This is an Author's Original Manuscript of an article published by Taylor & Francis in Innovations in Education and Teaching International on 09/08/2018, available online at http://www.tandfonline.com/10.1080/14703297.2018.1502090

University Student Retention: Best Time and Data to Identify Undergraduate Students at Risk of Dropout

José María Ortiz-Lozano [a]; Antonio Rua-Vieites [b]; Martí Casadesús-Fa [c]; Paloma Bilbao-Calabuig [d]

[a] *Faculty of Economics and Business Administration, Universidad Pontificia Comillas, Madrid, Spain;* [b] *Department of Quantitative Methods, Faculty of Economics and Business Administration, Universidad Pontificia Comillas, Madrid, Spain;* [c] *Department of Business Management and Product Development, Universitat de Girona, Girona, Spain;* [d] *Department of Management, Faculty of Economics and Business Administration, Universidad Pontificia Comillas, Madrid, Spain.*

Corresponding Author: José María Ortiz Lozano, E-mail: jmortiz@comillas.edu, Universidad Pontificia Comillas. C/Alberto Aguilera 23, 28015 (Madrid, Spain).

Dr. José María Ortiz Lozano is the Director of the Registrar's Office at Universidad Pontificia Comillas. He performs studies in the field of quality management and applies multivariate analysis techniques to study management.

Dr. Antonio Rua Vieites is a lecturer at Universidad Pontificia Comillas. He lectures and performs quantitative methods studies in the fields of management and sociology.

Dr. Marti Casadesus is a professor at the University of Girona. His studies focus on quality management and have been published in journals such as *Total Quality Management*, *International Journal of Quality & Reliability Management*, *The TQM Magazine*, and *International Journal of Operations & Management*.

Dr. Paloma Bilbao-Calabuig is a lecturer at Universidad Pontificia Comillas. She lectures and performs studies in the field of Corporate Governance and Sustainability and has published in journals such as *Human Ecology Review*.

This is an Author's Original Manuscript of an article published by Taylor & Francis in Innovations in Education and Teaching International on 09/08/2018, available online at http://www.tandfonline.com/10.1080/14703297.2018.1502090

University Student Retention: Best Time and Data to Identify Undergraduate Students at Risk of Dropout

Student dropout is a major concern in studies investigating higher education retention strategies. However, studies investigating the optimal time to identify students who are at risk of withdrawal and the type of data to be used are scarce. Our study consists of a withdrawal prediction analysis based on classification trees using both sociodemographic and academic data from 935 first-year students at an engineering school in Spain. We build prediction models using information collected at three different moments throughout the first semester of the students' first university year. Our results echo those of previous studies supporting the need for an early first-year intervention to prevent non-completion. In addition, academic performance data serve as a good predictor. Finally, academic monitoring throughout the first semester improves the prediction accuracy, challenging the demand for "as soon as possible" identification of students who are at risk of dropout.

Keywords: higher education, student withdrawal, student dropout, classification trees, retention strategies

Introduction

Student withdrawal from higher education (HE) leads to reputational prejudices against educational institutions and, occasionally, consequent income losses (Berge & Huang, 2004). High withdrawal rates lead to the perception that institutions fail to provide adequate resources to help students with difficult academic situations (Cabrera, Bethencourt, Pérez, & Alfonso, 2006). Therefore, these rates erode universities' brand image (Berge & Huang, 2004; Cabrera et al., 2006) to an extent that state education bodies consider these rates a prime criterion for HE evaluation (Thomas, 2011, 2012; Wimshurst, Wortley, Bates, & Allard, 2006).

The European HE context, i.e., our study context, is no exception, and student dropout has become a major challenge. The significance of this challenge varies across countries. Several national HE policies focus on attaining high completion and time-to-degree rates. In other policies, maintaining low dropout rates is a priority. Furthermore, the European Commission aims to improve the knowledge and skills demanded by the labour market and implement a more productive and socially equal environment (Vossensteyn et al., 2015).

Student dropout is caused by a complex set of factors and is context specific; thus, research efforts should focus on the measures implemented by both public and private

educational institutions to reduce dropout (Thomas, 2011; Thomas & Hovdhaugen, 2014). In contrast to studies investigating the causes of student dropout, although student retention poses a significant challenge for universities (Bernold, Spurlin, & Anson, 2007; Thomas, 2011, 2012), studies investigating the efficiency of retention strategies remain scarce (Brooman & Darwent, 2014; Thomas, 2011).

