

Modeling study duration considering course enrollments and student diversity

Niels Seidel

FernUniversität in Hagen, Research Cluster D²L², Universitätsstr. 11, 58084 Hagen, Germany
niels.seidel@fernuni-hagen.de

ABSTRACT

Students are self-determined to choose degree programs and courses at their own pace. However, this variety of choices can lead to a long duration of study, especially in part-time distance learning. Hence, this paper aims to explore data on course enrollments of students pursuing bachelor's and master's degrees in Computer Science and Mathematics at a European distance-based German university to uncover predictors for study duration. Distance students have highly diverse backgrounds, which might also be represented in their enrollment behavior and duration of study. Thus, it is vital to analyze this behavior to identify bottlenecks and adjust instructions. We employed a Multiple Regression Analysis with a Genetic Algorithm for model selection to uncover predictors that lengthen or shorten the study duration. For model selection, we considered demographic data, modes of study, enrollment behaviors, and individual courses. We used the method to find predictors within the data of 1898 students who graduated in at least one of the five study programs offered by the Faculty of Mathematics and Computer Science between 1999 and 2019. The enrollment behavior strongly predicts the duration of study compared to demographic and study-behavior predictors. Individual courses are good predictors for specific study programs.

Keywords

course enrollment, study duration, multiple linear regression

1. INTRODUCTION

In accordance with the Humboldtian model of higher education, students are self-determined to choose degree programs and courses at their own pace. However, this variety of choices can lead to a long duration of study, especially in part-time distance learning. At many universities, a long duration of study was not seen as a problem as long as sufficient capacity was available. In the OECD countries, the proportion of young people between 25 and 34 years of age with a university degree rose by a total of 14 percentage points

between 2000 and 2013 [27]. In Germany for instance, the number of students increased by 1,146,262 (63%) between 1998/99 and 2021/22¹. At the same time, the capacities for teaching did not increase at the same level. In this paper, we take a look at the duration of study as a driver for high student number.

Study programs are designed so that students graduate within a defined amount of time. If students exceed this regular period of study, it has consequences for the students, the teachers, and the respective faculty. Additional semesters cost a student time and effort to repeat courses and exams. This can also be accompanied by financial costs for tuition fees and a late entry into professional life or a higher career level. In addition, there are psychological burdens. For instructors, longer study durations mean an increase in the amount of supervision required due to the need to retake courses and exams. This is apparent in the supervision ratio which is defined as the number of students per teacher. From a faculty perspective, long-term students need to be considered for capacity planning. Just like the number of new program enrollments, the number of graduations is part of the target agreements or key performance indicators considered by the university management and ultimately the federal or state ministries of education. Today, higher educational organizations are placed in a very highly competitive environment. The analysis, presentation, and data mining is one approach to tackle challenges in the organization of study programs.

The causes of protracted studies are not necessarily due to a lack of motivation, performance, or effort on the part of students. Behavioral factors in the choice of one or more courses of study, as well as the distribution of the workload over the semesters, can have a major influence on the time to degree. Past studies have shown that differences in enrollment behavior are related to student diversity factors [2]. The manifestations of these factors vary by country/culture, university type, institution, and program respective subject domain. Further factors for a long study duration are under the influence of teachers. Repetition of courses and exams can be an indicator of high difficulty, but also of inadequate instructional design or exams with low pass rates. Thus, it is vital to analyze enrollment behavior to identify bottlenecks and adjust instructions. Other reasons for a slowed down study progress result from organizational bottlenecks, such

N. Seidel. Modeling study duration considering course enrollments and student diversity. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 621–628, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

© 2022 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.
<https://doi.org/10.5281/zenodo.6852956>

¹See <https://www.datenportal.bmbf.de/portal/de/K254.html> (accessed 2022/05/08).

as overfilled courses, missing or too late reexaminations, annual instead of semesterly course offers and examinations.

