECD, 2018). While there is no direct information on socioeconomic background, some indirect information is available and can be used. If known, students' home addresses can be used to gather socioeconomic data on the student, the university entrance grade can provide information about previous academic performance, and the first and last name can provide information on student immigration background.

German universities distinguish between a semester address and a home address. Accordingly, it is possible to determine whether the student has moved from her home for the purpose of studying, is commuting over long distances, or is studying in her hometown (Dekker et al., 2009). Median income available at the level of zip code can be used as a proxy for income background (Danilowicz-Gösele et al., 2014). The type of health insurance is also used as a proxy for socioeconomic background. Here one can distinguish between private and publicly insured students (Danilowicz-Gösele et al., 2014). Students with private health insurance are primarily children of parents who are self-employed, civil servants, or employees with an income above a certain threshold (in 2017 €57,600 per year). Thus, students with private health insurance are typically from families with a higher socioeconomic background.

Immigration background (with or without German citizenship) has been shown to be correlated with educational success in Germany (OECD, 2016). As a rule, however, institutions of higher education typically only know a student's citizenship, place of university entrance qualification, and place of birth. Thus, international students can be included in the group of foreign-educated students and students that have been partly or wholly educated in Germany but do not hold German citizenship. Non-German citizens born abroad are considered first-generation immigrants. However, second-generation immigrants with German citizenship cannot be directly identified from university administration data.

Since it is known, however, that second and third-generation immigrants underperform in the German educational system, it is important to be able to identify them. For this reason, first names and family names of students are examined to determine their ethnic origin. Germans born in Germany, whose first and surnames reveal an immigration background, are considered migrants of the second or third generation. The method of Humpert and Schneiderheinze (2002) is a common method for determining a subject's country and region of origin from the combination of first and surnames (Berger et al., 2004). Based on the methodology of Humpert and Schneiderheinze (2002), a name-database containing around 200,000 first names and another database containing around 600,000 surnames (Michael, 2007; Michael, 2016) are used. There is a probability for each country-name combination (using 145 countries) that indicates the likelihood that the student's family migrated from the given country. For gender-specific names, the information of the gender is also included; gender-neutral names, as well as names for which the gender-specificity depends upon the country, are marked as well.

Using the information in the database, the probability of an immigration background is determined from the distribution of first and surnames in represented countries; the region/country of origin is determined in a second step. Since most names are common in more than one country, the 145 countries are aggregated into 11 regions. In accounting for the main countries and regions of origin for immigrants into Germany, we distinguish between the following 11 regions (Statistisches Bundesamt, 2015):

- North America
- Central and South America
- Northern and Western Europe
- Southern Europe
- Eastern Europe
- North Africa
- Rest of Africa
- Western Asia
- Eastern and South-Eastern Asia
- Southern Asia
- Australia, New Zealand, and Melanesia

Some of the regions above, such as the Americas, are uncommon regions of origin for foreign students in Germany. Thus, even though the countries in those regions are very heterogeneous, the high level of aggregation does not present a problem for the analysis of German student data. In Germany, the most frequently represented countries among students with an immigration background are Turkey, Italy, Croatia, Russia, and China (Statistisches Bundesamt, 2015; Heublein & Burkhart, 2013, p. 23). For this reason, in addition to the regions given above, these countries will be considered separately.

The validity of the imputation was checked in two different ways. Firstly, the group of non-German students with known citizenship was used. Of the 4,004 foreign citizens in the sample, more than 94% of the first and surname combinations were correctly assigned. Secondly, the imputed immigration background from 1,598 first names was compared with the migration information in the German Socio-Economic Panel (GSOEP). In the questionnaire, the respondents report their first name and, if applicable, immigration background. Applying our

Table 2: Ethnic composition of the student population

	State University				Private University of Applied Sciences			
Region	Students with foreign nationality	Domestic students with immigration background	Immigration background	Proportion of student body	Students with foreign nationality	Domestic students with immigration background	Immigration background	Proportion of student body
North America	8	46	54	0.20%	0	41	41	0.25%
Central & South America	27	133	160	0.60%	15	88	103	0.64%
Northern & Western Europe	103	1,102	1,205	4.52%	88	778	866	5.35%
Southern Europe	615	324	939	3.52%	341	255	596	3.68%
Eastern Europe	433	419	852	3.19%	87	322	409	2.53%
North Africa	296	137	433	1.62%	90	113	203	1.25%
Rest of Africa	136	116	252	0.94%	39	61	100	0.62%
Western Asia	748	697	1,445	5.41%	812	1,008	1,820	11.24%
Eastern & Southeast Asia	392	323	715	2.68%	142	65	207	1.28%
Southern Asia	116	165	281	1.05%	40	201	241	1.49%
Australasia	3	2	5	0.02%	0	1	1	0.01%
Countries								
Italy	174	153	327	1.23%	102	103	205	1.27%
Russia	93	154	247	0.93%	33	143	176	1.09%
Turkey	620	641	1,261	4.73%	761	1,000	1,761	10.88%
China	278	274	552	2.07%	123	26	149	0.92%
Germany	23,757	0	-	71.07%	14,537	0		71.07%
Summary								
Number of students	26,686				16,192			
Non-identified names	234				147			
Immigrants	7,721				11,510			
Germans	18,574				4,684			
Immigration Rate	**28.93%**				**28.93%**			

Notes:

Number of students:	Number of undergraduate SU (PUAS) students between 2000 and 2017 (2007 and 2017).
Non-identified:	First and second name not in the database.
Immigrants:	Students with foreign nationality, foreign place of birth, or, most likely, a foreign name.
Germans:	Students with German citizenship and no apparent immigration background.

imputation method to the GSOEP information resulted in correctly labeling 82% of existing and non-existing immigration backgrounds. Note that in the second test—using the GSOEP data—only the subject's first name was used which is expected to lower the accuracy of the imputation. Excluding the subject's surname lowered the imputation's accuracy in the first test from 94% to 88%.

The imputed immigration data for both universities are summarized in Table 2. At both universities, 29% of the students are first or second-generation immigrants, and the distribution of countries of origin is similar at both universities. The only difference is that the proportion of Chinese and Turkish students is higher at the PUAS.

3.2. DATA DESCRIPTION

Tables 3a and 3b show a summary of the data for both universities. In each of the columns, the data is summarized with respect to the year of enrollment. Thus, the descriptive statistics in column (6) refer to students who enrolled in 2012 and either dropped out or graduated by 2017 or earlier. There are marked differences between universities. The number of students enrolled at the SU is usually substantially higher than the number enrolled at the PUAS. Moreover, enrollment at the PUAS is limited to only one study program. At the SU, however, of the 20,707 enrolled students between 2007 and 2017, 11,193 students were enrolled in two or more study programs, 10,467 in three or more programs, and 2,770 in four or more programs. Thus, at the SU, students might be counted more than once if they enrolled in different programs. An example illustrates this: students who plan to become schoolteachers study two majors, e.g., German and Mathematics; consequently, they are enrolled in two different departments and would be counted twice. For this reason, the type of study program is used as a predictor at the PUAS and not at the SU. Furthermore, there are also differences regarding university entrance requirements. Generally, the prerequisites for studying at a PUAS are less restrictive than at a university; this is true for both the grade of the university entrance qualification (for instance, there might not be a numerus clausus) and the type of university entrance qualification. As a result, the composition of the student body is different.

First, looking at the SU, women are overrepresented in most of the years, which is likely explained by a large education department at the SU (cf. Table 3a). Age at enrollment is between 21 and 22.6 years. Between 24% and 29% of the students do not have an immigration background. The percentage of foreign-born students is between 7% and 11%. There does not appear to be a time trend with regard to migration. The vast majority of students live in a city other than the home city of the university in question. The average grade for the university entrance exam is between 2.6 and 2.9. Between 5% and 8% of the students have private health insurance, and the average number of failed exams is between 0.44 and 0.75.

Comparing the descriptive statistics for the PUAS in Table 3b to the descriptive statistics for the SU in Table 3a reveals substantial differences. Male students are overrepresented at the PUAS, the age of enrollment is higher, and there are more foreign students. Fewer students have a regular university entrance degree. There is no information about the grade of the entrance degree, nor do we have data on the type of health insurance. The average number of failed exams ranges between 0.17 and 0.62 and is thus lower than at the SU.

Table 3a: Summary statistics: SU (mean and standard deviation)

Cohort	(1) 2007	(2) 2008	(3) 2009	(4) 2010	(5) 2011	(6) 2012
Gender (1=male; 0=female)	0.34	0.43	0.40	0.43	0.51	0.46
Age at enrollment	21.24	21.84	21.86	21.93	22.28	22.60
	(3.15)	(3.75)	(3.56)	(3.72)	(4.38)	(4.67)
First generation immigrant (1=yes; 0=no)	0.08	0.11	0.10	0.09	0.07	0.08
Second generation immigrant (1=yes; 0=no)	0.19	0.17	0.19	0.18	0.17	0.19
City of entrance qualification (1= city of university; 0=else)	0.14	0.19	0.18	0.18	0.21	0.22
General university entrance qualification (1=yes; 0=no)	0.97	0.95	0.95	0.94	0.94	0.94
University of applied sciences entrance qualification (1=yes; 0=no)	0.00	0.01	0.00	0.01	0.01	0.01
Restricted university entrance qualification (1=yes; 0=no)	0.01	0.01	0.01	0.02	0.01	0.01
Foreign university entrance qualification (1=yes; 0=no)	0.03	0.04	0.03	0.03	0.04	0.03
Grade of university entrance qualification	2.87	2.85	2.79	2.71	2.68	2.61
	(0.82)	(1.00)	(0.97)	(0.87)	(0.92)	(0.89)
Health insurance (1=private; 0=public)	0.06	0.08	0.06	0.05	0.07	0.06
# of enrolled study programs	3.06	2.62	2.67	2.71	2.32	2.20
	(1.85)	(1.87)	(1.76)	(1.97)	(1.68)	(1.51)
Lateral entrants (1=yes; 0=no)	0.17	0.25	0.30	0.39	0.38	0.44
# of semesters at prev. university	1.63	2.47	2.65	3.15	2.63	3.02
	(4.45)	(5.32)	(5.01)	(5.09)	(4.50)	(5.13)
Average grade per semester	2.46	2.49	2.45	2.50	2.51	2.49
	(0.55)	(0.59)	(0.56)	(0.56)	(0.58)	(0.58)
Average CPs per semester	12.91	17.18	18.33	19.80	15.38	15.22
	(16.44)	(26.92)	(29.29)	(30.12)	(22.38)	(23.87)
No exam taken	0.19	0.20	0.22	0.19	0.24	0.31
# of exams per semester not participated in	0.18	0.40	0.38	0.44	0.43	0.45
	(0.62)	(1.56)	(1.44)	(1.28)	(1.16)	(1.35)
# of failed exams per semester	0.44	0.63	0.62	0.75	0.59	0.64
	(1.02)	(1.78)	(1.88)	(1.90)	(1.24)	(1.77)
Obs.	2,637	1,846	2,215	2,170	2,860	2,674

Note: Performance data refers to data from the first semester.