Previous studies confirm that the early identification of students at risk of dropping out is an effective strategy because it facilitates early intervention actions, such as tutoring, counselling and mentoring, all of which prevent student withdrawal (Bland, Taylor, Shollen, Weber-Main, & Mulcahy, 2009; Cabrera et al., 2006; Herzog, 2006; Larose et al., 2011; Lowis & Castley, 2008). By insisting on "early intervention," scholars emphasise that action must be taken with first-year students (Thomas, 2012; Vinson et al. 2010; Wilson et al., 2016). Logically, the early identification of students at risk of dropping out should also occur during that year.

However, despite its relevance for institutional decision-making, the exact moment of identification has not been extensively studied. For earlier identification, institutions could use pre-university data, including both academic and sociodemographic data, and the time allowed for the intervention could be the entire first academic year. A later intervention could rely on the students' university academic results, which could compel institutions to "wait and see" before intervening while also providing the institutions an opportunity to optimise their intervention resources and enhance the efficiency of their actions (McFarlane, 2016). Our study addresses the following research questions to help universities address this issue and to support universities in the design of intervention strategies: When is the best time and what are the best data to identify undergraduate students at risk of dropping out?

Our study fits within the field of Educational Data Mining (EDM), which is a relatively new research field aimed to build knowledge and establish methodologies based on massive amounts of data obtained from educational contexts (Baradwaj & Pal, 2012; Dekker, Pechenizkiy, & Vleeshouwers, 2009). EDM promotes advances in data-based decision-making in education, helping practitioners analyse and transform large amounts of data into rich, easily manageable and reliable knowledge (Lin, 2015). EDM-approached studies in HE have focused on decision situations, such as curriculum design and the prediction of student academic performance, including dropout, to enhance HE quality (Dekker et al., 2009; Lin, 2015; Quadri & Kalyankar, 2010).

Given our prediction-centred research goal and the amount of available data, we chose the classification trees technique over the other EDM methods, such as clustering, association analysis, Bayesian and neural networks, and logistic regression. Classification trees outperform other EDM predictive techniques in the context of dropout prediction. Classification trees allow the use of any type of data scale. This approach is simple to apply, and the results are easy to interpret as shown in previous dropout prediction studies (Baradwaj & Pal, 2012; Dekker et al., 2009; Lin, 2015; Quadri & Kalyankar, 2010; Ramaswami & Bhaskaran, 2010; Vandamme, Meskens, & Superby, 2007) and studies within the particular engineering education context (Mendez, Buskirk, Lohr, & Haag, 2008; Pal, 2012).

Spanish HE engineering was our study context. This is an interesting context for

two reasons. First, the demanding engineering curricula and the students' lack of the required mathematical knowledge, problem-solving abilities and study and learning skills contribute to dropout (Bernold et al., 2007; Cole, High, & Weinland, 2013; Engelbrecht, Harding, & Du Preez, 2007; Fantz, Siller, & DeMiranda, 2011; Forsman, Van den Bogaard, Linder, & Fraser, 2015; Mendez et al., 2008; Raelin et al., 2014; Sancho-Vinuesa, Escudero-Viladoms, & Masià, 2013; Van den Bogaard, 2012). Second, the Spanish HE bodies recognise the urgent need to improve the dropout rates in the country with the goal of relieving the dramatic unemployment rates and redefining the economic model (Arce, Crespo & Míguez-Álvarez, 2015).

Causes of withdrawal and retention strategies

Previous studies have extensively explored and classified the causes of university withdrawal (Adam & Gaither, 2005; Cabrera et al., 2006; Herzog, 2006; Tinto, 2006; Vivian, 2005; Yasmin, 2013; Yorke & Thomas, 2003). Several authors refer to the following three general dimensions (Berge & Huang, 2004; Swail, 2004): students' social causes, students' cognitive causes, and education institutions' causes. Other authors identify pre-entry scenario causes, including poor information leading to inappropriate university programme choice and a gap between the student's expectations and the actual education experience or the pre-university academic preparation (Thomas, 2011). The following causes are associated with the students' personalities: poor levels of self-efficacy, autonomous learning, or social integration (Brooman & Darwent, 2014). Finally, dropout may stem from the students' social interactions as follows: academic performance, family traits, in-class behaviour, and social-life participation (Esteban-García et al., 2016; Tinto, 2006).

