Bachelor Thesis Analytics: Using Machine Learning to Predict Dropout and Identify Performance Factors

https://doi.org/10.3991/ijai.v1i1.11065

Jalal Nouri (✉), Ken Larsson
Stockholm University, Stockholm, Sweden
jalal@dsv.su.se

Mohammed Saqr
University of Eastern Finland, Joensuu, Finland

Abstract—The bachelor thesis is commonly a necessary last step towards the first graduation in higher education and constitutes a central key to both further studies in higher education and employment that requires higher education degrees. Thus, completion of the thesis is a desirable outcome for individual students, academic institutions and society, and non-completion is a significant cost. Unfortunately, many academic institutions around the world experience that many thesis projects are not completed and that students struggle with the thesis process. This paper addresses this issue with the aim to, on the one hand, identify and explain why thesis projects are completed or not, and on the other hand, to predict non-completion and completion of thesis projects using machine learning algorithms. The sample for this study consisted of bachelor students' thesis projects (n=2436) that have been started between 2010 and 2017. Data were extracted from two different data systems used to record data about thesis projects. From these systems, thesis project data were collected including variables related to both students and supervisors. Traditional statistical analysis (correlation tests, t-tests and factor analysis) was conducted in order to identify factors that influence non-completion and completion of thesis projects and several machine learning algorithms were applied in order to create a model that predicts completion and non-completion. When taking all the analysis mentioned above into account, it can be concluded with confidence that supervisors' ability and experience play a significant role in determining the success of thesis projects, which, on the one hand, corroborates previous research.

On the other hand, this study extends previous research by pointing out additional specific factors, such as the time supervisors take to complete thesis projects and the ratio of previously unfinished thesis projects. It can also be concluded that the academic title of the supervisor, which was one of the variables studied, did not constitute a factor for completing thesis projects. One of the more novel contributions of this study stems from the application of machine learning algorithms that were used in order to – reasonably accurately – predict thesis completion/non-completion. Such predictive models offer the opportunity to support a more optimal matching of students and supervisors.

Keywords—Thesis, bachelor, completion, machine learning, retention, performance, learning analytics.

1 Introduction

The thesis has become a central element as well as a formal requirement of graduate programmes for more than a century and continues to be an essential component in most universities [1, 2]. The thesis is supposed to establish the competences of developing critical thinking, empirical research literacy, and synthesis of knowledge and the assessment of the veracity of information. In most universities around the world, the bachelor thesis is a mandatory last step towards the first graduation in higher education and thus constitutes a central key to both further studies as well as employment that requires higher education degrees.

Students who obtain a graduate degree realise a wide array of benefits that include personal, economical and long-lasting career advantages[1, 2]. These students are more likely to be employed, the jobs they secure are better paid than those of their counterparts, and they more likely to have stable jobs. Further benefits extend to their families in the form of increasing chances of high-quality education, better parenting, and health insurance benefits [3-5]. The benefits of higher education also extend to governments and society at large, who derive a range of direct and indirect benefits, such as better return on investment in education, lower crime rates, improved tax revenues and less dependence on welfare programmes [3, 4]. Therefore, ensuring that students enrolled in graduate programmes obtain their degrees in a timely fashion is in the best interest of governments, higher education institutions and students alike [3, 6].

However, the thesis is a challenging endeavour that requires skills, aptitude, and determination for successful, timely completion [1, 2, 7-9]. As expected, a considerable number of students struggle with the thesis process, resulting in delays, disruptions, and non-completion of their degrees [10-12]. Non-completion results in a vast waste of faculty time and institutional resources, a devastating personal experience for students that costs precious time, loss of money and energy, and a societal loss of high-skilled workers [1, 3, 11]. Thus, it is reasonable to assume that non-completion of higher education degrees should be viewed as a substantial problem that requires serious attention and proactive planning [1, 3, 11, 13-15].

Previous research related to thesis projects has identified some variables that influence the performance of students undertaking thesis projects; variables that, in particular, point out the relation between the student candidate and the supervisor [1, 11, 16]. The specific student variables that have been indicated as influencing thesis completion are students' attitudes and motivation [2], the students' average entry grade [14], and the students' communication and language skills [15]. Among the supervisor variables, it has been shown that the supervisor's experience, research output and workload constitute factors of thesis success [15, 17]. However, the review of the literature leads to a conclusion that there are few studies explicitly focusing on bachelor thesis projects. Studies on completion of thesis projects mostly concern the doctor-

ate thesis [18, 19] while studies on undergraduate non-completion tend to focus on the whole programme, not the thesis specifically [20-22]. Furthermore, most studies have used a qualitative approach to investigate factors for thesis completion; single factors have been looked at in an isolated way with a primary focus on student variables and on completion factors (and not on non-completion and supervisor variables) [23-25] and there are few contemporary studies on completion and non-completion of bachelor thesis.