As the institutions are different, the variables are likely to have a different impact on prediction accuracy. This does not only apply to the demographic variables but also to the performance data, which has the highest explanatory power and is available after completion of the first semester. Of particular importance are earned credit points per semester, the average score of successfully completed exams, the number of successfully completed exams, and the successful completion of exams deemed most important for the student's respective study program.

Table 3b: Summary statistics: PUAS (mean and standard deviation)

Cohort	(1) 2007	(2) 2008	(3) 2009	(4) 2010	(5) 2011	(6) 2012
Gender (1=male; 0=female)	0.66	0.68	0.65	0.64	0.65	0.62
Age at enrollment	22.27	23.95	24.40	24.36	24.01	23.97
	(2.99)	(3.33)	(3.22)	(3.04)	(2.76)	(2.39)
First generation immigrant (1=yes; 0=no)	0.13	0.15	0.13	0.10	0.08	0.09
Second generation immigrant (1=yes; 0=no)	0.20	0.15	0.18	0.17	0.18	0.18
City of entrance qualification (1=city of university; 0=else)	0.26	0.34	0.34	0.30	0.30	0.32
General university entrance qualification	0.56	0.49	0.48	0.49	0.51	0.49
University of applied sciences entrance qualification (1=yes; 0=no)	0.40	0.44	0.48	0.46	0.44	0.44
Restricted university entrance qualification (1=yes; 0=no)	0.01	0.00	0.01	0.02	0.02	0.05
Foreign university entrance qualification (1=yes; 0=no)	0.04	0.06	0.03	0.02	0.02	0.02
Lateral entrants (1=yes; 0=no)	0.32	0.34	0.34	0.34	0.34	0.34
Average grade per semester	2.37	2.32	2.32	2.28	2.24	2.28
	(0.55)	(0.51)	(0.53)	(0.53)	(0.51)	(0.53)
Average CPs per semester	12.78	16.25	18.77	19.11	19.98	19.69
	(11.15)	(10.84)	(10.71)	(11.17)	(11.67)	(11.63)
No exam taken	0.35	0.18	0.11	0.09	0.09	0.09
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
# of exams per semester not participated in	0.40	0.50	0.21	0.25	0.25	0.23
	(0.57)	(0.80)	(0.42)	(0.46)	(0.43)	(0.49)
# of failed exams per semester	0.17	0.30	0.62	0.56	0.50	0.54
	(0.34)	(0.45)	(0.79)	(0.69)	(0.62)	(0.68)
Obs.	193	1,175	1,423	1,343	1,358	1,563

Note: Performance data refers to data from the first semester.

4. Empirical Strategy

We now present the empirical strategy behind the development of the EDS. Instead of relying on a single method, the EDS model is composed of multiple evaluation methods (classifiers). The methods are used alongside each other to evaluate their respective predictive powers. We combine the methods by means of the AdaBoost algorithm (Schapire & Freund, 1997; Schapire & Freund, 2012). The methods used for the analysis are logit regression models, neural network models, and decision tree algorithms. The regression models and the AdaBoost were computed in Stata, the decision trees in Weka (Frank et al., 2016), and the neural network in MemBrain (Jetter, 2017).

First, a prediction model (parameters, weights, rules, and point estimates) is developed using the training data. The aim of the model is to identify potential dropouts as early as possible by classifying student observations as graduates or dropouts and then checking the precision of the prediction. Subsequently, the results of the individual methods are merged using the boosting algorithm first developed by Schapire and Freund (1997; 2012).

4.1. LOGIT MODEL

As a starting point for our analysis, we estimate a logit model

$$P(y_{it} = 1 | x_i, z_{it}) = \Lambda(\beta_0 + \beta_1 x_i + \beta_2 z_{it}),$$

with *i* and *t* denoting student and semester, respectively, and Λ representing the logistic function. The dependent variable y_{it} is a binary variable indicating graduate (0) and dropout (1). Demographic information x_i is time-invariant, while the performance data z_{it} varies over time. Section 5 discusses the results of the logit model using student performance and demographic data from the time of enrollment up to the sixth and fourth semester for the SU and the PUAS, respectively. The logit model affords some advantages in that the coefficients are easier to interpret, making it easier to understand the importance and magnitude of the explanatory variables on the likelihood of dropping out.

4.2. NEURAL NETWORK

The backpropagation algorithm is used for the multilayer perceptron (MLP). In summary, the architecture of the MLP can be described by about 31 neurons (depending on the semester and university) in the input layer, 16 neurons in the first, and 8 neurons in the second hidden fully-connected layer and one neuron in the output layer. We select the logistic function as the activation function for all neurons. The training process is briefly described below (Mucherino et al., 2009).

The neurons of the input layer become initialized with the training data set, which consists of the external inputs (determinant variables) and the actual outcome y_{it} (dropout or graduate). All other neurons existing in the hidden layers are set randomly between minus one and one. In the supervised learning process, the network predicts student outcomes from the training data. The network then uses the assigned prediction weights and probability estimates to forecast student outcomes $\widetilde{y}_{it}$. An advantage of supervised learning is that the prediction algorithm is assigned an error term e_t, the difference between the actual study outcome explained by the training data and the predicted outcome from the neural network. The error or loss function is the sum of squared errors.

$$\boldsymbol{e_t} = \sum_{\boldsymbol{i}} (\tilde{y}_{it} - y_{it})^{\boldsymbol{2}}$$

The error function has the advantage that it is continuously differentiable and, thus, simplifies the weight adjustment process during the training phase. Backpropagation optimizes the weights such that the neural network can learn how to correctly assign inputs to outputs by minimizing the error function at every step.

4.3. BAGGING RANDOM FOREST

Predictions for the outcome variable across observations are determined by decision tree algorithms. An overview of the most frequently used algorithms can be found in Schapire and Freund (2012) and Sammut and Web (2017). In the present paper, we use the C4.5 algorithm for decision trees (Hall et al., 2009). The C4.5 recursively performs the process of tree building, using information gain to guide the attribute selection process. In addition, this algorithm uses an enhancement of attribute selection and branching.

Since decision trees are a very flexible nonparametric machine learning algorithm, they tend to overfit the data. To decrease the variance and to improve the precision of the estimates, we use the bagging (bootstrap aggregation) meta-learning algorithm. Random forest is a method for generating multiple versions of the tree by bootstrapping on the training sample and averaging these to get an improved classifier (Breimann, 1996; 2001). While bagging constructs a large number of (possibly similar) trees with bootstrap samples, the random forest algorithm additionally chooses a random subset of predicting variables before each node is split. This will lead to different, uncorrelated trees from each sample.[2] We applied bagging on the test data before estimating a random forest, therefore bagging with random forest (BRF).

4.4. META-ALGORITHM ADABOOST

To combine the predictive powers of the neural network, regression model, and BRF, we use a boosting algorithm. Boosting algorithms evaluate the influence of the individual methods (weak classifiers) and merges the results into a single (strong) classifier. Here the adaptive boosting (AdaBoost) algorithm developed by Freund and Schapire (1997) is applied. The AdaBoost algorithm was originally used to solve character recognition problems, but it also achieved good

[2] From all tested decision trees (i.a. C4.5, M5p, CART, Decision Stump, RepTree) with all tested meta-learning algorithms (i.a. bagging, random subspace, random committee, classification via regression, random forest), the BRF and C4.5 perform best. Results are available upon request.

results in solving various classification problems. It is a general method for improving the classification accuracy. The basic idea is to combine the results obtained from various methods into an efficient decision-making rule so that, in our application, dropout behavior can be forecasted with better accuracy. On the basis of the calculated forecasts, these methods (described above) are initially weighted equally. In each repetition of the algorithm, the individual weights are adapted according to the distribution in such a way that the resulting classifier has the smallest possible error value. The prediction of the AdaBoost is calculated as the sum of the weighted predictions and has better prediction accuracy as compared to using any single method. Moreover, since the proposed EDS can be implemented at any German university and over any given number of semesters, the AdaBoost avoids the need to choose a single best working method.

4.5. Choice of identification threshold

Each forecasting method estimates a dropout-probability for each student that is between 0 (graduate) and 1 (dropout). Thus, the EDS needs a threshold beyond which potential dropouts are defined to be at risk. The choice of threshold is important when implementing an EDS. An EDS has little value in itself unless it is used as a basis for interventions aimed at, for instance, lowering dropout rates. Thus, the EDS could be used to inform students about the risk of failure. In practical terms, university administration assigns a threshold delineating students as at-risk and thus in need of intervention. The lower the threshold, the higher the rate of correctly predicted dropouts, but at the same time, the rate of correctly identified students decreases, as many students that will not drop out are treated as potential dropouts. This may have a negative impact on the student body's acceptance of the EDS. This is a serious trade-off which is described in Swets (1988) and discussed in a study closely related to our paper by Gleason and Dynarski (2002) and Bowers et al. (2013). Knowles (2015, p. 23) summarizes: "Where do the indicators draw the line between false-alarm and true classification of students and is the resulting student group the group that schools should serve?" One possible solution is to assign the threshold using the average dropout rate of students enrolled in previous terms. Note that deviations in the dropout rate between cohorts establish a margin of error when distinguishing probable dropouts from actual dropouts. To test the predictive power of our EDS on previous cohorts, we set this threshold such that the number of identified dropouts coincides with the number of known dropouts in the test cohort for each semester. This allows distinguishing between the two causes of deviations from the true dropout rate, namely inadequate data and forecasting error resulting from the chosen method. If the EDS is used on current students, the threshold should be based on the average dropout rate of previous cohorts.