Hence, this paper aims to explore data on course enrollments of students pursuing bachelor’s and master’s degrees in Computer Science and Mathematics at a German distance-based university to uncover predictors for study duration. Our study aims at providing initial insights into enrollment processes of German distance learning students. In particular, we are going to focus on one research question (RQ): *(RQ1) What predictors significantly influence the duration of study?* To answer this question we employed a Multiple Linear Regression Analysis with a Genetic Algorithm for model selection to uncover predictors that lengthen or shorten the study duration. For model selection, we considered demographic data, modes of study, enrollment behaviors, and individual courses. We used the method to find predictors within the data of 1898 students who graduated in at least one of the five study programs offered by the Faculty of Mathematics and Computer Science between 1999 and 2019.

Identifying predictors associated with time to graduation can help educators design better degree plans, and students make informed decisions about future enrollments. Distance students have highly diverse backgrounds, which might also be represented in their enrollment behavior and duration of study. Thus, it is vital to analyze this behavior to identify bottlenecks and adjust instructions.

2. RELATED WORKS

There exist various studies focusing on enrollment data. In this section, we provide an overview of the background and intent for the analysis of these data and shed a light on the data and methods used.

Most of the research on enrollment data relates to educational institutions in the Anglo-American world. Among the cited literature in this paper, only two papers refer to African [34, 13] and three to European institutions of higher education [7, 33, 3]. The majority of the works come from traditional universities compared to distance learning universities as referred by [33] or MOOCs [31].

The intentions for analyzing enrollment data range from descriptive analysis, prediction to the preparation of interventions. [33] identify factors contributing to students continuing for the duration of their distance learning studies and completing their degree. The motivation for enrollment to computer science degree programs has been explored by Duncan et al. [15]. Age, gender, and demographic trends in motivation (goals, opportunities, and assurance of goal achievement) for enrollment have been analyzed and significant motivation differences regarding gender and age have been reported. Sahami et al. [30] explored the phenomena of performance decline using the computer science enrollments data from Stanford University and found that despite increased enrollments, student performance remains stable. Analysis is conducted on different scales such as courses [7], study programs [35, 34] and faculties [35, 34]. [24] and [9] for instance, analyzed changes in the enrollment and study progress before and after policy changes. [35] focuses on students’ experiences of guidance in relation to their study progress and perceptions of their learning outcomes. The impact of co-enrollment was studied by [8] and [37]. The prediction of dropout (e.g. [10, 22, 7]), study performance

(e.g. [11, 14, 17, 25, 29, 38, 39]), and future enrollments (e.g. [20, 23, 36]) gained a lot of attention in the last years. Prediction of time to degree were employed by [18] and [21]. [6] identify potential predictors of academic success including the time to graduation for Ph.D. students. Age, sex, employment institution, mentor experience, and tuition subsidy had no influence on the time to graduation and completion rate. [35] predicted slow study progress from self-report data using Binary Logistic Regression. [24] identified factors affecting time to bachelor’s degree attainment. Dahdouh et al. used association rules mining over course enrollments for recommendations of further study paths [12]. The rules are used for recommending suitable courses to students based on their behavior and preferences. [7] investigated bottlenecks of learning progress in order to support the student advisory services, while [28] make use of enrollment data to prepare re-enrollment campaigns.

Data collected from university information systems has been proved to be the source of helpful information (e.g. [24, 9, 4]) for improving study processes and educational decisions. However, due to a strong data protection culture, some European universities tend to interpret the European data protection regulations (GDPR) very strictly. Even within institutions researchers do not get access to personal data and are also not allowed to link anonymized data. Student performance data such as grades are considered particularly sensitive. Another common source for investigating enrollment behavior comes various forms of self-reports including surveys (e.g. [26, 29, 13, 13, 31, 19]). The used variables cover a broad range that reflects cultural and institutional conditions. For example, the housing situation was studied in countries where campus universities are found [5].

