Sandra Schulz (Hrsg.): Proceedings of DELFI Workshops 2019 Berlin, Germany,
September 16, 2019

Mining Academic Data to Support Students' Advisors: A Preliminary Study

Lennart Egbers[1], Agathe Merceron[1] and Stephan Wagner[1]

Abstract: Many universities take measures to reduce the number of students dropping out. To support students' advisors better becomes crucial. Besides their knowledge that they acquire through experience, which is a very important human factor in that process, advisors usually know very little about how students get along in their studies. In this paper, we present preliminary work to support advisors better when meeting students. The current investigation includes two main parts called "overview" and "typical completing behaviours". The overview part contains visualizations giving general information about how students manage the degree as well as information contrasting students who complete the degree and students who drop out. Typical completing behaviours are obtained through clustering. In this work, data from 2276 students have been analysed.

Keywords: Students' advisors, drop-out students, time to graduation, interactive dashboard, typical completing behaviours.

1 Introduction

Many universities take measures to reduce the number of students dropping out. One of such measures taken in some universities in Germany is a mandatory meeting between a student and the degree advisor if a student is behind in her/his third semester of studies. To be behind is usually defined as having earned a third or less of the credits of the first two semesters. Besides the knowledge that they acquire through experience, which is a very important human factor in that process, advisors know very little about how students get along in their studies. For example, most of the bachelor degrees are designed to take full-time six semesters. However, many students work beside their studies and the majority of them need more semesters. Usually, advisors do not know how many students need six, seven or more semesters to complete successfully their degree. When they study longer, how do students shift the courses over the semesters? Do they enrol in many courses and repeat them, or do they enrol in a few courses each semester? All this information is not known and advisors, as well as program directors, would find it useful to be more knowledgeable on how students get along during their studies.

In this paper, we present preliminary work to support advisors better when meeting students. Presently, the investigation includes two parts called "overview" and "typical

[1] Beuth University of Applied Sciences, Fachbereich VI – Informatik und Medien, Luxemburger Str. 10, 13353 Berlin, {s71948, merceron, s68772}@beuth-hochschule.de

DOI: 10.18420/delfi2019-ws-102

completing behaviours". The overview part contains visualizations giving general information about how students manage the degree as well as information contrasting students who complete the degree and students who drop out. This information includes: how many semesters do students take to complete the degree? How do enrolments and marks distribute over the courses? How do students dropping out compare with students completing the degree?

For the second part, completing students have been clustered according to their marks and their number of enrolments in all compulsory courses of the degree. The four resulting clusters give insight on how students succeed. In this contribution, the data from a six-semester bachelor including 2276 students have been analysed.

The rest of the paper is organized as follows. The next section contains related works. Section three presents the context of this work. Section four is devoted to the analysis of the data. The paper ends with a conclusion and future works.

2 Related works

On one hand, a lot of work has been done to predict students at risk of failing their study program or to predict the mark of students at the end of their degree. Some of these works show the relevance of the results for advisors. From the many works predicting students at risk, the finding made by [DPV09] is for us particularly interesting: Dekker at al. have found that the strongest course predictor of success used by advisors so far had not been identified as such by the algorithms. This finding led advisors to revise their approach. [As17] predict the mark at the end of a four-year bachelor degree. Interestingly, the high school certificate is not a predictor of the mark of the bachelor degree, which came as a surprise for the leaders of the degree program; further, they have found that students tend to have the same kind of marks (good, average or low) in all courses during the four years. [Zi15] show that marks of a three-year bachelor degree are essential to predict marks in the follower master degree. While analysing how students' paths differ from the ideal one as designed in the curriculum, [CMS12] have found that students who take longer to complete the degree have lower marks than students who stick to the time foreseen in the curriculum.