When setting the threshold, one is faced with a tradeoff: lowering the threshold will reach more potential dropouts at the cost of a higher misidentification rate; alternatively, raising the threshold will increase the accuracy of the EDS while decreasing the identification rate of at-risk students. In Section 5, we present the results using two different thresholds. The "true" threshold is based on the actual number of dropouts so that the number of identified at-risk students matches the number of dropouts. The "average" threshold is based on the average dropout rate for the 2010-2012 cohorts.

4.6. PERFORMANCE

Following signal detection theory and diagnostic systems accuracy theory (Swets, 1988; Zweig & Campbell, 1993), the performance of a machine learning method can be described by its forecasting accuracy, specificity, recall, and precision (Ting, 2011; Powers, 2011). Similar to binary or binomial classification, the task is to classify elements of a given set into two groups. These can be arranged into a 2x2 contingency table or confusion matrix:

Confusion matrix

	Prediction is dropout	Prediction is graduate
Student is dropout	True positive (t_p)	False negative (f_n)
Student is graduate	False positive (f_p)	True negative (t_n)

For our purposes, a correctly predicted graduate is a student who is correctly rejected as an at-risk student, i.e., a true negative. Consequently, a correctly predicted dropout is correctly identified as an at-risk student, i.e., a true positive. Derived from the confusion matrix, we define our measures of forecasting quality as follows:

Accuracy: $\frac{t_p+t_n}{t_p+f_p+f_n+t_n}$

Precision: $\frac{t_p}{t_p+f_p}$

Recall (sensitivity or true positive rate): $\frac{t_p}{t_p+f_n}$

Specificity (true negative rate): $\frac{t_n}{t_n+f_p}$

Since the aim of the EDS is to identify students at risk, in the present study, besides accuracy, both recall and precision are of particular relevance. Specificity measures the accuracy with which graduates are identified, and, therefore, is not as informative for the purpose of the present study.

Recall, also known as sensitivity or true positive rate, measures how many of the at-risk students are identified, while the precision, also known as the positive predictive value, measures how many of the identified students are, in fact, at risk. Since the true identification threshold is set such that the predicted dropout rate equals the known dropout rate in the test cohort, it follows that the number of false negatives equals the number of false positives, thus $f_p = f_n$. As a result, precision and recall are identical with the true identification threshold. In the next section, we focus on accuracy and recall only.

Following Bowers et al. (2013), we further illustrate the diagnostic quality of our classifiers by plotting the Receiver Operating Characteristics (*ROC*) curve. The *ROC* curve represents specificity and recall in a coordinate system, where recall is plotted on the y-axis and one minus the specificity on the x-axis. Hence the *ROC* curve depicts relative trade-offs between true positives and false positives. For example, the best possible prediction method would yield the point, $(x, y) = (0,1)$, signifying 100% recall (no false negatives) and 100% specificity (no false positives). A random guess is represented by the 45° line (50% false negatives and 50% false positives). The closer the *ROC* curve is to the upper left corner, the higher the overall accuracy of the test (Zweig & Campbell, 1993). In addition, to present the classification model´s performance in one single scalar value, we use the area under the *ROC* curve (*AUC ROC*). The *AUC ROC* is between 0 and 1, although areas under 0.5 are below the 45° reference line and imply less accurate predictions than random guessing (Bradley, 1997).

5. RESULTS: FORECASTING STUDENT DROPOUT

5.1. REGRESSION RESULTS

In order to get a better understanding of the data and the factors that are related to dropout, we show and discuss the results of the pooled sample logit model. The models discussed are strictly descriptive. Table 4a shows the results of the logit models using the student performance and demographic data from the first four semesters of the SU (cf. Table 4a, columns 1 to 5). Note that we only use the training data and report the odds ratios; we keep the specifications of the models simple, as we want to point out correlations in the data between the dependent and explanatory variables so as to find good predictors for student dropout. More sophisticated modeling to identify causal effects is beyond the scope of the current paper. However, universities considering whether or not to implement an EDS need to be aware that the results do not have a causal interpretation (Zafar et al., 2017). The goal of the present paper is to combine various prediction methods and to build a self-adjusting turnkey application. The

Table 4a: Effects of performance and demographic variables on dropout prediction (SU)

Dependent variable: student drops out(1=yes; 0=no); logistic regression (odds ratio)					
	(1) Enrollment	(2) 1st semester	(3) 2nd semester	(4) 3rd semester	(5) 4th semester
Gender (1=male; 0=female)	1.612** (0.000)	1.309** (0.000)	1.313** (0.000)	1.288** (0.000)	1.200* (0.035)
Age at enrollment	1.076** (0.000)	1.048** (0.000)	1.070** (0.000)	1.046** (0.000)	1.077** (0.000)
First generation immigrant (1=yes; 0=no)	1.462** (0.000)	1.083 (0.222)	1.054 (0.300)	1.113 (0.959)	1.072 (0.860)
Second generation immigrant (1=yes; 0=no)	1.222** (0.000)	1.074 (0.410)	1.082 (0.661)	1.004 (0.429)	1.019 (0.661)
City of entrance qualification (1=city of university; 0=else)	1.337** (0.000)	1.084 (0.170)	1.143+ (0.076)	1.199* (0.036)	1.332** (0.005)
Univ. of Appl. Sciences entrance qualification (1=yes; 0=no)	2.033** (0.003)	1.611+ (0.085)	1.132 (0.683)	0.878 (0.714)	0.888 (0.776)
Restricted university entrance qualification (1=yes; 0=no)	0.822 (0.432)	1.173 (0.612)	0.891 (0.738)	0.897 (0.786)	0.903 (0.828)
Foreign university entrance qualification (1=yes; 0=no)	1.055 (0.695)	0.869 (0.387)	0.944 (0.769)	0.887 (0.589)	0.968 (0.898)
Grade of university entrance Qualification	1.368** (0.000)	1.056+ (0.081)	1.000 (0.996)	0.947 (0.220)	0.944 (0.281)
Health insurance (1=private; 0=public)	0.875+ (0.097)	0.694** (0.000)	0.775* (0.042)	0.670** (0.004)	0.718* (0.044)
# of enrolled study programs	0.879** (0.000)	0.934** (0.000)	0.978 (0.292)	1.048+ (0.070)	1.087** (0.005)
Lateral entrants	0.399** (0.000)	0.570** (0.000)	0.474** (0.000)	0.498** (0.000)	0.462** (0.000)
# of semesters at prev. university	1.089** (0.000)	1.061** (0.000)	1.059** (0.000)	1.047** (0.000)	1.030+ (0.052)
Average grade current semester		1.658** (0.000)	1.381** (0.000)	1.193** (0.006)	1.352** (0.000)
Average CPs current semester		0.948** (0.000)	0.932** (0.000)	0.936** (0.000)	0.942** (0.000)
No exam taken current semester		17.142** (0.000)	5.988** (0.000)	3.470** (0.000)	4.180** (0.000)
# of exams current semester not participated in		1.464** (0.000)	1.272** (0.000)	1.183** (0.000)	0.993 (0.892)
# of failed exams current semester		1.380** (0.000)	1.252** (0.000)	1.175** (0.000)	1.274** (0.000)
Constant	0.231** (0.000)	0.177** (0.000)	0.120** (0.000)	0.262** (0.000)	0.175** (0.000)
Performance previous semesters:					
Previous average grades & CP			YES	YES	YES
Previous # of exams			YES	YES	YES
Prev. # of not participated. exams			YES	YES	YES
Previous # of failed exams			YES	YES	YES
Important Exams		YES	YES	YES	YES
AIC	15,657.54	11,988.42	7,448.05	5,757.76	4,178.26
N	12,728	12,728	9,228	8,015	6,693

*Notes: +p<0.1. *p<0.05. **p<0.01. Standard errors in parentheses.*

binary dependent variable has a value of 0 for graduation and 1 for dropout. Recall that "dropouts" are students who leave the university without a degree; information on whether the student continues her studies at another university immediately after dropping out or at a later date is not available. The number of observations in Table 4a drops by 47% from the first (12,728) to the fourth semester (6,693) due to students dropping out. It follows that the coefficients in the columns are not directly comparable, as the size and composition of the sample change every semester.

We look first at the fit of the regression model as described by the Akaike information criterion (*AIC*). Using only the demographic information available at the time of enrolment, the *AIC* is 15,658 (Table 4a, column 1). Incorporating the performance data from the first semester reduces the *AIC* to11,988 (Table 4a, column 2). The *AIC* drops to 7,448 in the second semester and 4,178 in the fourth semester. Thus, as expected, the fit of the model improves with progressive semesters.

Note that the estimates in column (1) only reflect the demographic variables, i.e., information that is available at the time of enrollment. At enrollment, males have a 60% higher chance of dropping out than females. Age at enrollment shares a positive correlation with dropping out. At the time of enrollment, immigrants have a higher dropout risk as compared to native students (baseline category), and first-generation immigrants have a higher dropout risk than second-generation immigrants. The rate of dropping out at the time of enrollment is 22% higher for first-generation immigrants and 46% higher for second-generation immigrants. Students with a high school degree that affords them entrance into a university of applied sciences (Fachhochschulreife) are less likely to graduate as compared to students with a general university entrance qualification (Allgemeine Hochschulreife). The effect from the grade of the high school degree (Abiturnote) is negative and statistically significant.[3] The coefficient on the dummy variable for private health insurance is only marginally significant. And, it is significantly more likely that lateral entrants graduate at a SU.

Most of the demographic variables lose statistical significance when controlling for the performance data available after the first semester. Gender, for instance, has a sizable and significant effect in column (1), the effect becomes smaller over time and in column (5), using the information from the 4th semester, the effect is substantially diminished. In semesters 5 and 6 (not reported here), the effect is even smaller and insignificant. Note that immigration status is no longer significant once achievement data becomes available; the magnitude of deviation in precision caused by imprecisely estimated effects is also reduced. Thus, the rich student data

[3] In the German grading system (school and tertiary education), the grading scale ascends from highest to lowest in achievement, i.e., a 1 is excellent and a 5 indicates failure.