Previous studies have shown that educational institutions exert efforts to address dropout causes (Berge & Huang, 2004; Bernold et al., 2007; Thomas, 2012; Van den Bogaard, 2012; Yorke, 2016; Zepke & Leach, 2010). These studies have exposed the relationship between retention strategies undertaken by institutions and students' academic success. However, "[i]t is (…) difficult to translate this knowledge into activities that impact on student persistence and success and institutional outcomes" (Thomas, 2012). Despite the lack of a more obvious practical benefit, these previous studies have provided a profound understanding of how retention strategies can address the causes of withdrawal.

Several retention strategies focus on social causes. These strategies aim to improve students' sense of engagement and belonging to the institution, which is a very important part of their lives. When students feel that they are a part of the institution, retention rates and academic success improve (Thomas, 2012; Yorke, 2016; Zepke & Leach, 2010). Such strategies consist of preparing students for their academic experience, i.e., by providing better information regarding the programme's content and having programme managers that are more involved in the recruitment process, and developing a culture of engagement and belonging within the institution (Thomas, 2011, 2012). Co-curricular interventions are also included in this group of strategies because these interventions intensify student involvement and preparation for university life (Wilson et al., 2016). These strategies often involve tutoring, advising and mentoring actions, all of which have a positive impact on first-year student withdrawal prevention (Bland et al.,

2009; Cabrera et al., 2006; Larose et al., 2011).

Other retention strategies aim to reduce students' cognitive causes of dropout. These strategies focus on the first-year curriculum content and assessment design (Wilson et al., 2016) and collaborate to alleviate the social causes of dropout, supporting the hypothesis that academic performance traits most strongly determine student withdrawal (Esteban-García et al., 2016).

A third approach integrates the two previous groups of strategies into an "institutional" retention approach searching for a more consistent and accurate way of acting (Cabrera et al. 2006; Wilson et al., 2016). Examples of such strategies include integrated co-curricular and curricular design and effective recruiting actions.

When to take action

Several scholars demand a "holistic" approach to student retention because they believe that interventions focusing on individuals with a high withdrawal risk place the responsibility of change on the students (Esteban-García et al., 2016; Thomas, 2012; Tinto, 2006). The holistic perspective encourages institutions to undertake transformation (e.g., via the creation of a culture of integration, engagement, and retention) to benefit all students, including those identified as "not at risk".

However, "[i]t is not always clear from the research whether the challenges of transition into higher education are common for all students, or just specific groups" (Thomas, 2011). In fact, numerous studies have empirically supported the target-oriented approach as a way to reduce university dropout (Brooman & Darwent, 2014; Bland et al., 2009; Cabrera et al., 2006; Larose et al., 2011; Sancho-Vinuesa et al, 2013). Furthermore, tutors do not always have the required time to work properly with many tutees (McFarlane, 2016), which could also support a targeted retention strategy.

Several sociological models help HE institutions predict student withdrawal to facilitate intervention and improve student academic performance (Lowis & Castley, 2008). Studies have shown that the sooner students at a higher risk are subject to retention actions, the more efficient the strategies are (Herzog, 2006; Lowis & Castley, 2008; Vivian, 2005). The first university year is widely recognised as critical for student transition in HE (Wilson et al., 2016). The question of when is the optimum moment in the first year to make such inferences and take other actions remains mainly unsolved despite certain evidence indicating that the "best" moment is between the first 6-7 weeks and the first long break (Thomas, 2012; Vinson et al. 2010; Wilson et al., 2016).

Several studies report a long and continuous "execution" period for retention strategies to be efficient (Brooman & Darwent, 2014; Lowis & Castley, 2008; Sancho-Vinuesa, 2013). This longer retention process combines individual academic follow-up in the form of assessed assignments and exams that occur throughout the specific course with other student-centred actions, such as tutoring. Giving students individual feedback on their performance improves not only their cognitive skills but also their engagement because they have had time and information to develop higher self-efficacy and more independent learning during their transition period. Students gain a sense of control over the programme content and feel more confident. These strategies prevent dropout and

increase academic success (Brooman & Darwent, 2014). This lengthier retention strategy should continue throughout the first semester of the first university year, and feedback and contact actions should address different aspects of the programme content (Brooman & Darwent, 2014).

Lengthier retention strategies can benefit from both academic and sociodemographic data. Most studies are based on students' sociodemographic traits, but empirical evidence has also revealed that academic performance (including pre-entry academic results and, if applicable, entry grades) can be a sound basis for withdrawal prediction (Aina, 2013; Lin, 2015; Lowis & Castley, 2008; Stoessel, Ihme, Barbarino, Fisseler, & Stürmer, 2015).