Subgroups including their intersections have been rarely considered [35, 3, 2]. [3] for instance, identified differences in study success and early dropout between minority and majority students in economics which can be attributed to differences in high school education, but not on academic and social integration. [2] considered dimensions underpinning students’ study philosophy towards teaching, learning, and study for different groupings and subgroup interactions (e.g. age, sex, ethnicity, study discipline, academic performance). The definition of student profiles [7, 13] is an approach coming from social science which can be helpful to distinguish and explain patterns of subgroups.

The analytical methods used for enrollment analyses include frequent item mining [1, 12], sequence mining [1, 9], Clustering [34], Social Network Analysis [37], Latent Profile Analysis [13], and Linear/Logistic Regression [24, 35, 34, 6]. For example, Elbadrawy et al. [16] used sequence mining via the so-called Universal discriminating Pattern Mining framework capable of mining enrollment patterns from groups of low and high-performing students to enable educators for better degree planning. [26] applied an investment theory to predict the degree of commitment. The application of a Multiple Linear Regression by [5] and [24] underlines its advantages with regard to traceability, explainability, and the possibility of deriving interventions.

3. METHODS

3.1 Data

The data set contains 1489 bachelor students and 1014 master’s students who enrolled in 1999 to 2016 and finished the degree until 2019. The collected data include student en-

rollments to courses during their studies, information about completion of the degree, and a list of courses required to complete the degree. In addition, the enrollment data do not contain information on whether a student finished a course successfully since different departments carry out the oral and written examinations at the Faculty. University data protection rules restrict the use and analysis of the exam results. By enrolling in the program, students gave their consent to the processing of the data used in this analysis. To further ensure data privacy the unique identifiers of the students have been pseudonymized in order to prevent linking with other datasets and to prohibit the identification of individual students. However, the identification of individuals cannot be ruled out, the data set will only be provided on request instead of being published.

The available data includes demographic data as well as information on the enrolled programs and courses. From this information, four diversity dimensions will be categorized with regard to (i) demographics, (ii) study behavior, (iii) enrollment behavior, and (iv) course impact. While the first three categories are related to the students, the latter refers to organizational and didactical aspects mainly influenced by the responsible teachers.

The demographic data available contains the age at program admission, gender, and the completion of previous bachelor's or master's degrees. The age ranges between 14 and 69 in all study programs. Detailed demographic information per program are listed in Tab. 1.

From the program enrollment data, we derive study behavior information. Students at the Faculty of Mathematics and Computer Science can enroll in up to three programs at the same time. These programs are prioritized by the student (cf. Program priority). For each program, students can decide whether to study full-time or part-time which has an effect on the expected study duration. The duration of study in part-time study is half as long as the full-time duration. Furthermore, Master's programs distinguish consecutive study after completing a related bachelor's degree and non-consecutive study. A second degree is stated if a student already achieved a degree on the same level (e.g. a second bachelor's degree). Listener status describes the opportunity to join a program as a guest or listener without the obligation to achieve a degree.

The enrollment behavior is described by the number and variety of course enrollments per semester and in total. For the first three semesters first-time and re-enrollments are counted separately (e.g. *Enrollments 1st semester*, *Repetitions 2nd semester*). For the number of unique courses, we distinguish between courses offered at the Faculty (*Different Faculty courses*) and those offered at another faculty (*Different other courses*). Semesters without any enrollments are described as *semesters off*.

Furthermore, 25 % of the most frequently enrolled course have been dummy-coded for each student representing the fourth diversity dimension.

3.2 Multiple Linear Regression

For each study program, the student data was represented in a Learner Profile including the before mentioned data about the demographics, study behavior, and enrollment behavior as well the binary information about the most frequently enrolled courses.

Outliers regarding the total number of different enrolled courses, the total course repetitions, and the repeating enrollments have been removed. Values above the mean plus three times the standard deviation have been considered as outliers. Finally, 884 B.Sc. and 1014 M.Sc students remained in the dataset.

