Jakob Lindner, BSc

Development of a Transparent System

for

Early Dropout Detection in Higher Education

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme:

Information and Computer Engineering

submitted to

Graz University of Technology

Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Institute of Interactive Systems and Data Science

Graz, Mai 2021

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date, Signature

Supervisor

Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Acknowledgements

At this point, just before my graduation, I would like to thank all those who have accompanied, motivated and supported me on all or parts of my educational path. It would be appropriate to mention them all here. This list of names would certainly be long and incomplete, as it covers a time span of almost 3 decades and, regrettably, I do not personally know many supporters due to their indirect involvement. So, I cannot begin to express my deepest thanks to them all.

Accompanying me throughout, sometimes more active sometimes less active, were my parents Gabriele and Armin Lindner. They helped me to overcome some ditches and conquered some obstacles for me. Without them, I would certainly not be in this position today. I'm deeply indebted to them.

For the direct support during the creation of this thesis, I would firstly like to thank Philipp Leitner, who played through our weekly jour fixes over 8 month a decisive role in keeping me moving forward.

I also want to thank Denis Helic, my supervisor, who offered me the opportunity of writing a highly interesting thesis and without whom this thesis would not exist.

I am also indebted to Florian Gundl, who supported me with his expertise and perspective as a sociologist and thus enabled a more holistic approach to the topic of this thesis.

I'd also like to acknowledge Ulrich Heublein, an educational researcher, who I came to know of because of his research and who took time to share his opinion on my ideas even though I was just a strange caller to him.

I would also especially like to thank my partner Victoria Obersriebnig, who supported me during the writing of this thesis and was often patient with me.

Finally, I would like to thank all my friends who have accompanied and supported me through this work.

Kurzfassung

Neben der allgemeinen Verbesserung der Studienbedingungen ist die gezielte Unterstützung von abbruchsgefährdeten Studierenden ein wirksames und kosteneffizientes Mittel zur Senkung der Abbruchsquoten an Universitäten. Für eine derartige Maßnahme ist eine Erkennung bzw. Identifizierung von abbruchsgefährdeten Studierende seitens der jeweiligen Universität unumgänglich.

Bisherige System zur frühzeitigen Erkennung von potentiellen Studienabbrechern stützen sich häufig auf Daten, die eine Universität erst beschaffen müsste oder sind algorithmisch so konzipiert, dass die Interpretierbarkeit und Nachvollziehbarkeit der Prognosen gering ist. Ersteres erschwert zeitnahen Einsatz, letzteres verhindert den Einsatz im Öffentlichen Bereich, in dem Transparenz zwingend erforderlich ist.

In dieser Arbeit wurde ein System zur frühzeitigen Erkennung von Studienabbrechern in einzelnen Studiengängen entwickelt. Das System wurde in einer Fallstudie entworfen, implementiert und evaluiert, wobei dafür nur Prüfungsdaten aus dem Informatikstudium der Technischen Universität Graz, einer österreichischen Universität, herangezogen wurden. Algorithmisch wurde das System mittels logistischer Regression umgesetzt.

Somit benötigt das entwickelte System für seine Funktion nur bereits erhobene Daten, die sich im direkten Einflussbereich einer Universität befinden und seine Ergebnisse sind zudem sehr transparent.

Bei der Vorhersage zukünftiger Studienabbrecher erreichte das System Trefferquoten von über 90%.

Damit ist es für den praktischen Einsatz an Hochschulen prädestiniert und ermöglich Universitäten, ihre Studierenden noch besser zu unterstützen und ihre Abbruchsquoten zu senken.

Abstract

For reducing dropout rates at universities, as a complement to the general improvement of study conditions, targeted support for student at risk of dropping out is an effective measure, both in terms of impact and in terms of costs. A prerequisite for universities to offer targeted support to students at risk of dropping out is the timely and reliable identification of such students.

Previous systems for the early identification of dropouts often rely on data that must first be collected from a given university or are algorithmically implemented in such a way that interpretability and comprehensibility are low. The former makes timely application difficult; the latter prevents application in the public domain where transparency is imperative.

In this thesis, a system for detecting students at risk of dropping out at the level of study programs was developed. The system was designed, implemented, and evaluated in a case study using only exam data from the Computer Science program of the University of Technology Graz, an Austrian university. Further, only logistic regression was used for the algorithmic implementation of the dropout predictions. Thus, the developed system requires only data in the direct possession of a university for its function and its results are also highly transparent. In predicting future dropouts, the system was able to achieve over 90% recall. This predestines it for practical application and equips universities with a viable tool to better support their students and reduce their dropout rates.

Table of Contents

1. Introduction

Student attrition or student dropouts is a widely recognized phenomenon and poses especially in the higher education sector a though challenge.

Consequences of student dropouts make an appearance on a wide range of different levels. For further argumentation, these levels are divided into the following areas: the individual, the institutional and on the social levels.

Students and their direct social environment are affected as individuals. According to the findings of the Austrian Student Survey 2019 students in Austria had an average monthly budget of 1216€ in 2019 . The contributions of a student’s social environment composed of 221€ monetary income from family and 128€ of benefit in kind from their parents (Unger et al., 2020, p. 390). In total this sums up to a financial aid of 349€ per month on average linked to the direct social environment of a student, which naturally induces corresponding expectations towards the respective student. Not fulfilling those and one’s own expectations can cause psychological stress. Studies do show, that dropping out negatively influences self-esteem (Hoeschler & Backes-Gellner, 2014). For the respective individual dropping out sometimes means ending up with unfulfilled educational aspirations and accompanying low self-esteem. Low self-esteem is again linked to versatile individual problems like substance abuse, aggression, membership in deviant eating disorders (Leary et al., 1995).

With an university student population size from slightly bigger than $300 \cdot 10^3$ in Austria in 2019, the privately raised financial support summed up to a total amount of more $100 \cdot 10^6$€ per month (Unger et al., 2020, p. 13). Even with an conservative assumed dropout rate of 20% by only considering early dropouts this is indisputable a non-negligible financial burden for private households (Unger et al., 2009, p. 30). If one excludes the knowledge, which a drop-out acquires up to its leaving, then this money, raised for the purpose of a completed study, must be considered lost.

On the institutional level universities are influenced by dropouts as a whole and internally at various different levels. One could categorize the consequences in academic and financial ones. Students not completing their study means the loss of academic input, potential researchers and employees, and thereby, naturally occurring synergistic effects. Thus, significant opportunity costs occur. Provided the termination was not to the advantage of the reproduction of the institution and thus in general beneficial (Bourdieu, 1988). A degree not only ensures the academic qualification, but the compliance of the receiving individual to the dominant culture of the academic field and the respective university (Ulriksen et al., 2010, p. 216). Financially the costs of care for students who later drop out are lost. Lately , dropout rates receive more attention as an indicator of the extent to which the educational mandate is being fulfilled by a given university (Gaebel et al., 2012, p. 36). Austrian legislation makes indirectly use of this indicator by linking public funding of university to their performance. Public funds recently are calculated partly based on the number of successfully completed studies (*Universitätsfinanzierung NEU*, 2017). As a result, there is increasing incentive for universities to reduce dropout rates.

Society does not remain unaffected either by dropouts. It is subject to socioeconomic consequences effecting returns to education and economic growth overall (Larsen et al., 2013, p. 36). The responsible mechanisms are indeed complex and scientific literature diverges in this regard. For the sake of comprehensibility, the nature of the arguments in this context are presented in the form of few plausible examples. The consequences are differentiated in direct and indirect ones.

An unsuccessful student directly decreases the state’s income. A higher education level is related to a higher income. Together with a progressive tax system, as Austria and many countries have, this leads to significant higher taxes and social security fees paid. On the other side, the state also has expenses for the educational system and loses the financial contribution of the individual during the time of study. According to the OECD's figures published in 2019, the state of Austria gained a net profit of approximately $322 \cdot 10^3$ \$ in the year 2016 for every university male graduate (OECD, 2019, p. 131). That is about $270 \cdot 10^3$ € lost for Austria with every unnecessary male dropout. It should be mentioned, since this is the net profit, the economical contribution of a lower degree was already taken into account.

The indirect consequences are even more complex and thus harder to anticipate. An educational gradient can be observed in various areas of society. One example is the health system. People with lower education tend to live unhealthier and cause more health expenses (Cutler & Lleras-Muney, 2010).

Also, the unemployment rate is significant lower among higher degree holders (Arbeitsmarktservice Österreich, 2019). This again means more income and less expenses for society. According to a recent research report of the institute for advanced studies in Vienna, labor market conditions are worsening for non-academics and improving for academics. This is reflected by better labor market chances today and in the future (Binder et al., 2017, p. 75).

High dropout numbers also contribute to a shortage in skilled labor, which inhibit economic growth. The Austrian Institute of Economic Research states that academics are the main generator of knowledge. Knowledge is the most important production factor in modern economies. It maintains competitiveness and helps to solve social problems. They also find that an academic degree resulted in a total return of investment of 5 - 7% in the year 2010. This makes education a serious option for investments. For comparison federal bonds only yield 3% in return in the same year (Janger et al., 2017).

For all these reasons, this paper aims to help to reduce dropout rates at universities.

Due to these serious effects of university dropouts, various efforts are being made to counteract this problem. These efforts are naturally made on every adversely effected level. Sometimes more and sometimes less systematic. Clear signs that the extent and severity of dropout related consequences has been recognized and that this issue has become a focus of society are the aforementioned profound changes in the law on university funding in Austria.

All initiatives to reduce dropout rates have in common that they must fit into the existing system. The education system creates a framework that makes certain initiatives possible and certain more difficult or even prevents them. In this way, measures are selected, and the room of maneuver is restricted. This sometimes does not meet the existing aspirations. Especially since the change in the system of university funding creates an even stronger incentive to intervene and manage dropout rates at the university level. Therefore, quickly realizable solutions are needed. However, it is not only about effectiveness and efficiency, but also about feasibility under the current restrictions.

There can be no doubt that the identification of possible dropouts is beneficial to managing dropout rates. It facilitates targeted intervention and thereby enhances effective and efficient influencing of the dropout rate. This makes it a key application of Educational Data Mining (EDM) methods, which was already understood as early as in 2009 (Baker & Yacef, 2009).

In the course of this work, a system for the early detection of dropouts is developed. This system operates under the current requirements of the Higher Education Area at university level in Austria. It is cost-efficient, adaptable, accurate and implementable at any point of time for any Austrian university. For this purpose, it uses only data in the direct possession of the respective university. This restriction opens the option of direct timely implementation without having to rely on other parties or decision makers.

Due to the legal situation and the method of analysis, the number of internal university decision makers required for implementation is also small. This enables rapid implementation and quick access to results relevant for the respective stakeholders.

Thus, this thesis introduces and provides an actionable approach for Austrian universities to act according to their aspirations and needs in terms of managing dropout rates to fulfilling their educational mandate.

The development and evaluation of the proposed approaches is based on data generated under real conditions. It consists of the examination data of the students of the study program Computer Science at Graz University of Technology (TUG). The data of both Bachelor and Master students enrolled between 2005 and 2015 are present.

To the best of the authors knowledge, there is no comparable system for Austrian universities.

Previous approaches for investigating dropouts focused purely on understanding the behavior of dropouts and not on detecting them. Usually, data of various data sources was used, the acquisition of which therefore required considerable effort. Qualitative data, but also quantitative data, were usually tailored to the respective approach and were collected specifically for it. One such an example is the article “Prognose des Studienabbruchs” (Brandstätter et al., 2006). For this study data of students from during their high school and from during their university time was collected. The number of students for whom such data could be obtained tends to be small.

The approach discussed in this thesis works with already existing administrative data of a university under its own control. This results in the advantages described above with which the needs already mentioned can be met.

2. Background

Let there be a set of objects where each object is assigned to a class. Classification is the problem of predicting for each of these objects its class based on the object’s properties. Binary classification is a special case of this problem. The objects whose classes are to be determined belong to exactly two classes.

In this work, students are to be categorized as *graduates* or *dropouts*. Therefore, the aim is to construct a binary classifier.

The way such a classifier is implemented depends very much on the overall context in which it is to operate. The context includes the problem to be solved, the data used and, above all, the qualitative understanding of the dynamics and the influencing factors that lead to a certain class membership of an object. For this reason, in addition to the discussion of binary classifications, this chapter also takes a closer look at the decision-making process that leads to dropping out and the factors that influence this process.

2.1 Evaluation of a binary classifier

If a binary classifier is defined and predictions can be made, the quality of the predictions need to be assessed.

Mathematically, a binary classifier defines a function f that maps the set of objects O to a set of classes $C = \{negative, positive\}$.

$$f: \begin{cases} O \to C \\ x \mapsto f(x) \end{cases}$$

The content of the set C can be defined arbitrarily. To keep to the common terminology, the classes are called $negative$ and $positive$. Depending on whether the class assigned by the function matches the actual class of the object, a correct or incorrect prediction results. A result can be incorrectly or correctly negative or positive. This means that the results can be divided into one of four categories: false negative, true negative, false positive and true positive. The significance of a classifier's predictions can be quantified by using the frequency of results in these categories. For this purpose, the frequencies are usually summarized in a confusion matrix. Table 1 represents such a confusion matrix for a binary classifier.

Table 1 - confusion matrix of binary classifier

		actual class	
		actual class positive	**actual class negative**
predicted class	**predicted class positive**	$count\ of\ true\ positives\ \coloneqq\ t_p$	$count\ of\ false\ positives\ \coloneqq\ f_p$
	predicted class negative	$count\ of\ false\ negatives\ \coloneqq\ f_n$	$count\ of\ true\ negatives\ \coloneqq\ t_n$

The frequency counts alone do not indicate the reliability of a classifier. For a reasonable evaluation, they must be put into context with each other. Depending on the requirements of the respective application, different key figures are relevant for the evaluation of a classifier (Powers, 2011). For the evaluation of the binary classifier, which is to be constructed in the course of this work, a set of different key figures may be considered.

In fact, performance of dropout indicators is reported inconstantly across the research community. The use of different key figures makes it difficult to compare results between different papers. Here, the recommendation from the community is followed and the following key figures are used (Bowers et al., 2013):

Table 2 - Key figures for evaluation of a binary classifier

$$accuracy \coloneqq \frac{t_p + t_n}{t_p + f_p + f_n + t_n}$$

$$Precision\ /\ positive\ predictive\ value(PPV) \colon= \frac{t_p}{t_p + f_p}$$

$$Recall\ /\ Sensitivity\ /\ true\ positive\ rate(TPR)\ \coloneqq \frac{t_p}{t_p + f_n}$$

$$Specificity\ /\ true\ negative\ rate\ \ (TNR)\colon= \frac{t_n}{t_n + f_p}$$

$$(1\ -\ Specificity)\ /\ false\ positive\ rate\ (FPR)\ \coloneqq \frac{f_p}{t_n + f_p}$$

Each key figure measures different qualities of binary classifier under test:

- **Accuracy:** Proportion of correctly predicted objects**.**
- **Precision or PPV:** Proportion of predicted positives that are actual positives. The empirical probability of a positive predicted object to be actual positive.
- **Recall or TPR:** Proportion of correctly predicted positives.
- **Specificity or TNR:** Proportion of predicted negatives that are actual negatives. The empirical probability of a negative predicted object to be actual negative.
- **(1 – Specificity) or FPR:** Proportion of incorrectly predicted negatives.

Since this thesis aims to construct a system for managing dropout rates, if the student presented to a classifier is considered a dropout, this is indicated by the $positive$ class and if the respective student is considered a graduate, this is indicated by the $negative$ class. The key figures above then measure the performance regarding indicating dropouts.

2.2 Decision threshold

Certainly, all the above-mentioned key figures are of interest to the university administration for a dropout detection system. However, TPR and FPR are of particular relevance. This is because the university can exert influence here to adapt the classifier to its needs.

Classifying systems often estimate probabilities and make their prediction regarding the class of an object accordingly. In the case of a dropout prediction system, this means that a student is assigned a probability between 0 and 1. 0 means an estimated probability of 0% and 1 means a probability of 100% for a student to be a dropout. Which probability is interpreted as belonging to a certain class is determined by the decision threshold. To select the most likely forecast, the decision threshold is usually set at 50%. If an object is estimated to have a probability greater than or equal to 0.5, the classifier would predict a dropout.

However, the decision threshold can be set arbitrarily. The value of the decision threshold directly influences TPR and FPR and here the university can have a direct influence corresponding to the policies to be adapted through the choice of the threshold.

If the objective is to detect dropouts as reliably as possible, the decision threshold can be lowered to maximize TPR. However, TPR and FPR are usually counter-dependent on each other. If TPR of a binary classifier increases, the respective classifier often assigns more objects to the positive class, which also increases FPR. If FPR rises more actual graduates will be treated as dropouts. This could decrease acceptance among students and unnecessarily increase costs for the university. A widely adapted method to choose an optimal threshold appropriate for the task on hand and treat the trade-off between TPR and FPR are ROC curves (Bradley, 1997). For these reasons, the binary classifier of this work will also be evaluated with this method later.

2.3 Choice of algorithm

A binary classifier is, as already mentioned, realized by a function that maps the objects to be classified to two classes. There are various algorithms for the definition and optimization of such a function. An algorithm applied to a given problem certainly needs to meet technical requirements, but it must also need contextual requirements.

In this thesis a system for predicting university dropouts is to be constructed. This should provide university with a basis for better and more targeted management of dropout rates. Finally, when applied, it should assist universities in the implementation of steering mechanisms regarding dropouts. However, as an institution, a university is subject to certain constraints regarding the implementation of steering mechanism in terms of good governance. Hence, any system proposed, or any algorithm supporting implementation of steering mechanism such as here, also must meet those constraints or contextual requirements. Key characteristics of good governance that are particularly important are openness, transparency and accountability (Bundschuh-Rieseneder, 2008).

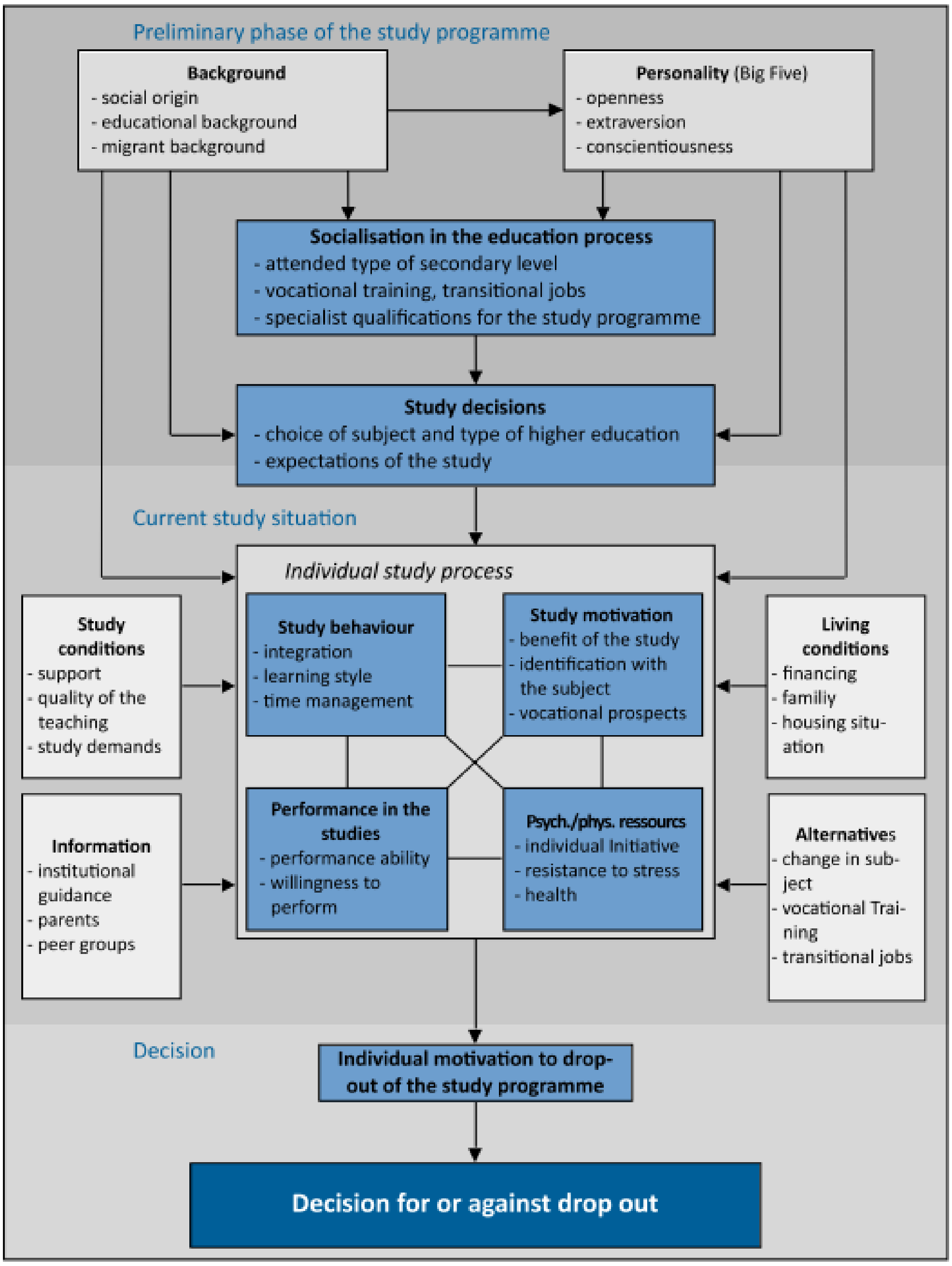
Any algorithm used in this thesis therefore must have a high interpretability and explainability so that any decision made based on the algorithm meets those requirements. This means the use of black-box algorithms should be avoided and white-box algorithms should be preferred.

An often used and widespread algorithm with these properties is logistic regression. In various disciplines it is valued for its interpretability and therefore preferred over more complex algorithms (Dreiseitl & Ohno-Machado, 2002). As a standard algorithm, the methods of interpretation of models realized with it are well researched and furthermore offers a good trade-off between accuracy and interpretability (Barredo Arrieta et al., 2020). For these reasons, the early dropout detection system in this paper is implemented with logistic regression.

2.4 Making the decision to dropout

To understand the potential and limitations for distinguishing graduates from dropouts, e.g. identifying dropouts, it is necessary to understand the reasons for a dropout to occur.