On the other hand, substantial work has also been done to recommend students which courses to take in the next semester, see for example [Ba18]. Less work has been done to support specifically human students' advisors when they meet students to advise them. Schwendimann et al. have conducted an extensive review of learning dashboards published in the literature [Sch17]. In this work, they define a learning dashboard as "a single display that aggregates different indicators about learner (s), learning process(es) and/or learning context(s) into one or multiple visualizations". From the 55 dashboards that were analyzed in that survey, the majority was directed to teachers or students. Only four dashboards were designed for teachers and administrators, who are not specifically advisors, and one for learning & design managers. A more recent survey on learning

dashboards focuses on dashboards for learners [Ji18]. The research most related to our aim is the work by [Mi18]. The authors make available to the advisor a dashboard showing the current marks of the students, its comparisons with peers, a planning module and a diagram showing different study lengths based on historical data. This dashboard supports the conversation between the advisor and the student and has proven to be useful. An ultimate aim of our work is to take into account conjointly marks and study length as done in [Ba18] in a planning module.

3 Context

The data of 2276 students of a six-semesters Bachelor study-program since its creation in 2005 as collected by the central administration of a German University have been analysed. The data contain for each student his/her date of entry in the University, high school certificate mark when present, every single course s/he enrolled in, when and the mark earned as well as the graduation date for completing students.

To earn a degree, a student has to pass successfully every single course. A student has a finite number of attempts to pass a course, usually three. Some universities put a limit on the number of enrolments for a course, like four enrolments. If s/he fails the third attempt on a single course or does not pass a course within four enrolments, a student gets exmatriculated. This is one reason for dropping out. The most common reason for dropping out is, however, students quitting the degree by themselves. It is well known that many students abandon the study program during the first semester. Other students, though, abandon later. Some universities are introducing special advice for students in their third semester targeting those that are behind. Being behind is, in our case, defined as having earned one third or less of the 60 credits from the first two semesters. An outcome from such special advice should be a personal study plan leading to the successful completion of the program.

There are 11 grades (1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 5.0), with 1.0 being the best and 5.0 the worst. Grade 5.0 means failed, and the student has to repeat the exam if it was not the third and, therefore, the last attempt. For all other grades, the course is passed. In our study, the grades were aggregated for better pseudonymization, so that only the values 1.3, 1.7, 2.3, 3.3, 4.0, 5.0 remained. It is possible for a student to enrol in a course and not to take part in the exam. This is coded as NT in the data.

Each degree program has a study plan which describes all courses and the semester each course belongs to. However, there is no obligation to adhere to this plan. There are two different types of courses: mandatory and elective. Mandatory courses form the basis of the study. Elective courses serve as a specialization and can be chosen from a pool of offers. If an elective course is not passed, another one can be chosen as an alternative.

Students are quite free: they do not have to complete the degree in six semesters and receive no warning if they take longer. They might take holiday-semesters, not enrolling

in any course at all, and resume their studies afterwards. Most of the universities have no tuition fees, only some modest administration fees that cover health insurance and public transport in the city. Some students take advantage of this fact and remain enrolled as a student though they do not enrol in any course and, in reality, do not study anymore.

4 Methods and results

After thorough preparation and cleaning, data have been explored and analysed from various angles. The two following sub-sections present the most important results.

4.1 Overview

To uncover information about students dropping out, one needs first to define them. The general definition of dropout is not being enrolled in the institution anymore. However, given the German context, this general definition is not constrained enough. As seen in the context section, a student might still be enrolled in the university but, in reality, has dropped out of the degree. To identify such students as dropping out, we explored the completing students: do they take holiday-semesters and, if yes, how many in a row?

It turns out that 8.76% of the graduates took holiday semesters. For most of them (84%), the break lasted only one semester. 8.7% took two semesters. Only 4.3% took three holiday semesters in a row. Longer breaks only occurred very sporadically. Therefore, students who have not enrolled in any course for more than two semesters in a row are considered as dropouts in this study.

Data exploration did not show any strong correlation between marks in high school certificate and marks at the end of the degree. Data exploration has shown also that marks in elective courses tend to be better than marks in mandatory courses, and that there is no student dropping out because of three failed attempts in an elective course. Therefore, support for students' advisors focuses on mandatory courses.