The mentioned variables have been selected from the Learner Profile and used to produce model formulas. These formulas are passed to a fitting function. The variables in the formula correspond to the data in the Learner Profile. The duration of study was defined as the dependent variable. The remaining variables were used as independent variables in the formulas. By default, an intercept is included in all models. Due to the initial use of a large number of variables, it is necessary to find a simpler model based on fewer variables. Instead of trying all candidates for a suitable model with an unapplicable brute force approach, the candidate set is explored by a Genetic Algorithm (GA). A GA can readily find the best models without fitting all possible models. For the GA the formula is encoded as a sequence of binary values. This sequence forms a population that will undergo an evolution by adapting certain bits to form a new generation. The genetic algorithm keeps track of a population of models and their size. Asexual reproduction, sexual reproduction from parental generations, and immigration are the three methods used to create the next generation of models.

As decision criterion the Akaike Information Criterion (AIC) is used. It is defined by

$$AIC = -2l(\hat{\beta}_M, \sigma^2) + 2|M + 1| \quad (1)$$

with $l(\hat{\beta}_M, \sigma^2)$ as the maximum value of the log-likelihood and M as the number of variables present in the current model. On the one hand, one can see that the AIC value is negatively directed, which is why the goal of model selection is to minimize this value. On the other hand, a high number of variables is penalized. Thus, a too complex model is prevented. The models are fitted to every generation by using the AIC values to calculate each model's fitness, w . The i th model's fitness is calculated as follows:

$$w_i = \exp(-(AIC_i - AIC_{best})) \quad (2)$$

where AIC_{best} is the best AIC in the current population of models. Lower AIC means higher fitness. Inference was aided by point and interval (95 % CI) estimates, the goodness of fit measures, AIC, and p values.

In order to measure and compare the goodness of a fitted model we compute the Cragg-Uhler *Pseudo-R*². *Pseudo-R*² is defined as one minus the ratio of the residual deviance and the intercept (null deviance):

$$R^2 = 1 - \frac{ResidualDeviance}{NullDeviance} \quad (3)$$

R^2 describes the deviation of the current model between 0 and 1, whereas 0 means total deviation and 1 a complete congruence.

4. RESULTS AND DISCUSSION

Appendix A.1 provides an overview of the fitted models and the number of predictors with regard to the four categories of diversity dimensions. Except for the B.Sc. CS the values for R indicate a good model fit. For the smaller number of graduates in the M.Sc. in Mathematics a very good was

Table 1: Demographic information about the students who graduated in the Computer Science (CS) and Mathematics programs

Program name	B.Sc. CS	M.Sc. Practical CS	M.Sc. CS	B.Sc. Mathematics	M.Sc. Mathematics
Time range	1999-2019	2003-2019	2003-2019	2000-2018	2003-2018
N (mean \pm sd)					
Women	454	16	120	130	9
Men	634	153	690	180	33
Total	686	169	803	198	42
Age at admission (years mean \pm sd)					
Women	31.77 \pm 6.14	31.56 \pm 5.39	32.53 \pm 7.84	28.39 \pm 7.88	31.78 \pm 10.44
Men	30.44 \pm 6.24	29.86 \pm 5.42	31.61 \pm 6.82	30.79 \pm 7.88	30.45 \pm 7.93
Total	30.69 \pm 6.23	30.02 \pm 5.42	31.74 \pm 6.98	30.29 \pm 7.89	30.74 \pm 8.41
Time to degree (semesters mean \pm sd)					
Women	14.15 \pm 6.19	10.19 \pm 4.81	5.58 \pm 3.31	11.06 \pm 3.7	5.89 \pm 1.05
Men	11.89 \pm 6.18	8.5 \pm 3.82	5.89 \pm 3.56	10.22 \pm 4.83	7.91 \pm 3.74
Total	12.31 \pm 6.24	8.66 \pm 3.94	5.85 \pm 3.52	10.4 \pm 4.61	7.48 \pm 3.44

achieved. The AIC and BIC measures are not suitable to make comparisons between the programs but relate to the model complexity. The model of the bachelor of CS appears to be the most complex with 29 predictors. Here again, the M.Sc. in Mathematics stands out with simpler model of 8 predictors.