The introduction of thesis management systems, such as SciPro from Stockholm University [26] and Thesis Writer (TW) from Zurich University of Applied Sciences [27], offer an accurate recording of many aspects and interactions of the thesis process. These recordings along with the logs of using the system enable the use of learning analytics techniques for the pursuit of the factors and indicators behind thesis completion. Learning analytics have been used successfully to early map the indicators of successful course completion, inform course design, provide insights and feedback to teachers and students, as well as improve education outcome [28] and as such, learning analytics methods could offer valuable insights to students, educators and administrators. Examples include the early prediction of a troubled thesis.

It is against such a background this study takes as a departure point to better understand factors that influence completion – and in particular – non-completion of bachelor thesis projects by investigating some student and supervisor variables. We do this supported by the introduction of thesis management systems, such as Daisy and SciPro from Stockholm University, which record large amounts of data related to the thesis process, and consequently pave the ground for using learning analytics techniques to improve our understanding of thesis management and students' success factors [29]. More specifically, this study aims to, on the one hand, identify factors that contribute to thesis completion and non-completion, and on the other hand, predict completion and non-completion of thesis projects using machine learning algorithms. Early prediction of non-optimal thesis projects and insights from successful completers might help us introduce proactive interventions that salvage students at risk of non-completion and decrease the time to complete thesis projects.

1.1 Factors for completion of thesis projects

The literature review has led to the identification of two groups of factors that influence thesis outcomes: the student candidate and the supervisor. Below we give an account of what is known about these two groups of factors.

1.2 The thesis candidate

Rennie et al. proposed the term 'thesis-blocking' in 1987 while using a grounded theory approach to investigate the problem of thesis delay (Rennie & Brewer, 1987). They suggested that there are more ways for thesis blocking than completing it in a timely fashion. According to their research, successful thesis completion requires the candidate's conformity and acceptance of the thesis process, the willingness of the candidate to manage their idealism and cope with the overwhelming nature of the

project. Failure to resolve a candidate's negative feelings, the hesitation to approach their supervisors, are the causes for many candidates to be stuck in the middle of the path [2]. House and Johnson's findings point to the applicants' average entry grade as a decisive predictive factor of successful, timely completion [14], a finding that was corroborated by Jiranek [15] and Wright and Cochrane [30].

Nevertheless, some studies have shown that entry grade is not a significant predictor of completion [30, 31]. For instance, in a study by Pascarella and Terenzini [32], the background characteristics including entry grades (R2c .009) showed only to explain a small part of retention; it is academic and social integration that explain the persistence (R2c .127). The result suggests that the academic organisation has a potentially substantial effect on the student's study outcome.

Age of the candidate seemed to show different association with time to completion or grade of the thesis; the inconsistency continued when researchers considered the disciplines in the analysis [15, 17, 30, 33].

Qualities and approach of the candidate were also reported as considerable predictors; for instance the ability of the candidate to cope with the demanding nature of the study and the flexibility to adapt to the process [2], motivation to finish in a timely fashion, engagement with coursework, prior research work, prior coursework in a relevant field and choosing the appropriate courses that are most relevant to the thesis topic during the programme [16, 34]. Other factors also include communication skills and language proficiency skills [15], self-reliance and independence [35]. Family responsibilities and children seem to impact both genders in different ways. However, a right balance and proactive planning along with institutional support could mitigate the impact and assist the candidates [1, 11, 15, 17, 33, 35]. Research has also emphasised the ability to access laboratory and scientific resources, along with the availability of students' guidance and support services for a successful process [1]. Contrary to the common belief, part-time older candidates seemed to fare better than their counterparts in their approach to research, organising their duties and being independent [30].

1.3 The supervisor

The supervisor role is instrumental in every step in the development of the thesis; the role starts by guiding the research proposal topic, supervising the plan, overseeing or participating in the implementation of the research or the project. The supervisor guides the thesis document writing process, rectifies flaws, suggest directions and approves the final document in its final form. The supervisor' role extends to the arrangement of the defence examination and preparing the candidate for the event [36]. As expected, since the supervisor has control over each step in the process and has to be satisfied by the quality of the work the candidate is producing, they can intentionally or unintentionally delay the process.

In some cases, the supervisor may decide to terminate the process, as in cases where the candidate is deemed unfit to do the presumed work. Rennie and Brewer liken the supervisor's negative role to the writer's block phenomenon [2]. They suggested that both phenomena share essential features, the main problem being the writ-

er's internalisation of the negative feedback received from supervisors and poor management of duties and time constraints.

A healthy student–supervisor relationship is helpful to the success of the thesis project. The thesis is an embedded social exercise more than most of the other educational compositional projects, therefore collaborating with the supervisor, regular productive meetings and the ability to reach a shared understanding are central to the success of the project [1, 11, 16, 37, 38]. A relationship in which the supervisor exerts a moderate control over the process and greater affiliation was found to produce the best outcome regarding time and completion rates [39]. Supervisor experience and research output is a considerable factor that might be a predictor factor in the positive direction [15]. Furthermore, supervisor support is an indispensable element in all aspects of the development of the thesis process through all the stages, from the inception and the choice of the research topic to the tackling of obstacles and final presentation of the thesis [1, 15, 33, 37, 40].