Table 4b: Effects of performance and demographic variables on dropout prediction (PUAS)

Dependent variable: student drops out (1=yes;0=no); logistic regression (odds ratio)					
	(1) Enrollment	(2) 1st semester	(3) 2nd semester	(4) 3rd semester	(5) 4th semester
Gender	1.576^{**}	1.279^{**}	1.064	0.968	0.865
(1=male; 0=female)	(0.000)	(0.002)	(0.556)	(0.808)	(0.387)
Age at enrollment	1.045^{**}	1.030^{**}	1.031^{**}	1.035^{*}	1.072^{**}
	(0.000)	(0.001)	(0.006)	(0.016)	(0.000)
First generation immigrant	1.602^{**}	0.904	0.795	0.827	0.763
(1=yes; 0=no)	(0.976)	(0.002)	(0.001)	(0.000)	(0.045)
Second generation immigrant	1.003	0.777^{**}	0.692^{**}	0.569^{**}	0.712^{*}
(1=yes; 0=no)	(0.001)	(0.387)	(0.139)	(0.323)	(0.248)
City of entrance qualification	1.197*	1.006	0.868	0.892	0.856
(1= city of university; 0=else)	(0.045)	(0.932)	(0.138)	(0.342)	(0.306)
Univ. of appl. sciences entrance	2.136^{**}	1.314^{**}	1.122	1.194	1.040
qualification (1=yes; 0=no)	(0.000)	(0.000)	(0.228)	(0.139)	(0.795)
Restricted university entrance	3.148^{**}	2.143^{**}	1.865^{*}	1.980^{+}	1.922
qualification (1=yes; 0=no)	(0.000)	(0.000)	(0.035)	(0.083)	(0.199)
Foreign university entrance	5.365^{**}	3.388^{**}	1.414	0.788	0.320^{*}
qualification (1=yes; 0=no)	(0.000)	(0.000)	(0.280)	(0.544)	(0.017)
Lateral entrants	1.468^{**}	1.345^{**}	1.212^{**}	1.142^{*}	1.110
	(0.000)	(0.000)	(0.000)	(0.046)	(0.202)
Average grade current semester		2.141^{**}	1.357^{**}	1.352^{**}	0.999
		(0.000)	(0.001)	(0.010)	(0.995)
Average CPs current semester		0.974^{**}	0.901^{**}	0.893^{**}	0.915^{**}
		(0.008)	(0.000)	(0.000)	(0.000)
No exam taken current semester		17.281^{**}	10.948^{**}	3.620^{**}	2.645^{*}
		(0.000)	(0.000)	(0.002)	(0.038)
# of exams current semester not		1.333^{**}	1.281^{**}	1.140^{*}	1.330^{**}
participated in		(0.000)	(0.000)	(0.027)	(0.000)
# of failed exams current semester		1.339^{**}	1.082^{*}	1.108^{*}	1.230^{**}
		(0.000)	(0.040)	(0.024)	(0.000)
Constant	-0.055	-0.065^{+}	0.461^{**}	0.670^{**}	0.704^{**}
	(0.108)	(0.070)	(0.000)	(0.000)	(0.000)
Type of study program	YES				
Previous performance:					
Previous average grades & CP			YES	YES	YES
Previous without exam			YES	YES	YES
Prev. # of exams not participated in			YES	YES	YES
Previous # of failed exams		YES	YES	YES	YES
Important exams			YES	YES	YES
AIC	3,983.58	6,108.78	3,719.80	2,525.38	1,703.21
N	7,077	7,077	6,329	5,847	5,448

*Notes: +p<0.1. *p<0.05. **p<0.01. Standard errors in parentheses.*

available at the time of enrollment is only valuable for identifying at-risk students at the very beginning of their studies, since as early as after the first semester, performance data picks up the most relevant information (Stinebrickner & Stinebrickner, 2012; 2014). One exception is the dummy variable for private health insurance, where a significant correlation is still estimated to be in the 4th semester. Controlling for academic performance, students who have private health insurance are more likely to graduate than those who have public health insurance. As stated above, students with private insurance are more likely to come from high-income families or have parents who are civil servants. Thus, even controlling for academic performance, family background partly explains dropping out even in later semesters. The performance variables (Average Grade, No Exam, Not Participated, and Failed Exam) are negatively associated with study success (columns 2-7). Not surprisingly, failed exams and non-participation in exams are good predictors for dropouts. Note that the explanatory power of the performance indicators is decreasing in successive semesters, but academic performance in previous semesters continues to have explanatory power in later semesters (results are not reported in the Tables). This is even true for performance variables in the first semester. Thus, students who do not drop out after having performed poorly in the first semester still face a higher probability of not finishing their studies. In addition, the number of credit points (CP) is also a statistically significant predictor of the dependent variable.

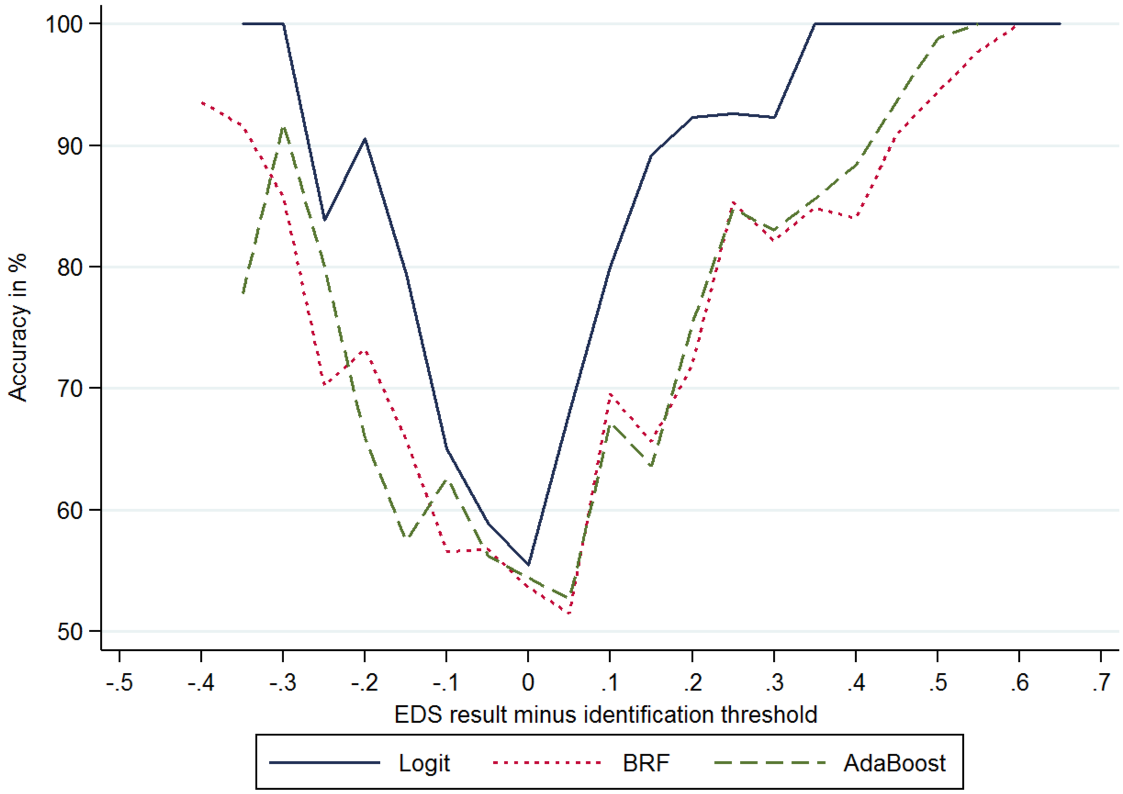
Table 4b shows the results of the logit estimation at the PUAS. The number of observations drops from 7,077 in the first semester to 5,448 students in the fourth semester. Similar to the SU, the model fit improves with consecutive semesters. A presumption is that the tuition fees—that are absent at the SU—accelerate the decision to drop out. The results are comparable with the results from the SU—especially with regard to the strength and direction of the coefficients on the performance-related data.

5.2. ACCURACY OF CLASSIFIERS

Before we describe the results for the different classifiers, Figure 1 shows the forecast accuracy of the logit, BRF, and AdaBoost models for the SU. The results for the PUAS are very similar and not reported. Each method estimates a dropout probability for each student between 0 (graduate) and 1 (dropout). Forecasted dropouts with probabilities close to 0 or 1 are accurate. Forecasts close to the identification threshold are uncertain. Figure 1 illustrates the accuracy. As expected, the proportion of correct predictions is decreasing as the identification threshold approaches the true threshold. This is true for all classifiers; however, the AdaBoost outperforms the logit and the random forest, albeit not over the entire range of observations.

Furthermore, the predictive accuracy is tied to the underlying dropout rate, which determines the threshold. As explained above, we work with two thresholds. First, to test the accuracy of our procedure, we use the rate of actual dropouts per cohort in each semester to define the true threshold. Second, to simulate the performance of the EDS in a more realistic setting, we use the number of dropouts in the previous cohorts to define the average threshold. Whether the threshold is set using the dropout rate per cohort in each semester or the average dropout rate of

Figure 1: Accuracy of the EDS



previous cohorts (some threshold must be set when implementing the EDS for use in intervention at universities), is of consequence. Accuracy is increasing as the dropout rates increase or decrease, as it becomes easier to classify observations correctly. Thus, the accuracy of the EDS depends on the dropout rate of the university and the cohort. Since dropout rates tend to decrease over time—as dropouts leave their university studies earlier than graduates—the accuracy of prediction is also expected to increase with time. And in fact, accuracy does increase with successive semesters at both universities.

5.2.1. Logit model

Table 5a summarizes the forecasting quality measures of the logit model. As expected, the quality of prediction increases over time. This applies to all quality measures. For instance, recall (how many of the at-risk students are identified) at the SU increases from about 71% in the first semester to 80% in the fourth semester. At the PUAS, recall for the 1st and 4th semesters was 69% and 78%, respectively.

Table 5a: Performance of the logit model based on the dropout rate of the test cohort

Logit	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy[a]	63.47	76.15	81.80	86.55	89.56	66.53	82.67	88.47	91.36	93.76
Recall[a]	61.67	70.80	74.05	78.71	79.94	48.95	68.68	74.30	76.68	78.26
Number of graduates	1,112	1,039	1,027	1,015	992	976	976	974	969	961
Number of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	724	830	883	916	922	733	860	902	918	926
Incorrectly predicted dropouts	388	209	144	99	70	243	116	72	51	35
Correctly predicted dropouts	626	514	411	377	279	233	261	211	171	126
Incorrectly predicted graduates	389	212	144	102	70	243	119	73	52	35
Correctly predicted graduates[a]	65.11	79.88	85.98	90.25	92.94	75.10	88.11	92.61	94.74	96.36
Incorrectly predicted dropouts[a]	34.89	20.12	14.02	9.75	7.06	24.90	11.89	7.39	5.26	3.64
Correctly predicted dropouts[a]	61.67	70.80	74.05	78.71	79.94	48.95	68.68	74.30	76.68	78.26
Incorrectly predicted graduates[a]	38.33	29.20	25.95	21.29	20.06	51.05	31.32	25.70	23.32	21.74

Notes: [a] In percent.