Given the relevance of the debate regarding the best time and data for retention strategies, our research design aimed to evaluate the effectiveness of identifying students at a higher risk of dropout at different moments during the beginning of the first academic year applying different data. The purpose was to discover whether institutions can improve their withdrawal prevention strategies using time and type of data as key parameters.

Methods

To identify students at a risk of dropout, we adopted the classification trees technique. Classification trees build predictive models by selecting a set of explanatory variables, i.e., predictors, that best predict the observation types represented on the target variable values. The models include a tree consisting of nodes and a target variable distribution function assigned to each node. To form a tree, an iterative algorithm creates splits of observations in each node forming homogenous, exhaustive and mutually exclusive groups. This segmentation process choses the predictor that produces the most discriminative division. This process is accomplished by checking the impurity level, i.e., the probability that an observation from the sample belongs to one of the values identified in the target variable (Bacallao & Bacallao, 2010; Breiman, Friedman, Stone, & Olshen, 1984).

In our study, student dropout was the target variable, and the students' socio-demographic and academic data were the predictors (variables containing the students' data). We performed a three-model analysis. In the first model, we processed student data collected during the induction week (induction-week model); the second model was derived from data obtained during the first 6-7 academic weeks (first-6-7-weeks model); and the third model used data collected at the end of the first semester when the definite academic results were disclosed (end-first-semester model). We analysed 935 first-year students enrolled in the Electromechanical Engineering Degree programme of the Universidad Pontificia Comillas in Madrid. The data were collected over four academic years, i.e., 2010/2011 to 2013/2014. We used both sociodemographic and academic performance variables as predictors (see Appendix 1).

We used Classification and Regression Trees (CART) and Quick, Unbiased, Efficient, Statistical Tree (QUEST) algorithms to build the trees. These algorithms check the impurity level of each discriminative division based on entropy criteria. We followed the general rule of a minimum of 100 students to divide a node and a minimum of 50

students to create a node to limit the growth of the tree (Mercado, 2012).

Results

The induction-week model resulted in a one-level tree (three nodes). Of the 935 students in the original dataset, only 19.1% withdrew. However, 39.9% of the students with an "admission test grade" of 7.57[1] or less dropped out. Thus, "admission test grade" is a dropout predictor in the model as follows: a student scoring 7.57 or less has a 0.399 probability of dropping out during the first university year (compared with the probability of 0.191 in the total group). The tree's predictive ability is measured by the rate of correct predictions (60.9%) as follows: the rate of students predicted as "risky" who withdrew and the estimated probability of the model to make incorrect predictions (0.325), i.e., to identify "risky" students who did not drop out and "not-risky" students who dropped out.

After the first 6-7 academic weeks, students complete several mid-term tests; thus, more academic performance data were available for our processing. We obtained a three-level tree (seven nodes) and two academic performance predictors. According to this model, the individuals most likely to drop out are those with an "admission test grade" of 6.57 or less (probability of 0.54 to drop out) and those with an "admission test grade" between 6.57 and 8.01 who did not pass their mid-term Chemistry test (probability of 0.391 to drop out). In this case, the model's predictive ability is 70.9% of correctly predicted students, and the estimated probability of incorrect prediction is 0.297.

To build the end-first-semester model, we added the first official academic results (available at the end of the first semester) to the set of possible predictors. The resulting model was a four-level tree (eleven nodes) identifying students with the highest risk of dropout among those forming nodes 1, 5 and 9 (Figure 1).

[1] The HE admission grade in Spain is the weighted average grade of the two final school years' results and the score obtained on the official university admission exam. Academic grades range from 0 to a maximum of 10, and '5' is the passing grade.