Supervisors overwhelmed by research work, teaching or multiple students could lead to them having less time for students and minimise their interaction with candidates [11, 17]. Furthermore, the supervisor is constructive, and on-time feedback, commitment to plans and encouragement and communicability were reported by students as the most desirable factors that helped them complete their thesis projects [41].

2 Method

2.1 Sample and context

The sample for this study consisted of bachelor students' thesis projects (n=2436) during the period between 2010 and 2017 at the Department of Computer and Systems Sciences, Stockholm University, Sweden. Since it takes approximately 350 days for students to complete a thesis project, data from the year 2018 were excluded as they contained many projects likely to be completed after the data extraction.

The dropout rate for the thesis project at the department is approximately 30% for the period studied. We have included all bachelor thesis projects that adhere to the present curriculum for thesis projects. Table 1 below presents descriptive statistics regarding both student and supervisor variables.

Table 1. Descriptive statistics

Thesis Project	n	%	M	SD
Number of students doing thesis projects	2436			
Number of supervisors	155			
Completed Vs Incomplete Students (dropouts)				
Students with a completed thesis	1078	70.11		
Student with an incomplete thesis	728	29.88		
Students method courses grade	2436		1.82	1.86
Students average thesis grade	1708		3.38	0.46
Students average study grade	1708		3.38	0.75
Students average days to complete	1708		343.58	308.49
Supervisor average thesis grade	155		3.54	0.48
Supervissor No. of scientific publication	155		38.98	34.1
Supervisor No. of incomplete thesis projects	155		11.62	8.33
Superisor no of started thesis projects	155		29.32	15.72
Supervisor ratio incomplete thesis projects	155	39		0.18
Supervisor average days to complete	155		400.06	137.66

2.2 Data collection

A challenge in data collection for learning analytics is to avoid amplifying errors from different standards in data sources, especially if some sources are external and out of control. In this study, to minimise this risk for all data sources we used data that are under the control of the university.

Data collection was performed in several iterative steps. Using SQL (structured query language) queries, we extracted data from two different data systems used by the department to record data about the thesis projects. From these systems, we collected thesis project data concerning both students and supervisors. Informed by factors identified by previous research [14, 15], and taking into account additional variables that were available in the systems that record thesis data. We focused in general on three groups of factors that influence the academic thesis process, namely:

- Student's previous performance in the bachelor programme
- Supervisor's thesis project performance and experience
- Supervisor's research output.

More specifically, we extracted the following variables:

- Thesis project: start and completion date, from this the number of days to completion was calculated.
- The students: the grade of the thesis, method course and average grade in the study before the bachelor thesis.
- The supervisors: academic title, number of publications, year of publication, from this average number of publications per year was calculated, number of complete/incomplete thesis projects, number of started thesis projects, and average days of supervisors to complete thesis projects were calculated from the projects.

All data was anonymised by converting personal identifiers to fictive IDs. The researchers who did the analysis did not know the identity of the subjects. The data was subsequently prepared for statistical and predictive analytics by removal of extreme- and null values and through the computation of relevant variables.

Ethical approval for this study was obtained through the Regional Board of Ethical Vetting in Stockholm. Consent for participating in this research was also obtained from the selected supervisors in the sample. Six supervisors and their associated thesis projects were excluded due to no consent for using their data was received.

2.3 Data analysis

The analysis was performed using SPSS, and R. Spearman correlation test to investigate the correlation between incomplete thesis projects (dropouts) with student and supervisor variables. Multiple independent sample t-tests were performed in order to explore differences between completers and non-completers with regards to student and supervisor variables. The Shapiro−Wilk test of normality was employed and confirmed that the assumptions for the t-tests were satisfied.

For the predictive analytics, seven supervised machine learning classifiers were applied: Naive Bayes, Logistic Regression, kNN, Neural Network, Deep Learning, Decision Tree, and Random Forest in order to predict completers and non-completers of thesis projects. These classifiers were chosen because they are frequently used for predicting dropout, and each has demonstrated good and comparable performance in predicting at-risk students and dropout [42, 43]. The data set was split into a training and testing set. The training set consisted of 70% of the total data set, and the testing set the remaining 30%. After implementation of the predictive models, features were ranked using the information gain ratio. To prevent overfitting and increase robustness, 10-fold cross-validation was performed, where performances were measured from multiple iterations of cross-validation and averaged over iterations. To measure the prediction performance of the different models, the area under the receiver operating characteristic curve (AUC) was obtained, along with measures for precision and recall.

In addition, a clustering of the thesis projects was performed through the k-means clustering algorithm as well as through factor analysis using the principal component method with varimax rotation.