The four diversity dimensions have a different influence on the models. In general, it can be said that demographic factors and study behavior predicting the study duration less than the enrollment behavior. The effect of individual courses depend on the study program.

The fitted linear regression models for predicting variables influencing the duration of study in each of the five study programs are presented in the Appendix A.2. The size of the coefficients expresses the number of semesters by which the study is extended or, if negative, shortened. For example, if a student takes 3 courses in the Bachelor CS in the 3rd semester, the duration of study is shortened by 3 times -0.38, i.e. by 1.14 semesters. Binary represented values like gender or taking a certain course correspond to factor 1. As stated before, age has no significant impact on study duration. Note, that the coefficient for the age is multiplied by the number of years. As a result, this apparently small coefficient may predict the study duration of elderly students. Also the gender impact is comparatively small, but recognizable with opposite direction in the CS bachelors and Practical CS masters' programs. For the same two programs the existence of a past degree predicts the time to degree. While the length of study for students in the Bachelor CS is shortened by the experience gained in another program, the length of study is lengthened for students in the Master Practical CS. Studying in multiple programs at the same time can be beneficial for the overall study duration. This can be explained by the fact that examination credits from one study program can be credited in the thematically related study programs of the faculty. Thus, a successfully completed examination can be used in several study programs. However, for the M.Sc. CS additional activities on other programs is at the expense of the duration of study. The enrollment to courses of other faculties extends the time needed for completion. As expected, the total course repetition and the variety of chosen Mathematics-related or CS-related course strongly predict study duration. A single semester off lengthens the study duration by more than one semester.

5. SUMMARY AND OUTLOOK

In this paper, we explored data on course enrollments of students pursuing bachelor's and master's degrees in Computer Science and Mathematics at a European distance-based German university to uncover predictors for study duration. We tried to consider the highly diverse backgrounds of distance-learning students that are represented in a restricted and pseudonymized dataset consisting only of information on current and past study programs and enrolled courses. From this information, Learner Profiles have been created. These profiles contained measures that are potentially suitable for describing influencing factors for the duration of study. Instead of predicting the time of study completion for future cohorts, we used them to describe and analyze the past student (and teacher) behavior. We find it is vital to analyze this behavior to identify bottlenecks and adjust instructions as well as the organization of study programs.

We employed a Multiple Regression Analysis with a Genetic Algorithm for model selection to uncover predictors that lengthen or shorten the study duration. For the models, we considered demographic data, study behavior, enrollment behaviors, and individual courses. We used the method to find predictors within the data of 1898 students who graduated in at least one of the five study programs offered by the Faculty of Mathematics and Computer Science between 1999 and 2019. With regard to RQ1 the enrollment behavior strongly predicts the duration of study compared to demographic and study-behavior predictors. Individual courses are good predictors for specific study programs.

As a next step, we want to identify changes in the fitted models over time. The considered time range of almost 20 years included many changes of regulations, tuition fees, and teaching staff. Similar to the work of [24] and [9] we want to trace predictors over time in order to recognize relevant trends for teachers and faculty managers. With this regard, we also would like to continue our past research about student course recommenders [32].

Acknowledgments

This research was supported by the Research Cluster "Digitalization, Diversity and Lifelong Learning – Consequences for Higher Education" (D²L²) of the FernUniversität in Hagen, Germany.