A wide range of influencing factors could potentially contribute to the phenomenon of dropouts (Larsen et al., 2013, p. 40). Moreover, the path to dropout is not a direct one whose destination is already determined at the outset. Studying is a process that is constantly evaluated by both the institution and the student and can lead to a decision to drop out at any time. As early as 1975, Tinto's 'student integration model', which massively influenced research into dropouts, reflected these circumstances (Tinto, 1975). With the state of knowledge regarding dropouts advancing since then, the influence of many proposed potentially influencing factors is empirically proven. This resulted in Heublein's model 'dropout from higher education process', which incorporates the current state of research (Heublein, 2014). Heublein’s model is shown in Figure 1.



DZHW 2014

Figure 1 - Heublein's model: 'Dropout from higher education process' (Heublein, 2014)

In Figure 1 the decision-making process for dropping out is divided into 3 phases: the preliminary phase of the study program, the current study phase and the decision phase.

The factors in the preliminary phase determine the student related conditions for the following phase and are mainly individual trades of a given student such as the sociodemographic background, psychological characteristics and academic preconditioning.

In the studying phase both factors on the part of the university and factors on the part of the student are decisive. Factors emanating from the university are usually referred to as external factors and factors inherent in the student are usually referred to as internal factors.

Discrepancies between internal and external factors are the source of motivation to make the decision to drop out in the decision phase.

The model thus gives an impression of the complexity of the interrelationships that lead to such a decision. Due to this very complexity, there are still open questions regarding dropouts. However, there is a broad consensus that the decision to drop out is rarely based on only one of the factors shown in Figure 1. Instead, it is assumed that for each dropout several factors and their complex interaction are responsible. There is also consensus that a decision to discontinue a study is not made spontaneously but after careful and long consideration (Heublein, 2014). Since it is assumed that isolated and temporarily limited adversities do not lead to dropouts, it can be assumed that dropouts can be better understood by aggregated measures than by raw data.

Lastly it should be mentioned that data on which this work is based - which will be discussed in more detail later - is limited purely to the performance in the study. Therefore, only statements about performance data can be made in this work. However, it is likely that any negative influences will eventually be reflected in the performance data. Thus, based on performance data a dropout can most likely be predicted, but the responsible factor cannot be identified.

3. Data

In this chapter, the data on the basis of which the dropout detection system is to be developed is examined. First, the raw content of the data is briefly described. Then, the data will be exploratively analyzed and pruned according to their significance for detecting dropouts. Finally, variables for describing students are extracted and likewise examined for their significance for detecting dropouts.

3.1 Data description

The data provided by the Technical University of Graz (TUG) consists of two datasets. The *student dataset*, containing all students enrolled in Computer Science from the years 2005 to 2015 or study year 2005 to 2014 and the *exam dataset* with the examination data for every Computer Science student in that time frame. Together the student dataset and the exam dataset form the dataset used in this thesis. In the following the content of both files or rather the two datasets are described in a qualitative manner.

3.1.1 The student dataset

The student dataset consists of 1861 entries or students. Every student in the dataset is described by 11 variables or features as shown in Table 3.

Table 3 - Variables describing a student

	column name	**description**
1	PERSONENID	UIN for a given student
2	GESCHLECHT	Gender
3	SFORM	School type in which university qualification certificate was obtained
4	STUDIDF	UIN for Computer Science Bachelor (CSB)
5	ZULAD	Date of enrollment into CSB
6	CLOSDAT	Date of closure of CSB
7	ABSCHLUSS	Whether a degree for CSB was obtained or not
8	STUDIDF_MASTER	UIN for Computer Science Master (CSM)
9	ZULAD_MASTER	Date of enrollment into CSM
10	CLOSDAT_MASTER	Date of closure of CSM
11	ABSCHLUSS_MASTER	Whether a degree for CSM was obtained or not

Every student has *a unique identification number (UIN)* by which he can be identified throughout the whole dataset. Also available is the *Gender* of every student. This enables a gender differentiated examination. The given *school type* provided classifies the type of school which issued the university qualification certificate. This is particularly interesting for an Austrian dataset, as there are also special types of schools only existent in Austria. In addition to general education, graduates of these institutions also have specific expertise in the subject area of the school. For this dataset which is related to Computer Science the school type *Höhere technische Lehranstalt (HTL)* could be of interest. Some schools of this type have a strong computer science education. Their graduates therefore potentially have a great deal of prior knowledge in the field of CS.

Every study has its own unique identification number at the TUG. The UID of the Computer Science Bachelor (CSB) and the Computer Science Master (CSM) is part of the description of a student. For every student, the university must track, which study the student is studying. This is presumably done with the UID for studies and here students appear to be in the student dataset because their data entity incorporates the UIDs of CSB. Further every student has an enrollment date, an exit date and two binary fields for tracking if a degree was obtained for both CSB and CSM. The enrollment date is the date a given student enrolled in the respective study. The exit date is the date a given student exited the study.

3.1.2 The exam dataset

The exam dataset consists of 80.011 entries or exams. Every exam is described by 14 features or variables as shown in Table 4.

Table 4 - Variables describing an exam

	column name	description
1	PERSONENID	UIN for a given student
2	STUD_IDF	UIN of study respective exam was linked to
3	FACH_ART	Status of respective course in respective curriculum
4	FACH_TXT	Text elucidating FACH_ART
5	PF_F033521	Flag if course is mandatory in CSB
6	PF_F066921	Flag if course is mandatory in CSM
7	LV_NR	UID for the respective course
8	LV_SEM	Winter or summer semester with the year
9	LV_TITEL	Name of the course the exam was taken in
10	SWS	Hours per semester week spent for the course in credit hours. Predecessor of ECTS
11	LV_TYP	Type of course
12	NOTE	Grade of the respective exam from 1 – 5 or 'E','O','U'
13	NOTE_TXT	Grade written out as text
14	PRFG_DATUM	Date the exam was taken

Every exam is assigned to a student which has written the exam. This is accomplished by an UIN of the respective student describing the exam.

A student can study multiple study programs at the same time and has the possibility to assign an exam to one of his studies. A study is linked to a given exam by the study's UID describing the exam. The study which the exam is linked to can be change with an official procedure. It is not apparent whether such a change would be reflected in the dataset. There is nothing that would indicate that. So, it is quite possible that an exam was taken in order to finish CS but was linked to a different study at first and relinked to CS at the end to finally graduate. Specific restrictions in individual studies create an incentive for this. If such restrictions are perceived as disruptive, they can often be avoided by this practice. So, it is difficult if not impossible to determine which study objective or study is actually being pursued. Every course the exam was taken in has a certain status in the curriculum of the study it was assigned to. The status for example could define if a course is mandatory in the given curriculum or not. Every exam is described by status of the course in the curriculum it was linked to. This information could again be corrupted, as the respective exam could have been assigned to a degree program which is not the study pursuit mainly. Every exam is also described by two flags which indicate whether the given exam is mandatory in CSB or CSM. Further analysis reveals that this flag is only correctly set if the respective exam is assigned to CS. Every course has a UID. An exam is also described by the UID of the course it was taken in. A string field records the year and the term, winter or summer term, an exam was taken in. Also, in a separate variable the name of the course is recorded. Further the official estimation of the hours per week during the semester needed to complete a course are given in the data. The course type is also recorded. It allows for example to differentiate between courses where attendance is mandatory or not or where assignments need to be hand in or not. Further the grade of a given student in a given exam is recorded in the respective exam entry. There are eight grades. From 1 to 5, one for noting successful participation without further differentiation, one counterpart for noting participation without success and one for noting invalidity in case of cheating. This is done once per number and once per text. Every exam also has a date of when it was written.

3.2 Statistics and Data Selection

The former section described the content of the datasets in a purely qualitative manner and thus provided a general overview. This section aims to describe the data in a quantitatively emphasized manner. Further the content of the datasets is to be examined in context to each other.

3.2.1 The student dataset

As already mentioned, the student dataset consists of 1861 students and every student in the dataset is described by 11 variables. To get a first impression of the quality and completeness of the data, the number of unique values and the number of missing values of every variable are summarized in Table 5.

Table 5 - Student number unique and missing values

	column name	number unique values	number missing values
1	PERSONENID	1861	0
2	GESCHLECHT	2	0
3	SFORM	27	0
4	STUDIDF	1	0
5	ZULAD	652	6
6	CLOSDAT	468	804
7	ABSCHLUSS	1	1630
8	STUDIDF_MASTER	1	1629
9	ZULAD_MASTER	181	1629
10	CLOSDAT_MASTER	46	1782
11	ABSCHLUSS_MASTER	1	1794

The table shows that there are 1861 students entered and every student has an UID. Gender is binary and there is no entry missing. There are 27 different type of school entered and for every student the school type was recorded. The UID of CSB and CSM are entered or are left empty. Since empty field hold no value the number of unique values is 1. It is noteworthy that every student has the UID of CSB, but 1629 of 1861 are missing the UID of CSM. Furthermore, 1629 students are also missing the entry of the enrollment for CSM. It could therefore be assumed that all students who were enrolled in CSB during the period in question were selected for the compilation of this dataset. Additionally, the information of these students regarding CSM was included in the dataset. With this insight, the data can be used to determine which student started a master’s degree program after their bachelor’s degree at TUG. It also follows that students who obtained a bachelor’s degree elsewhere and then completed CSM at TUG do not appear here. In contrast, students who enrolled in CSB at TUG, dropped out of this bachelor’s program unsuccessfully, completed a comparable bachelor’s program elsewhere and enrolled in CSM at TUG are very much covered by the data set.

The presence or absence of enrollment date or closing date, together with the content of the ABSCHLUSS fields, provides information about the study status of a given student. That is, whether a student finished his studies with or without a degree or whether he is still studying. In other words, this data can be used to determine whether a student's study status is graduate, dropout or ongoing.

Underlying Population

Next the assumption stated above that only students who were enrolled in CSB during a specified time period were included in the dataset, should be verified. In order to do so an auxiliary data field named *date combination* is introduced.

The data field is used to assign a 4-digit number $c = \mathrm{d}_1\mathrm{d}_2\mathrm{d}_3\mathrm{d}_4$ to each student. Each digit can either be 1 or 0 and is determined with a boolean expression. For a given student s it holds:

$$d_1 = s.date\_enrollment\_BCS \neq None$$
$$d_2 = s.date\_closure\_BCS \neq None$$
$$d_3 = s.date\_enrollment\_MCS \neq None$$
$$d_4 = s.date\_closure\_MCS \neq None$$

To write down the expressions here, the given student s is treated as an object and the needed values of the data fields are accessed in the commonly known syntax.

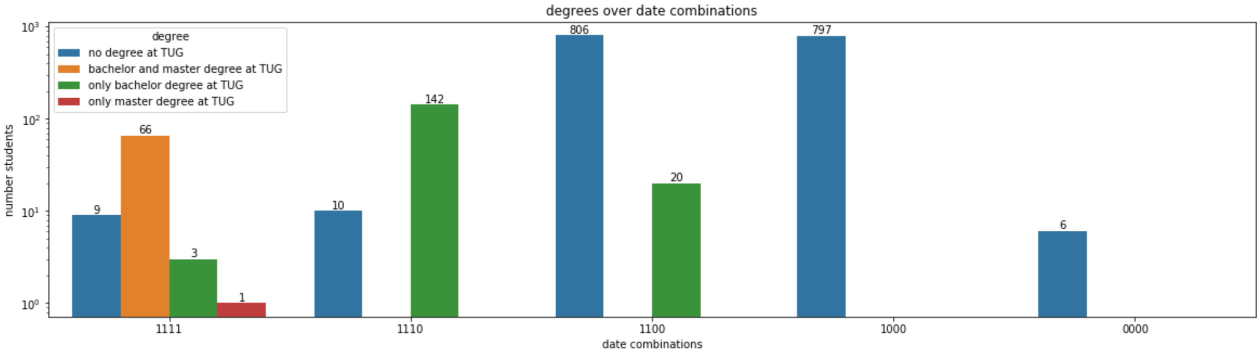


Figure 2 - all available date combinations in student dataset

Figure 2 shows the number of students over *date combination*. The ordinate is scaled logarithmically to improve the readability. The absolute numbers of students per category are plotted above each bar.

The assumption is corroborated by the fact that all students present have an enrollment date for CSB, but no student has only an enrollment date for CSM. That can be observed in the fact that no occurring category has a leading zero.

In addition, it is also evident that one student unsuccessfully dropped out of CSB to later successfully complete CSM at TUG.
It appears that six students have not entered a date at all. However, no exams from those six students were recorded in the exam dataset either.

Comparison of student cohorts

To obtain a degree in a particular study program, students must fulfil the requirements of the respective curriculum. A curriculum and the requirements a given student is subjected to could change over time. This could compromise comparability. To examine comparability of the students over time student cohorts are formed for further evaluations. All students who enrolled in the same study year form a student cohort or cohort. Students of a given student cohort are exposed to the same study conditions and they necessarily study the same curricular.

In Figure 3 and Figure 4 the student cohorts of CSB and CSM are analyzed.

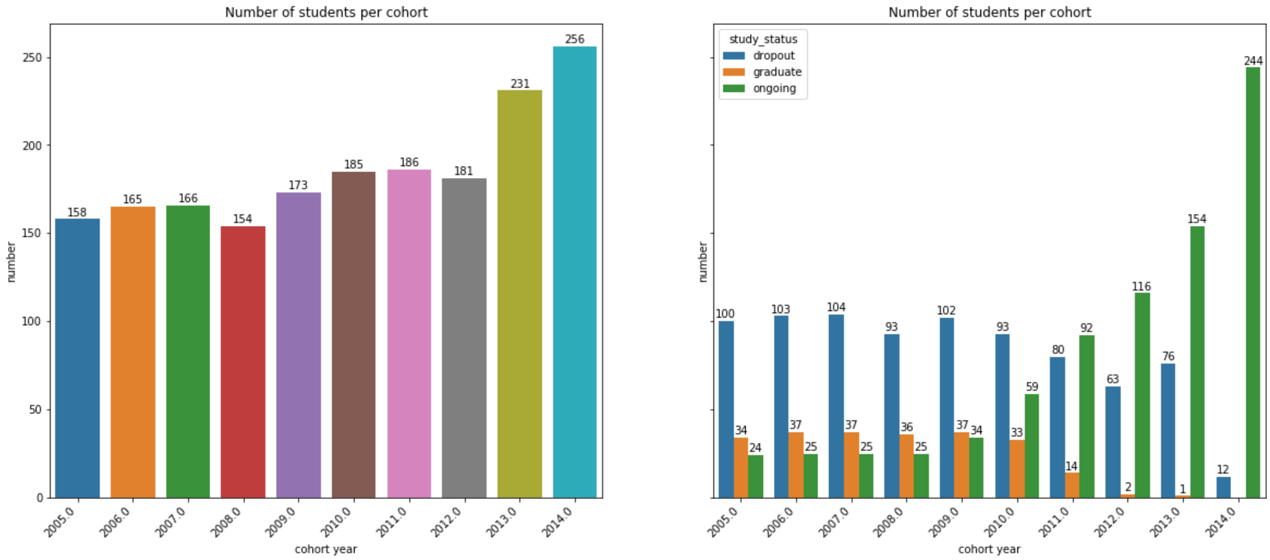


Figure 3- bachelor students and respective study status per year

In Figure 3 on the left side the number of students of each student cohort in CSB is displayed. On the right side the number of students in each student cohort is differentiated by study status at the time of data extraction in the year 2015.
For the total number of students in the individual cohorts there is a tendency to increase with the years. More and more people are studying. According to

Statistik Austria the total number of students increased roughly by $60 \cdot 10^3$ students or 30% in the period in question(Statistik-Austria, n.d.).

The proportions of the different study status are very similar in the first years when comparing the years with each other. From 2009 on, the share of ongoing students increases, while the share of dropouts and the share of graduates decreases. This is not surprising. Students who enrolled in a later year had less study time until the dataset was created. According to legal requirements in Austria, a bachelor's degree program should be designed so that it can be completed in 6 semesters or 3 years. In addition, students in the bachelor's program are granted 2 tolerance semesters. They can therefore study the bachelor's program for 8 semester or 4 years under the same regulations. Assuming a study duration of 4 years, only students who enrolled in 2010 or earlier had a realistic chance of completing their studies. It is also noticeable that the proportion of dropouts is also relatively high for short study durations. That means some of the dropouts occur relative early.

It can be noted that the structure of the student cohorts in terms of study status from 2005 to 2009 seems to remain the same. The structure of cohort 2010 changes slightly, but still resembles the structure of the former years. However, the structures of the cohorts from 2011 to 2015 are irregular. Subsequently, it is therefore assumed that the cohorts 2005 to 2010 are a reflection of the underlying overall system. The cohorts of the remaining years are not representative due to the limited study period.

A look at the Figure 4 for CSM gives a different picture.

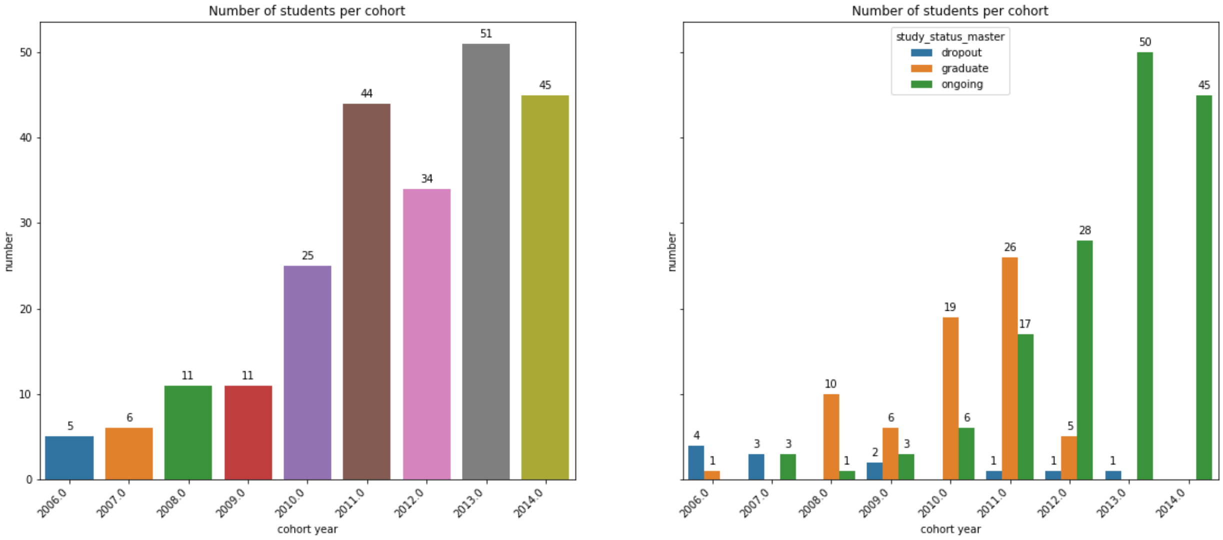


Figure 4 - master, students and respective study status per year

Figure 4 shows the same as Figure 3 but for CSM. One can observe that the sizes of the student cohorts are smaller than for CSB.

This makes sense when you consider that a relevant bachelor's degree is the prerequisite for a master's degree. These higher prerequisites reduce the amount of people who could study CSM and thus the amount of people who study it.

But the observation made can only be partly and not exclusively attributed to the higher entry barrier. The selection that takes place through the method of creating the data set also contributes to this. Completing a bachelor's degree takes time. Students who, as in this dataset, enrolled in 2005 at the earliest can only begin their master's studies after the time they needed to complete the bachelor's degree. Since the time window for which data are available appears to be the same for CSM as for CSB, the course of study for students in the master's program is simply barely captured by the dataset. The German higher education system is similar to the Austrian system. Data analyses of data from German universities, which reflect the overall situation more completely, also show lower enrollment numbers in masters than for their bachelor's counterparts (Heublein, 2014). The already mentioned suspicion suggests that this is likely due to higher enrollment requirements. To study CSM, one needs a relevant bachelor's degree.

In the diagram the number of students also appears to be trending upward over the years Also, the changing proportion sizes respective to the study status over the years can again be observed. But these observations are again subject to the bias of the dataset. Furthermore, it can be observed that the success rate in terms of obtaining a degree for CSM is clearly higher compared to the success rate in CSB. The above-mentioned data analyses from Germany draws the same conclusions.

Here, the students included in the diagram are subject to two performance-based selection processes, which facilitate success in CSM. Once through the required completion of the bachelor's degree. Every master student experiences this selection. And a second time, because a fast completion of the bachelor increases the probability to appear in the diagram. A fast student is more likely to enroll in the master's program during the observation period. This selection takes again only place due to the nature of the dataset and also causes a biased representation of the overall situation.

Figure 5 is intended to analyze the proportions of the different study status in the individual cohorts. For this purpose, the percentage share of each study status in a cohort is shown.

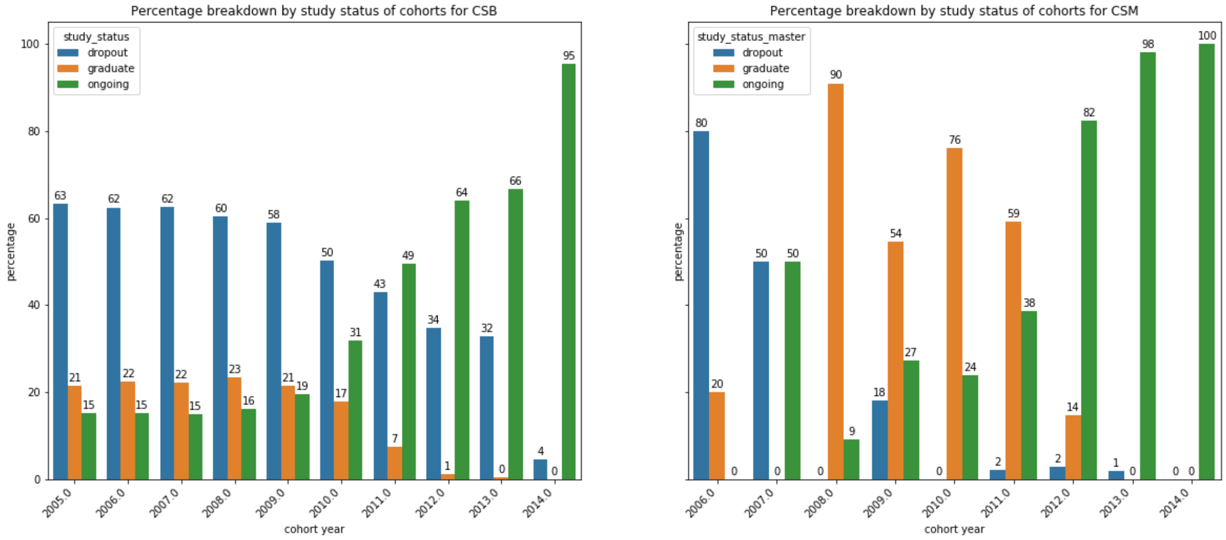


Figure 5 - proportions for different study status in percentage

On the left picture in Figure 5 the proportion sizes of CSB and on the right picture the proportion sizes of CSM are displayed. To benefit readability the display of the proportion sizes is limited to 2 digits.