This is an Author's Original Manuscript of an article published by Taylor & Francis in Innovations in Education and Teaching International on 09/08/2018, available online at http://www.tandfonline.com/10.1080/14703297.2018.1502090

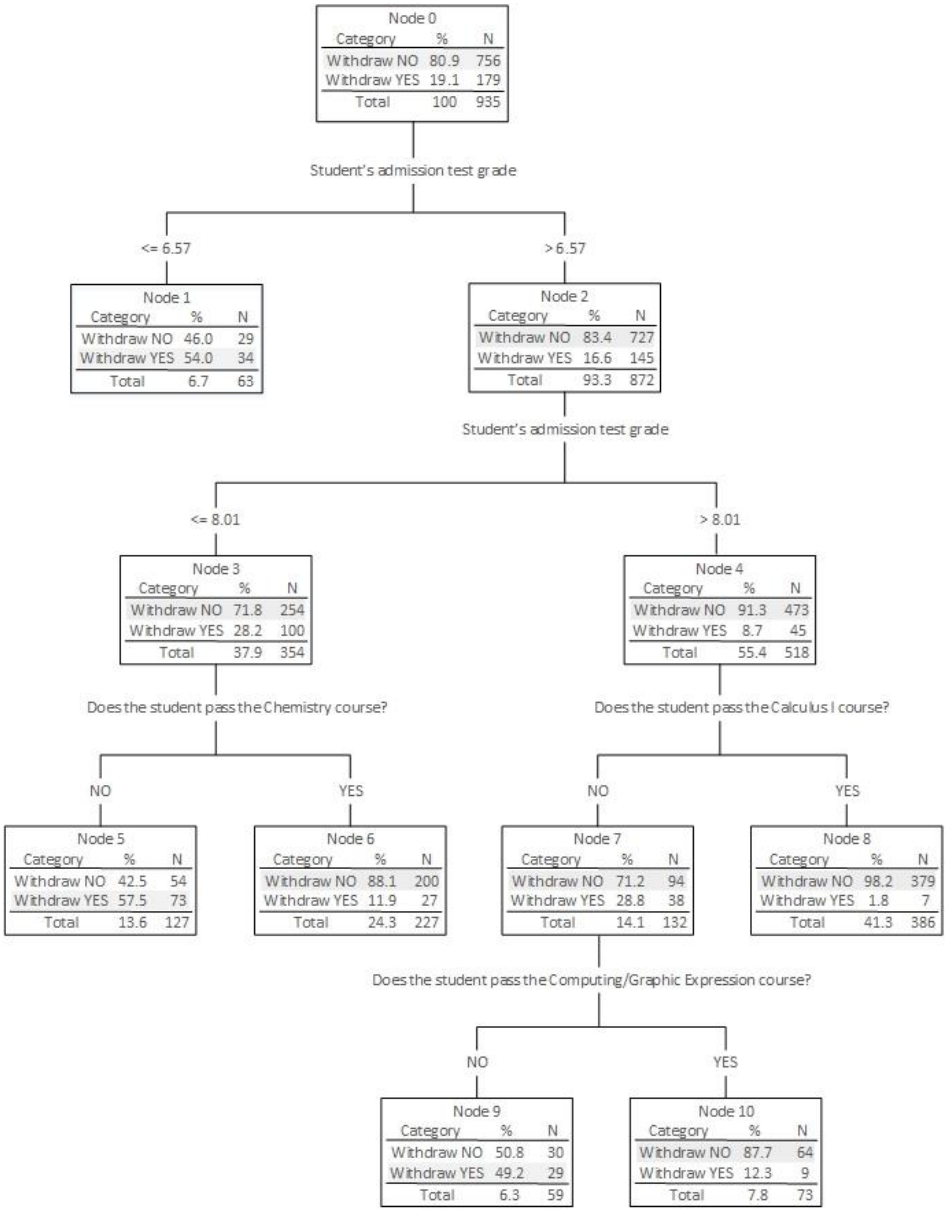

Figure 1. End-first-semester model classification tree

Compared with the two previous models, this model maintains the "admission test grade" variable as a dropout predictor in all three models as the first student discrimination variable (first split of each tree). The end-first-semester model also reveals that the Chemistry academic results are good predictors of student withdrawal. Finally, our third model enriches the previous model because it helps identify students who are at risk among those with an "admission test grade" above 8.01. Such students, who could have been classified as "not risky" by the first-6-7-weeks model, can be identified as "risky" if their results in the "passes the Calculus I course" and "passes either the Computing or Graphic Expression courses" predictors are considered.

All four predictors are academic performance variables. This model's predictive ability is 76% (correct prediction) and 0.213 (estimated probability of incorrect prediction).