References

- [1] Z. Abdullah, T. Herawan, N. Ahmad, and M. M. Deris. Extracting highly positive association rules from students' enrollment data. *Procedia-Social and Behavioral Sciences*, 28:107–111, 2011.
- [2] M. Alauddin and A. Ashman. The changing academic environment and diversity in students' study philosophy, beliefs and attitudes in higher education. *Higher Education Research & Development*, 33(5):857–870, 2014.
- [3] I. J. M. Arnold. Ethnic minority dropout in economics. *Journal of Further and Higher Education*, 37(3):297–320, 2013.
- [4] K. E. Arnold and M. D. Pistilli. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. *Proceedings of the 2nd LAK Conference*, pages 267–270, 2012.
- [5] S. Beekhoven, U. D. Jong, and H. V. Hout. The impact of first-year students' living situation on the integration process and study progress. *Educational Studies*, 30(3):277–290, 2004.
- [6] B. Benzon, K. Vukojevic, N. Filipovic, S. Tomić, and M. G. Durdov. Factors that determine completion rates of biomedical students in a PhD programme. *Education Sciences*, 10(11):1–8, 2020.
- [7] A. Böttcher, V. Thurner, T. Häfner, and S. Ottinger. Adaptierung von Beratungsangeboten auf der Basis von Erkenntnissen aus der Analyse von Studienverlaufsdaten. In *9. Fachtagung Hochschuldidaktik Informatik (HDI)*, pages 57–64, 2021.
- [8] M. G. Brown, R. Matthew DeMonbrun, and S. D. Teasley. Conceptualizing Co-enrollment: Accounting for student experiences across the curriculum. In *LAK '18 8th International Conference on Learning Analytics and Knowledge, March 7-9, 2016*, pages 305–309, Sydney, 2018. ACM.
- [9] D. Canales Sánchez, T. Bautista Godínez, J. G. Moreno Salinas, M. García-Minjares, and M. Sánchez-Mendiola. Academic trajectories analysis with a life-course approach: A case study in medical students. *Cogent Education*, 9(1):2018118, dec 2022.
- [10] H. E. Caselli Gismondi and L. V. Urrelo Huiman. Multilayer Neural Networks for Predicting Academic Dropout at the National University of Santa - Peru. In *2021 International Symposium on Accreditation of Engineering and Computing Education (ICACIT)*, pages 1–4, 2021.
- [11] R. Costa-Mendes, T. Oliveira, M. Castelli, and F. Cruz-Jesus. A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. *Education and Information Technologies*, 26(2):1527–1547, 2021.
- [12] K. Dahdouh, A. Dakkak, L. Oughdir, and A. Ibriz. Large-scale e-learning recommender system based on Spark and Hadoop. *Journal of Big Data*, 6(1), 2019.
- [13] M. De Clercq, B. Galand, and M. Frenay. One goal, different pathways: Capturing diversity in processes leading to first-year students' achievement. *Learning and Individual Differences*, 81:101908, 2020.
- [14] E. Demeter, M. Dorodchi, E. Al-Hossami, A. Benedict, L. Slattery Walker, and J. Smail. Predicting first-time-in-college students' degree completion outcomes. *Higher Education*, 2022.
- [15] A. Duncan, B. Eicher, and D. A. Joyner. Enrollment motivations in an online graduate cs program: Trends and gender- and age-based differences. In *Annual Conference on ITiCSE*, pages 1241–1247, 2020.
- [16] A. Elbadrawy and G. Karypis. UPM: Discovering Course Enrollment Sequences Associated with Success. In *Proceedings of the 9th LAK Conference*, LAK19, pages 373–382, New York, NY, USA, 2019. Association for Computing Machinery.
- [17] A. Gambini, M. Desimoni, and F. Ferretti. Predictive tools for university performance: an explorative study. *International Journal of Mathematical Education in Science and Technology*, pages 1–27, jan 2022.
- [18] T. Hailikari, R. Sund, A. Haarala-Muhonen, and S. Lindblom-Ylänne. Using individual study profiles of first-year students in two different disciplines to predict graduation time. *Studies in Higher Education*, 45(12):2604–2618, 2020.
- [19] T. Hailikari, T. Tuononen, and A. Parpala. Students' experiences of the factors affecting their study progress: differences in study profiles. *Journal of Further and Higher Education*, 42(1):1–12, 2018.
- [20] N. A. Haris, M. Abdullah, N. Hasim, and F. Abdul Rahman. A study on students enrollment prediction using data mining. In *ACM IMCOM 2016: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, 2016.
- [21] S. Herzog. Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*, 2006(131):17–33, 2006.
- [22] B. Jeon and N. Park. Dropout Prediction over Weeks in MOOCs by Learning Representations of Clicks and Videos. *CoRR*, abs/2002.0, 2020.
- [23] M. S. Kiran, E. Siramkaya, E. Esme, and M. N. Senkaya. Prediction of the number of students taking make-up examinations using artificial neural networks. *International Journal of Machine Learning and Cybernetics*, 13(1):71–81, 2022.
- [24] W. E. Knight. Time to Bachelor's Degree Attainment: An Application of Descriptive, Bivariate, and Multiple Regression Techniques. *IR Applications*, Volume 2, September 8, 2004. 2004.
- [25] M. F. Musso, C. F. R. Hernández, and E. C. Cascallar. Predicting key educational outcomes in academic trajectories: a machine-learning approach. *Higher Education*, 80(5):875–894, 2020.