In the left diagram, the regular and irregular cohorts are clearly visible. As can be seen from the new figures, in the cohorts with sufficient study time, about 60% of students drop out, about 20% obtain a degree, and about 15-20% are still studying.

In the right image for CSM the shape of the diagram changes. By rescaling the chart, it appears somewhat random, the cause of which cannot be directly explained here. Fluctuations are also already favored by the low student numbers. One can also see that the percentage of dropouts is usually immensely lower than in CSB. The proportion of students who are still studying, on the other hand, is a bit, and the proportion of graduates is considerably higher.

In summary, the 2005 to 2010 cohorts capture the entire bachelor's population. The data for the remaining years are incomplete due to the limited study period. The entire CSM data are incomplete because they capture only a portion of the master's population. Therefore, only limited conclusions about the master can be made with these data.

In summary, it can be stated that comparability is only given between cohorts 2005 - 2010 for CSB.

Total cohort

The previous study focused on the final configuration of the individual cohorts at the time of the creation of the dataset with regard to study status.

The next step is to investigate the composition of the student cohorts over time. This requires a meaningful and appropriate measure of time. The chosen measure should be relevant to the subject under consideration and in the overall context.

At TUG, as at many other universities, one year is divided into two semesters. The various study curricula are divided into semesters and the organizational process is aligned according to this division. Therefore, a time measure in semesters is chosen.

The composition of student cohorts changes precisely when a student leaves the cohort. That is, when a student withdraws from the study under investigation.

Therefore, only graduates and dropouts are relevant for the considerations that should be made here. Ongoing students remained in the cohorts until the dataset was created. It is not possible to determine with certainty whether they would have graduated or dropped out. They have no influence on the composition of the cohort and are left out of the following considerations.

For graduates and dropouts, the study duration or the semester in which they left the cohort and changed it can be calculated using the respective enrollment and deregistration date.

In order to present only unbiased information, this is only done with data from the CSB. Furthermore, the data is limited to cohorts with a sufficient observation period. As already described, this applies to the cohorts from 2005 to 2010.

Figure 6 is the result of plotting the number of students over the respective study duration. The diagram consists of 6 sub-diagrams. Each sub-chart represents the change in a given student cohort over time.

Although the distributions across cohorts in Figure 6 are very similar, the distribution of dropouts and graduates differs in a given cohort in all student cohorts used.

Two broader clusters or accumulations can be observed in the distributions of dropouts. It is bimodal. This bimodality allows for a subdivision of dropouts into early dropouts and late dropouts. The center of the first cluster is in the second or third semester, the center of the second cluster is in the seventh, eighth or ninth semester, depending on the cohort. The maximum of the first cluster seems for all years more pronounced than that of the second, i.e. it is higher. Only year 2005 seems to be an exception. In general, this gives the impression that the majority of dropouts usually occur in the first few semesters. Hence the cohort of early dropouts seems bigger. This is coherent with the observations made above, that for years with a short observation period the dropouts are significant. Early dropouts cause and bear fewer costs than late dropouts. From a cost perspective, these circumstances are pleasing. However, it is questionable what led to the dropouts and how the students concerned would have developed if they had continued studying.

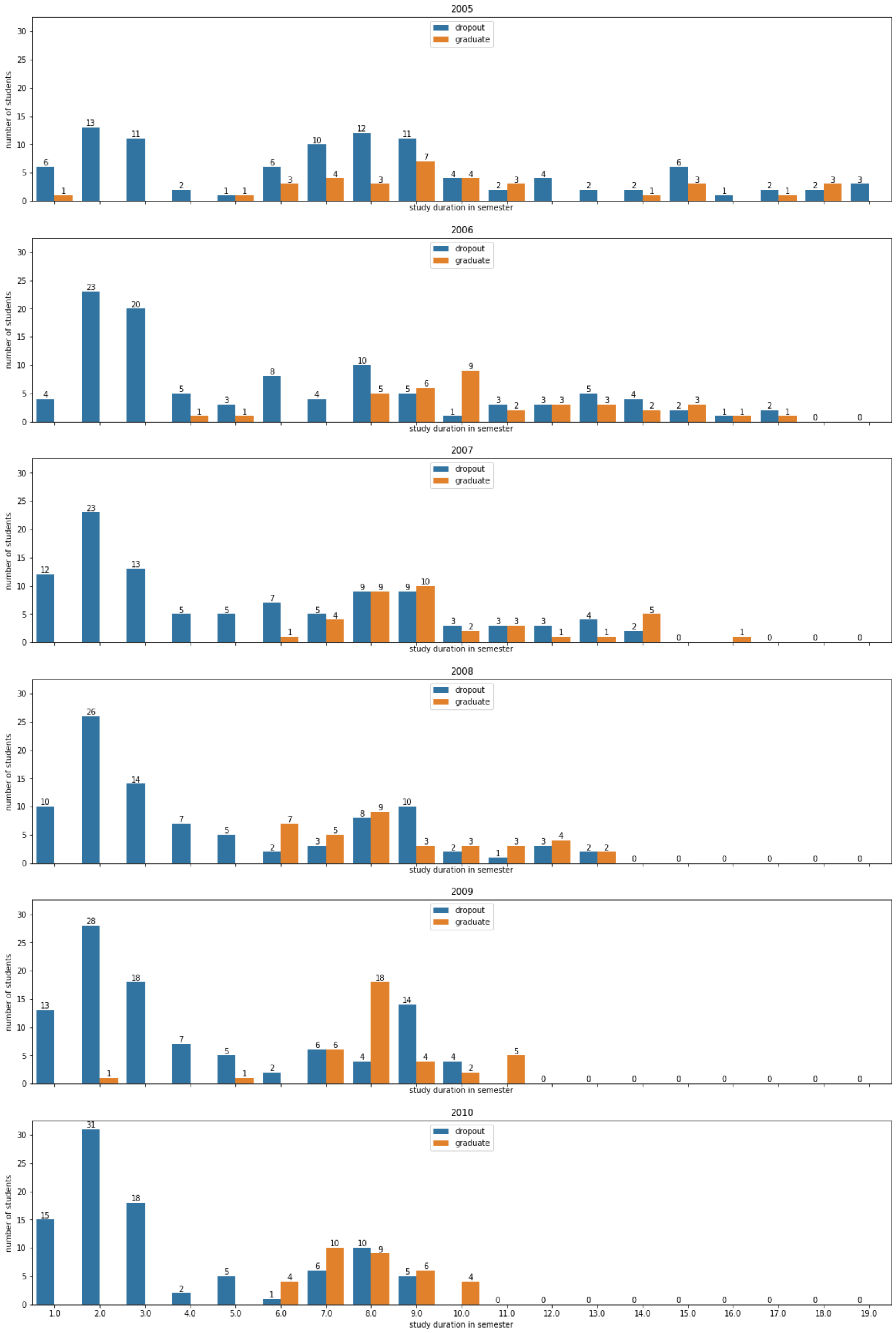


Figure 6 – Absolute change in relevant student cohorts over time

The distribution of graduates, on the other hand, is unimodal and shows only one cluster or accumulation. Its center lies around the eighth semester. The distribution is asymmetric and as expected prolonged to the right. Students tend to take longer to complete their studies than the intended 6 semesters.

The axes of the individual sub-diagrams are scaled identically. If one compares the cohorts year by year with increasing number of years, one can see that the number of missing data points also increases at the end of the time series. This is again due to the limited observation period. In the previous diagrams, this circumstance was already apparent in the student cohorts of later years, as they already lacked numerically stronger semesters. If, for the purpose of comparability, one wants data for each semester from each student cohort, then not only the year but also the semester number must be limited in the evaluation. Turning to the cohort of 2010, the year with the shortest observation period, one sees that the limiting number of semesters is 10. This reduces the maximum number of semesters for the cohorts from 19, as in Figure 6, to 10 semesters. If one limits the semesters to 10 and converts the absolute student numbers into percentages in relation to the total number of students in the respective cohort, Figure 7 is the result.

The total number of students in Austria increases over the years. Likewise, the total number of students per cohort tends to increase with the years. Therefore, the representation in percentages in Figure 7 facilitates the comparison between the cohorts by excluding this factor. The observations made above are also valid here.

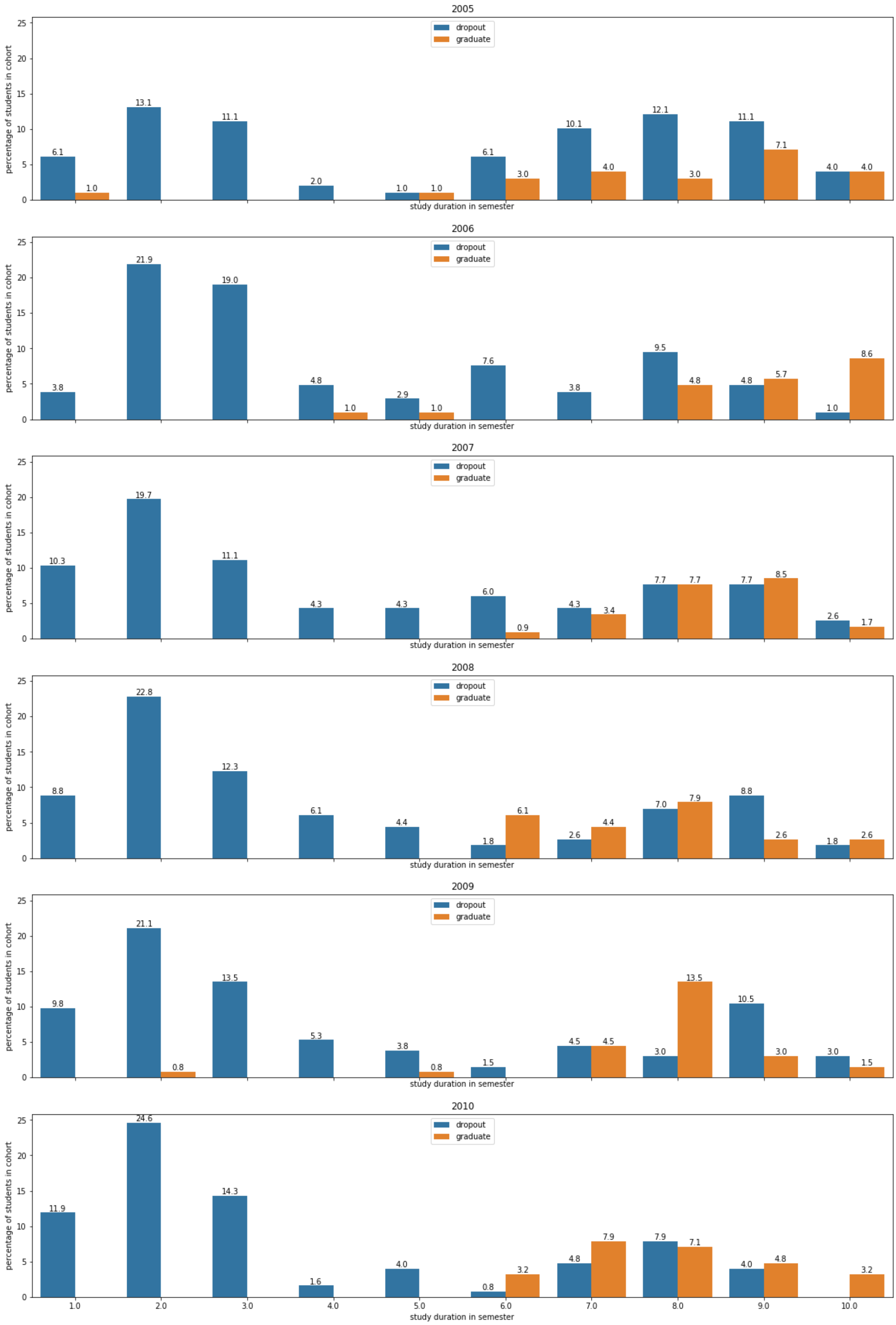


Figure 7 - Change in percentage in relevant student cohorts over time

Due to the restrictions made with regard to years and semesters, the data basis in the cohorts are semantically the same. The cohorts are quite similar in comparison to each other, but still do vary. A trend is discernible, but only blurred due to these deviations. The improvement, similarity and dissimilarity can be better observed in Figure 8.

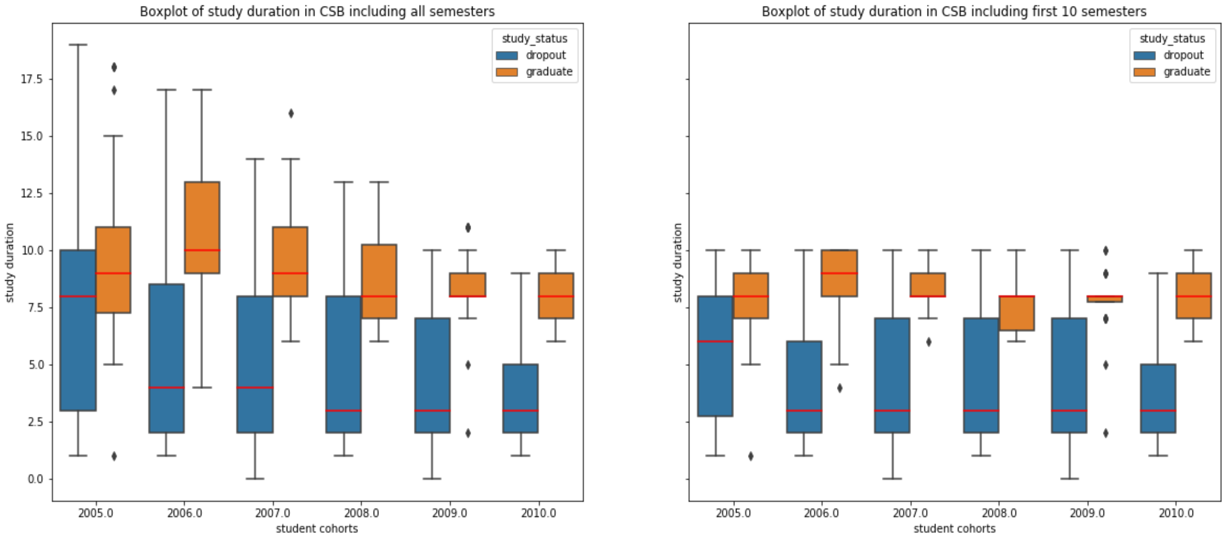


Figure 8 - Boxplot of study duration with and without limited semester

The data basis of the left-hand diagram in Figure 7 is formed by all available semesters of the respective cohort. The number of semesters of the cohorts on the right side is limited to 10. The similarity of the individual distributions clearly increases. The spread is of similar width and the medians are also closer together. For the majority of cohorts, it is clear from the position of the median that the cohort of early dropouts includes the majority of dropouts. Only the cohort of 2005 is an exception. The widths of the distributions of the graduates are different, but the medians are similar. These width differences could also be due to natural fluctuations.

Each cohort is exposed to the conditions of the underlying study. The changes over time in the respective cohort are also a result of these conditions. If one considers the individual cohorts as individual samples from the CSB study, then looking at the aggregation of all cohorts into one could prove interesting in order to get an impression of the general course of studies.

Before aggregating all cohorts into a total cohort, it is appropriate to validate the equality of the cohorts to justify this process. There are two opposing hypotheses:

H_0 ≔ cohorts are taken from a common base population and are pairwise similar.

H_1 ≔ cohorts are not taken from a common base population and are not pairwise similar.

Table 6 is the contingency table of the cohorts with regard to the distributions of study duration.

Table 6 - Contengency table for all cohorts about study duration

		study duration										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
cohort	**2005**	7	13	11	2	2	9	14	15	18	8	**99**
	2006	4	23	20	6	4	8	4	15	11	10	**105**
	2007	12	23	13	5	5	8	9	18	19	5	**117**
	2008	10	26	14	7	5	9	8	17	13	5	**114**
	2009	13	29	18	7	6	2	12	22	18	6	**133**
	2010	15	31	18	2	5	5	16	19	11	4	**126**
	Total	**61**	**145**	**94**	**29**	**27**	**41**	**63**	**106**	**90**	**38**	**694**

If one calculates the expected frequencies under H_0 on basis of Table 6, Table 7 results.

Table 7 - Expected frequencies for all cohorts about study duration

		study duration									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
cohort	**2005**	8.7	20.7	13.4	4.1	3.9	5.8	9	15.1	12.8	5.4
	2006	9.2	21.9	14.2	4.4	4.1	6.2	9.5	16	13.6	5.7
	2007	10.3	24.4	15.8	4.9	4.6	6.9	10.6	17.9	15.2	6.4
	2008	10	23.8	15.4	4.8	4.4	6.7	10.3	17.4	14.8	6.2
	2009	11.7	27.8	18	5.6	5.2	7.9	12.1	20.3	17.2	7.3
	2010	11.1	26.3	17.1	5.3	4.9	7.4	11.4	19.2	16.3	6.9

The contents of Table 6 and Table 7 resemble each other, but some corresponding values lie close together, some rather not. The Chi-squared of Homogeneity performed on these tables enables a conclusive statement to be made. The test yields $p - value = 0.35$. With chosen $\alpha = 0.05$ it follows $p - value = 0.35 > \alpha = 0.05$ and therefore H_0 cannot be rejected. So H_0 is assumed in reasonably good belief and the aggregation of all cohorts is considered legitimate.

In the visual inspection of the distributions in Figure 8, irregularities stand out. The median of dropouts in cohort 2005 does not seem to correspond to the general pattern. This, too, is to be investigated. For this purpose, the average study durations of the cohorts are to be compared by means of a t-test. In the pairwise comparison, the following two hypotheses are tested:

$$H_0 \coloneqq mean(cohort_x) - mean(cohort_y) = 0$$
$$H_1 \coloneqq mean(cohort_x) - mean(cohort_y) \neq 0$$

The comparison is performed for the overall average study duration and for the average study duration for graduates and dropouts. Due to the larger samples resulting from the calculation of the average over all semesters, the differentiation by means of study status is possible, because the amount of data per calculation is sufficient for a statement. The level of trust with $\alpha = 0.01$ is $1 - \alpha = 1 - 0.01 = 99\%$. It is chosen that high because the findings must invalidate the indications above. The results of the tests are summarized in Table 8, Table 9 and Table 10.

It should be noted that the increase in the family-wise error rate due to multiple hypothesis tests is not controlled in the test scenario. The use of Bonferroni correction or similar measures is deliberately avoided, as these methods aim to reduce the probability of a type 1 error, and in the process often increase the probability of a type 2 error. The objective in this test scenario for selecting meaningful data must be to reduce the probability of a type 2 error.

Table 8 - t-test results, pairwise comparison of overall average study duration

		cohort					
		2005	**2006**	**2007**	**2008**	**2009**	**2010**
cohort	**2005**	H_0	H_0	H_0	H_0	H_0	H_1
	2006	H_0	H_0	H_0	H_0	H_0	H_0
	2007	H_0	H_0	H_0	H_0	H_0	H_0
	2008	H_0	H_0	H_0	H_0	H_0	H_0
	2009	H_0	H_0	H_0	H_0	H_0	H_0
	2010	H_1	H_0	H_0	H_0	H_0	H_0

Table 9 - t-test results, pairwise comparison of dropout's average study duration

		cohort					
		2005	2006	2007	2008	2009	2010
cohort	2005	H_0	H_1	H_0	H_1	H_1	H_1
	2006	H_1	H_0	H_0	H_0	H_0	H_0
	2007	H_0	H_0	H_0	H_0	H_0	H_0
	2008	H_1	H_0	H_0	H_0	H_0	H_0
	2009	H_1	H_0	H_0	H_0	H_0	H_0
	2010	H_1	H_0	H_0	H_0	H_0	H_0

Table 10 - t-test results, pairwise comparison of graduate's average study duration

		cohort					
		2005	**2006**	**2007**	**2008**	**2009**	**2010**
cohort	**2005**	H_0	H_0	H_0	H_0	H_0	H_0
	2006	H_0	H_0	H_0	H_0	H_0	H_0
	2007	H_0	H_0	H_0	H_0	H_0	H_0
	2008	H_0	H_0	H_0	H_0	H_0	H_0
	2009	H_0	H_0	H_0	H_0	H_0	H_0
	2010	H_0	H_0	H_0	H_0	H_0	H_0

In Table 8 and Table 9, the individual sample sizes on which the test was carried out correspond to the convention with over 30. In Table 8, on the other hand, the majority fall short of this convention, which affects the significance.

For cohorts for which it could not be determined in the pairwise comparison that they have different averages, $H_0$is entered. For those for which different averages could be determined, H_1 is entered.

Only comparisons with the 2005 cohort show different averages. This confirms the assumption suggested by Figure 8 that cohort 2005 breaks with the general pattern. Other comparisons do not reject H_0. So again, for those comparisons H_0 is assumed in reasonably good belief.

The test showed that 2005 is out of the ordinary, now the question is why. Compared to other cohorts, this cohort has fewer early dropouts and more late dropouts. In the chronological sequence, it comes first, the other cohorts which come after are all similar. It could therefore be that a significant change in CSB has taken place that has only targeted late cohorts. This change could also have been brought about deliberately. Especially as this change, as already briefly mentioned, seems advantageous from a certain perspective. Encouraging early dropouts and discouraging late dropouts reduces wasted capital for all stakeholders, provided that future graduates are not negatively affected. However, this hypothesis cannot be confirmed or rejected with the available data. Nevertheless, because the confidence level is chosen high and because of the chronological sequence, cohort 2005 is not used for the investigation of the general course of studies. In summary, the aggregation of all cohorts except cohort 2005 is considered legitimate.

If the cohorts are combined using the above considerations, Figure 9 results. Figure 9 shows the general course of studies in the CSB or the course of study of the total cohort. Included are the cohorts [2006, 2010] with a limited observation period of [1, 10] semesters.