This is an Author's Original Manuscript of an article published by Taylor & Francis in Innovations in Education and Teaching International on 09/08/2018, available online at http://www.tandfonline.com/10.1080/14703297.2018.1502090

Our results reveal the relevance of academic performance variables in identifying undergraduate students who are at a risk of dropout. Although pre-university academic performance (admission test grade) is a discriminant predictor in all models' first split (the unique discriminant in the induction-week model), delaying the moment to identify students who are at risk allowed us to incorporate new predictors. These new predictors helped detect students who are at risk who would have otherwise been ignored. Thus, we worked with additional data, which contributed to improving the predictive ability of our models. Our results suggest that first, institutions should consider their curricular courses performance data as dropout predictors and that second, institutions should consider a longer-term and continuous retention strategy versus a single-decision strategy implemented at the very beginning of the students' university life.

Discussion

According to the study results, the first year is the appropriate time to identify students who are at a risk of dropout (Thomas, 2012; Vinson et al. 2010; Wilson et al., 2016), and the specific moment chosen for identification is a key issue in the general retention strategy. Institutions must achieve an equilibrium between the urgency and resources of their retention strategies. A very early identification of students who are at risk leaves room for numerous and more intensive interventions. However, as shown by our results, a slower pace in the measures taken prevents institutions from excluding certain students who are at risk from their interventions (Brooman & Darwent, 2014; Lowis & Castley, 2008; Sancho-Vinuesa, 2013). A slower pace also allows institutions to avoid increasing tutors' workload and decreasing the quality of their work, both stemming from dealing with an excessive number of students (McFarlane, 2016). Our study does not measure the extent to which the moment of the identification contributes to dropout prevention. However, we hypothesise that the sooner the students who are at risk are identified, the more efficient retention strategies can be, which has been suggested by previous studies (Herzog, 2006; Lowis & Castley, 2008; Vivian, 2005).

Our findings also reveal that academic results are appropriate data for dropout prediction (Aina, 2013; Lin, 2015; Lowis & Castley, 2008; Stoessel et al., 2015). Academic results have been shown to be better dropout predictors than sociodemographic data (Esteban-García et al., 2016). We provide evidence that accurate identification directed towards dropout prevention is based on students being academically assessed at the beginning of their university life. Institutions could consider their curricula content and assessment design tools to improve their formulation of withdrawal prevention strategies, which is consistent with the hypothesis that "[d]ata itself will not improve study success but enable targeted interventions" (Heublein et al., 2008 cited by Vossensteyn et al., 2015; Thomas, 2012).

The EDM approach used in this study confirms that large student databases are precious resources for decision-making in the search of HE quality (Dekker et al., 2009; Lin, 2015; Quadri & Kalyankar, 2010). The resulting classification trees in our study enable the decision maker to establish early warning systems to identify specific students who are at risk (Lin 2015) and refine these systems as student databases are updated. However, this methodological approach has the following two important limitations: first, when the dataset is relatively small, the results are inaccurate (Mercado, 2012), and

second, institutions must cope with extracting large amounts of data from very diverse sources and processing the data in logically designed datasets that merge all relevant information.

Our study adopted a student-centred action approach. Thus, the application of our models does not allow institutions to implement the retention strategies proposed by studies supporting a holistic approach (Esteban-García et al., 2016; Thomas, 2012; Tinto, 2006). Additionally, because our study was framed within the specific Spanish HE engineering context, without access to valuable data indicating the students' engagement level, we do not propose a "one size fits all" prediction model. Nevertheless, customised to local study programmes, contexts and resources, our methodology could be replicated. Students could be asked whether they wish to know about their withdrawal risk level and receive intervention action and how much credibility they would give such an academic strategy. A negative response to any – or all – of these questions could limit the applicability of our model, and then, the issue would fall into the field of universities' quality management, which could be a topic for further studies.

References

Adam, A. J., & Gaither, G. H. (2005). Retention in higher education: A selective resource guide. *New Directions for Institutional Research, 125*, 107-122.

Aina, C. (2013). Parental background and university dropout in Italy. *Higher Education*, 65, 437–456.

Arce, M. E., Crespo, B., & Míguez-Álvarez, C. (2015). Higher Education Drop-out in Spain–Particular Case of Universities in Galicia. *International Education Studies*, 8(5), 247–264.

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 63-69.

Berge, Z. L., & Huang, Y.-P. (2004). A Model for Sustainable Student Retention: A Holistic Perspective on the Student Dropout Problem with Special Attention to e-Learning. *Deosnews, 13*(5).

Bernold, L. E., Spurlin, J. E., & Anson, C. M. (2007). Understanding Our Students: A Longitudinal-Study of Success and Failure in Engineering With Implications for Increased Retention. *Journal of Engineering Education, 96*(3), 263-274.