- [26] S. Noxel and L. Katunich. Navigating for Four Years to the Baccalaureate Degree. AIR 1998 Annual Forum Paper. Technical report, Ohio State University, 1998.
- [27] OECD. *Education at a Glance 2020*. 2020.
- [28] J. C. Ortagus, M. Tanner, and I. McFarlin. Can Re-Enrollment Campaigns Help Dropouts Return to College? Evidence From Florida Community Colleges. *Educational Evaluation and Policy Analysis*, 43(1):154–171, 2021.
- [29] H. Prabowo, A. A. Hidayat, T. W. Cenggoro, R. Rahutomo, K. Purwandari, and B. Pardamean. Aggregating Time Series and Tabular Data in Deep Learning Model for University Students’ GPA Prediction. *IEEE Access*, 9:87370–87377, 2021.
- [30] M. Sahami and C. Piech. As CS Enrollments Grow, Are We Attracting Weaker Students? pages 54–59, 2016.
- [31] E. Schneider and R. F. Kizilcec. ”Why Did You Enroll in This Course?”: Developing a Standardized Survey Question for Reasons to Enroll. In *Proceedings of the First ACM Conference on Learning Scale Conference, L@S ’14*, pages 147–148, New York, NY, USA, 2014. Association for Computing Machinery.
- [32] N. Seidel, M. C. Rieger, and A. Walle. Semantic Textual Similarity of Course Materials at a Distance-Learning University. In T. W. P. And, P. B. And, S. I. H. And, K. K. And, and Y. Shi, editors, *Proceedings of 4th CSEDM Workshop co-located with the EDM 2020 Conference*. CEUR-WS.org, 2020.
- [33] J. Simons, S. Leverett, and K. Beaumont. Success of distance learning graduates and the role of intrinsic motivation. *Open Learning*, 35(3):277–293, 2020.
- [34] F. Siraj and M. A. Abdoulha. Uncovering hidden information within university’s student enrollment data using data mining. In *2009 Third Asia International Conference on Modelling & Simulation*, pages 413–418. IEEE, 2009.
- [35] T. Skaniakos, S. Honkimäki, E. Kallio, K. Nissinen, and P. Tynjälä. Study guidance experiences, study progress, and perceived learning outcomes of Finnish university students. *European Journal of Higher Education*, 9(2):203–218, 2019.
- [36] J. Ward. Forecasting enrollment to achieve institutional goals. *College and University*, 83(3):41, 2007.
- [37] K. A. Weeden, B. Cornwell, and B. Park. Still a Small World? University Course Enrollment Networks before and during the COVID-19 Pandemic. *Sociological Science*, 8:73–82, 2021.
- [38] S. K. Yadav and S. Pal. Data mining: A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv:1203.3832*, 2012.
- [39] M. Yağcı. Educational data mining: prediction of students’ academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1):11, 2022.