Table 5b: Performance of the logit model based on the average dropout rate

Logit	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy[a]	63.19	72.75	76.36	81.33	81.73	66.87	80.53	86.96	91.11	93.05
Recall[a]	69.06	83.61	87.57	90.81	91.69	48.95	72.37	78.17	79.37	81.37
Number of graduates	1,112	1,039	1,027	1,015	992	976	976	974	969	961
Number of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	688	699	736	785	779	738	817	872	909	913
Incorrectly predicted dropouts	424	340	291	230	213	238	159	102	60	48
Correctly predicted dropouts	743	631	499	442	323	233	275	222	177	131
Incorrectly predicted graduates	272	95	56	37	26	243	105	62	46	30
Correctly predicted graduates[a]	61.87	67.28	71.67	77.34	78.53	75.61	83.71	89.53	93.81	95.01
Incorrectly predicted dropouts[a]	38.13	32.72	28.33	22.66	21.47	24.39	16.29	10.47	6.19	4.99
Correctly predicted dropouts[a]	73.20	86.91	89.91	92.28	92.55	48.95	72.37	78.17	79.37	81.37
Incorrectly predicted graduates[a]	26.80	13.09	10.09	7.72	7.45	51.05	27.63	21.83	20.63	18.63

Notes: [a] In percent.

As explained above, both accuracy and recall are based on an identification threshold that, in this paper, matches the actual rate of dropouts in the test cohort (true threshold). Therefore, the values reported here are the optimal values of accuracy and recall, conditional on the given variables and chosen method.

If the true number of dropouts is unknown—as is the case when implementing the EDS at universities—the identification thresholds have to be based on the dropout rates of previous cohorts (average threshold, cf. Table 5b). This is expected to reduce the forecast performance. In our case, accuracy (recall) for the SU ranges between 73% (84%) in the first semester and 82% (92%) in the fourth semester. There are high recall rates because the number of dropouts in the test cohort is below average. However, the high fraction of correctly identified dropouts (recall) comes at the cost of a substantial increase in false-positives, i.e., incorrectly predicted-dropouts (accuracy).

Table 6a: Performance of the BRF based on the dropout rate of the test cohort

BRF	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy[a]	65.02	78.41	82.43	86.95	88.96	63.36	82.52	88.79	91.36	93.58
Recall[a]	63.55	73.83	75.14	79.96	79.08	44.54	68.95	75.35	77.13	78.26
Number of graduates	1,112	1,039	1,027	1,015	992	976	976	974	969	961
Number of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	738	848	887	916	917	708	857	903	917	924
Incorrectly predicted dropouts	374	191	140	99	75	264	118	70	51	35
Correctly predicted dropouts	645	536	417	383	276	212	262	214	172	126
Incorrectly predicted graduates	370	190	138	96	73	268	119	71	52	37
Correctly predicted graduates[a]	66.37	81.62	86.37	90.25	92.44	72.54	87.81	92.71	94.63	96.15
Incorrectly predicted dropouts[a]	33.63	18.38	13.63	9.75	7.56	55.46	12.09	24.65	22.87	21.74
Correctly predicted dropouts[a]	63.55	73.83	75.14	79.96	79.08	44.54	68.95	75.35	77.13	78.26
Incorrectly predicted graduates[a]	36.45	26.17	24.86	20.04	20.92	27.46	31.32	7.29	5.37	3.85

Notes: [a] In percent.

5.2.2. Bagging with random forest and neural network

Next, we use machine learning methods to predict study outcomes. In line with similar analyses found in the literature, the forecasting results from the logit regression model, neural network, and random forest methods that we use are all quite similar (Tables 6a-d). Furthermore, we also

Table 6b: Performance of the BRF based on the average dropout rate

BRF	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd Sem.	4th Sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy[a]	65.35	74.84	77.24	80.79	81.43	63.71	82.45	87.52	91.44	92.87
Recall[a]	71.43	86.36	89.01	90.40	91.40	44.33	76.05	80.28	80.72	81.37
Number of graduates	1,112	1,039	1,027	1,015	992	976	976	974	969	961
Number of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	665	694	728	774	773	714	829	873	910	911
Incorrectly predicted dropouts	447	345	299	241	219	265	91	56	43	30
Correctly predicted dropouts	725	627	494	433	319	211	289	228	180	131
Incorrectly predicted graduates	290	99	61	46	30	262	147	101	59	50
Correctly predicted graduates[a]	59.80	66.79	70.89	76.26	77.92	73.16	84.94	89.63	93.91	94.80
Incorrectly predicted dropouts[a]	40.20	33.21	29.11	23.74	22.08	55.67	23.95	19.72	19.28	18.63
Correctly predicted dropouts[a]	71.43	86.36	89.01	90.40	91.40	44.33	76.05	80.28	80.72	81.37
Incorrectly predicted graduates[a]	28.57	13.64	10.99	9.60	8.60	26.84	15.06	10.37	6.09	5.20

Notes: [a] In percent.

Table 6c: Performance of the neural network based on the dropout rate of the test cohort

Neural network	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy[a]	62.53	72.75	81.54	85.27	86.35	66.67	82.49	88.70	80.00	94.76
Recall[a]	60.69	47.80	70.09	73.28	72.78	49.13	68.51	74.44	92.79	80.42
Number of graduates	1,112	1,039	1,027	1,015	992	976	976	974	969	961
Number of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	714	937	901	923	904	707	826	870	892	897
Incorrectly predicted dropouts	398	102	126	92	88	233	114	68	41	28
Correctly predicted dropouts	616	347	389	351	254	225	248	198	164	115
Incorrectly predicted graduates	399	379	166	128	95	233	114	68	41	28
Correctly predicted graduates[a]	64.21	90.18	87.73	90.94	91.13	75.21	87.87	92.75	95.61	96.97
Incorrectly predicted dropouts[a]	35.79	9.82	12.27	9.06	8.87	24.79	12.13	7.25	4.39	3.03
Correctly predicted dropouts[a]	60.69	47.80	70.09	73.28	72.78	49.13	68.51	74.44	80.00	80.42
Incorrectly predicted graduates[a]	39.31	52.20	29.91	26.72	27.22	50.87	31.49	25.56	20.00	19.58

Notes: [a] In percent.

Table 6d: Performance of the neural network based on the average dropout rate

Neural network	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy[a]	63.38	73.14	76.86	78.78	79.34	66.74	81.34	87.54	82.44	92.51
Recall[a]	69.06	80.99	84.86	87.06	83.67	48.69	74.03	80.45	91.65	81.82
Number of graduates	1,112	1,039	1,027	1,015	992	976	976	974	969	961
Number of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	647	703	745	760	772	710	791	840	874	871
Incorrectly predicted dropouts	465	336	282	255	220	235	94	52	36	26
Correctly predicted dropouts	701	588	471	417	292	223	268	214	169	117
Incorrectly predicted graduates	314	138	84	62	57	230	149	98	59	54
Correctly predicted graduates[a]	58.18	67.66	72.54	74.88	77.82	75.53	84.15	89.55	93.68	94.16
Incorrectly predicted dropouts[a]	41.82	32.34	27.46	25.12	22.18	25.00	10.00	5.54	3.86	2.81
Correctly predicted dropouts[a]	69.06	80.99	84.86	87.06	83.67	48.69	74.03	80.45	82.44	81.82
Incorrectly predicted graduates[a]	30.94	19.01	15.14	12.94	16.33	50.22	41.16	36.84	28.78	37.76

Notes: [a] In percent.

confirm the superior performance of BRF. This method outperformed the others in terms of forecasting accuracy by 0.88 - 2.93% (SU) and 0.88 - 1.03% (PUAS) (Tables 6a and 6b).

Using the average identification threshold, the accuracy of the BRF (neural network) is 75% (73%) in the first semester and increases to 81% (79%) in the fourth semester. Since the dropout rate in the test cohort is below average, we expect recall to be high. Our expectations are confirmed, as recall is 86% (81%) in the first semester and 91% (83%) in the fourth semester, when using the BRF (neural network).

To further illustrate the diagnostic quality of our classifiers, we plot the ROC curves and present the *AUC ROC* in Figures 2a and 2b. First, all methods perform substantially better than a random guess. Second, prediction power improves with more information in higher semesters; the *AUC ROC* increases. In addition, the results of the methods differ slightly by university and semester. This is our motivation for combining the predictive power of neural networks, BRF, and the logit model by using the AdaBoost algorithm in Section 5.2.3.