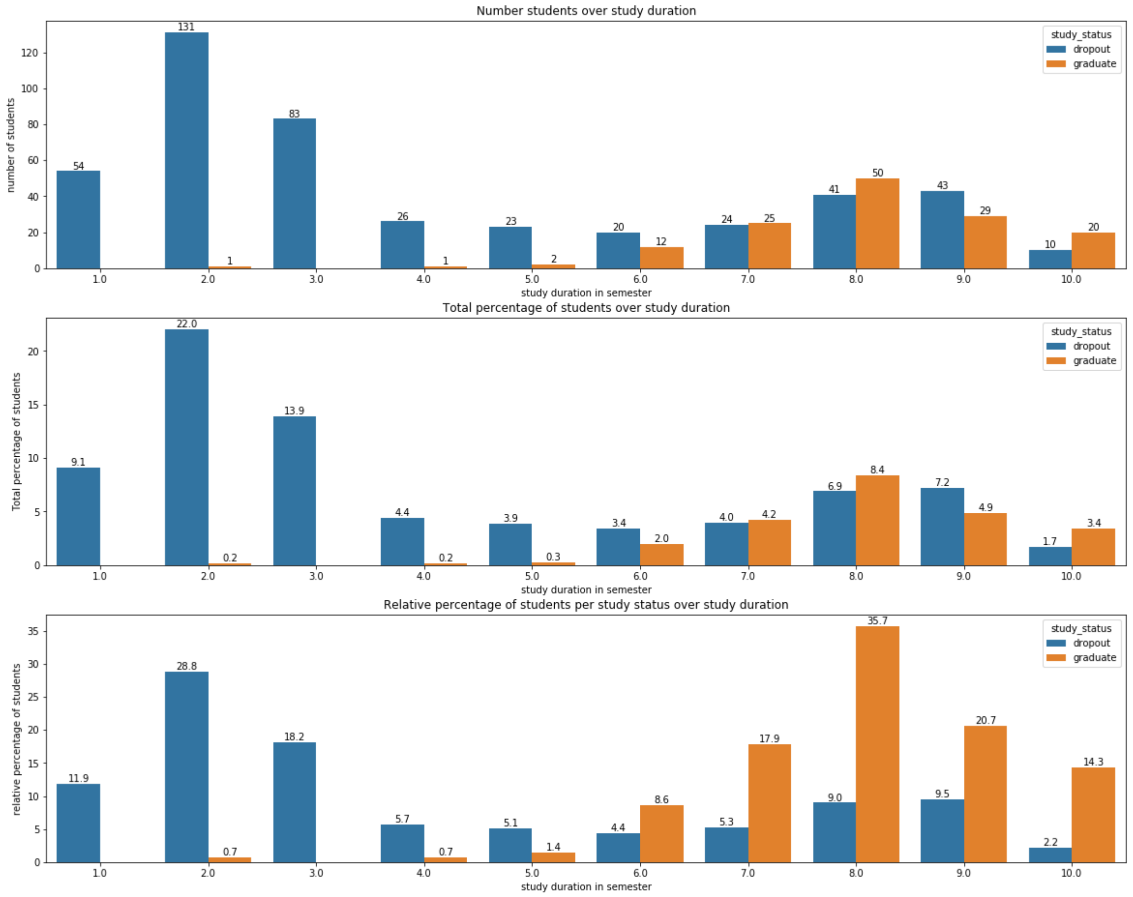


Figure 9 - general course of study for CSB

Figure 9 consists of three sub-diagrams. The top sub-chart represents the change in the total cohort in absolute numbers. The middle chart represents the change in the total cohort in absolute percentages. The percentages are based on the total number of students in the cohort. The bottom sub-chart represents the change in the total cohort in relative percentages. The percentages of dropouts are based on the number of dropouts, the percentages of graduates are based on the number of graduates in the cohort.

Figure 10 summarizes the data flow from the total student dataset to the total cohort and its further differentiation. The cited decisions to select and cull data are justified above. Key figures for the data segments or groups of students identified as relevant are compiled in Table 11.

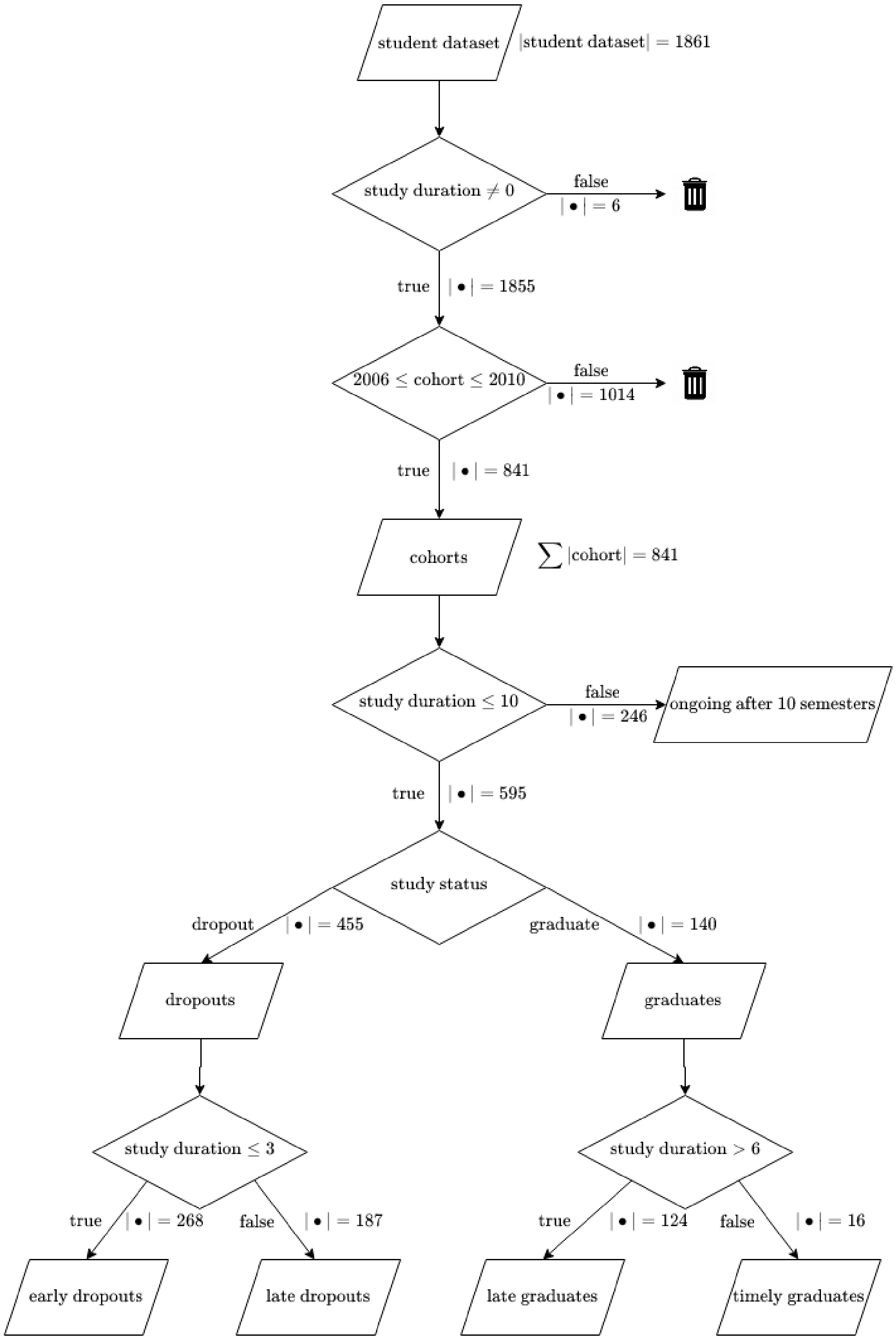


Figure 10 - Data flow students

In addition to dividing dropouts into early and late dropouts, graduates are divided into timely and late graduates. Subdivision of dropouts is based on the two modes in the distribution of dropouts during study shown in Figure 9. The subdivision of graduates is based on the legal framework in Austria according to which a bachelor's program should be designed in such a way that it can be completed in 6 semesters.

Table 11 - Key figures for viable students for analysis

		percentage		study duration	
	number	**absolute**	**relative**	**mean**	**variance**
total	841	100	-	6	14.5
graduates	140	16.6	100	8	1.8
late grad.	124	14.7	88.6	8.4	1
timely grad.	16	1.9	11.4	5.5	1.2
dropouts	455	54.1	100	4.1	7.7
early drop.	268	31.9	58.9	2.1	0.5
late drop.	187	22.2	41.1	7.1	3.5
still ongoing	246	29.3	100	-	-

The key indicators in Table 11 were calculated based on the total cohort or the viable students for analysis. The composition of students is stereotypical for CSB, so therefore are the derived key figures. However, the included students cannot be further differentiated at this point of the analysis. Students who lack the serious intend to graduate in CSB may by still present in the data. And if so, the proportion of them is unknown. Furthermore, no distinction is made between dropouts who leave the higher education sector and students who change their studies. Both categories appear here as dropouts.

Therefore, these figures must be interpreted with caution.

3.2.2 The exams dataset

After the quantitative examination of the student dataset and the selection with regard to the significance of certain data, the exam data will now be examined. As already mentioned, the exam dataset consists of 80.011 exam entries and every exam in the dataset is described by 14 variables. To get an impression of the quality and completeness of the data, the number of unique values and the number of missing values of every variable are summarized in Table 12.

Table 12 - number unique and missing values for exam dataset

	column name	number unique values	number missing values
1	PERSONENID	1696	0
2	STUD_IDF	129	0
3	FACH_ART	16	20957
4	FACH_TXT	14	20963
5	PF_F033521	1	49872
6	PF_F066921	1	78762
7	LV_NR	3380	0
8	LV_SEM	49	0
9	LV_TITEL	2269	0
10	SWS	31	0
11	LV_TYP	22	0
12	NOTE	8	0
13	NOTE_TXT	8	0
14	PRFG_DATUM	3485	0

Table 12 shows that exams of 1696 different students are recorded in the exam dataset. The corresponding number of missing values shows that a student is assigned to each exam. To a total of 129 different studies exams are assigned to. This means that the students recorded were not only enrolled in CSB and CSM, but presumably in considerably more studies. This complicates the analysis because it is not clear which exams were taken with the intend of completing CSB and CSM. As mentioned above, there are good reasons for a student to enroll for CSB or CSM, for example, and take exams, only to eventually drop out and complete another degree program with the exams taken. Or to do it the other way around and take the exams relevant to CSB, for example, with another degree program and then enroll in CSB and complete the degree program in an exorbitantly short time on the basis of the exams already done.

Students in Austria are subject to very few restrictions on enrollment in studies, assignment and reassignment of examinations to studies. This means that any scenario in these matters is conceivable and quite possible. This raises two questions that are highly relevant to what can be said about CS based on the given data. Which degree program does the respective student actually intend to complete? So, what is his main study? In this dataset, a student could, in the worst case, be exam-active but not have any exam assigned to his or her main study. And if there are such or other unfavorable cases, is it possible to identify them? According to the responsible developers, and also in order to be able to implement the legal requirements, the student alone is responsible for the assignment. This assignment remains in effect even if the examination is credited for another degree program. The value of the study id is never updated. So, it is not possible to identify the main study on the basis of this data set.

There are 16 different status for the respective course in the respective curriculum and a large number of missing values. Since the value of this fields are based on the study the exam was assigned to by the corresponding student it holds no valuable information regarding CS. The fields which seem to provide information on whether the subject of the respective examination is a compulsory subject in CSB or CSM also show an interestingly high number of missing values. Closer inspection reveals that these fields have different values for exam entries of the same course. Hence, it is not a reliable source for whether an exam was written in a mandatory course of CS or not. Furthermore, the table shows that the dataset includes 3380 different course IDs or courses in which examinations were taken. If one looks at the number of course names, the difference to the number of course ids is notable. The mapping between course name and course ID can therefore not be unique. The course number serves as UID and the name can be chosen arbitrarily by the course responsible. So, a given course name could be used multiple times over different courses. Every exam entry has a course ID and a course name. The field to record the term, as it includes the year, has 49 different values. The number of different values for SWS is 31, resulting in 31 different levels of workloads for courses. There are 22 different kinds of courses and no missing values for this field. Every exam has one of the eight grades possible. The data when a given exam was taken is available for all entries.

3.2.3 Total cohort

With the goal in mind to understand the factors which determine success and failure of students during studying CS and the knowledge gained during previous analyses, it is evident that not all exam data will facilitate reaching that goal.

The previous analysis of the student data showed that the provided data is only representative for the CSB. Hence valid statements can only be made for CSB. Also, only a portion of students is representative due to the varying observation period. From the significant data that resulted from considering these constraints, the total cohort emerged. The resulting goal is now to understand the dynamics that determine success or failure in CSB. If this is the objective, then for the same reasons only examination performance of the total cohort can contain valuable information. Therefore, only exam entries of students in the total cohort are relevant for further analysis. Exam data of the students who were left out would only lead to distortions and disruptive information.

Selection of relevant exam entries

First the right exam data must be selected. The total cohort consists of 841 students. After the first 10 semesters, there are 140 graduates, 455 dropouts and 246 students still studying. Of these 841 students, exams are recorded for 785 students and none for 56. Of these 56 students, 55 are dropouts and 1 is still studying after the 10th semester. Since the exam behavior is examined here, students without exams will be disregarded for the following considerations.

Of the total 80011, 47188 examination entries are from the 785 remaining relevant students. This includes exam entries of all kinds. With regard to CSB and taking into account the admission and deregistration date of a given student, 3 semantically different types of exam entries can be identified. Namely, exams taken before enrollment in CSB, exams taken during CSB and exams taken after CSB.

For the analysis of CSB, the value and the significance of the information provided by the exam entries depends on the category the respective exam is in. Exams taken during the CSB are naturally the most relevant for the analysis of CSB. Exams taking after CSB are irrelevant and can be discarded since they cannot hold valuable information about CSB. The significance of exams taken prior to enrollment in CSB is uncertain.

Considering only exams before and during CSB and limiting the observation period to up to 10 semesters relative to the respective student, one ends up with 38.839 relevant exam entries for the total cohort. As a result, a further 15 students, all dropouts, are discarded, because they did only exams after CSB. Table 13 summarizes the resulting composition of the total cohort.

Table 13 – number of students present in relevant exam entries of the total cohort

	graduates	**dropouts**	**ongoing**	**Total**
students	140	385	245	770
absolute percentage	18.2	50	31.8	100

Now 18.2% are graduates, 50 % are dropouts and 31.8% are ongoing after the 10th semester.

About mandatory and non-mandatory courses

If one looks at the public available current CSB curriculum of the TU Graz, it is noticeable that the proportion of non-mandatory subjects is low compared to mandatory courses (Curriculum CSB, 2020).

To graduate in CSB one must earn 180 ECTS. The bachelor thesis account for 9 credits of those. So, 171 of must be earned by completing courses.

Courses for 10 credits can be chosen freely by the student and courses for 14 credits need to be chosen from a certain set of courses. Thus, courses worth 147 credits are mandatory. All in all, about 86% of all courses are compulsory, about 8.2% are elective and about 5.8% can be chosen without restriction. As already mentioned in the data here credit hours are used instead of ECTS. Credit hours can be converted to ECTS by multiplying the number of credit hours with 2.

These 14% of non-compulsory courses are most likely not decisive for successful graduation. Due to the freedom of choice, practically all obstacles could be circumvented.

According to information from Graz University of Technology, similar circumstances also applied in the curricula under the influence of which the data examined were produced. It follows that the compulsory subjects are decisive for a successful graduation. Corresponding differentiation of the courses is therefore key for successful interpretation.

It is not clear from the dataset which courses are mandatory and which courses are non-mandatory. Fields which convey information of the course type are, as already mentioned, dependent on the study an exam entry was assigned to.
With the flexibility with which students can assign exams to studies the information is highly unreliable.
However, examinations assigned to CSB are correctly marked as mandatory subjects in CSB. To be able differentiate courses in mandatory and non-mandatory for CSB regardless of exam assignment, all subjects ever marked as mandatory in CSB are counted as mandatory for CSB. For CSB 101 courses can be identified as mandatory.

So, in the following considerations exams are differentiated in exams of courses that are mandatory or non-mandatory courses by using the 101 courses identified as mandatory. Table 14 summarizes the composition of the relevant exam entries for the total cohort.

Table 14 - exams differentiated depending on mandatory and non-mandatory courses

	number exams			**percentage**			**Exams per student**
	total	**mandatory**	**non-mandatory**	**Absolute**	**mandatory**	**non-mandatory**	
graduates	10576	8252	2324	27.2	78	21.9	75.54
dropouts	11818	4915	6903	30.4	41.6	58.4	30.7
ongoing	16445	11430	5015	42.3	69.5	30.5	67.12
all	38839	24597	14242	100	63.3	36.7	38839

Table 14 shows the proportion of exams done by students depending on the study status. The respective shares are also further broken down into exams of mandatory and non-mandatory courses.

Graduates are responsible for 27.2% of all written exams, dropouts for 30.4% and ongoing students for 42.3%. By taking the different size of these groups shown in Table 13 into account, the number of exams per student on average can be calculated. On average graduates write 75.54 exams per person, dropouts 30.7 and ongoing students 67.12. The average performance in terms of written exams is for dropout clearly lower than for graduates and thus dropouts are far away from having the necessary workload done for completing CSB. However, as shown in Table 11, the average study duration of dropouts is half as long as the average study duration of graduates. Hence, the difference in exams written may not be so blatant if this comparison is done for single semester. Ongoing students on the other hand show only a slightly smaller output as graduates.

The efforts of graduates are with 78% clearly focused on mandatory courses. Ongoing students also focus their efforts with 69.5% on mandatory courses while dropouts do the main part of their workload with 58.4% in non-mandatory courses.

The data shows that graduates generate more output and that this output is mainly focused on mandatory exams. Ongoings' output is comparable to graduates' output. They focus mainly on mandatory exams and are doing similar amounts of exams, but in both domains less than graduates do.

If the average number of exams taken by graduates is used as a comparative measure, it is noticeable that the average number of exams written by dropouts is too low to be able to complete any degree.

About temporal order

Above it was already established that the value of information an exam entry could offer regarding CSB depends on when the exam took place relative to the enrollment and optionally existing closure date of the respective student.

Exam entries of exams which took place after CSB are discarded because of this. Recordings of exams which took place during CSB are kept. The meaning of exams which took place before CSB is unclear.

The knowledge acquired before, could be beneficial for success in CSB. Just already knowing the institutional processes could increase the likelihood of success. Also, mandatory courses may have already been taken. This would reduce the workload in advance and thus increase the probability of success. Table 13 – number of students present in relevant exam entries However, prior experience could also be an indicator of lower chances of success. This is because students with previous experience already were enrolled in an other study. The proportion of students who finished a degree and start another bachelor is most likely small. So, students with previous experience could have shown failure by not completing the other degree or keep study another study while enrolling in CSB. Thus, they effectively divide their work power, shrinking it for both studies or they enroll in CSB with the intention to not complete it. Any case decreases the chance of success in CSB.

To understand the relation of probability of success to exams taken prior to enrollment in CSB, students are classified in categories based on in which categories their exams fall. Students who only did exams before their enrollment in CSB are in category -1, students who did exams before and during CSB are in category -1/0 and students who only did exams during CSB are in category 0. Figure 11 results if the number of students and the percentage per study status in each category is plotted over these categories.

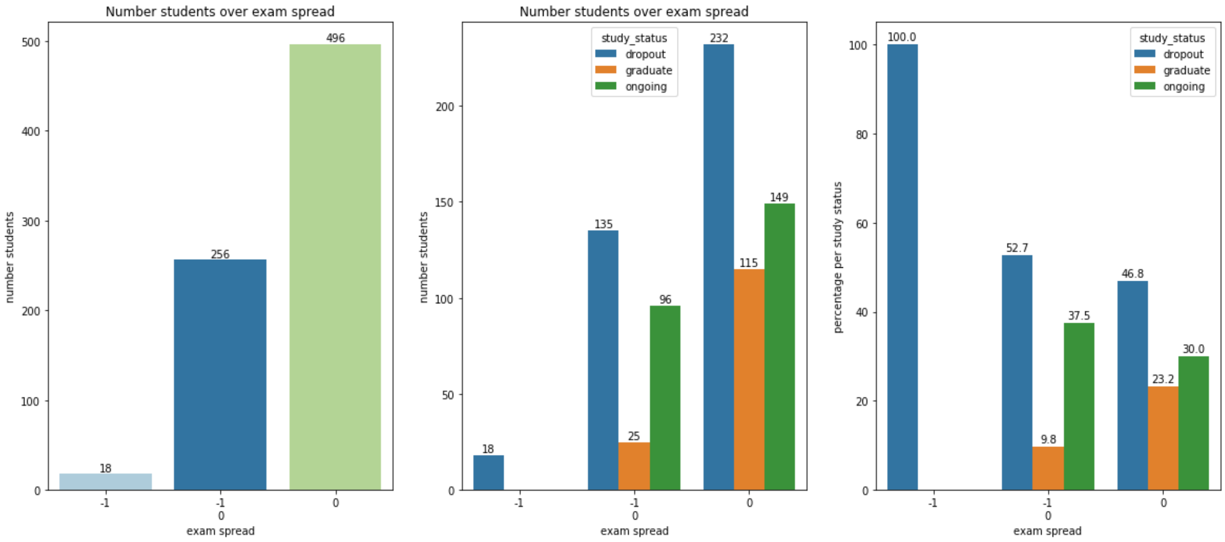


Figure 11 – students over exam spread

The left sub-diagram shows the absolute number of students over the categories. In the middle diagram those numbers are differentiated by study status. The right sub-chart shows the proportion of students per study status in the respective category.

In the left diagram one can see that 18 students, all of which are dropouts, are in category -1, 256 students are in category -1/0 and 496 students are in category 0. From the right sub-diagram, it is clear, the success rate for students with previous studying experience is with 9.8% low compared to the 23.2% success rate of students with no previous studying experience at TU Graz. The portion of dropouts with experience is 52.7% and the portion of dropouts without experience is 46.8%. For ongoing students with experience, it is 37.5% with experience and 30% without experience. The empirical proportion of dropouts and ongoing students is bigger for students with experience compared to students without experience. A performed Chi-squared test confirms the difference of the proportions with $p-value = 3.6 \cdot 10^{-5}$ and a resulting rejection of the null hypothesis for equality.

So, the probability of success is lower for students who already have study experience at the TU Graz.

To gain better understanding of the probability of success, next the study behavior before CSB and during CSB should be compared.

Therefore, exams are separated in exams before and exams during CSB. Table 15 is dedicated to exams which were wrote prior to enrollment in CSB and Table 16 is dedicated to exams which were done while being enrolled in CSB.

Table 15 - exams before CSB differentiated by study status

	number exams			percentage			
	total	mandatory	non-mandatory	absolute	mandatory	non-mandatory	exams per student
graduates	725	321	404	7.5	44.3	55.7	29
dropouts	5813	1742	4071	60.4	30	70	38
ongoing	3079	962	2117	32	31.2	68.8	32.1
all	9617	3025	6592	100	31.5	68.5	35.1

9617 exams of the total 38839 are taken before CSB. Hence, 24.8% or about one quarter of the relevant exams taken by the total cohort is done before CSB.