3 Results

3.1 Statistical analysis

After performing the descriptive analysis presented in Table 1, a correlations tests (Spearman's) was performed in order to study the correlation between incomplete thesis projects (dropouts) with student and supervisor variables (see full correlation matrix in Table 2). This analysis revealed that non-completion (dropout) is significantly (albeit weakly) correlated with grades on method course ($r=0.14$, $p<0.01$),

students' average grade in their study programme at the university (r=0.19, p<0.01), the average time it takes for supervisors to complete thesis projects (r=-0.17, p<0.01), the number of scientific publications published by supervisors (r=0.06, p<0.05), the ratio of incomplete thesis projects of supervisors (r=-0.35, p<0.01), and the total number of incomplete thesis projects of supervisors (r=-0.20, p<0.01). As can be noted, the ratio and total amount of unfinished thesis projects by supervisors presented the strongest correlations with thesis dropout. One can also note that the experience of teachers measured in the number of thesis projects supervised did not render a significant correlation with thesis non-completion (r=-0.14, p>0.05). Furthermore, it is noted that the supervisor's experience measured in the number of scientific publications shows a weak and almost non-existing correlation with non-completion of thesis projects (r=-0.07, p<0.07).

Table 2. Correlations with incomplete thesis projects (dropout)

Variable	1	2	3	4	5	6	7	8
Incomplete thesis projects	1.00							
Method course grade	**0.14	1.00						
Student average study grade	**0.19	**0.05	1.00					
Supervisor average day to day	**-0.17	-0.19	-0.06	1.00				
Supervisor no of scientific publications	**0.06	**-0.05	**-0.07	**-0.05	1.00			
Supervisor no. of started thesis projects	-0.14	0.03	**-0.05	**0.22	**-0.15	1.00		
Supervisor incomplete thesis projects %	**-0.35	**-0.05	**-0.04	**0.47	**-0.26	**-0.15	1.00	
Supervisor no of incomplete thesis projects	**-0.20	-0.20	**-0.05	**0.48	**-0.26	0.02	**0.08	1.00

** p significant at 0.01

Multiple independent t-tests were also performed in order to explore differences between completers and dropouts with regards to many student and supervisor variables. See Table 3 for a full presentation of the t-test results. Based on these tests the following can be concluded:

- There is significant difference between completers (M=2.02, SD=1.90) and non-completers (M=1.45, SD=1.71) regarding how well they performed during the preparatory and mandatory method course, t(-8.30)= 41.74, p<0.05;
- There is significant difference between completers (M=3.49, SD=0.69) and non-completers (M=3.18, SD=0.77) regarding their average grade during their studies in the programme they are seeking to graduate in, t(-11.42)= 9.09, p<0.05;
- There is significant difference between completers (M=3.57, SD=0.46) and non-completers (M=3.49, SD=0.47) in terms of their supervisors' average thesis grade, t(-5.09)= 2.40, p<0.05;
- There is significant difference between completers (M=387.23, SD=137.54) and non-completers (M=439.45, SD=166.80) in terms of their supervisors' average time to complete a thesis, t(9.40)= 25.67, p<0.05;
- There is significant difference between completers (M=0.34, SD=0.16) and non-completers (M=0.48, SD=0.18) in terms of their supervisors' ratio of incomplete thesis projects, t(22.25)= 5.45, p<0.05;

Significant differences were, however, not revealed concerning the total number of scientific publications published by supervisors or the total number of thesis projects supervised by the supervisors.

Table 3. Differences between completed and non-completed thesis projects (t-test)

		M	SD	F	t	p
Method course grade	Complete	2.02	1.90	41.74	-8.3	<0.01
	Incomplete	1.45	1.71			
Student average Study Grade	Complete	3.49	0.69	9.09	-11.42	<0.01
	Incomplete	3.18	0.77			
Supervisor average thesis grade	Complete	3.57	0.46	2.40	-5.09	<0.01
	Incomplete	3.49	0.47			
Supervisor average day to complete	Complete	387.23	137.54	25.67	9.4	<0.01
	Incomplete	439.45	166.80			
Supervisor no. of scientific publications	Complete	3.1	0.72	0.27	-1.66	0.1
	Incomplete	2.67	1.02			
Supervisor no. of started thesis projects	Complete	29.06	15.43	1.13	0.98	0.33
	Incomplete	29.62	15.9			
Supervisor percentage of incomplete thesis projects	Complete	0.34	0.16	5.45	22.25	<0.01
	Incomplete	0.48	0.18			

P significant at 0.05

3.2 Predictive analytics: classification and cluster analysis

3.3 Prediction of completers and non-completers

Then predictive analytics was performed using several machine learning models (Naive Bayes, Logistic Regression, Deep Learning, Decision Tree and Random Forest) in order to predict the completion/non-completion variable using the features described in Table 1. The performance across the models showed AUC values between 0.56 and 0.79 with modest gains for the Deep Learning and Logistic Regression model (see Table 4).