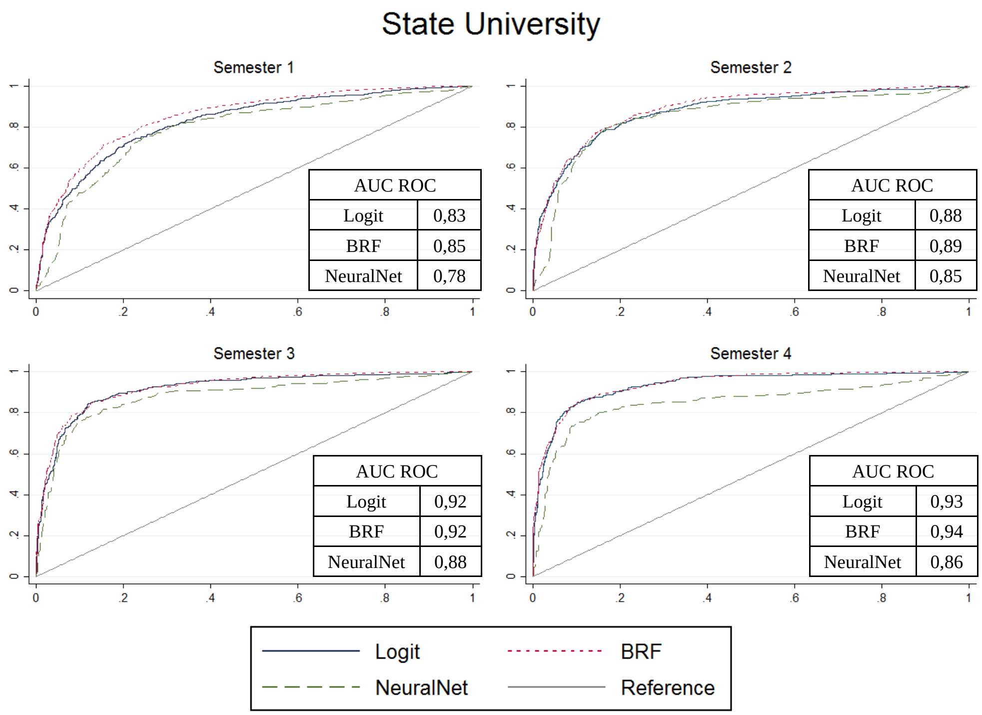


Figure 2a: *ROC* curves—State University

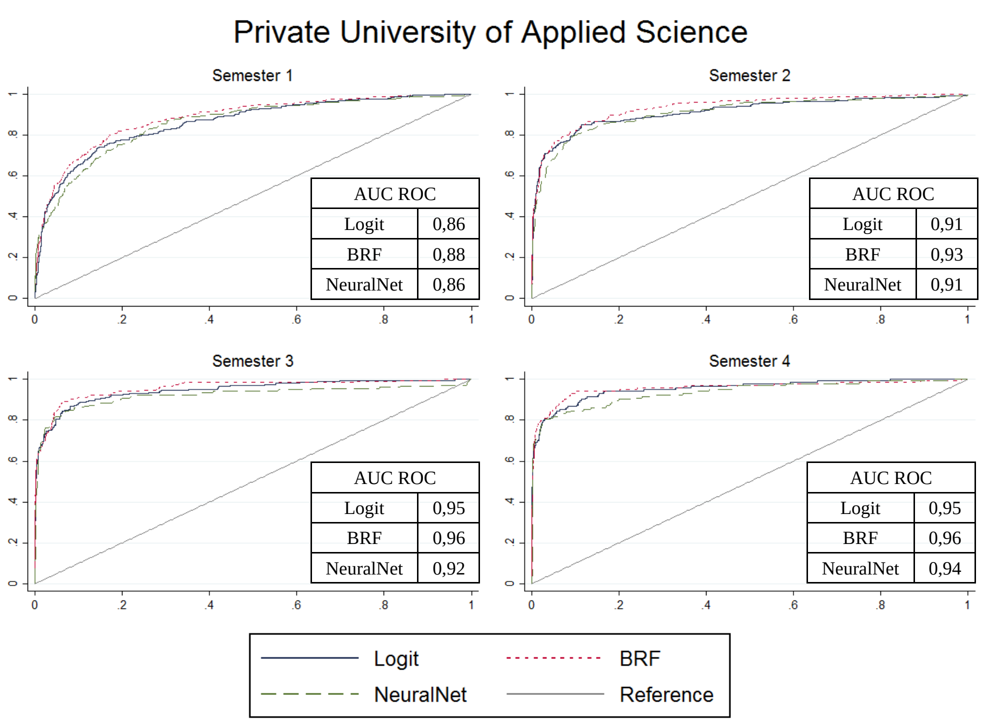


Figure 2b: *ROC* curves—Private University of Applied Science

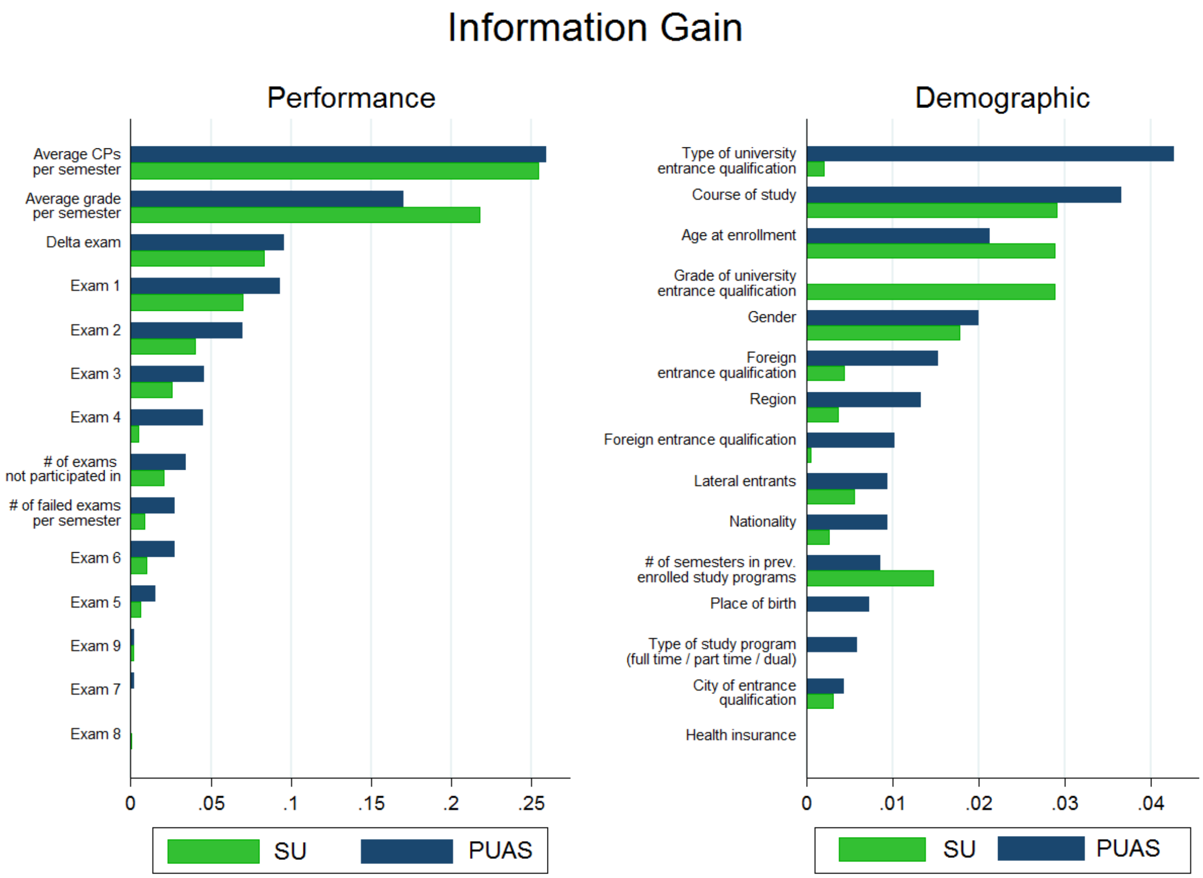


Figure 3: Information gain: BRF, first semester

While not the focus of our study, we use the values from the information gain in the random forest, using data from the first semester, to assess the relative importance of the input variables. In Figure 3, we differentiate between demographic variables (right) and performance variables (left) and the two universities.

It is apparent that performance data is better at predicting dropouts than demographic data at both universities. This confirms the results from the logit model. In particular, the pace of study (avg. CP per semester), the average grade (avg. grade per semester), as well as the most important exam, all have a high degree of explanatory power. Comparing SU and PUAS, the five most important predictor variables are identical for both universities and the difference in information gain is small.

A substantial yet expected difference between the two universities is that the variable 'type of entrance degree' is almost irrelevant at the SU with a value of 0.008, while it is the most important demographic variable at the PUAS with an information gain of 0.043.

Table 7a: Performance of the AdaBoost based on the dropout rate of the test cohort

AdaBoost	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy[a]	67.65	78.53	82.43	87.62	89.63	67.17	84.49	89.70	81.95	95.51
Recall[a]	65.81	73.83	74.95	80.58	79.94	49.78	72.10	76.69	93.50	83.22
Number of graduates	1,112	1,039	1,027	1,015	992	976	976	974	969	961
Number of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	771	850	888	923	923	711	839	876	896	901
Incorrectly predicted dropouts	341	189	139	92	69	230	101	62	37	24
Correctly predicted dropouts	668	536	416	386	279	228	261	204	168	119
Incorrectly predicted graduates	347	190	139	93	70	229	101	62	37	24
Correctly predicted graduates[a]	69.33	81.81	86.47	90.94	93.04	75.64	89.26	93.39	96.03	97.41
Incorrectly predicted dropouts[a]	30.67	18.19	13.53	9.06	6.96	24.47	10.74	6.61	3.97	2.59
Correctly predicted dropouts[a]	65.81	73.83	74.95	80.58	79.94	49.78	72.10	76.69	81.95	83.22
Incorrectly predicted graduates[a]	34.19	26.17	25.05	19.42	20.06	50.00	27.90	23.31	18.05	16.78

Notes: [a] In percent.

Table 7b: Performance of the AdaBoost based on the average dropout rate

AdaBoost	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy[a]	67.28	75.35	78.07	82.13	82.18	67.10	83.18	88.95	85.85	93.45
Recall[a]	73.20	86.91	89.91	92.28	92.55	49.34	77.35	83.83	92.88	85.31
Number of graduates	1,112	1,039	1,027	1,015	992	976	976	974	969	961
Number of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	688	699	736	785	779	712	803	848	881	876
Incorrectly predicted dropouts	424	340	291	230	213	232	82	43	29	21
Correctly predicted dropouts	743	631	499	442	323	226	280	223	176	122
Incorrectly predicted graduates	272	95	56	37	26	228	137	90	52	49
Correctly predicted graduates[a]	61.87	67.28	71.67	77.34	78.53	75.74	85.43	90.41	94.43	94.70
Incorrectly predicted dropouts[a]	38.13	32.72	28.33	22.66	21.47	24.68	8.72	4.58	3.11	2.27
Correctly predicted dropouts[a]	73.20	86.91	89.91	92.28	92.55	49.34	77.35	83.83	85.85	85.31
Incorrectly predicted graduates[a]	26.80	13.09	10.09	7.72	7.45	49.78	37.85	33.83	25.37	34.27

Notes: [a] In percent.

5.2.3. AdaBoost

Table 7 summarizes the forecast accuracy of the AdaBoost, our preferred classifier. It shows the results for the SU and the PUAS; there are noticeable differences in the levels of forecast accuracy, recall, and precision between the two institutions. However, for both institutions, prediction accuracy increases as early dropouts leave the university. Thus, not surprisingly, regular updates from end-of-semester performance data improve the prediction results.