The content of Table 15 reveals that the share of exams written before CSB is with 7.5% the lowest for graduates. Exams written by dropouts account with 60.4% for the main portion. The share of exams written of ongoing students lies with 32% in-between.

Taking the number of students with experience per study status shown in Figure 11 into account, graduates wrote 29, dropouts 38 and ongoing students 32.1 exams on average.

Assuming number exams written as a measure of experience, graduates seem to be least experienced. Therefore, and together with the lower success rate for students with experience, knowledge about the institutional process does not seem to pose a significant advantage for successful graduation.

With 44.3% graduates account for the biggest share of mandatory exams before CSB. Dropouts account for 30%, ongoing students for 31.2%. With 55,7% graduates are account for the smallest share of non-mandatory exams. Dropouts account for 70% and ongoing students for 68.8%. Graduates' efforts focus the most on mandatory courses compared to non-graduates. Nevertheless, the main focus of every group lies on non-mandatory exams. So previous study experience is primarily not focused on CSB.

The situation is different for exams taken during CSB.

Table 16 - exams during CSB differentiated by study status

	number exams			percentage			
	total	**manda-tory**	**non-manda-tor**	**abso-lute**	**manda-tory**	**non-manda-tory**	**exams per student**
graduates	9851	7931	1920	33.7	82.2	19.5	70.36
dropouts	6005	3173	2832	20.5	52.8	47.2	16.62
ongoing	13366	10468	2898	45.7	78.3	21.7	54.56
all	29222	21572	7650	100	73.8	26.2	38.86

29222 exams of the total 38839 are taken during CSB. Hence, 75.2% or about three quarter of the relevant exams taken by the total cohort is done during CSB. The content of Table 16 shows similar patterns as Table 14.

The numbers of exams per student are calculated by taking the number of students in -1/0 and 0 in Figure 11 into account. Graduates have on average the highest number of written exams, ongoing have the second highest number and dropout the lowest. The difference in amount of written exams between those groups is strongly distinct here.

Graduates and ongoing students focus strongly on mandatory, while ongoing dropouts spread their efforts evenly in those domains. If one assumes a limited amount of personal energy a given student can spend on pursuing a degree, this tendency of dropouts could be a limiting factor for success.

Study behavior before and during CSB clearly differ. But not only in different proportions in different exam categories but also in their temporal development and overall composition.

The data about the study behavior during CSB is complete and representative. All students starting to study CSB are represented in the dataset and all exams of them are recorded. The observation period is clearly defined and limited to 10 semesters. This and the similarities between single semester over cohorts ensure comparability and meaningful results upon examination.

The study behavior depicted in the dataset before CSB is not limited to a defined period of time but extends irregularly over 32 study semesters in the past. The recorded previous university experience is therefore very diverse and the study behavior heterogeneous. Since only students who later enrolled in CSB are recorded, it is also not possible to determine here whether the past study behavior over time is relevant or significant regarding enrollment to CSB. Further, students who appear here without prior university experience, could also have been studying at a different university and therefore do have experience. Such information is not included in the data.

Therefore, the available data for pre-CSB are not representative for the total population of CSB and therefore the temporal course remains disregarded.

About student retention in the data

In the previous subchapters, the exam data was carefully selected with the aim of understanding the exam behavior of students regarding their study success. Already before the selection not every student of the total cohort had an exam entry assigned. The selection of certain exams resulted in further students ending up having no exam assigned. The process of culling and selection subsets of exam data responsible for this is summarized in Figure 12.

As one can see in Figure 12, first only exams which were written by students of the total cohort were selected. Then all exams written after the CSB closure date of the student who took the respective exam were discarded. Also, all exams written after the 10th semester or outside the defined observation were discarded as well. After this, out of the total 80011, 38839 exams remained. The numbers of students present in the remaining exam entries are summarized in Table 13.

Next exam entries were differentiated into exams written before and during CSB. In the previous subchapter, exams taken before CSB did proof to be insignificant regarding success in CSB.

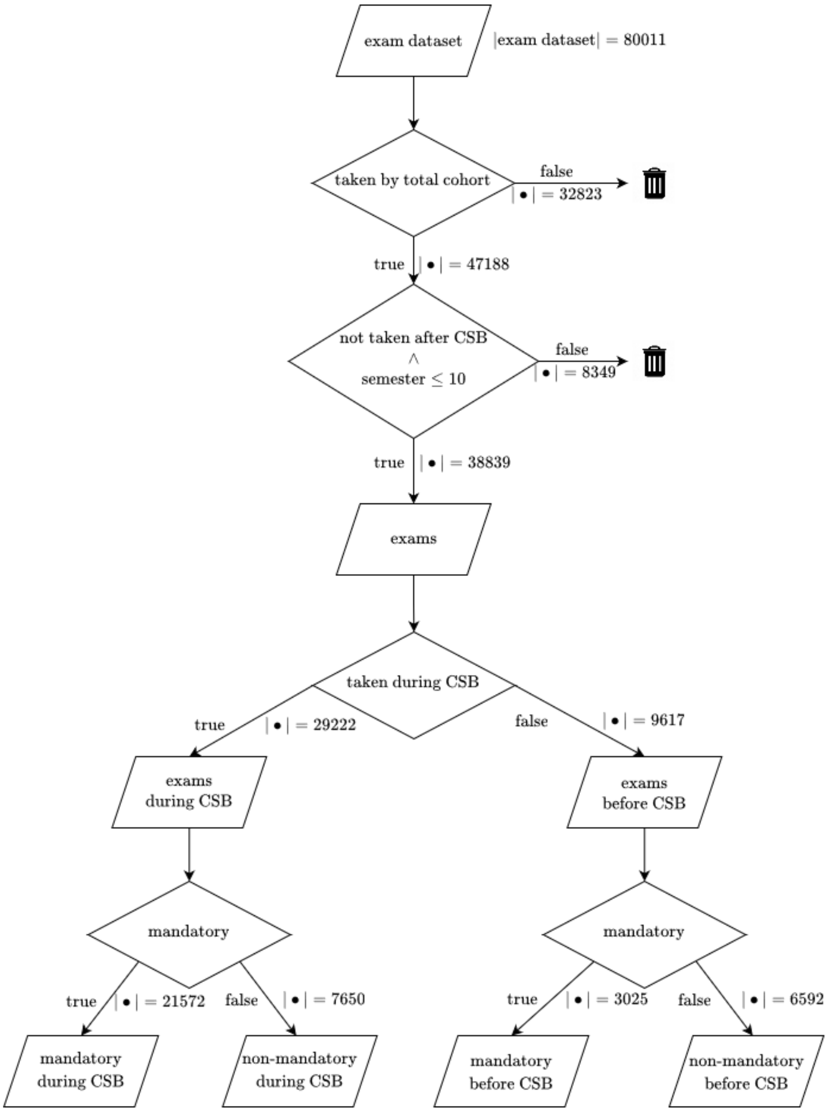


Figure 12 - exam selection and differentiation

So, the focus lies now on exams taken during CSB. Figure 12 also incorporates following considerations.

In the set of exams taken during CSB, again only a portion of the students are present. 18 students, all dropouts, only have exams written before CSB assigned to and are thus not present in exams taken during CSB. This can also be observed in Figure 11.

As already stated, a distinction between exams of mandatory and non-mandatory courses is necessary, since due to the curricular structure of the CSB, mandatory subjects are decisive for study success. If exam entries during CSB are further divided into two subsets of mandatory and non-mandatory exams, numbers of present students change once again.

The number of students present in the subsets of exam entries are summarized in Table 17.

Table 17 - numbers of student present in subsets of exam entries

	number of students present			
subset of exams	graduates	dropouts	ongoing	Total
Exams during CSB	140	367	245	752
Mandatory during CSB	140	323	239	702
Non-mandatory during CSB	138	212	215	565

3.2.4 Study behavior of the total cohort during CSB

In this section the study behavior of students in relation to exams written is to be analyzed in more detail. Since the aim of this work is to distinguish dropouts from graduates the examination focus on that. To do this, two questions must first be answered. Which exams entries should be used for the analysis and how should the chosen exam entries be evaluated and put into context to each other over time?

When deciding which exam entries should be used, it is best to remember the underlying objective. If one wants to understand the dynamics that determine academic success, the analysis must turn towards the study behavior of graduates. For there are many ways to fail completing a degree, but there are only a few to finish it. Essentially, dropouts do not have to meet any requirements, whereas graduates have to meet specific requirements like having done all mandatory exams at some point.

In Table 16 one can see that graduates focus their efforts mainly on mandatory exams and less on non-mandatory exams. Since graduates do not show a lot of efforts towards non-mandatory exams, their efforts are constantly focused on mandatory exams. Further, non-mandatory exams are only a small proportion of the CSB and as already mentioned most likely do not pose any serious obstacles for completing the degree.

Consequently, one can conclude that non-mandatory exams do not hold valuable information about study success. Further analysis therefore concentrates on mandatory exams during CSB.

For the analysis of study behavior exam entries of any given student must be put into context or be compared to the exam entries of any other given student. If all exam entries of one student are compared to all exam entries to another student, the result may not be meaningful, as students with different study durations would also be compared. Certainly, comparing a set of exams written in fewer semesters than the exams written which they are compared to would result in a distorted picture of reality. Thus, such a method of comparison should be avoided.

Another approach would be to compare only exam entries from a given semester. The semester of an entry here means the study semester of the respective student in which the entered exam took place. Thus, only academic performance at the same point in study would be compared with each other. The results of such a comparison would certainly be more reasonable but would suffer from the disadvantage that the previous study performance would be disregarded in such a comparison. However, the performance already achieved is undeniably decisive for a successful interpretation of the results of such comparisons. Consider two students who have already taken different numbers of mandatory exams up to a certain comparison semester. One student will have already done more necessary work to complete CSB than the other. The student who has already done more will possibly write fewer exams in the semester in which the two are compared, because the necessity of doing so is no longer given for him. His efforts, for example, could be mainly directed towards his bachelor’s thesis. A comparison of the exam entries of only that one semester between those two students would give the impression that the student having done less would be seem more productive and thus seem to be more likely to graduate. The fact that in truth the other student is more likely to graduate because he is closer to meet the requirements for graduation would remain unnoticed. While the results of such comparisons would be less tendentious, than results obtained with the former approach, they would still show a distorted picture of reality. If possible, comparisons of this kind would therefore be better avoided.

The last approach described here would be to make the comparisons based on those exam entries that apply to exams taken up to a given semester of comparison or observation point. This would include the overall performance of each student up to a certain semester of comparison in any comparison. So, this approach does not share adverse tendencies of the previous two approaches. Results obtained with it should give an accurate picture of reality at a given observation point. Therefore, further analyses will be conducted while this approach is pursued.

It should be noted that a different approach with the same qualities would be to take all exam entries of a student and only consider students with the same study duration for comparison. However, the duration of study is only known when a given student withdraws from CSB. Thus, it is also known whether a student is a dropout or a graduate. A comparison including the study duration would therefore include information that would anticipate the result and is therefore not particularly suitable. Unless one were to try to predict the study duration and indirectly conclude the future study status. But this is outside the scope of this paper.

An observation point, of the last-mentioned approach, is a semester up to which all exams written are used for analysis at a given observation point. Since observation period is limited to the 10th semester and the analysis is conducted on the base of exams entries applying to exams written during CSB there are a total of 10 observation points.

At any observation point a certain number of students is still enrolled in CSB. The number of students still enrolled is derived from students present in the subset of exam entries under analysis. Here, those are the students who are present in the exam entries applying to exams of mandatory courses taken during CSB.

For analysis at a given observation point only exam entries of those students are considered who are still enrolled at the given observation point. Otherwise, the observation period of compared exam entry subsets of students would differ and results obtained would be distorted. In the upper diagram of Figure 13 the number of students considered for a given observation point is shown for all possible observation points.

Students may write no exam during any semester. So, it is possible that some students are enrolled in CSB but did not write any exam up to a given observation point. The subset of those students' exam entries is empty and offers no information than the absence of just that.

In the lower chart of Figure 13 the numbers of students are displayed, who wrote at least one exam up to the respective observation point.

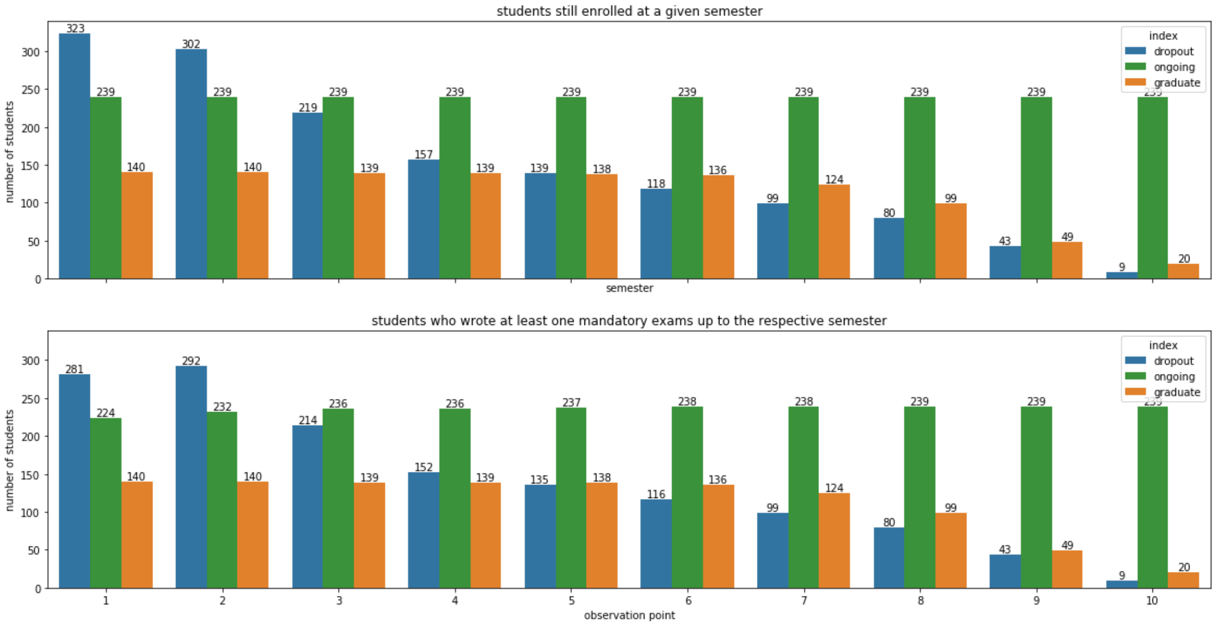


Figure 13 - number of students still enrolled or exam active up to an observation point

In other words, Figure 13's upper chart shows the students still studying CSB at a given semester and the lower chart shows the number of those students who already wrote at least one exam up to the respective semester so their exam entries can be compared to that of other students.

Graduates have mandatory exams written at every observation point, whereas dropouts and ongoings do not. The number of students studying CSB decreases steadily, but the number of students with at least one exam written at a given observation point increases for dropouts and ongoings. Hence one can observe that there are students who enroll, do not write mandatory exams for some semesters, and start doing so later. Stream of data for students is not constant, but intermittent. Students despite having the same enrollment date start studying at a different point in time, which complicates the analysis.

Figure 13 also illustrates a phenomenon that permeates the dataset: performance-based selection. As the semester number increases, the proportion of dropouts in the remaining cohort diminishes and that of graduates increases. Therefore, the higher the semester number, the higher the probability that a given student is a graduate.

With decreasing number of students who are writing exams, the number of written exams also decreases. This can be observed in Figure 14.

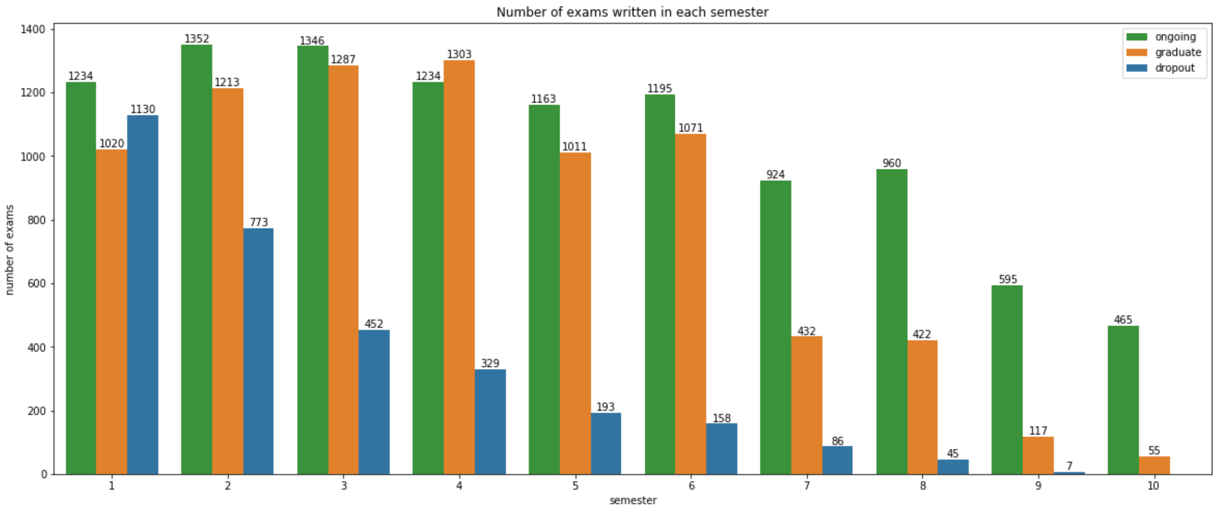


Figure 14 - number of mandatory exams written during CSB each semester

Considering Figure 13 and Figure 14 one can see that the exam activity also declines. The number of written exams per student decreases with increasing semester.

Time course of study

Before turning to analysis of the exam entries the general course of study should be evaluated under consideration of only the students present in the set of mandatory exams.

It should be noted that all graduates must be present in the set of mandatory exams. This is confirmed by Table 17. All students who are not present in the mandatory exams are not graduates. Since mandatory courses are the main component of CSB, all students who are present in the set of mandatory exams are more likely to have the intention of obtaining a degree and all who are not, are less likely to have serious intention of doing so.

The data of which the course of study shown in Figure 9 is derived contains students who are not present in the set of mandatory exams. Therefore, students are included who likely have no serious intention of completing CSB. Key figures in Table 11 are also derived from this data.

The general course of study should give an impression of the very nature of a study. So, this representation of the nature of CSB should be also given by only including students likely to have the intention to graduate in CSB. Table 17 shows the number of students present in the set of mandatory exams but lacks further information. Table 18 contains the description of Table 11 only including students present in the set of mandatory exams taken during CSB.

Table 18 – key figures of students present in mandatory exams written during CSB

		percentage		study duration	
	number	**absolute**	**relative**	**mean**	**variance**
total	702	100	-	6.6	13.9
graduates	140	19.9	100	8	1.8
late grad.	124	17.7	88.6	8.4	1
timely grad.	16	2.3	11.4	5.5	1.2
dropouts	323	46.01	100	4.6	7.8
early drop.	166	23.6	51.4	2.2	0.4
late drop.	157	22.4	48.6	7.1	3.3
still ongoing	239	34.05	100	-	-

If the values from Table 18 are compared to the former values of Table 11 one can see the portion of graduates increased from 16.6% to 19.9%. Also, the proportion of dropouts decreased from 54.1% to 46.01%, making the situation appearing less grim than before. However, the study duration of dropouts slightly increased form 4.1 semester to 4.6 semester. This increases the lost investment of dropouts. The portion of ongoing students after the 10th semester increased from 29.3% to 34.05%.

The development of the cohort studying more likely with the intention to complete CSB is also different from the development of the total cohort shown in Figure 9. The updated diagram regarding the temporal development of the cohort's composition can be seen in Figure 15.

The stereotypical shape of the study course remains unchanged. So, it is inert to the presence of students with no observable efforts towards completing CSB. However, as the tabular description changed, so did the illustration.

In the middle sub-chart, one can see that the only mode of the graduates in the 8th semester is even more pronounced and stands out even more clearly. The students who were present in Figure 9 and are not present in Figure 15 are all dropouts. In the lowest sub-chart showing the relative percentage one can see that values in the semesters of the late dropouts increased and so percentages of the early dropouts decreased.

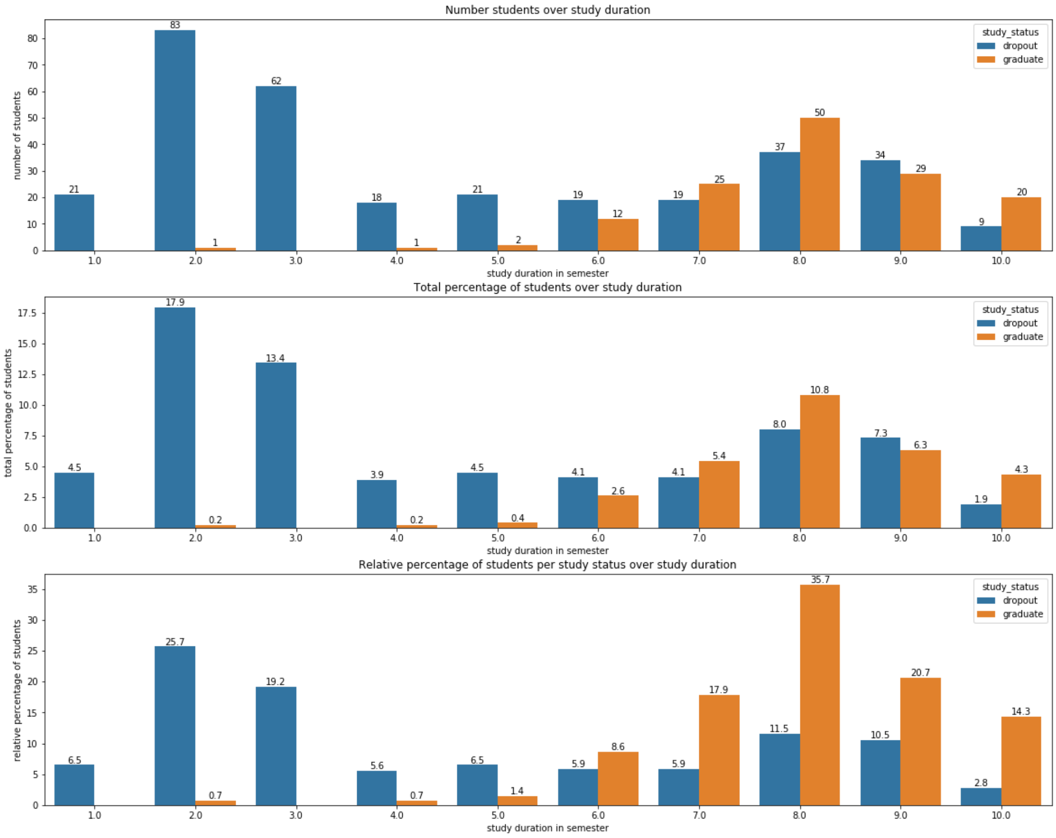


Figure 15 – Change over time in cohort of students present in exams written during CSB

Another link should be added. Figure 15 shows when any student closes the CSB and stops studying it. It shows the change of the number of students studying CSB and thereby it shows the change of the numbers displayed in Figure 13.