APPENDIX

A. LINEAR REGRESSION MODELS

A.1 Overview of Linear Regression Models

Table 2: Overview of the fitted Linear Regression Models of students' time to degree

Program	B.Sc. CS	M.Sc. CS	M.Sc. Practical CS	B.Sc. Mathe	M.Sc. Mathe
Model goodness					
R ²	0.67	0.81	0.88	0.82	0.92
AIC	1380.39	611.59	2306.13	324.20	115.11
BIC	1433.69	656.85	2388.48	353.61	129.11
Number of predictors					
Demographic-related	3	0	3	0	0
Study-related	3	2	1	2	2
Enrollment-related	11	10	11	5	6
Course-related	13	2	4	8	1
Total	29	13	17	14	8

* p<.1, ** p<.01, *** p<.001

A.2 Coefficients of the best fit Linear Regression Models

Table 3: Coefficients of the best fit Linear Regression Models of students' time to degree

Coefficient	B.Sc. CS	M.Sc. CS	M.Sc. Pract. CS	B.Sc. Mathe	M.Sc. Mathe
(Intercept)	10.90***	4.11**	1.93***	9.49***	3.27***
Age	0.04	-	0.01*	-	-
Male	-1.02*	-	0.19	-	-
previousDegreesMaster	-1.61*	-	0.81*	-	-
Fulltime study	-1.06*	-	-0.25*	-	-
Programme priority	-0.79	1.67	-	-3.04*	3.27***
Second degree	0.68	1.28*	-	1.73**	-
Semesters off	0.95***	1.44*	1.13***	1.48***	3.27***
Total course repetitions	0.23***	0.49***	0.46***	0.30***	0.49***
Different CS courses	-0.05	0.2***	0.18***	0.10**	-
Different other courses	0.04	-	0.26***	-	0.20***
Enrollemnts 1st semester	-0.17*	-0.19*	-0.61***	-	-0.23*
Enrollments 2nd semester	-0.05	-0.28*	-0.46***	-	-
Enrollments 3rd semester	-0.38**	-0.38***	-0.34***	-	-
Repetitions 1st semester	-0.37	-0.89***	-0.71***	-1.00**	-
Repetitions 2nd semester	-0.35	-0.75***	-0.57***	-	-1.03***
Repetitions 3rd semester	-0.12	-0.8**	-0.17*	-1.72***	-0.62*
Course 1144	-	-	-	-5.22***	-
Course 1145	-	-	-	4.64***	-
Course 1202	-	-	-	-1.45*	-
Course 1358	-	-	-	-	-1.81*
Course 1359	-	-	-	-	2.38*
Course 1361	-	-	-	-1.54*	-
Course 1584	1.49**	-	-	-	-
Course 1613	-0.07	-	-	-	-
Course 1618	0.49	-	-	-	-
Course 1657	-1.43	-	-	-	-
Course 1658	0.34	-	-	-	-
Course 1661	0.63	-	-	-	-
Course 1666	-	-	0.27**	-	-
Course 1671	0.14	-	-	-	-
Course 1678	-0.46	-	-	-	-
Course 1793	0.21	-	-	-	-
Course 1801	0.06	-	-	-	-
Course 1814	-	-	0.33***	-	-
Course 1853	-	0.54*	-	-	-
Course 1866	0	-	-	-	-
Course 1895	0.01	-	-	-	-
Course 1896	1.21*	-	-	-	-

* p<.1, ** p<.01, *** p<.001