If the identification threshold for the SU test cohort is determined based on the average dropout rate instead of the true dropout rate, the AdaBoost accuracy (recall) rate is 75% (87%) in the first semester and 82% (93%) in the fourth semester. Comparing the performance measures with the true threshold and the average threshold, it turns out that, as before, accuracy is lower and recall is higher when using the average threshold, as the dropout rate of the test cohort is below average, which leads to an over-identification of potential dropouts.

When implementing the EDS at universities, the selection of relevant information is an important issue. While machine learning can certainly handle large data sets, data cleaning is a resource-intensive task. Thus, for reasons of efficiency and, maybe more importantly, ethical concerns, it is worthwhile to think about restricting the required variables. The use of demographic data might raise concerns among the students if they are labeled as potential dropouts based upon their gender or ethnicity and irrespective of their academic potential. Moreover, data privacy rules might aggravate implementation, in particular, when the system uses demographic information. This issue cannot be fully resolved in this paper, but it remains an important question, whether demographic variables substantially improve the performance of an EDS. Here, we briefly address it by discussing the value added for the performance of the EDS over time.

First, we focus on the relevance of using information collected at the time of enrollment—namely, demographic data. In the first semester, before having taken any exams, about 21% of all dropouts in the sample left the PUAS, and 28.5% left the SU (cf. Table 7). The forecast accuracy is about 68% for both institutions but with distinct differences in the dropout detection rate. At the PUAS, successful students are better predicted than at-risk students, while at-risk students are better predicted than successful students at the SU (this pattern is consistent throughout all semesters). Alternatively, one could use performance data only (cf. Table 8).

At both universities, forecasts are only marginally enhanced when using both demographic and performance data as opposed to just performance data. Thus, the use of student demographic data is only beneficial if no performance data are available, as performance data and the demographic data are correlated. Forecasts using performance data from the first semester are only marginally enhanced by the addition of demographic data. The additional forecast accuracy gained from the demographic data is reduced with each new update from the student

Table 8: Performance of the AdaBoost using only academic performance data

AdaBoost	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy[a]	-	76.60	81.42	86.61	88.52	-	83.64	90.45	92.44	94.76
Recall[a]	-	71.49	73.51	79.12	77.94	-	69.89	78.20	79.02	80.42
Number of graduates	-	1,039	1,027	1,015	992	-	976	974	969	961
Number of dropouts	-	726	555	479	349	-	380	284	223	161
Correctly predicted graduates	-	833	880	915	915	-	836	881	890	897
Incorrectly predicted dropouts	-	207	147	100	77	-	109	58	43	28
Correctly predicted dropouts	-	519	408	379	272	-	253	208	162	115
Incorrectly predicted graduates	-	206	147	100	77	-	104	57	43	28
Correctly predicted graduates[a]	-	80.17	85.69	90.15	92.24	-	88.94	93.92	95.39	96.97
Incorrectly predicted dropouts[a]	-	19.92	14.31	9.85	7.76	-	11.60	6.18	4.61	3.03
Correctly predicted dropouts[a]	-	71.49	73.51	79.12	77.94	-	69.89	78.20	79.02	80.42
Incorrectly predicted graduates[a]	-	28.37	26.49	20.88	22.06	-	28.73	21.43	20.98	19.58

Notes: [a] In percent.

performance data following the end of a semester. This is important information when planning, for instance, interventions based on the forecasting system. Only if successful interventions take place right at the beginning of the student's career, even before students take the first exams, is demographic data an important source of information. Once performance data is available after the first semester, rich demographic data adds only little additional information to the forecasting model. After the first semester, the percentage of correctly predicted dropouts at the SU is 71% when using academic performance data only and 74% when using demographic and achievement data.

5.2.4. Robustness of performance with respect to assignment of test data

As described above, data from students having matriculated in 2010 (SU) and 2012 (PUAS) was used as test data. We selected these years to ensure that the vast majority of students had either completed their studies or dropped out. In later years, the proportion of students still enrolled increases, as Table 9 shows. The differences between the SU and the PUAS are remarkable. Students at the SU are enrolled for much longer than students at the PUAS. The proportion of students who enrolled in a (3 year) bachelor program in the 2010 academic year and who were still enrolled during the 2017/18 Winter Term is less than 3% at the PUAS and almost 12% at

Table 9: Proportion of students in an academic year who were still enrolled in the Winter Term of 2017/2018

Year	SU	PUAS
2009	8.23%	2.63%
2010	11.74%	2.42%
2011	16.26%	4.76%
2012	20.56%	6.94%
2013	30.57%	16.07%
2014	48.38%	35.72%
2015	58.79%	64.12%
2016	76.96%	83.39%
2017	97.72%	95.27%

the SU. The difference is at least partly explained by student fees. While student fees are in place for most study programs at the PUAS, there have been no fees at SUs since the Winter Term of 2011/12. Hence aside from administration fees, enrollment is free and offers some benefits for students.

In Table 10, we summarize the accuracy and recall results from using the BRF on our test data. Testing with data from later years tends to improve performance, in particular, recall; this is especially so for early semesters at the SU. This is explained by the composition of semester cohorts in later years. Recall that only dropouts and graduates are included in the data, and the frequency of dropouts is decreasing with progressive semesters. Therefore, later semester cohorts are invariably defined as having fewer dropouts. Other students of that cohort are still matriculated and not included in the data. Thus, our predictor performs well, as it only has to predict students at the tails of the distribution: early dropouts and fast graduates.

Table 10: Performance of BRF—sensitivity to test data

BRF	State University					Private University of Applied Sciences				
	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.	Enroll-ment	1st sem.	2nd sem.	3rd sem.	4th sem.
Accuracy(2010)[a]	65.02	78.41	82.43	86.95	88.96	61.75	82.03	88.00	90.28	92.89
Accuracy(2011)	59.35	81.07	83.99	87.63	88.71	62.31	80.46	87.45	90.37	93.92
Accuracy(2012)	69.07	83.02	86.90	89.25	90.19	63.36	82.52	88.79	91.36	93.58
Accuracy(2013)	74.41	86.53	88.65	89.57	89.86	64.80	82.78	88.93	92.99	95.98
Recall(2010)	63.55	73.83	75.14	79.96	79.08	42.02	65.69	69.08	67.97	69.91
Recall(2011)	66.88	81.73	82.01	84.80	83.22	39.90	61.54	69.23	70.93	77.94
Recall(2012)	76.79	86.18	87.39	88.54	87.06	44.54	68.95	75.35	77.13	78.26
Recall(2013)	84.09	91.15	91.33	91.28	89.50	50.32	71.07	75.00	80.43	84.92

Notes: [a]The year denotes the matriculation year of the test cohort. All results in percent.

6. CONCLUSIONS

University attrition is an important issue for education policy. Student attrition is costly for all parties involved; resources spent on educating students and the effort and time spent by the student in the university system are both of limited economic value when not accompanied by a graduating certificate. Thus, it is in everybody's interest to optimize (prevent or speed up) student attrition through diagnosis and intervention. Due to a very liberal enrollment policy, the absence of a protocol for monitoring student progress, and the generous financial incentives for students studying and predicting dropout phenomenon at German universities is interesting from a scientific as well as a policy point of view. There is a considerable relevance for federal and state policy; however, at the university level, there is also considerable relevance, as resources are allocated between universities and within universities based on enrollment and the graduation of students. This paper develops and tests a forecasting system for the early detection of university dropouts. The forecasting system is based on administrative data available at German universities; it is self-adjusting and can be used to identify students at risk and to allocate students into support interventions at universities.

In addition to using traditional regression analyses to predict dropouts, we also employ machine learning algorithms that do not rely on complex model building and self-adjust to newly available data. Instead of relying on a single method, we use the AdaBoost algorithm to combine the various methods employed. This reduces the disadvantages inherent in using any single method as well as those caused by the heterogeneity of study programs and student body compositions at the different universities.

In the present paper, we use data from a state and private university to develop and test an EDS model. The predictive power of our preferred method, the AdaBoost, is strong. The accuracy of the results improves with increasing semesters. However, our analysis shows that time—progressive semesters—plays an important role in the prediction accuracy of the EDS. Using only demographic data available at enrollment, our early detection system already correctly predicts 67% of dropouts at the SU; prediction accuracy increases to 80% in the fourth semester. The corresponding numbers for the PUAS are 50% at the time of enrollment and 83% in the fourth semester. Moreover, using the rich demographic data available does not substantially improve prediction accuracy once performance data becomes available. Thus, demographic data is only relevant for the EDS at enrollment and the first semester. In future research, this issue needs to be analyzed in more detail.

The advantage of the EDS presented is that after having identified students at risk, it can serve as a basis for an early intervention system to either prevent dropouts or to even speed up a student's decision to drop out. In this way, the public and private costs associated with attrition can be reduced by implementation of an EDS as a starting point for allocating intervention support to students at-risk and for testing the effectiveness of student intervention.

References

Arcidiacono, P., Aucejo, E., Maurel, A. & Ransom, T. (2016) College attrition and the dynamics of information revelation. *NBER Working Papers - National Bureau of Economic Research*.

Arnold, K. & Pistilli, M. (2012) Course signals at Purdue: using learning analytics to increase student success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 267-70.

Asif, R., Merceron, A., Ali, S.A. & Haider, N.G. (2017) Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 177-94.

Baker, R.S.J.d. (2010) Statistical data mining tutorials. In B. McGaw, P. Peterson & E. Baker, eds. *International Encyclopedia Of Education*. UK: Elsevier. 112-18.

Baker, R.S.J.d. & Yacef, K. (2009) The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-16.

Barber, R. & Sharkey, M. (2012) Course correction: using analytics to predict course success. *LAK '12 Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 259-62.

Barefoot, B.O., Gardner, J.N., Cutright, M., Morris, L.V., Schroeder, C.C., Schwartz, S.W., Siegel, M.J. & Swing, R.L. (2005) *Achieving And Sustaining Institutional Excellence for the First Year of College*. San Francisco, CA.: Jossey-Bass.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T. & Popelinsky, L. (2012) Predicting drop-out from social behaviour of students. *Proceedings of the 5th International Conference on Educational Data Mining* , 103-09.

Bean, J.P. (1983) The application of a model of turnover in working organizations to the student attrition process. *The Review of Higher Education*, 6, 129-48.