About measuring a students' progress

To graduate the objectives of the curriculum must be met. These objectives define the amount of work needed to be done for graduation. The amount of work is specified in the form of courses that must be completed positively. Each of these courses is assigned a share of the total workload by means of a key figure. In contemporary curricula, ECTS are used for this purpose. In the dataset here, however, the previous measure, namely hours per week in a semester or credit hours is still used.

A student's progress in the curriculum can therefore be measured by the sum of credit hours of the courses where exams have already been passed.

Figure 16 shows the sum of credits hours of all positive graded mandatory exams for every student at every possible observation point.

For completing CSB or any bachelor study 180 ECTS must be earned. As already mentioned in the current CSB curriculum mandatory courses account for 147 ECTS. With a conversion factor of 2 between credit hours and ECTS this corresponds to 73.5 credit hours in mandatory exams necessary to graduate from CSB. Since for reasons stated above only mandatory exams are analyzed here and the proportion of mandatory courses fairly stayed the same for CSB according to TUG officials one could except graduates having approximately 73.5 credit hours earned at graduation. However, the mandatory courses were only identified as such with some degree of accuracy. Accuracy is compromised, for example, by the fact that not all mandatory courses remain necessarily mandatory across all curricula. For these and various other reasons, the observed sums of credit hours may deviate from the expected sums of credit hours. As the deviations should be within limits, progress in studies should still be measurable.

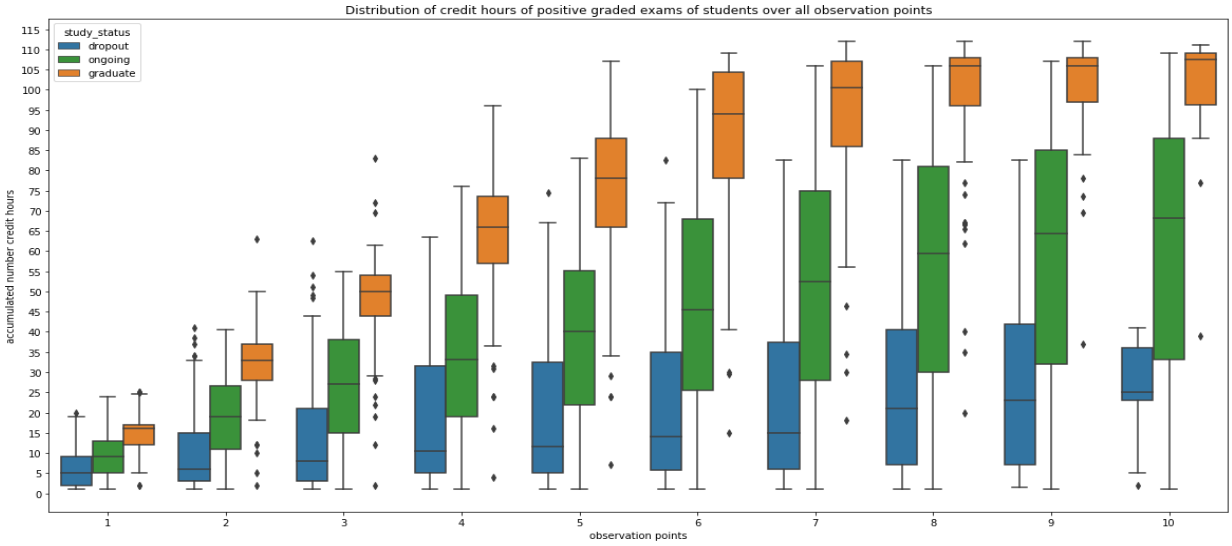


Figure 16 - Distribution of sum of credit hours of mandatory exams grades positive per student for all observation points

One can see in Figure 16 that the distributions of sum of credit hours is strongly related to the study status.

If the distributions of a given study status are compared to the distributions of the other study status, one finds that spread of the distribution of graduates is the smallest and the median the biggest over all observation points. The median increases significant and steadily over the observation points. The spread of dropouts' distribution is usually the widest and the median is the smallest over all observation points. The median increases slightly with the observation points. The distributions of ongoings lie between those of graduates and dropouts but overlap more with those of dropouts. At any observation point the distributions of the different study status do overlap. However, if only graduates and dropouts are considered the distributions barley overlap. Quite the contrary is the case. The 75th percentile of dropouts' distribution is significant smaller than the 25th percentile of the graduates' distribution over all observation points. This is also valid for the first observation point, where only exams from the first semester are considered. Despite the short observation period the difference of the distributions is already given. It follows that the accumulated sum of credit hours allows an easy distinction between graduates and dropouts at every observation point in the first 10 semesters.

The method of display in Figure 16 offers a good visual representation of the distributions under analysis, but it does not allow differentiation between individual students. The temporal development of the accumulated credit hours of individual students can therefore not be observed.

Figure 17 is intended to provide an overview that facilitates differentiation between the individual students. Figure 17 displays the temporal course of the accumulated credit hours for every student as a single line in the line-chart. Lines of graduates are colored orange; lines of dropouts are colored blue. The end of the line, when a student deregisters from CSB, is highlighted red.

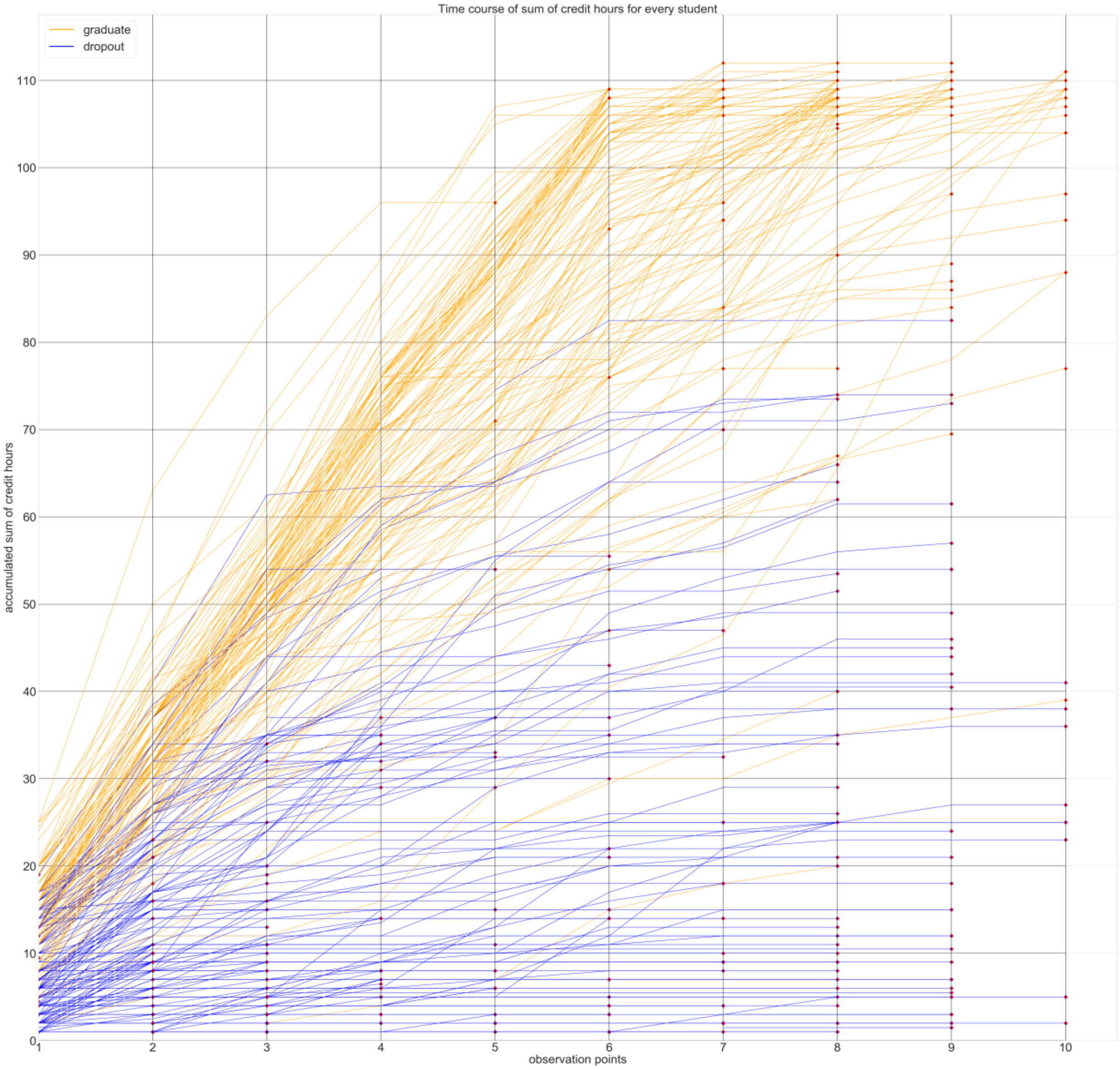


Figure 17 – Temporal development of accumulated credit hours of mandatory exams for every student

It is notable that graduates finish their studies with different numbers of credit hours. This could be due to several different causes.

Exams could have been written at another university, e.g. in an exchange semester. Nothing in the dataset suggest that such exams are included in the data files. Despite that the differentiation between mandatory and non-mandatory is, judged by the figure on hand, working well, it is not necessarily correct for all courses. Some courses may well be misclassified. Especially since the CSB curriculum was changed several times during the observation period. Also, exams could have been written at TUG, but before CSB. With the available information, these hypotheses can neither be confirmed nor refuted.

Besides some outliers, there are two clusters of lines observable. One where lines of graduates and one where line of dropout are clustering together. As before in Figure 16 accumulated credit hours separate dropouts and graduates.

There are two types of outliers. Students who behave like graduates at the first few observation points only to dropout at a later observation point and students who behave like dropouts to graduate in the end.

To identify dropouts behaving like graduates as dropouts early on is impossible with the feature used here and is most likely impossible with other features derived from the exam data. After all, progress in credit hours is the only thing that is necessary formally for graduation.

Graduates who behave like dropouts at first are another matter. Since credit hours of mandatory exams earned before CSB are not included, these outliers could be result of this. To shed light on this matter Figure 18 shows the temporal course of credit hours including mandatory exams taken before CSB.

Graduates still graduate with different numbers of mandatory exams. However, one can observe that some of those outliers indeed disappear. But also, the imaginary border between the clusters of dropouts and ongoings becomes more blurred. At early observation dropouts and graduates cannot be distinguished anymore. The included information is more disruptive than helping for the separation between dropouts and graduates.

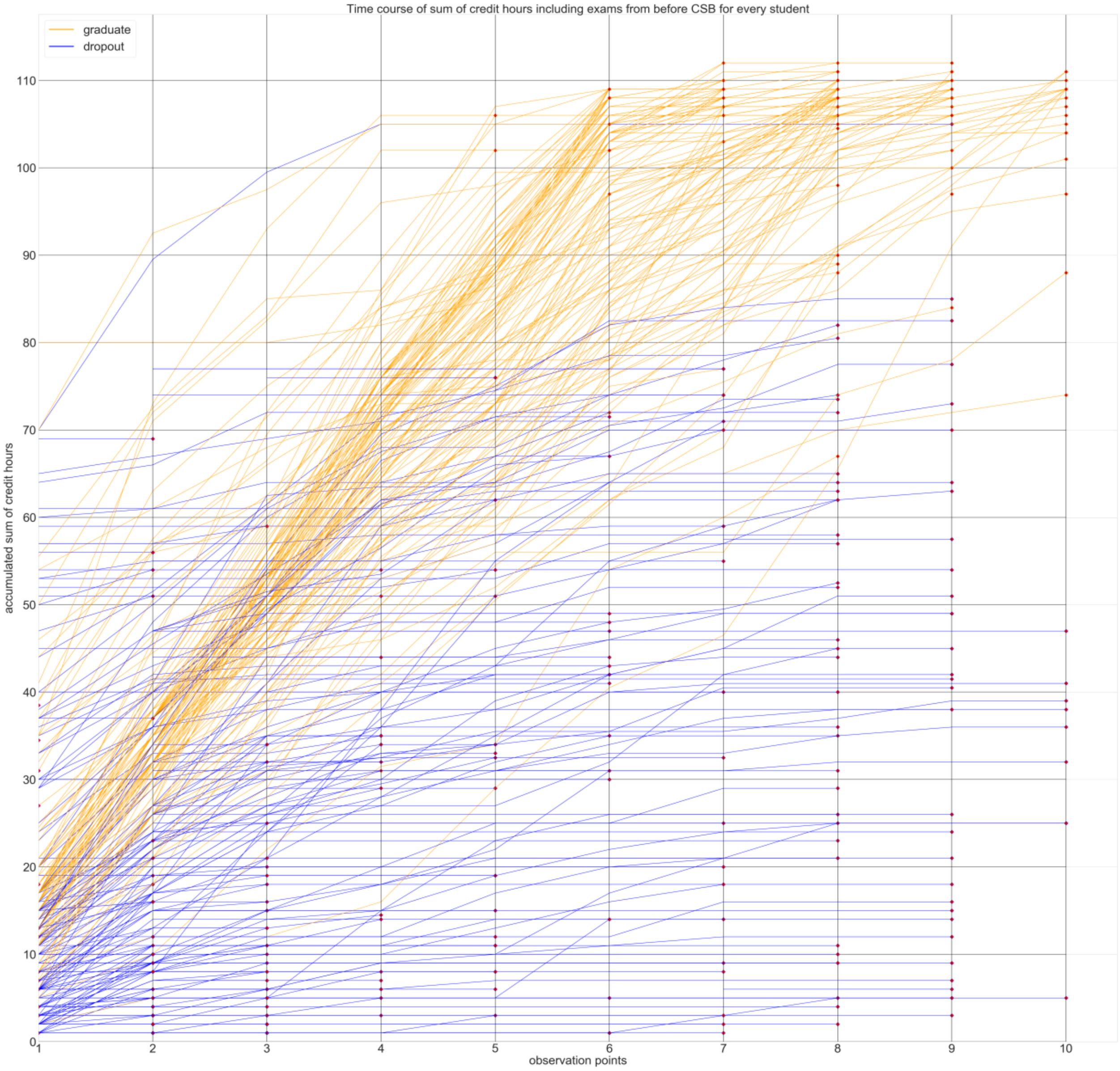


Figure 18 - Temporal development of accumulated credit hours of mandatory exams including exams written before CSB for every student

About average increase of credit hours

The previous study of the aggregate amount of credit hours revealed that graduates progress faster in their studies than dropouts. The increase of credit hours is greater. This can be observed in Figure 17 by the slopes. Those slopes are to be examined here. For this, the average credit hours per semester are calculated for each student at every observation point. Figure 19 is the resulting boxplot if the average credit hours per semester are plotted over the observation points for all students.

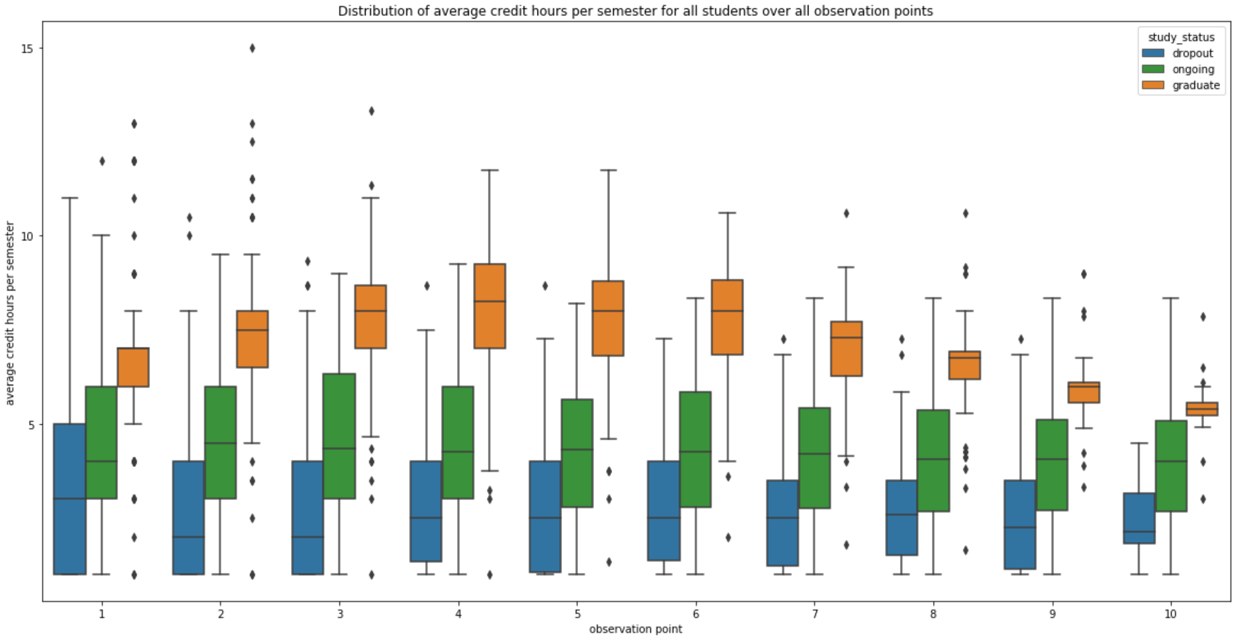


Figure 19 - average semester hours per semester

As already announced, in Figure 19 one can observe that the average credit hours per semester are related to the study status.

The increase of credit hours per semester tends to be the highest for graduates and the lowest for dropouts. The increases of ongoings tend to lie in between those of graduates and dropouts. Graduates' increase increases for the first observation point to the 5th and decreases afterwards. The progression in curriculum appears to be faster in the earlier semesters to slow down in the later semesters. The progression of dropouts and ongoings fairly stay the same over the observation points. However, the low border of dropouts' distributions increases over observation points. Meaning that the students progressing the slowest in CSB in terms of mandatory exams drop out first.

About grades

After examining the study progress, i.e. the amount of work already done, the work's quality is now to be examined. For this purpose, the grades of the exams taken are to be analyzed.

There are only grades from 1-5 in the mandatory exams written during CSB. The other 3 possible values for grades mentioned in Table 12 are not present in this subset of exam entries.

First, the average grades of the students are to be examined. Figure 20 shows the average grades of the students over all observation points in accordance with the established comparison approach.

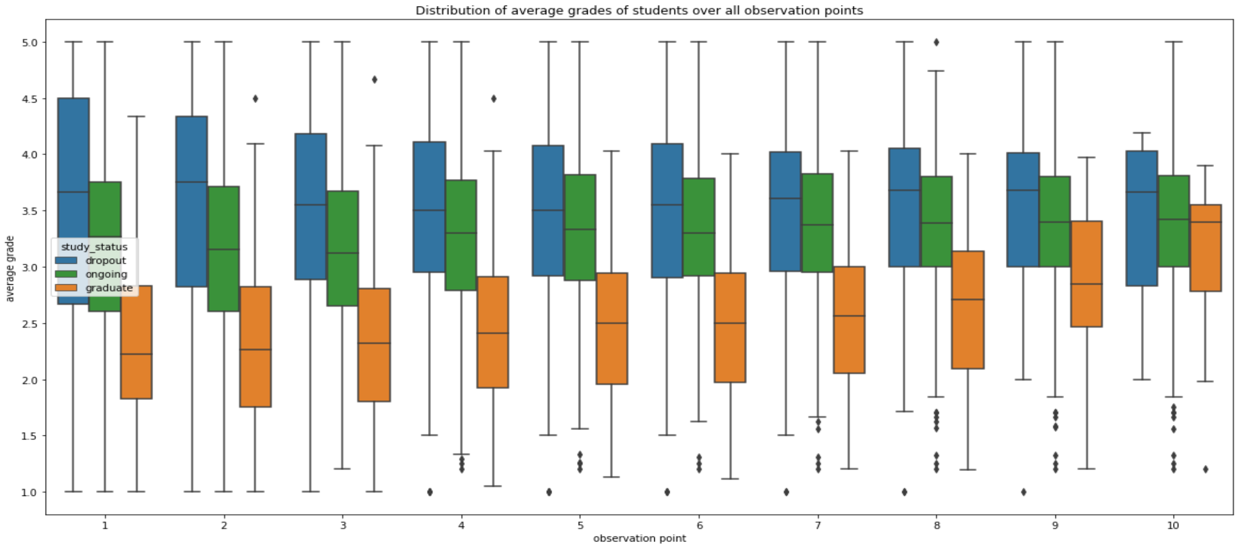


Figure 20 - Average grades of students over observation points

One can observe that distributions of average grades is clearly related to the study status.

The centers of the distribution of average grades of dropouts tend to be the highest compared to those of ongoings and graduates. The median is quite steady over the observation points and lies between 3.5 and 3.75. The spread tends to be the widest and narrows down with increasing observation point number. Only the last observation point is an exception. There, the spread increases again.

The centers of graduates' distributions of average grades tend to be the lowest. The median of the distributions steadily increases with the observation points. It increases from ~2.3 at the first observation point to ~3.4 the last observation point. The spread of the distributions narrows down with the observation points.

Centers of ongoings' distributions tend to lie in between those of dropouts' and graduates' distribution. The median is quite steady as the median of dropouts' distributions. The spread of ongoings' distributions narrows down with the observation points.

The distribution of graduates, dropouts and ongoing students do overlap. If only dropouts and graduates are considered, the overlap is smaller but still there. The size of the overlap is dependent on the observation point. It tends to be bigger at the first few observation points and the last few observation points and to be smaller in the middle observation points.

As illustrated in Figure 13, the number of students considered at each observation point strongly decreases with increasing observation point number. This is because only students who have not yet deregistered from CSB are considered. The increase of the medians of graduates shows that graduates who graduate later do have higher or worse average grades. This phenomenon is therefore another face of performance base selection only among graduates.

The decreasing spread of all distributions means that remaining students' averages lie closer together. The exception at the last observation for dropouts is likely caused by the fact, that only 9 dropouts are present. With less students considered the effect of each student on the shape of boxplot increases.

Overall, the irregularities are minor and little dependent on the respective observation point but strongly dependent on the study status.