Earlier work on dashboards has shown that teachers want dashboards with few but essential visualizations that do not overload them [An 16]. Assuming that this lesson also holds for advisors, at this stage, an interactive dashboard containing three visualizations has been developed: (i) an overview diagram about study length, (ii) an overview diagram showing the number of semesters till dropping out, and (iii) an overview diagram about the overall performance of students in all mandatory courses, as well as the contrasting performance of dropping-out students and completing students.

Figure 1, the first diagram mentioned above, shows how long it takes students to complete their degree. The standard period of study for the degree program examined here is six semesters. The figure shows that most graduates need between six and eight semesters to graduate. Figure 2 shows the diagram giving the number of semesters spent in the program

until students drop out. Though the highest bar is for the first semester, a number of students drop out after two, three or even more semesters.

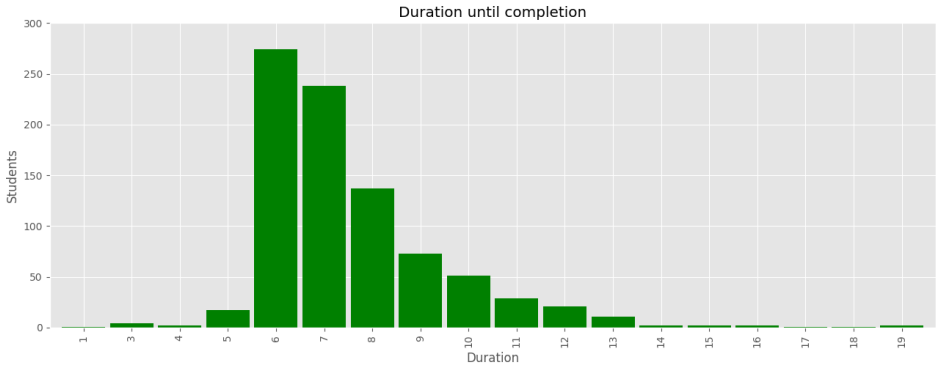

Figure 1: Number of semesters till graduation

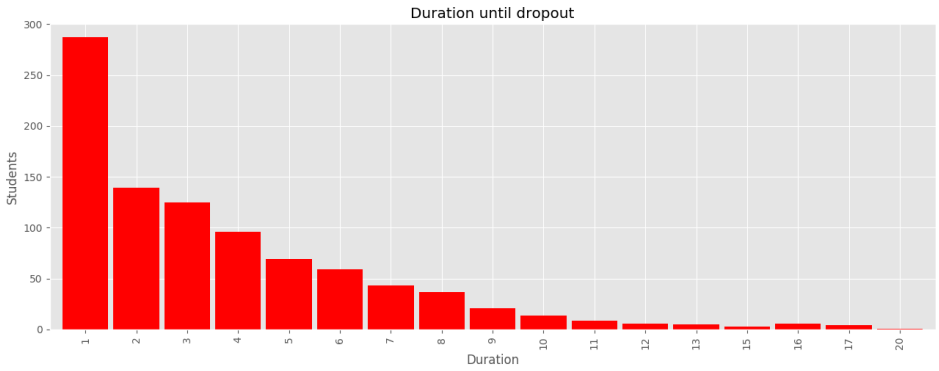

Figure 2: Number of semesters till dropout

The last visualization shows the distribution of grades per course per semester as described in the study plan. Semesters can be selected individually. One can also choose whether only completing students, only dropout-students or all students are taken into consideration. In Figure 3, the first semester is selected and two diagrams are shown, left with the choice “completing students” and right with the choice “dropout-students”. Note that a student may be counted several times in a column. For example, a student who does not sit the exam at the end of the semester fails the next exam but pass the next time with the mark 3.0 will be counted three times, one time as NT, one time as 5.0 and one time as 3.0. From bottom to top the different colours show the share of 1.3 (red), then 1.7 (blue) and so on till 5.0 (green) and finally NT (pink). One can see that the courses Mathematics I (biggest column on the left, the name is given by the course code in the diagram) and Programming I (second biggest column on the right) have the biggest share of NT and of

the mark 5.0 (failed) in both diagrams. This suggests that these courses are perceived as more difficult by all students. One notices too that the share of NT and 5.0 is in all courses is higher for dropout students than for completing students.