Table 4. Prediction accuracy and ROC

Model	Accuracy	AUC
Logistic regression	75.4%	0.79
Deep Learning	71.6%	0.77
Naïve Bayes	71.4%	0.74
Decision Tree	71.0%	0.56
Random Forest	70.8%	0.76

The logistic regression model proved to perform best concerning accuracy and AUC, with almost 91% accuracy in predicting the actual completers. However, the actual non-completers could only be predicted with a 42% accuracy (see Table 5). The Deep Learning model's performance, on the other hand, was more balanced and

it was able to identify 72% of the actual students that did not complete the thesis project and 71% of the actual completers (see Table 6).

As can be seen from Table 7, the features with most weight were the ratio of unfinished thesis projects of supervisors, students' average grade during university studies, and the average time it takes for supervisors to complete a thesis project.

Table 5. Prediction of completers and non-completers using Logistic regression

	True Completer	**True Nin-completer**	**Class Precision**
Predicted Completer	310	90	71.50%
Predicted Non-completer	32	64	66.67%
Class Recall	90.64%	41.56%	

Table 6. Prediction of completers and non-completers using Deep learning

	True Completer	**True Nin-completer**	**Class Precision**
Predicted Completer	244	43	85.02%
Predicted Non-completer	98	111	53.11%
Class Recall	71.35%	72.08%	

Table 7. Weights of selected features

Features	**Weights**
supervisor ratio incomplete thesis projects	1.0
Student average study grade	0.52
supervisor average days to complete	0.42
Student method course grade	0.37
Supervisor average thesis grade	0.21
Supervisor no. of scientific publications	0.03
Superivisor no. of started thesis projects	0.0

3.4 Cluster and factor analysis of complete and incomplete thesis projects

Then the k-means clustering algorithm was applied to see if the thesis projects can be grouped into some distinct clusters that share a similarity. A four-cluster solution demonstrated best silhouette scores. Among the 2436 student thesis projects in the sample, 1152 thesis projects were classified into the first cluster, 623 thesis projects were in the second, 421 thesis projects were in the third, and 240 thesis projects were in the fourth, respectively (see Table 8).

Table 8. K-means cluster analysis

Cluster	No of thesis projects	Characteristics (Variables)
1	813	Non-completers are on average 100% less, completers are on average 45% more, superivisor ratio of incomplete thesis projects 37% smaller, average supervisor days to complete 21% less.
2	704	Non-completers are on average 226% more; completers are on average 100% less, supervisor ratio of imcomplete thesis projects 27% larger, average supervisor days to complete 44% more.
3	666	Non-completers are on average 100% less, completers are on average 44% more, supervisor no. of scientific publications is on average 40% less.
4	253	Non-completers are on average 56% less, completers are on average 24% more, supervisor no. of scientific publication is on average 201% more.

What can be concluded from the cluster analysis is that one non-completer cluster is explained by supervisors having a high ratio of unfinished thesis projects and taking, on average, more days to complete thesis projects. For the completers, there are three different clusters respectively characterised by: 1) supervisors' ratio of unfinished thesis projects are smaller as well as their average days to complete thesis projects; 2) less scientific publications of supervisors; and 3) more scientific publications of supervisors. The cluster analysis thus points at three critical factors, namely: previous performance of supervisors measured in average time to complete thesis projects, and the ratio of incomplete/complete thesis projects, as well as the research output of supervisors measured in the total number of published scientific publications.

The cluster analysis was complemented with factor analysis using the principal component method with varimax rotation. The factor analysis resulted in four factors with eigenvalues greater than 1 that explained thesis outcomes. When excluding factors that did not contain items with loadings above 0.6, three factors remained that have been named: supervisor performance; supervisor research output; and student performance. Table 9 reports the factor loadings, eigenvalues and explained variance for each of the factor.

Table 9. Results of factor analysis

Items	Supervisor performance	Supervisor research output	Student performance
Supervisor no. of Incomplete thesis projects	0.850		
Supervisor average thesis grade	-709		
Supervisor no. of started thesis projects		0.653	
Supervisor no. of scientific publications		0.624	
Student method course grade			0.795
Student average study grade			0.679
Eigenvalue	2.38	1.90	1.24
% of variance	24%	19%	12%

Consequently, the results from the factor analysis demonstrate that non-completion of thesis work is dependent on supervisor performance in terms of the number of thesis projects they complete and the grades of these, how many thesis projects super-

visors have experienced, supervisors research output, and students grades on prepatory methodology courses and their average study grade.

4 Discussion

The bachelor thesis is, almost worldwide, a necessary last step towards the first graduation in higher education and thus constitutes a central key to both further studies in higher education as well as employment that requires higher education degrees. In light of this, non-completion results in a vast waste of faculty time and institutional resources, a devastating personal experience for students that costs precious time, loss of money and energy, and a societal loss of high-skilled workers [1, 3, 11].