An average grade does not reveal its composition, e.g. the grades it is calculated of. Now the composition of the average is to be examined more closely in Figure 21. Therefore, the proportion of exams graded a certain grade in the set of exams of each student is calculated for each grade. The calculated proportions of all students are shown in Figure 21. Since the grade averages do not change significantly with increasing observation points, all exams of a student are considered to calculate the respective grade shares and the temporal dimension is unconsidered.

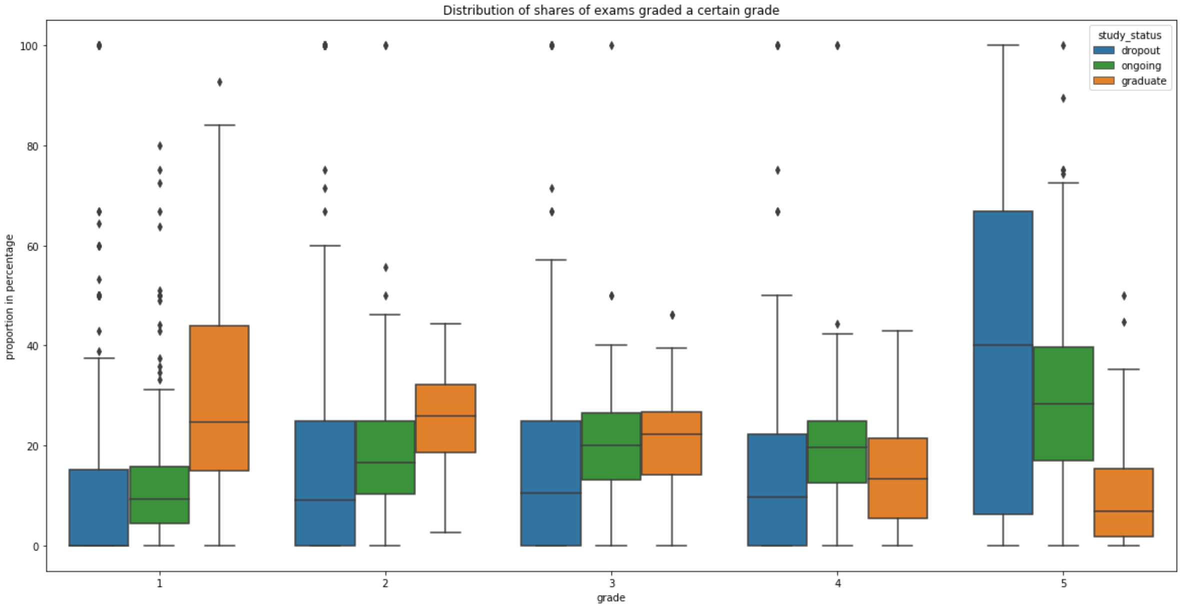


Figure 21 -proportions of grades for all students

One can see that graduates clearly tend to get more exams graded 1 then ongoings and dropouts. The distribution of proportions for grade 1 for graduates only slightly overlap with those for the other study status. With the increase or worsen of grades the distributions of shares for the different study status do overlap more and the behavior tends to be more similar. At grade 5, the only negative grade, the distributions differ once again. Dropouts clearly tend to get a great share of exams graded negative. Their distributions for the lower grades end at 0, the distribution for grade 5 does not. So, there are students who even only get negative grades. Graduates clearly tend to get the smallest share of exams graded 5. Dropouts and graduates can be differentiated best, when only proportions of exams graded 1 and 5 are considered.

To examine the constancy of students' performance in terms of grades, the standard deviation of grades is calculated for every student. Figure 22 shows the standard deviation for all students differentiated by study status over the observation points in a boxplot.

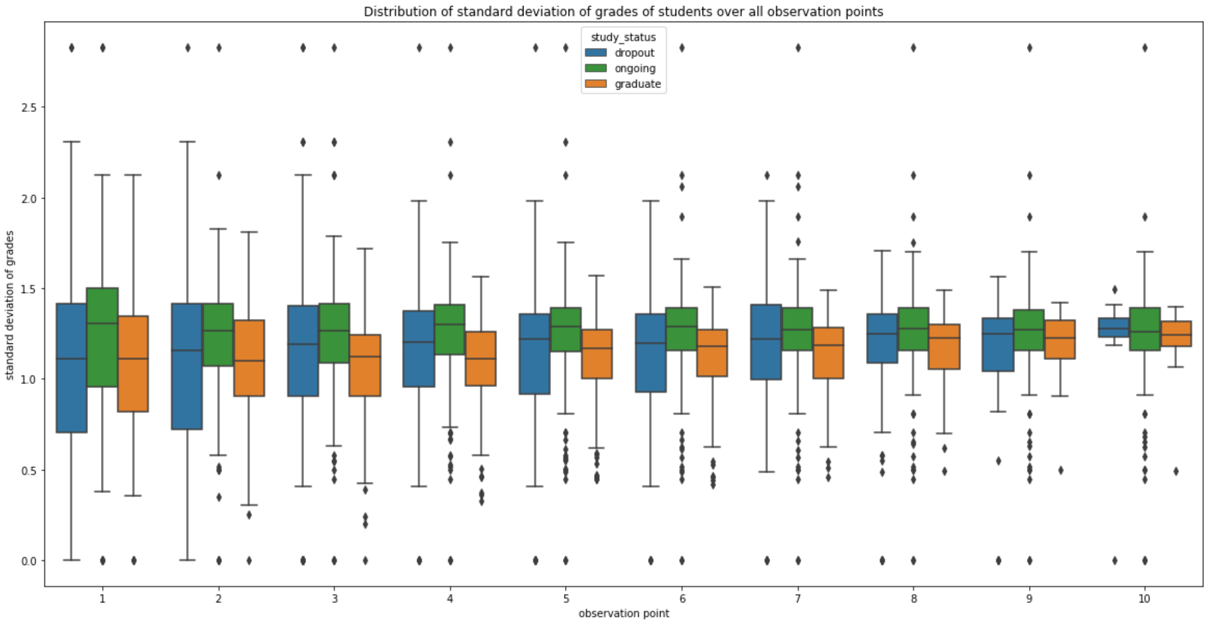


Figure 22 - standard deviation of the grades per student

One can see that ongoing students tend to have the most varying grades. This is reflected best by having the highest medians over nearly all observation points. The median of dropouts is slightly bigger than that of graduates at all observation points.

Also, the spreads are wider or consistency varies more among dropouts than among graduates. Meaning graduates grades are more consistent regarding the quality of their output. Also, the spread at early observation points is wider and narrows down with increasing observation point number. So, it narrows down with decreasing number of students and increasing study experience.

About inactive semesters

Students are not obliged by the university to write exams. It is perfectly possible for a student to pause a semester, write no exam resulting in an inactive semester. Next the number of inactive semesters is to be examined. Figure 23 illustrates the distributions of the sum of inactive semesters over all observation points.

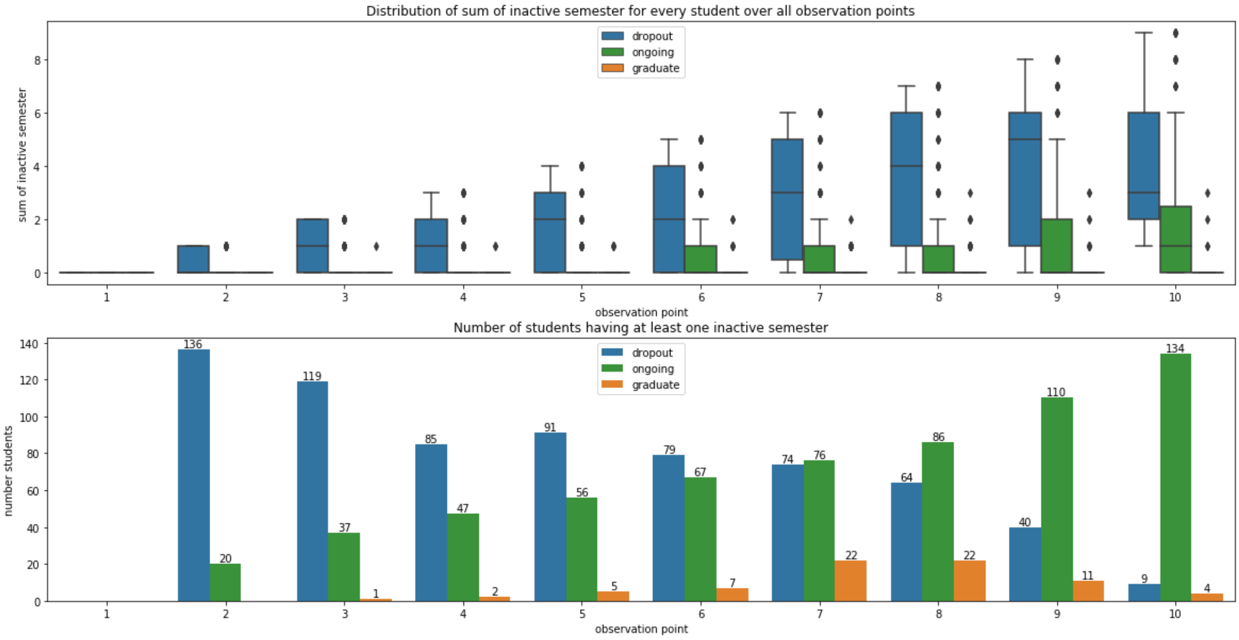


Figure 23 - Number of inactive semesters for every student

The upper chart of Figure 23 shows the distribution of the number of inactive semesters for every study status at every observation point. The lower chart shows the number of students having at least one missing semester at a given observation point.

One can observe that only a few graduates do have inactive semesters and if so, they tend to have them late. Caution is required if the inactive semesters are used to distinguish dropouts from graduates. As one can see in Figure 13 the number of students at later observation points is small. So, the number of graduates having at least one inactive semester at observation point 7 and 8 shown here is quite a significant proportion. So, it is more suitable for distinguishing dropouts and graduates in the early semester.

4. Predicting dropouts

After the discussion of the data, the gained knowledge should now be used to predict study success. Predicting student success is defined as predicting the prospective study status of a given student. Since the classifier predicting is constructed with the objective to provide support for managing dropout rates, the focus lies on predicting dropouts.

To obtain a classifier that can make predictions under realistic conditions, only the data which was considered valuable in the previous data culling process is used. Hence, for the reasons given above, only mandatory courses written during enrollment in CSB of the respective student are used.

The algorithm with the help of which the predictions are to be made has already been determined: logistic regression. The next step is to determine the architecture of the entire system.

4.1 System architecture

A system's architecture is dictated by the functional requirements and the situational requirements.

The prediction system to be constructed here should be functionally capable of making prediction as early as possible to facilitate timely countermeasures and not to miss too many dropouts. The availability of information to make predictions depends on the situation. Due to its nature, the university's data flow is clocked and divided into semesters. As the semesters progress, information gradually becomes available.

As more data gets available semester-by-semester, a presumed classifier has more data to base the predictions on. More exam data is available over a longer observation period and number of students decrease since some might already have quit studying.

So, data changes over time. Not only the amount of data available for students, but also the number of students on which the predictions of a classifier should base upon. It can be assumed that such serious changes in the data will affect the quality of the predictions. The proposed architecture in Figure 24 is intended to prevent this.

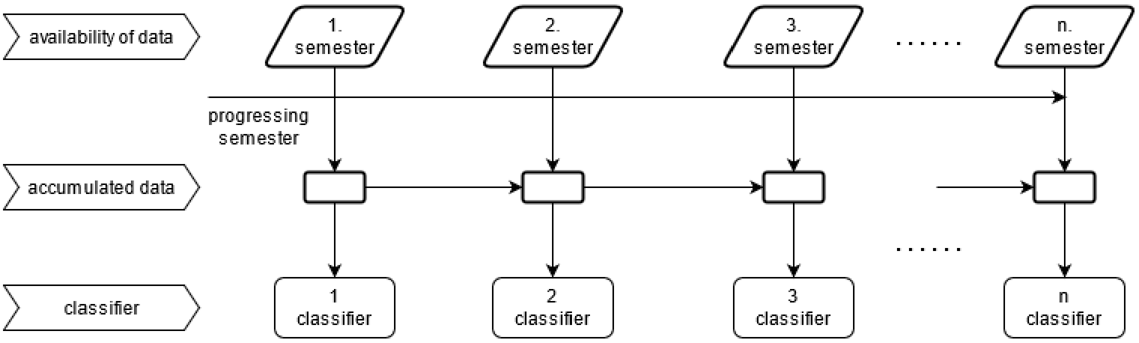


Figure 24 - System Architecture

Since data changes from semester to semester, the classifier changes as well, resulting in multiple classifiers. Every classifier gets exposed to the data up to a semester he is operating on and only considers students who did no quit studying in a former semester.

4.2 Setup

During the data analysis it became apparent that a limitation of the observation period to 10 semesters is appropriate for the data on which this work is based. This means that 10 classifiers are needed to implement a system with the architecture described in Figure 24. Due to the architecture, each classifier draws from a different set of data with different numbers of observations or students. The number of students presented to each classifier is shown in Figure 25.

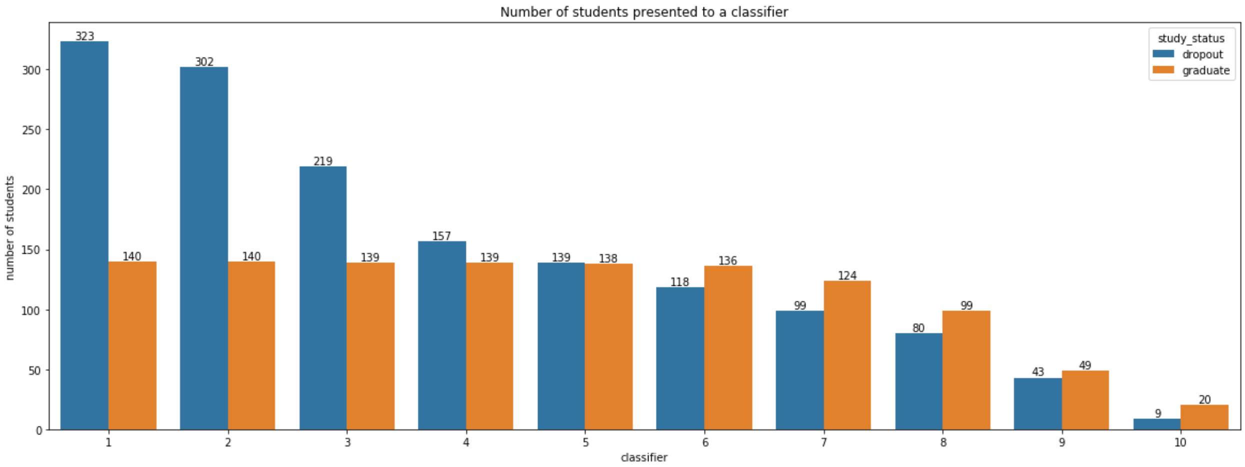


Figure 25 - number of students a given classifiers draws upon

Students presented to a classifier need to be described by distinctive features. The features used for this are the accumulated variables analyzed in the section *3.2.4 Study behavior of the total cohort during CSB.*

The previous analysis of the data showed that the accumulated variables could be well suited for differentiating between dropouts and graduates. To name them in summary, the features describing a given student are: 1. average grade, 2. variance of grades, 3. average credit hours per semester, 4. accumulated sum of credit hours, 5. number of inactive semesters, 6.-10. proportion of exams graded a certain grade from 1 -5.

The value ranges of these features are all numeric and can be understood in the figures in section *3.2.4 Study behavior of the total cohort during CSB.*

Here it is clearly visible what has already been mentioned above: The data only allows an analysis of dropout behavior based on performance data. The features used to describe a student very well reflect this fact.

Figure 1 shows the totality of the scientifically approved influencing factors.

One can see that many of the influencing factors for dropouts remain unnoticed since they are not included in the data. Although it can be assumed that these factors influence performance, it is not possible to determine which factor has which influence on the dropout rate based on the data.

5. Results and discussion

In this section, the system for detecting dropouts described above is realized and evaluated using the selected data. Then the results of the evaluation are presented and discussed. To enable the reader to place the results in the overall context of this work, an overview is first given in the form of a brief summary.

The results were obtained on the basis of the data selected in the course of this work. Which part of the total data basis is included in the selected data for prediction was decided by means of an elaborate data selection process.

The overall data on which this work is based includes students who enrolled into computer science bachelor (CSB) and computer science master (CSM) between years 2005-2015 or participated in the study years 2005 – 2014.

The total population of students who studied CSB in this time period appears to be completely recorded in the data set, whereas the total population who studied CSM appears to be incompletely recorded. To be able to make holistically valid statements with this thesis, the data for CSM were disregarded and only data for CSB were further considered. Hence, the analysis focuses exclusively on CSB.

When looking at the individual student cohorts, i.e. students who enrolled in the same academic year, irregularities became apparent. These could partly be attributed to the shortening observation periods for the cohorts resulting from the shortening relative distance between enrollment year and time of data collection for increasing study years. For this reason, the entire study period was limited to the first 10 semesters of each cohort, in which the main events take place due to the designed study duration of 6 semesters for bachelor’s degree programs. Thus, the 2011 to 2014 study cohorts were dismissed.

The study paths of the cohorts are very similar in comparison. Only cohort 2005 is an exception. The TUG administration stated that the bachelor’s and master's system was newly introduced in this very academic year. It can be assumed that the irregularity of cohort 2005 is due to this.

Therefore, the student cohorts 2006 to 2014 were selected.

The exam data available in the overall data set for students in the selected cohorts includes all exams taken by the respective students. Exams written after CSB were discarded due to their irrelevance to the analysis of CSB.

For the analysis of study success or failure, the focus on graduates turned out to be advantageous. Dropouts are extremely flexible in their failure, which results in a variety of multi-layered and diverse study paths whose analysis is correspondingly complex. For there are many ways to fail completing a degree, but there are only a few to finish it. Exams written prior to CSB were found to be insignificant to the success of graduates and were therefore also discarded.

Next, these remaining exams were differentiated into mandatory and non-mandatory for CSB. Due to the lack of availability of the corresponding curricula for the periods in question, this distinction was derived from the data set. Mandatory exams are the main component of CSB. Due to the resulting limited informative value of the non-mandatory exams, these were discarded. The procedure is also legitimized by the fact that all graduates do show appropriate performance in the mandatory exams and many dropouts do not. Exams in mandatory courses are therefore an excellent basis for differentiating between graduates and dropouts.

This remaining data correspond to the selected data.

To detect dropouts among the students in the dataset, 10 features, summarized in *Setup,* were derived from the data describing the students. The features suitability for the detection of dropouts was studied and confirmed in *Study behavior of the total cohort during CSB*.

For the realization of a classifier for the detection of dropouts, a system architecture was developed that tries to meet both the functional requirements of an early detection system and the situational requirements of the university as operational environment. The resulting system’s architecture is shown in Figure 24 - System Architecture It became apparent that, due to the time dependent change of data, a classifier for each of the 10 semesters covered in this thesis would be beneficial.

For the realization, all classifiers are trained and evaluated under the same conditions. All students are described by the 10 features. Only graduates and dropouts are considered, since it is yet not clear if ongoing students will graduate or dropout. Each classifier only is exposed to students still studying in the semester the respective classifier is responsible for.

Due to the data selection process, the data used is already of high quality. Nevertheless, there are some students whose set of exam entries is empty when features are calculated. So not all features for those students can be derived and those features are imputed with 0.

To gain independency of the specific configuration of the data, training and testing is done with 5-fold cross validation. As one can see in Figure 25 the proportions of dropouts and graduates is different for every classifier and the data is imbalanced for most of the classifiers. To avoid a biased model due to the data splits in the 5-fold cross validations, the splits are stratified. Since standard and min-max scalers have a negative effect on the results, scaling is not used. To avoid overfitting L2 regularization with $\lambda = 1$ is applied.

The obtain quantified results for analyzing the resulting classifiers, the key figures listed in Table 2 in section *2.1 Evaluation of a binary classifier* are calculated for each classifier. Since 5-fold cross validation is applied, 5 results per key figure are obtained for every classifier. Those results are averaged and summarized in Table 19. By averaging multiple key figures, key figures are hoped to be more reliable in presenting the true prediction power of such classifiers on such a dataset.

The results in Table 19 paint a clear and promising picture. Performance data indicates a high prediction power. This can be seen best in Table 20, where the averages and standard deviations of the key figures across all classifiers were calculated to facilitate interpretation.

The standard deviations are in the low single digits and due to value range of 0-100% low. This means that all key figures are relatively constant across all classifiers. However, standard deviations do differ slightly between key figures. Key figure with highest standard deviation is *recall*. Considering the single results for recall values in Table 19 it is evident that recall stays fairly steady for all classifiers and changes strongly to the positive for classifier 9 and 10. The increased standard deviation is a result of increasing prediction performance. Key figure with the second highest standard deviation is *specificity* and *false positive rate*. They both measure the same quality of a classifier and thus only *FPR* is considered.

A look at the single values of *FPR* reveals that the high standard deviation compared to the other figures is due to a positive trend. *FPR* is decreasing with increasing classifier number. Thus, with increasing semester graduates can be better distinguished from dropouts. The standard deviations of the other key figures are lower and the results of the single results for the key figures are therefore even more stable.

Since all key figures are relatively constant across all classifiers, the imbalance and change in data over the classifiers, observable in Figure 25 does not seem to compromise prediction quality.

Table 19 – Averaged results per classifier

Classifier	**Accuracy**	**Precision (Positive Predictive Value)**	**Recall (True Positive Rate)**	**Specificity (True Negative Rate)**	**False Positive Rate (1-Specifity)**
1	88.56	93.39	90.43	84.29	15.71
2	90.74	94.23	92.06	87.86	12.14
3	90.8	93.85	91.77	89.26	10.74
4	89.88	92.08	90.38	89.29	10.71
5	90.28	92.99	89.21	91.43	8.57
6	90.97	91.76	90.8	91.32	8.68
7	91.05	91.56	90.84	91.2	8.8
8	91.63	93.33	91.25	92	8
9	94.62	95	95	94	6
10	96.67	93.33	100	95	5

All means of all key figures shown in Table 20 lie around 90%, except FPR which lies around 10%. From the low standard deviations, it follows that all key figures are close to this means.

The average accuracy is 93.15%, the average recall is 92.17%, the average specificity is 90.57% and the corresponding average false positive rate is 9.44%. Comparing the key figures in Table 19 of the classifiers, according to the order of the classifiers, the values tend to improve, with a few exceptions. It seems that dropouts and graduates are better distinguished from each other as the study time progresses. Counteracting this tendency, which facilitates forecasting, is the decreasing amount of data available.