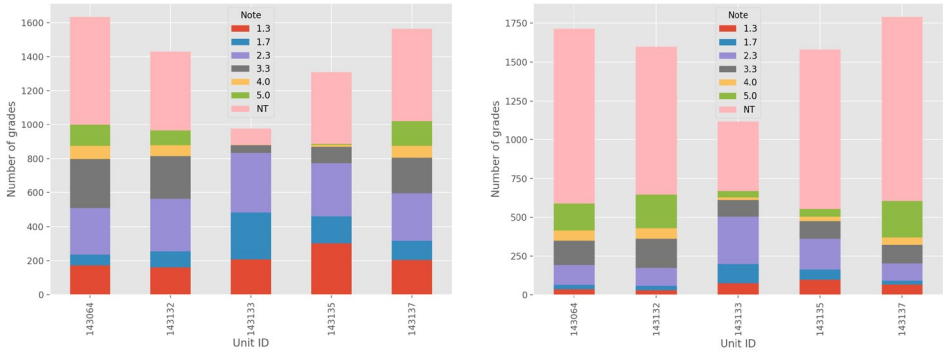

Figure 3: Distribution of grades per course in the 1st semester - completing left, dropout right

4.2 Typical completing behaviours

The aim of looking for typical behaviours among completing students is to give a kind of models on how other students achieved their studies to students seeking advice. These models or examples should be as near as possible to a student who needs advice.

To find out typical completing behaviours, the 787 completing students present in the dataset have been clustered. A student is represented by his/her marks and his/her number of enrolments in all compulsory courses. Marks and number of enrolments have the same order of magnitude and, for both, the smaller the better. Hence, these numbers have not been standardized for clustering. The aim of clustering is to group objects in clusters so that objects in one cluster are similar to each other and dissimilar to objects in other clusters. Clustering has been done using the classical algorithm K-means from the scikit-learn Python library. The K-means algorithm requires that the user chooses K, the number of clusters, see [Ha 12]. Using the well-known bent in the sum of squared errors' curve also called elbow curve [Ha 12] and the interpretability of the result, we fixed the number of clusters to four. The centres of the four clusters are depicted in Figure 4 (grades) and 5 (number of enrolments). The courses are represented by their codes with courses of the first semester on the left, followed by courses of the second semester and so on, till the last compulsory course in the 5th semester on the right.