Bacallao, J. & Bacallao, J. (2010). Imputación Múltiple en Variables Categóricas Usando Data Augmentation y Árboles de Decisión. *Investigación Operacional, 31(2)*, 133-139.

Bland, C. J., Taylor, A. L., Shollen, S. L., Weber-Main, A. M., & Mulcahy, P. A. (2009).

Faculty Success through Mentoring: A Guide for Mentors, Mentees, and Leaders. Plymouth, England: R&L Education.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Londres, Inglaterra: CRC press.

Brooman, S. & Darwent, S. (2014). Measuring the beginning: a quantitative study of the transition to higher education. *Studies in Higher Education*, *39*(9), 1523–1541.

Cabrera, L., Bethencourt, J. T., Pérez, P. Á., & Alfonso, M. G. (2006). El problema del abandono de los estudios universitarios. *Relieve*, *12*(2).

Cole, B., High, K., & Weinland, K. (2013). High School Pre-Engineering Programs: Do They Contribute To College Retention? *American Journal of Engineering Education (AJEE)*, *4*(1), 85–98.

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). *Predicting Students Drop Out: A Case Study*. *Educational Data Mining*, 41-50.

Engelbrecht, J., Harding, A., & Du Preez, J. (2007). Long-term retention of basic mathematical knowledge and skills with engineering students. *European Journal of Engineering Education*, *32*(6), 735–744.

Esteban-García, M.; Bernardo-Gutiérrez, Ana B.; Tuero-Herrero, E.; Cerezo-Menéndez, R. & Núñez-Pérez, J. C. (2016). El contexto sí importa: identificación de relaciones entre el abandono de titulación y variables contextuales. *European Journal of Education and Psychology*, *9*(2), 79-88.

Fantz, T. D., Siller, T. J., & DeMiranda, M. A. (2011). Pre-collegiate factors influencing the self-efficacy of engineering students. *Journal of Engineering Education*, *100*(3), 604.

Forsman, J., Van den Bogaard, M., Linder, C., & Fraser, D. (2015). Considering student retention as a complex system: a possible way forward for enhancing student retention. *European Journal of Engineering Education*, *40*(3), 235–255.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*, 131, 17-33.

Larose, S., Cyrenne, D., Garceau, O., Harvey, M., Guay, F., Godin, F., Deschenes, C. (2011). Academic Mentoring and Dropout Prevention for Students in Math, Science and Technology. *Mentoring & Tutoring: Partnership in Learning*, *19*(4), 419-439.

Lin, S-P. (2015). Using EDM for Developing EWS to Predict University Students Drop Out. *International Journal of Intelligent technologies and applied statistics*, *8*(4), 365-388.

Lowis, M., & Castley, A. (2008). Factors affecting student progression and achievement: prediction and intervention. A two-year study. *Innovations in Education and Teaching International*, *45*(4), 333-343.

McFarlane, K. J. (2016). Tutoring the tutors: Supporting effective personal tutoring. *Active Learning in Higher Education*, *17*(1), 77-88.

Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education*, *97*(1), 57–70.

Mercado, M. E. (2012). Las aplicaciones del análisis de segmentación. *Empiria. Revista de metodología de ciencias sociales*, *1*, 13–49.

Pal, S. (2012). Mining educational data to reduce dropout rates of engineering students. *International Journal of Information Engineering and Electronic Business*, *4*(2), 1.

Quadri, M. M. N., & Kalyankar, D. N. V. (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques. *Global Journal of Computer Science and Technology*, *10*(2), 2-5.

Raelin, J. A., Bailey, M. B., Hamann, J., Pendleton, L. K., Reisberg, R., & Whitman, D. L. (2014). The gendered effect of cooperative education, contextual support, and self-efficacy on undergraduate retention. *Journal of Engineering Education*, *103*(4), 599–624.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID Based Performance Prediction Model in Educational Data Mining. *International Journal of Computer Sciences Issues*, *7*(1), 10-18.

Sancho-Vinuesa, T., Escudero-Viladoms, N. & Masià, R. (2013). Continuous activity with immediate feedback: a good strategy to guarantee student engagement with the course. *Open Learning*, *28*(1), 51–66.

Stoessel, K.; Ihme, T. A.; Barbarino, M-L.; Fisseler, B. & Stürmer, S. (2015). Sociodemographic Diversity and Distance Education: Who Drops Out from Academic Programs and Why? *Research in Higher* Education, 56, 228–246.