Berger, M., Galonska, C. & Koopmans, R. (2004) Political integration by a detour? Ethnic communities and social capital of migrants in Berlin. *Journal of Ethnic and Migration Studies*, 30, 491-507.

Bound, J., Lovenheim, M.F. & Turner, S. (2010) Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. *American Economic Journal: Applied Economics*, 2, 129-57.

Bowen, W., Chingos, M. & McPherson, M. (2009) *Crossing the finish line: Completing college at America's public universities*. Princeton: Princeton University Press.

Bowers, A.J., Sprott, R. & Taff, T.A. (2013) Do we know who will drop out? A review of the predictors of dropping out of high school: precision, sensitivity and specificity. *The High School Journal*, 77-100.

Bradley, A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145-59.

Brandstätter, H. & Farthofer, A. (2002) Studienerfolgsprognose – konfigurativ oder linear additiv? [Predicting student success – Configurational or linear additive?]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 23, 381-91.

Brandstätter, H., Grillich, L. & Farthofer, A. (2006) Prognose des Studienabbruchs [Predicting Student Dropout]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 38, 121-31.

Breimann, L. (1996) Bagging predictors. *Machine Learning*, 24(2), 123-40.

Breimann, L. (2001) Random forests. *Machine Learning*, 45(1), 5-32.

Burrus, J., Elliott, D., Brennemann, M. & Markle, R. (2013) Putting and keeping students on track: toward a comprehensive model of college persistence and goal attainment. *ETS Research Report Series*.

Danilowicz-Gösele, K., Lerche, K., Meya, J. & Schwager, R. (2017) Determinants of students success at university. *Education Economics*, 25(5), 513-32.

Dekker, G.W., Pchenenizkiy, M. & Vleeshouwers, J.M. (2009) Predicting students drop out: a case study. In T. Barnes, M. Desmarais, C. Romero & S. Ventura, eds. *Proceedings of the 2nd International Conference on Educational Data Mining*. Cordoba, Spain. 41-50.

Feng, M., Heffernan, N. & Koedinger, K.R. (2006) Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. *International Conference on Intelligent Tutoring Systems*, 31-40.

Frank, E., Hall, M.A. & Witten, I.H. (2016) *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann.

GAEBEL, M., HAUSCHILDT, K., MÜHLECK, K. & SMIDT, H. (2012) Tracking learners' and graduates' progression paths. TRACKIT. *EUA Publications*.

GLEASON, P. & DYNARSKI, M. (2002) Do we know whom to serve? Issues in using risk factors to identify dropouts. *Journal of Education for Students Placed At Risk*, 25-41.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, R. & WITTEN, I. (2009) The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1), 10-18.

HEUBLEIN, U. (2014) Student drop-out from German higher education institutions. *European Journal of Education. Research, Development and Policy*, 49(4), 497-513.

HEUBLEIN, U. & BURKHART, S. (2013) Bildungsinländer 2011 - daten und fakten zur situain von von ausländischen studierenden [Educational natives 2011 – Data and facts describing the situation of international students]. Bonn.

HUMPERT, A. & SCHNEIDERHEINZE, K. (2002) *Stichprobenziehung für telefonische Zuwandererumfragen. Praktische Erfahrungen und Erweiterung der Auswahlgrundlage. [Sampling for telephone surveys of immigrants. Experience and broadening of the sampling frame]*. Münster: Waxmann.

JETTER, T. (2017) *Membrain NN* [Online]. Available at: https://www.membrain-nn.de/ [Accessed 03 April 2019].

KEMPER, L., VORHOFF, G. & WIGGER, B.U. (2018) *Predicting student dropout: A machine learning approach* [Online]. Available at: https://www.researchgate.net/publication/322919234_Predicting_Student_Dropout_a_Machine_Learning_Approach [Accessed 17 Juli 2019].

KNOWLES, J. (2015) Of needles and haystacks: building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, 7(3), 18-67.

KOTSIANTIS, S.B., PIERRAKEAS, C.J. & PINTELAS, P.E. (2003) Preventing student dropout in distance learning - using machine learning techniques. *Knowledge-Based Intelligent Information and Engineering Systems: 7th International Conference, KES 2003, Oxford, UK, Proceedings, Part II*, 267-74.

LARSEN, M.L., KORNBECK, K.P., KRISTENSEN, R.M., LARSEN, M.R. & SOMMERSEL, H.B. (2013) Dropout phenomena at universities: What is dropout? Why does dropout occur? What can be done by the universities to prevent or reduce it? A systematic review. *Danish Clearinghouse for Educational Research*.

MICHAEL, J. (2007) Anredebestimmung anhand des Vornamens [Determination of salutation by first name]. *c´t*, 17/2007, 182-83.

MICHAEL, J. (2016) Name Quality Pro (to be published). *(available from the author; mail to: namequality.pro@gmail.com)*.

MINAEI-BIDGOLI, B., KORTEMEYER, G. & PUNCH, W.F. (2004) Enhancing online learning performance: An application of data mining methods. *Proceedings of the Seventh IASTED International Conference on Computers and Advanced Technology in Education*, 173-178.

MUCHERINO, A., PAPAJORGJI, P.J. & PARDALOS, P.M. (2009) k-nearest neighbor classification. *Data Mining in Agriculture. Springer Optimization and Its Applications*, 34, 109-13.

NGHE, N.T., JANECEK, P. & HADDAWAY, P. (2007) A comparative analysis of techniques for predicting academic performance. *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, FIE´0, 37th Annual IEEE*, 7-12.

OECD. (2016) Immigrant background, student performance and students' attitudes towards science. In *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing.

OECD. (2018) *Equity in Education: Breaking Down Barriers to Social Mobility*. Paris: OECD Publishing.

PASCARELLA, E.T. & TERENZINI, P.T. (1979) Interaction effects in Spady’s and Tinto’s conceptual models of college dropout. *Sociology of Education*, 52, 197-210.

PEÑA-AYALA, A. (2014) Review: Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 1432-62.

POWERS, D.M.W. (2011) Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning* , 2(1), 37-63.

ROMERO, C. & VENTURA, S. (2010) Educational data mining: A review of the state of the art. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, 601-18.

SAMMUT, C. & WEBB, G. (2017) *Encyclopedia of Machine Learning and Data Mining*. New York: Springer US.

SANTANA, M., COSTA, E., NETO, B., SILVA, I. & REGO, J. (2015) A predicitive model for identifying students with dropout profiles in online courses. *Workshops Proceedings of EDM 2015: 8th International Conference on Educational Data Mining*, CEUR Workshop Proceedings 1446.

SARA, N.-B., HALLAND, R., IGEL, C. & ALSTRUP, S. (2015) High-school dropout prediction using machine learning: A Danish large-scale study. *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence*, 319-24.

SCHAPIRE, E. & FREUND, Y. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Science*, 55, 119-39.

SCHAPIRE, R.E. & FREUND, Y. (2012) *Boosting - Foundations and Algorithms*. Massachusetts: Institute of Technology.

SCHULZE-STOCKER, F., SCHÄFER-HOCK, C. & PELZ, R. (2017) Weniger Studienabbruch durch Frühwarnsysteme – Das Beispiel des PASST?-Programms an der TU Dresden. [Reducing the number of student dropouts using early warning systems - An example of the PASST? programme at TU Dresden]. *Zeitschrift für Beratung und Studium*, 26-32.

STATISTISCHES BUNDESAMT. (2015) Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund – Ergebnisse des Mikrozensus 2015 [Population and employment. Immigrants – Results from the Mikrozensus].

STINEBRICKNER, T. & STINEBRICKNER, R. (2008) The effect of credit constraints on the college drop-out decision: A direct approach using a new panel study. *American Economic Review*, 98, 2163-84.

STINEBRICKNER, T. & STINEBRICKNER, R. (2012) Learning about academic ability and the college dropout decision. *Journal of Labor Economics*, 32, 707-48.

STINEBRICKNER, T. & STINEBRICKNER, R. (2013) A major in science? Initial beliefs and final outcomes for college major and dropout. *Review of Economic Studies*, 81, 426-72.

STINEBRICKNER, T. & STINEBRICKNER, R. (2014) Academic performance and college dropout: using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*, 32, 601-44.

STRECHT, P., CRUZ, L., SOARES, C., MENDES-MOREIRA, J. & ABREU, R. (2015) A comparative study of classification and regression algorithms for modelling students' academic performance. *Proceedings of the 8th International Conference on Educational Data Mining*, 392-95.

SWETS, J.A. (1988) Measuring the accuracy of diagnostic systems. *American Association for the Advancement of Science*, 1285-93.

TING, K.M. (2011) Precision and recall. In C. Sammut & G. Webb, eds. *Encyclopedia of Machine Learning*. Springer US. 781 & 901.

TINTO, V. (1975) Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45, 89-125.

TRAPMANN, S., HELL, B., WEIGAND, S. & SCHULER, H. (2007) Die Validität von Schulnoten zur Vorhersage des Studienerfolgs - eine Metaanalyse [The validity of school grades for predicting study success - a meta-analysis]. *Zeitschrift für pädagogische Psychologie*, 21, 11-27.

WESTERHOLT, N., LENZ, L., STEHLING, V. & ISENHARDT, I. (2018) *Beratung und Mentoring im Studienverlauf - Ein Handbuch [Counseling and mentoring during your studies - A handbook.]*. Münster: Waxmann.

WIERS-JENSSEN, J., STENSAKER, B. & GROGAARD, J.B. (2002) Student satisfaction: towards an empirical deconstruction of the concept. *Quality in Higher Education*, 8, 183-95.

XENOS, M. (2004) Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education Journal*, 345-59.

YUKSELTURK, E., OZEKES, S. & TÜREL, Y.K. (2014) Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-Learning*, 118-33.

ZAFAR, M..B., VALERA, I., RDRIGUEZ, M.G. & GUMMADI, K.P. (2017) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *WWW '17 Proceedings of the 26th International Conference on World Wide Web*, 1171-80.

ZHANG, Y., OUSSENA, S., CLARK, T. & KIM, H. (2010) Use data mining to improve student retention in higher education - a case study. *Proceedings of the 12th International Conference on Enterprise Information Systems, Volume 1, DISI, Funchal, Madeira, Portugal, June 8 - 12, 2010*.

ZWEIG, M.H. & CAMPBELL, G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4), 561-77.