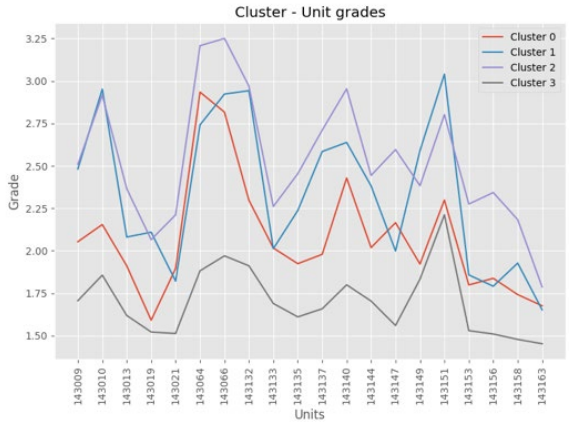

Figure 4: Centres of the four clusters: grades reached in each course

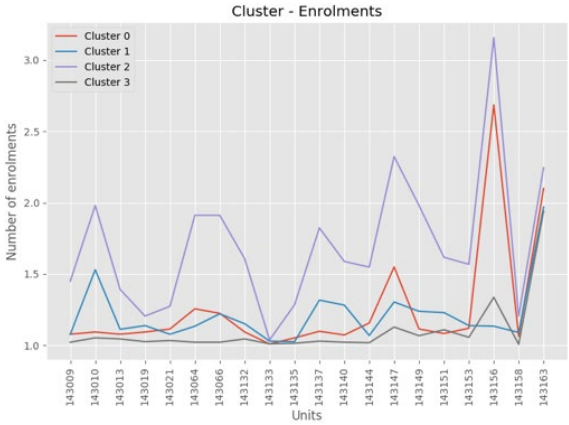

Figure 5: Centres of the four clusters: number of enrolments in each course

Cluster 3, the bottom black line, is the biggest cluster with 264 students. As can be seen in the plots, students in this cluster achieve the best marks in all compulsory courses, and, except for two courses, have the smallest number of enrolments, completing almost all courses with one semester. Cluster 2, the upper purple line, is the smallest cluster with 102 students and is almost the opposite of cluster 3. Students in this cluster need the most enrolments, the median in this cluster lies by 1.6, and except for two courses, have the least good marks. In between one notices cluster 1, 230 students, and cluster 0, 191 students. Students of cluster 0 tend to have better marks than students of cluster 1; in some courses, they need more enrolments, in other courses they need less. The boxplots for the number of enrolments (Figure 6) reveal that students of cluster 1 tend to have more enrolments than students of cluster 0. Focusing on the marks only, these results bear strong

similarity with those found in [As17]: students tend to have the same kind of marks (good, average or low) in all courses during the four years of their studies.

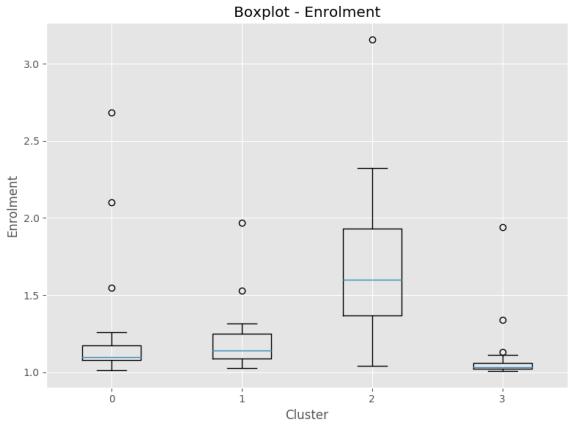

Figure 6: Boxplot per cluster with number of enrolments

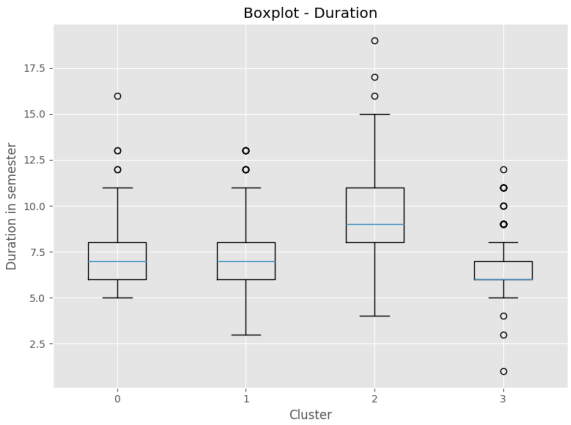

Figure 7: Boxplot per cluster with number of semesters till graduation

The number of enrolments per course does not say anything about the number of semesters necessary to graduate. Two students could have the same marks and enrol the same number of times, yet one could study part-time and needs 12 semesters to graduate and the other could study full time and graduate in six semesters only. Figure 7 shows that students in cluster 0 graduate the quickest, with a median of six semesters. The outliers needing less than five semesters are students who completed courses in another degree-program and could have those courses transferred in the degree-program studied here. It shows also that students of cluster 2 need the most time to graduate; this result reminds the finding of [CMS12]: students who take longer to complete the degree have lower

marks. Graduation time for students in cluster 0 and 1 is very similar; the median is seven semesters.