Swail, W. S. (2004). The art of student retention: A handbook for practitioners and administrators. *Educational Policy Institute. Texas Higher Education Coordinating Board 20th Annual Recruitment and Retention Conference Austin, TX June* (Vol. 21, p. 2004).

Thomas, L. (2011). Do Pre-entry Interventions such as 'Aimhigher' Impact on Student Retention and Success? A Review of the Literature. *Higher Education Quaterly*, *65*(3), 230-250.

Thomas, L. (2012). Building student engagement and belonging in Higher Education at a time of change. *Paul Hamlyn Foundation*, *100*.

Thomas, L. & Hovdhaugen, E. (2014). Complexities and Challenges of Researching Student Completion and Non-completion of HE Programmes in Europe: a comparative analysis between England and Norway. *European Journal of Education*, *49*(4), 457-470.

Tinto, V. (2006). Research and practice of student retention: what next? *Journal of College Student Retention: Research, Theory and Practice*, *8*(1), 1–19.

Vandamme, J. P. Meskens, N., & Superby, J. -F. (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics*, *15*(4), 405-419.

Van den Bogaard, M. (2012). Explaining student success in engineering education at Delft University of Technology: a literature synthesis. *European Journal of Engineering Education*, *37*(1), 59–82.

Vinson, D., S. Nixon, B. Walsh, C. Walker, Mitchell, E. & Zaitseva, E. (2010). Investigating the relationship between student engagement and transition. *Active Learning in Higher Education*, *11*(2), 131–143.

Vivian, C. (2005). Advising the At-Risk College Student. *The Educational Forum*, *69*(4), 336-351.

Vossensteyn, H., Kottmann, A., Jongbloed, B., Kaiser, F., Cremonini, L., Stensaker, B., … Wollscheid, S. (2015). Dropout and completion in higher education in Europe: main report.

Wilson, K. L.; Murphy, K. A.; Pearson, A. G.; Wallace, B. M.; Reher V. G.S. & Buys, N. (2016). Understanding the early transition needs of diverse commencing university students in a health faculty: informing effective intervention practices. *Studies in Higher Education*, *41*(6), 1023-1040.

Wimshurst, K., Wortley, R., Bates, M., & Allard, T. (2006). The impact of institutional

factors on student academic results: Implications for 'quality' in universities. *Higher Education Research & Development*, *25*(02), 131–145.

Yasmin. (2013). Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. *Distance Education*, *34*(2), 218-231.

Yorke, M. (2016). The development and initial use of a survey of student 'belongingness', engagement and self-confidence in UK higher education. *Assessment & Evaluation in Higher Education*, *41*(1), 154-166.

Yorke, M. & Thomas, L. (2003). Improving the Retention of Students from Lower Socio-economic Groups. *Journal of Higher Education Policy and Management*, *25*(1), 63-74.

Zepke, N. & Leach, L. (2010). Improving student engagement: Ten proposals for action. *Active Learning in Higher Education*, *11*(3), 167–177.

Appendix 1. Variables

Variable	Definition	Available at
Gender	Students' gender	Induction week
InterItinerary	National *versus* international (study abroad) track	
AccessUniv	Pre-university type of studies	
SubAccessUniv	Track followed in Spanish national *Baccalauréat*	
WorkExpPrevYear	Worked during the prior year	
MarkUnivAccess	Admission test grade	
PassInterChemistry	Passes mid-term Chemistry test	After the first 6-7 academic weeks
PassInterPhysics	Passes mid-term Physics test	
PassInterCalculusI	Passes mid-term Calculus I test	
PassInterGE	Passes mid-term Graphic Expression test	
PassInterIT	Passes mid-term Computing test	
PasInterAlg_Geom	Passes mid-term Algebra and Geometry tests	
PassChemistry	Passes Chemistry course	End of the first semester
PassPhysics	Passes Physics course	
PassCalculusI	Passes Calculus I course	
PassIT	Passes Computing course	
PasGE	Passes Graphic Expression course	
PassIT_GE	Passes either Computing or Graphic Expression courses (depending on which was taken during the	

	first semester).	

Note 1: Several additional sociodemographic variables were considered but not included in the final analysis because the preliminary analysis showed no evidence of a sufficient association level. These variables were "parents' educational background", "high-school of origin", and "state-owned *versus* private school of origin".