The results demonstrated that three general groups of factors influence students' performance during the thesis process regarding completion and non-completion. These were:

- Student's previous performance in the bachelor programme
- Supervisor's thesis project performance and experience
- Supervisor's research output

The statistical analysis revealed that non-completion of thesis projects most strongly correlated with the supervisor's ability to complete thesis projects measured in the ratio of unfinished thesis projects and average time to complete thesis projects. The independent sample t-tests pointed out significant differences between completed and incomplete thesis projects regarding students' previous performance (average grade for the whole bachelor programme), supervisors' average thesis grade, supervisors' average time to complete a thesis, and supervisors' ratio of unfinished thesis projects. The conducted factor analysis pointed out three factors: the supervisor's previous thesis performance; the supervisor's research output; and the student's previous performance. In total, 43% of the variance in the thesis outcomes (regarding completion) could be explained by the supervisor's previous performance and their research output, and 12% by the student's previous performance. The cluster analysis performed generated four clusters of thesis projects. Among these, one was a non-completer cluster that is explained by supervisors having a high ratio of unfinished thesis projects and taking, on average, more days to complete thesis projects. For the completers, we identified three different clusters in which the following variables were influential:

Cluster 1) Supervisors' ratio of unfinished thesis projects are smaller as well as their average days to complete thesis projects

Cluster 2) Less scientific publications of supervisors

Cluster 3) More scientific publications of supervisors.

Thus, taking all the above-mentioned analyses into account, we can with confidence conclude that the time supervisors take, on average, to complete thesis projects, and their experience of completing or not completing thesis projects (in terms of ratio of incomplete thesis projects), play a significant role in determining the completion and non-completion of thesis projects. No similar results have been found in previous

work, which to some extent has overlooked supervisor variables, and thus these findings can be considered as one of the main contributions of this paper. Furthermore, the analysis leads to the conclusion that the research output of supervisors is influential in determining success, but to a much lesser extent than the previously mentioned factors. This particular finding corroborates Jiranek (2010), who looked at supervisors of doctoral students. However, we could also conclude that the academic title of the supervisor, which was one of the variables studied, did not constitute a factor for completing thesis projects. Also, we could conclude that students' previous performance regarding average grade during the bachelor programme is shown to be one of the factors that influence completion and non-completion of thesis projects. However, this student-related factor was less influential than supervisors' average time to complete thesis projects and their ratio of unfinished thesis projects. The finding that students' previous grades are of less importance for thesis completion corroborates previous findings of Tinto [44], Pascarella and Terenzini [45], Astin [46], and Nouri et al. [47].

Another novel contribution of this study stems from the application of machine learning algorithms, which were used in order to predict thesis completion/non-completion. Using the set of features mentioned in previous sections, and especially when using the deep learning algorithm, it is possible to predict completers and non-completers of thesis projects reasonably accurately. For future work, we would suggest adding more student features/variables by using additional data collection methods such as questionnaires, in order to increase the performance of the predictive model — the limited amount of student variables is one of the limitations of this study.

As a final remark, it is argued that the insights gained from the statistical analysis, and the predictive models constructed through machine learning algorithms, can be used to support the matching of students and supervisors in order to increase graduation rates and the probability for thesis projects being completed.

5 References

[1] Ho, J. C., Wong, P. T. and Wong, L. C. What helps and what hinders thesis completion: A critical incident study, International Journal of Existential Psychology and Psychotherapy, vol. 3, 2010.

[2] Rennie, D. L. and Brewer, L. A grounded theory of thesis blocking, Teaching of Psychology, vol. 14, pp. 10-16, 1987. https://doi.org/10.1207/s15328023top1401_2

[3] Baum, S., Ma, J., and Payea, K. Education pays, The Benefits of Higher Education for Individuals and Society, 2013.

[4] Lance, L. Nonproduction benefits of education: Crime, health, and good citizenship, in Handbook of the Economics of Education. vol. 4, ed: Elsevier, 2011, pp. 183-282. https://doi.org/10.1016/b978-0-444-53444-6.00002-x

[5] Ma, J., Pender, M. and Welch, M. Education Pays 2016: The Benefits of Higher Education for Individuals and Society. Trends in Higher Education Series, College Board, 2016.

[6] Bourke, S., Holbrook, A., Lovat, T. and Farley, P., Attrition, completion and completion times of PhD candidates, in AARE annual conference, Melbourne, 2004.