Table 20 - mean and standard deviation of key figures for classifiers

	Accuracy	**Precision (Positive Predictive Value)**	**Recall (True Positive Rate)**	**Specificity (True Negative Rate)**	**False Positive Rate (1-Specifity)**
mean	91.52	93.15	92.17	90.57	9.44
std	2.38	1.1	3.14	3.08	3.08

The results displayed in Table 19 and Table 20 are based on a decision threshold of 0.5. As mentioned earlier in section *2.2 Decision threshold*, a given decision threshold does not necessarily lead to the best predictive results in general or to the best results for a university that wishes to manage its dropout rates.

As already mentioned above, two key figures are of particular relevance for dropout detection. Namely, *recall* and *false positive rate*. *Recall or true positive rate* reveals what percentage of existing dropouts are identified as such and *FPR* reveals what percentage of the graduates are falsely predicted as dropouts. Thus, *TPR* measures how well dropouts are identified and *FPR* measures how well graduates are identified. In other words, these two metrics measure the predictive power of the classifiers' predictions relevant to dropout management. The best possible classifier would have a *TPF* of 1 or 100% and a *FPR* of 0 or 0%. Meaning all dropouts and all graduates would be detected as such.

The choice of threshold influences *FPR* and *TPR* for a classifier. To explore the potential for improvement through the choice of decision thresholds and to look at the predictive power of the classifiers from the perspective of TPR and *FPR* metrics, Figure 26 was plotted.

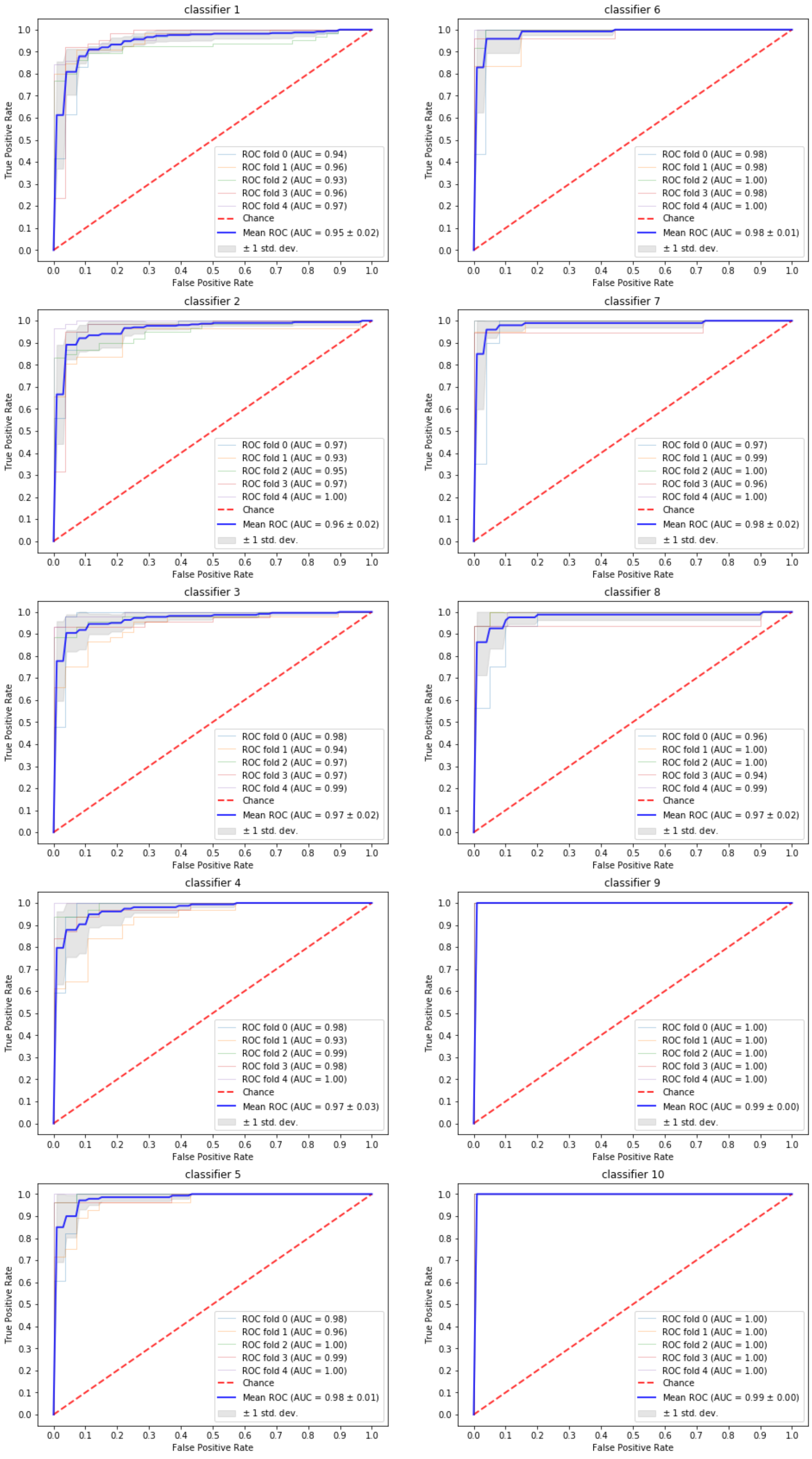


Figure 26 - ROC curve for every classifier

Figure 26 shows the ROC curve for every classifier. The data displayed was again obtained with training and testing with stratified 5-fold cross validation. The results of each fold are plotted and also the average result across all folds. Additionally, the AUC metric was calculated for comparison of the predictive power between the classifiers in the light of the trade-off between TPR and FPR.

The AUC measures the area under the ROC curve. It lies between 0 and 1. The higher its value the better the predictive power of a classifier and the weaker the relation between TPR and FPR. By comparing the average AUCs in Figure 26 over all classifiers one can see, that the AUCs are between 0.96 – 0.99. Hence, they are already close to the optimal value. Also, the AUCs are increasing with the classifiers' numbers. Meaning the prediction power is increasing with the semester. This indicated once more that dropouts and graduates are better distinguishable the more data is available.

However, above results are obtained without differentiation of student cohorts. So, temporal order was ignored. In an actual application of this system, however, the time order is predetermined and compliance with it is imperative.

Therefore, the proposed system is to be analyzed while considering the temporal order of student cohorts. This is done by using data of student cohorts 2006-2009 as training set and data from student cohort 2010 as test set. This results in different training and testing conditions than before. Training and test sets are not chosen randomly but based on their cohort year. The resulting training and test sets for every classifier are shown in Figure 27.

The division of the data into training and test data is now determined by the circumstances of the use case and not by considerations that should improve the learning environment for the algorithm. Figure 27 shows that the proportions of dropouts and graduates in the training data do not necessarily correspond to those in the test data. Split of training and test data is therefore not stratified. Also, the test dates are rigidly fixed. The key figures are therefore not averaged but determined only once.

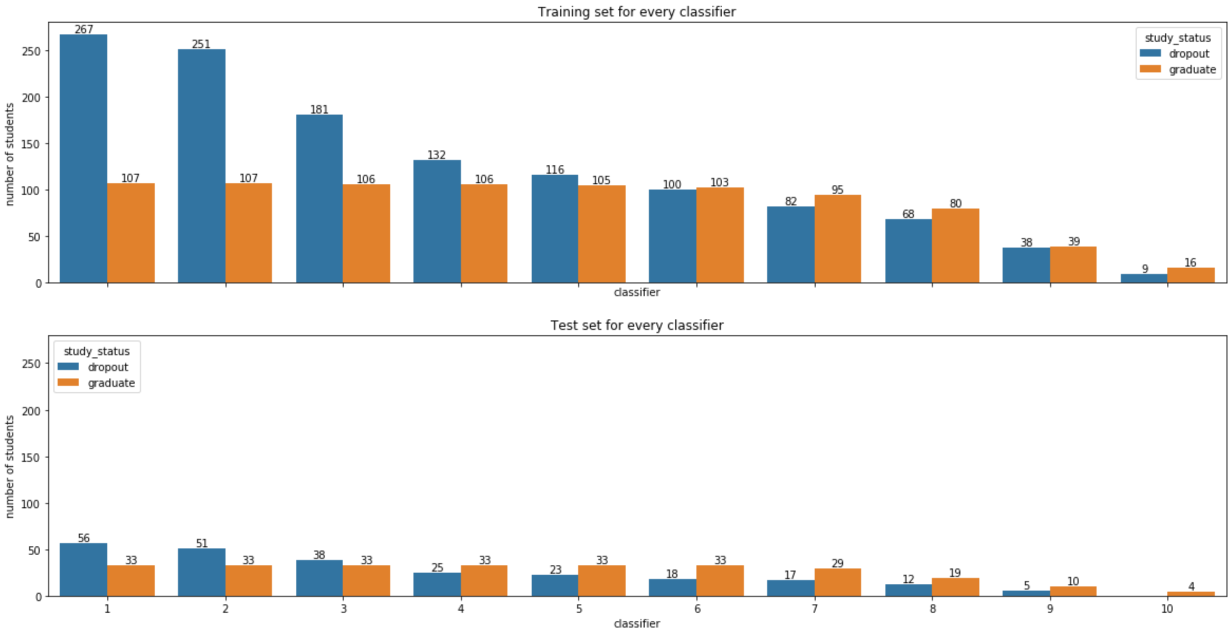


Figure 27 - Students presented to every classifier while considering temporal order of student cohorts

The same key figures as before are calculated and summarized for every classifier. The results differ, but still draw an optimistic picture. Table 21 contains the single values of every key figure for every classifier, Table 22 contains the average and the standard deviation for all key figures over all classifiers. In Table 21 *Precision* and *Recall* for classifier 10 are left empty as they are not obtainable since there are 0 dropouts in the test set for classifier 10. This can be seen in Figure 27.

The standard deviations displayed in Table 22 are still in the single digit but here in the upper instead as above in the lower.

The highest standard deviation results for *FPR*. A look at the single values for this key figure reveals that it varies between 6-30%. It is the highest for classifier 1 with 30%, then drops down to stay low, until it rises again for classifier 8, 9, 10. The key figure with the second highest standard deviation is accuracy. *Accuracy* fairly stays constant for all classifiers, except for classifier 1 and classifier 10. The result for classifier 10 is probably due to the sparse data basis. The accuracy for classifier 1 is lower here in Table 21 and also in Table 19 when temporal order was not considered. This is not particularly surprising, as only data from the first semester are available for prediction. The difference between graduates and dropouts is not yet distinctive, as the data analysis showed. The standard deviations of *precision* and *recall* are low. Those key figures stay steady across all classifiers.

Table 21 - Averaged results per classifier if temporal order is considered

Classifier	Accuracy	Precision (Positive Predictive Value)	Recall (True Positive Rate)	Specificity (True Negative Rate)	False Positive Rate (1-Specifity)
1	87.64	84.62	98.21	69.7	30.3
2	94.05	92.59	98.04	87.88	12.12
3	94.37	94.74	94.74	93.94	6.06
4	94.83	92.31	96	93.94	6.06
5	96.43	92	100	93.94	6.06
6	96.08	90	100	93.94	6.06
7	95.65	89.47	100	93.1	6.9
8	93.55	85.71	100	89.47	10.53
9	93.33	83.33	100	90	10
10	75	-	-	75	25

The means deviate from each other more than before. The *recall* is with 98.55% striking. This means on average 98.55% of the dropouts are predicted as such. With the low standard deviation, it follows that all classifiers detect dropouts reliable. However, what also stands out is the *FPR* with 11.91% on average. As already mentioned, *FPR* is strongly increased for classifiers 1 and 10. For classifier 10 this is not problematic, but for classifier 1 students are assessed after the 1st semester. Being classified as a dropout after the 1st semester could likely be demotivating for students. These undesirable potential effects must in any case be considered when applying the system and deriving policies.

Table 22 - mean and standard deviation of key figures for classifier if temporal order is considered

	Accuracy	Precision (Positive Predictive Value)	Recall (True Positive Rate)	Specificity (True Negative Rate)	False Positive Rate (1-Specifity)
mean	92.09	89.42	98.55	88.09	11.91
std	6.49	3.99	2	8.67	8.67

Since analysis of the results here resembles analysis of the results before it should be noted that drawing the ROC curve is not of any use here. Decision threshold can only be chosen on data already labeled, meaning students which status is already determined. This scenario is meant to evaluate the system under operational conditions where the study status of the student to predict is yet not available. In other words, in a practical application the system should give prediction regarding students still studying. It is not determined yet if a student will dropout or graduate. So, it is not possible to evaluate the system's performance and to choose more suitable thresholds base on such students or such an scenario.

The results show that the prediction of study success for single students based on exam data is possible. The system's architecture seems to meet the situational requirements and shows a good performance and gives reliable results. The fact that predictions regarding dropouts of this quality are possible proves a clear connection between students' exam performance and their dropout behavior.

The system performance was evaluated in two different scenarios. In the first scenario, all selected data was used equally to train and evaluate the system. Results obtained by the resulting system summarized in Table 19 and Table 20 show that dropout behavior is strongly related to exam performance. By not differentiating the data while learning, this system can be thought of establishing a general relation between dropout behavior and exam performance which is valid for the study CSB and time indifferent. Therefore, it is best suited to examine this connection further and in greater detail. Further work could address this issue. The knowledge gained could in turn contribute to the improvement of the dropout detection system. Since it maps the general connection, it is also suitable for establishing general policies.

In the second scenario, the data was used for training and evaluating the system in the order in which it was generated. Concretely, the system was trained with data from the 2006-2009 cohorts and evaluated with data from the 2010 cohort. This procedure corresponds to the use of the system as a dropout prediction system. The system learns predicting dropouts on former data and is used for predicting dropouts in future cohorts. Results of the approaches summarizes in Table 21 and Table 22 show that the application of this system is appropriate for this purpose. However, in case of application, the initial high FPR should be considered.

To test the proposed system, all classifiers were parameterized equally in both scenarios, including the decision threshold of 0.5. Results could improve if classifiers were to be optimized individually. The influence of different imputation methods was only briefly investigated. It could proof to be interesting to further analyze different imputation methods in future research.

The process of analyzing and selecting the data could also be further elaborated. One approach, for example, could be to subdivide and analyze the student cohorts not only according to enrolment in a specific academic year, but also according to enrollment in summer or winter semesters.

Finally, it should be noted that the correlation shown in the two scenarios is not necessarily a causal relationship. As already mentioned, a wide variety of factors can promote dropouts. These factors can make their appearance through the exam performance and their influence cannot be further differentiated here. A low performance in exams is not necessarily a reason for not graduating.

6. Conclusion

The aim of this work was to develop a system for the early detection of students at risk of dropping out. A system with which universities, especially Austrian universities, should be given a tool with which they can identify students at risk of dropping out to design and implement supportive measures at student level and thus reduce their dropout rates.

For this purpose, the university *Graz University of Technology* provided the entire pseudonymized examination data of students of the Computer Science program of the period from 2005 to 2015.

The developed system is deliberately based only on rudimentary examination data. Because of this, the system only needs to draw on data in the direct possession and sphere of influence of a given university. Since universities already record examination data, this allows for a quick implementation and application of the system, as no data collection or involvement of external stakeholders is necessary.

Variables derived from the examination data to make predictions whether a given student is at risk of dropping out or not are carefully chosen. All of them are accessible to direct human interpretation and have significance in their own right. Furthermore, logistic regression was chosen for the algorithmic implementation of making predictions regarding single students. The choice of variables and choice of algorithm makes the resulting overall system fully comprehensible and interpretable as are its predictions. As a result, the system can be used not only to make predictions regarding dropout risk but also to investigate the underlying dynamics of dropouts. Additionally, the question why a given student is assessed to be at risk of dropping out can be answered. This is what makes it possible to use the results of such a system in the implementation of regulatory steering mechanisms in a constitutional state.

The architecture of the system is such that it can be integrated into the organizational processes of the universities by providing or updating predictions regarding students after each semester.

The dynamics of dropouts differ from degree program to degree program. The architecture of the system is designed and intended for use on single degree programs. This allows the user to address the differences of each degree program during the application of the system by selecting significant examination data of the respective degree program on which the system is applied on.

Since the system was developed using data from the Computer Science bachelor’s degree program, data was analyzed and selected by their significance in this degree program for evaluation. This meant only using examination data from mandatory courses since those account for the main proportion in the curriculum and positive completion of them is necessary for graduation. Those requirements might not hold for all degree programs, although they are met in most engineering curriculums where dropout rates are severest.

In a simulated first application with the provided data, where those requirements are met, approximately 98% of dropouts were detected already in the first semester. However, this was based on a specific study year and results for different study years may differ. Also, approximately 30% of graduates were classified as dropouts after the first semester for the same study year.

The proportion of graduates wrongly classified as dropouts decreases strongly for predictions for higher semesters, but this clearly shows that no matter what measures are taken to manage dropout rates, both future dropouts and future graduates will always be affected. Any system user should keep in mind, that a distinction between dropouts and graduates is never exact. Both dropouts and graduates can either be supported by measures implemented to manage dropout rates and thus be motivated or restricted and thus be demotivated. The former certainly lowers the dropout rate, the latter is less likely to do so. Implementing measures for managing dropout rates inherently bear the risk of causing self-fulfilling prophesies, and poorly implemented measures will certainly cause them. Therefore, future research should address how the system can be applied beneficially in practice. Also, in terms of how such a system can be integrated into institutional processes so that no harm is done by naive usage. The system was evaluated with old data. It would certainly be interesting to apply the system to newer data and evaluate it in the course of future research. Application of the system on data of other study programs would also be highly interesting.

Proper application of the system makes it possible to detect dropouts. Predictions can be made after any semester. The system itself and its predictions are transparent and interpretable and can therefore be utilized institutionally.

Universities are thus enabled to offer individual support to students and to examine and better understand the systematics of dropouts on the level of degree programs.

List of References

All links were last visited in April 2021.

Arbeitsmarktservice Österreich. (2019). *Arbeitsmarktdaten im Kontext von Bildungsabschlüssen*. https://www.ams.at/content/dam/download/arbeitsmarktdaten/österreich/berichte-auswertungen/001_am_karten_2019.pdf

Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–17. https://doi.org/10.5281/zenodo.3554658

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*(October 2019), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Binder, D., Unger, M., Thaler, B., Ecker, B., Mathä, P., & Zaussinger, S. (2017). *MINT an öffentlichen Universitäten , Fachhochschulen sowie am Arbeitsmarkt - Eine Bestandsaufnahme*. https://irihs.ihs.ac.at/id/eprint/4284/1/2017-ihs-report-binder-mint-universitaeten-fachhochschulen.pdf

Bourdieu, P. (1988). *Homo Academicus*. Stanford University Press.

Bowers, A. J., Sprott, R., & Taff, S. (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*, *96*(2), 77–100. https://doi.org/10.1353/hsj.2013.0000

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Brandstätter, H., Grillich, L., & Farthofer, A. (2006). Prognose des Studienabbruchs. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologi*, *38*(3), 121–131. https://doi.org/10.1026/0049-8637.38.3.121

Bundschuh-Rieseneder, F. (2008). Good Governance: Characteristics, methods and the Austrian examples. *Transylvanian Review of Administrative Sciences*, *24*, 26–52. https://www.rtsa.ro/tras/index.php/tras/article/viewFile/91/87

Cutler, D. M., & Lleras-Muney, A. (2010). Understanding differences in health behaviors by education. *Journal of Health Economics*, *29*(1), 1–28. https://doi.org/10.1016/j.jhealeco.2009.10.003

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Gaebel, M., Hauschildt, K., Mühleck, K., & Smidt, H. (2012). *Tracking Learners' and Graduates' Progression Paths TRACKIT*. https://eua.eu/downloads/publications/tracking learners and graduates progression paths trackit.pdf

Curriculum CSB, 1 (2020). https://online.tugraz.at/tug_online/pl/ui/$ctx/wbSPO.downloadStudienVerlaufsplanPub?pStpStpNr=861&pVerlaufsplanDocNr=3025795

Heublein, U. (2014). Student Drop-out from German Higher Education Institutions. *European Journal of Education*, *49*(4), 497–513. https://doi.org/10.1111/ejed.12097

Hoeschler, P., & Backes-Gellner, U. (2014). *Shooting for the stars and failing: The effect of college dropout self-esteem*. *100*, 1–25. https://doi.org/10.5167/uzh-173560

Janger, J., Firgo, M., Hofmann, K., Kügler, A., Strauss, A., Streicher, G., &

Pechar, H. (2017). *Wirtschaftliche und gesellschaftliche Effekte von Universitäten*. https://www.wifo.ac.at/publikationen/studien?detail-view=yes&publikation_id=60794

Larsen, M. S., 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.

Leary, M. R., Schreindorfer, L. S., & Haupt, A. L. (1995). The Role of Low Self-Esteem in Emotional and Behavioral Problems: Why is Low Self-Esteem Dysfunctional? *Journal of Social and Clinical Psychology*, *14*(3), 297–314. https://doi.org/10.1521/jscp.1995.14.3.297

OECD. (2019). *Bildung auf einen Blick 2019: OECD-Indikatoren*. https://www.oecd-ilibrary.org/docserver/6001821mw.pdf?expires=1604666444&id=id&accname=guest&checksum=4CAB3A42AC9381EA6D77B286D562E1A1

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, *2*(1), 37–63. http://arxiv.org/abs/2010.16061

Statistik-Austria. (n.d.). Ordentliche Studierende an öffentlichen Universitäten 1955 - 2019. In *Statistik Austria, Hochschulstatistik*. https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bildung/hochschulen/studierende_belegte_studien/index.html

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Ulriksen, L., Madsen, L. M., & Holmegaard, H. T. (2010). What do we know about explanations for drop out/opt out among young people from STM higher education programmes? *Studies in Science Education*, *46*(2), 209–244.

https://doi.org/10.1080/03057267.2010.504549

Unger, M., Binder, D., Dibiasi, J., Engleder, J., Schubert, N., Terzieva, B., Thaler, B., Zaussinger, S., & Zucha, V. (2020). *Studierenden-Sozialerhebung 2019 – Kernbericht*. https://irihs.ihs.ac.at/id/eprint/5383/1/2020-ihs-report-unger-studierenden-sozialerhebung-2019.pdf

Unger, M., Wroblewski, A., Latcheva, R., Zaussinger, S., Hofmann, J., & Musik, C. (2009). *Frühe Studienabbrüche an Universitäten in Österreich*. https://doi.org/10.5167/uzh-68644

Universitätsfinanzierung NEU. (2017). https://www.bmbwf.gv.at/Themen/HS-Uni/Hochschulgovernance/Steuerungsinstrumente/Universitätsfinanzierung.html

List of Figures

List of Tables