5 Conclusion and future works

In this preliminary study, an overall data exploration and a clustering have been presented and first insights were gained. Further, a first prototype has been implemented. Students' advisors could have access to the most important visualizations through an interactive dashboard. That way, they can support students with more precise information. For example, if a student is struggling with mathematics, or might study more than six semesters, the advisor can reassure the student, as this is a quite normal situation. In this study program, we found that the biggest bulk of students' dropout occurs at the end of the first semester. However, this bulk contains less than half of all the students dropping out. As a consequence, it would be useful to target the advisory meeting in the second semester rather than in the third semester.

Clustering has put in evidence four typical completing behaviours. Interestingly, the findings show that students do not compromise between good marks and duration of studies: students with good marks study quicker while struggling students have less good marks and need longer to complete their studies. This suggests that a planning module could rather show enrolment paths from students in cluster 2 to struggling students seeking advice as examples to complete their degree.

All these results have been shown to three advisors / study program coordinators and their feedback was very positive. A next step is to find typical learning behaviours among all students, not only among the completing students, as has been done in this study. Technically, this step is more difficult because dropping out students have less data than completing students. Work in that direction is in progress. A future step is to investigate typical trajectories of students through courses so that advisors could help to recommend which courses to take next to struggling students.

Acknowledgement: We thank Haythem Boukhayatia, Steffen Burlefinger, Michael Harms and Jihed Mzoughi for their cooperation in data preprocessing and in designing and implementing a first prototype.

6 Bibliography

[An16] An, T-S., Dubois, F., Manthey, E., Merceron, A..: Digitale Infrastruktur und Learning Analytics in Co-Design. In Proceedings of the Workshop learning Analytics, co-located with the 13th e-Learning Conference of the German Society for Computer Science, Potsdam, Germany, 11.09.16, S. 8-17, 2016.

[As17] Asif, R.; Merceron, A.; Abbas A.S.; Haider N.G.: Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113:177-194, 2017.

[Ba18] Backenköhler, M.; Scherzinger, F.; Singla, A.; Wolf, V.: Data-Driven Approach Towards a Personalized Curriculum. Proceedings of the 11th International Conference on Educational Data Mining (Buffaloo, USA, June 19-21). EDM'18., 246-251, 2018.

[CMS12] Campagni, R.; Merlini, D.; Sprugnoli, R.: Analyzing paths in a student database. Proceedings of the 5th International Conference on Educational Data Mining (Chania, Greece, June 19-21). EDM'12., 208-209, 2012.

[DPV09] Dekker, G.W; Pechenizkiy, M.; and Vleeshouwers, J.M.: Predicting Students Drop Out: A Case Study. Proceedings of the 2nd International Conference on Educational Data Mining (Cordoba, Spain, July 1-3). EDM'09, 208-209. [DPV09], 2009.

[Ha18] J. Han, J.; Kamber M.; Pei, J.: Data Mining Concepts and Techniques. 3rd Ed. San Francisco: Morgan Kaufmann, 2012, p. 451-454, 486-487.

[Ji18] Jivet, J.; Scheffel, M.; Specht, M.; Drachlser, H.: License to evaluate: Preparing learning analytics dashboards for educational practice. Proceedings of the 8th international conference on learning analytics & knowledge LAK'18 (Sydney, Australia, March 5-9). ACM, 31-40, 2018.

[Mi18] Millecamp, M.; Gutierrez, F.; Charleer, S.; Verbert, K.; De Laet, T.: A Qualitative Evaluation of a Learning Dashboard to Support Advisor-Student Dialogues. Proceedings of the 8th international conference on learning analytics & knowledge LAK'18 (Sydney, Australia, March 5-9). ACM,56-60 [Mi 18], 2018.

[Sch17] Schwendimann, B.; Rodriguez-Triana, M.; Vozniuk, A.; Prieto, L.; Boroujeni, M.; Holzer, A.; Gillet, D.; and Dillenbourg, P. 2016. Perceiving learning at a glance: A systematic literature review of learning dashboard research. IEEE Transactions on Learning Technologies, Vol.10 (1), (2017).

[Zi15] Zimmermann, J.; Brodersen, K.H.; Heinimann, H.R.; Buhmann, J.M.: A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. Journal of Educational Data Mining, 7(3):151-176, 2015.