[7] Ferrer, F. P., Determinants of Performance in Thesis: Evidence from Selected Filipino Graduate Students, International Journal of Education and Research, vol. 2, pp. 189-202, 2014.
[8] Morton, K. R. and Worthley, J. S., Psychology graduate program retention, completion and employment outcomes, Journal of Instructional Psychology, 1995.
[9] Agu, N. and Oluwatayo, G. K. Variables attributed to delay in thesis completion by post-graduate students, Journal of Emerging Trends in Educational Research and Policy Studies, vol. 5, pp. 435-443, 2014.
[10] Kamler, B. and Thomson, P. The failure of dissertation advice books: Toward alternative pedagogies for doctoral writing, Educational Researcher, vol. 37, pp. 507-514, 2008. https://doi.org/10.3102/0013189x08327390
[11] Rauf, F. A. Challenges of Thesis Work: Towards Minimizing the Non-Completion Rate in the Postgraduate Degree Program, European Journal of Business and Management, vol. 8, pp. 113-124, 2016.
[12] . Wong, P. t. P Meaning Making and the Positive Psychology of Death Acceptance, International Journal of Existential Psychology & Psychotherapy, vol. 3, pp. 73-82, 2010.
[13] Chin, W. Y., Ch'ng, C. K., Jamil, J. M. and. Shaharanee, I. N. M Analyzing the factors that influencing the success of post graduates in achieving graduate on time (GOT) using analytic hierarchy process (AHP), in AIP Conference Proceedings, 2017, p. 040009. https://doi.org/10.1063/1.5012197
[14] House, J. D. and Johnson, J. J. Predictive validity of Graduate Record Examination scores and undergraduate grades for length of time to completion of degree, Psychological Reports, vol. 71, pp. 1019-1022, 1992. https://doi.org/10.2466/pr0.1992.71.3.1019
[15] Jiranek, V. Potential predictors of timely completion among dissertation research students at an Australian faculty of sciences, International Journal of Doctoral Studies, vol. 5, pp. 1-13, 2010. https://doi.org/10.28945/709
[16] Pitchforth, J., Beames, S. Y., Thomas, A., Falk, M. G., Farr, A. C., Gasson, S. et al., Factors affecting timely completion of a PhD: a complex systems approach, Journal of the Scholarship of Teaching and Learning, vol. 12, pp. 124-135, 2012.
[17] Van de Schoot, R., Yerkes, M. A., Mouw, J. M. and Sonneveld, H. What took them so long? Explaining PhD delays among doctoral candidates, PloS one, vol. 8, p. e68839, 2013. https://doi.org/10.1371/journal.pone.0068839
[18] Can, E., Richter, F., Valchanova, R. and Dewey, M. Supervisors' perspective on medical thesis projects and dropout rates: survey among thesis supervisors at a large German university hospital, BMJ open, vol. 6, p. e012726, 2016. https://doi.org/10.1136/bmjopen-2016-012726
[19] Van Ours J. C. and Ridder, G. Fast track or failure: a study of the graduation and dropout rates of Ph D students in economics, Economics of Education Review, vol. 22, pp. 157-166, 2003. https://doi.org/10.1016/s0272-7757(02)00029-8
[20] DesJardins, S. L., Kim, D.-O. and. Rzonca, C. S A nested analysis of factors affecting bachelor's degree completion, Journal of College Student Retention: Research, Theory & Practice, vol. 4, pp. 407-435, 2003. https://doi.org/10.2190/bgmr-3ch7-4k50-b5g3
[21] Graham, L. D. Predicting academic success of students in a master of business administration program, Educational and Psychological Measurement, vol. 51, pp. 721-727, 1991. https://doi.org/10.1177/0013164491513023
[22] Herzog, S. Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression, New directions for institutional research, vol. 2006, pp. 17-33, 2006. https://doi.org/10.1002/ir.185

[23] Ishitani, T. T. Studying attrition and degree completion behavior among first-generation college students in the United States, The Journal of Higher Education, vol. 77, pp. 861-885, 2006. https://doi.org/10.1080/00221546.2006.11778947
[24] Kovacic, Z. Early prediction of student success: Mining students' enrolment data, presented at the Informing Science + Information Technology Education Joint Conference, Cassino, Italy., 2010. https://doi.org/10.28945/1281
[25] Zwick, R. and Sklar, J. C. Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language, American Educational Research Journal, vol. 42, pp. 439-464, 2005. https://doi.org/10.3102/00028312042003439
[26] Hansen, P. and Hansson, H. Optimizing Student and Supervisor Interaction During the SciPro Thesis Process–Concepts and Design, in International Conference on Web-Based Learning, 2015, pp. 245-250. https://doi.org/10.1007/978-3-319-25515-6_23
[27] Rapp, C. and Ott, J. Learning Analytics in Academic Writing Instruction–Opportunities Provided by Thesis Writer (TW), in Bildungsräume 2017, C. U. Igel, Carsten; Wessner Martin, Ed., ed Bonn: Gesellschaft für Informatik, 2017, pp. 391-392.
[28] Ifenthaler, D., Mah, D.-K. and. Yau, J. Y.-K Utilizing learning analytics to support study success, ed: Springer, 2018. https://doi.org/10.1007/978-3-319-64792-0
[29] Rapp, C. and Kauf, P. Scaling Academic Writing Instruction: Evaluation of a Scaffolding Tool (Thesis Writer), International Journal of Artificial Intelligence in Education, pp. 1-26, 2018. https://doi.org/10.1007/s40593-017-0162-z
[30] Wright, T. and Cochrane, R. Factors influencing successful submission of PhD theses, Studies in higher education, vol. 25, pp. 181-195, 2000. https://doi.org/10.1080/713696139
[31] Siegel, L. A study of Ph. D. completion at Duke University, CGS Communicator, XXXVIII, vol. 1, p. 2, 2005.
[32] Pascarella, E. T. and Terenzini P. T., Predicting voluntary freshman year persistence/withdrawal behavior in a residential university: A path analytic validation of Tinto's model, Journal of educational psychology, vol. 75, p. 215, 1983. https://doi.org/10.1037/0022-0663.75.2.215
[33] Manathunga, C. Early warning signs in postgraduate research education: A different approach to ensuring timely completions, Teaching in Higher Education, vol. 10, pp. 219-233, 2005. https://doi.org/10.1080/1356251042000337963
[34] Maher, M. A., Ford, M. E. and. Thompson, C. M Degree progress of women doctoral students: Factors that constrain, facilitate, and differentiate, The Review of Higher Education, vol. 27, pp. 385-408, 2004. https://doi.org/10.1353/rhe.2004.0003
[35] Castro, V., Cavazos Jr, J., Garcia, E. E. and Castro, A. Y. The road to doctoral success and beyond, International Journal of Doctoral Studies, vol. 6, pp. 51-78, 2011. https://doi.org/10.28945/1428
[36] Retalis, S., Papasalouros, A., Psaromiligkos, Y., Siscos, S. and Kargidis, T. Towards networked learning analytics–A concept and a tool, presented at the fifth international conference on networked learning, Lancaster, UK, 2006.
[37] Koskenoja, M. Factors supporting and preventing master thesis progress in mathematics and statistics, International electronic journal of mathematics education, 2019. https://doi.org/10.29333/iejme/3986
[38] Styles, I. and Radloff, A. The synergistic thesis: Student and supervisor perspectives, Journal of Further and Higher Education, vol. 25, pp. 97-106, 2001. https://doi.org/10.1080/03098770020030533
[39] de Kleijn, R. A., Mainhard, M. T., Meijer, P. C., Pilot, A. and Brekelmans, M. Master's thesis supervision: Relations between perceptions of the supervisor–student relationship,

final grade, perceived supervisor contribution to learning and student satisfaction, Studies in Higher Education, vol. 37, pp. 925-939, 2012. https://doi.org/10.1080/03075079.2011.556717

[40] Lindsay, S. What works for doctoral students in completing their thesis?, Teaching in Higher Education, vol. 20, pp. 183-196, 2015. https://doi.org/10.1080/13562517.2014.974025

[41] de Kleijn, R. A., Mainhard, M. T., Meijer, P. C., Brekelmans, M. and Pilot, A. Master's thesis projects: student perceptions of supervisor feedback, Assessment & Evaluation in Higher Education, vol. 38, pp. 1012-1026, 2013. https://doi.org/10.1080/02602938.2013.777690

[42] Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R. and. Baron, J. D Early alert of academically at-risk students: An open source analytics initiative, Journal of Learning Analytics, vol. 1, pp. 6-47, 2014. https://doi.org/10.18608/jla.2014.11.3

[43] Kashyap, A. and Nayak, A. Different Machine Learning Models to Predict Dropouts in MOOCs, in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 80-85. https://doi.org/10.1109/icacci.2018.8554547

[44] Tinto, V. Dropout from higher education: A theoretical synthesis of recent research, Review of educational research, vol. 45, pp. 89-125, 1975. https://doi.org/10.2307/1170024

[45] Pascarella, E. T. and Terenzini, P. T. Predicting freshman persistence and voluntary dropout decisions from a theoretical model, The journal of higher education, vol. 51, pp. 60-75, 1980. https://doi.org/10.1080/00221546.1980.11780030

[46] Astin, A. W. Student involvement: A developmental theory for higher education, Journal of college student personnel, vol. 25, pp. 297-308, 1984.

[47] Nouri, J., Larsson, K., & Saqr, M. (2019). Identifying factors for master thesis completion and non-completion through learning analytics and machine learning. In proceeding of 14th European Conference on Technology Enhanced Learning (EC-TEL).

6 Authors

Jalal Nouri is associate professor at the Department of Computer and Systems Sciences, Stockholm University.

Ken Larson is currently PhD candidate, formerly a lecturer at the department of Computer and Systems Sciences of Stockholm University.

Mohammed Saqr is a senior researcher at the School of Computing, University of Eastern Finland.

Article submitted 2019-06-19. Resubmitted 2019-07-23. Final acceptance 2019-07-23. Final version published as submitted by the authors.