Early Dropout Predictors in Social Sciences and Management Degree students

José M. Ortiz-Lozano, Pilar Aparicio-Chueca, Xavier M. Triadó-Ivern & Jose Luis Arroyo-Barrigüete

Student dropout is a major concern in studies investigating retention strategies in higher education. This study identifies which variables are important to predict student dropout, using academic data from 3,583 first-year students on the Business Administration (BA) degree at the University of Barcelona (Spain). The results indicate that two variables, the percentage of subjects failed and not attended in the first semester, demonstrate significant predictive power. This has been corroborated with an additional sample of 10,784 students from three-degree programs (Law, BA, and Economics) at the Complutense University of Madrid (Spain), to assess the robustness of the results. Three different algorithms have also been utilized: neural networks, random forest, and logit. In the specific case of neural networks, the NeuralSens methodology has been employed, which is based on the use of sensitivities, allowing for its interpretation. The outcomes are highly consistent in all cases: both a simple model (logit) and more sophisticated ones (neural networks and random forest) exhibit high accuracy (correctly predicted values) and sensitivity (correctly predicted dropouts). In test set average values of 77% and 69% have been respectively achieved. In this regard, a noteworthy point is that only academic data from the university itself was used to develop the models. This ensures that there's no dependence on other personal or organizational variables, which can often be difficult to access.

Keywords: Prediction, university dropout, educational data mining, academic performance, neural networks

1. Introduction

University studies have become essential for societal and personal development, contributing to economic progress and civic responsibility (Kuh et al, 2008; Musso et al, 2020). However, students often encounter various challenges and crises throughout their academic journey, including the recent COVID-19 pandemic, which brought about sudden shifts to online education, student isolation, and increased screen dependency. As a result, some students may change their study programs, experience delays in

degree completion, or ultimately drop out of their studies (Su & Guo, 2021; Sarfraz et al., 2022; Teodorescu & Amalfi, 2022).

Understanding the determinants of academic success in university education has been a significant area of research (García-Ros & Pérez-González, 2009; Terenzini et al., 1981; Tinto, 1993). In light and post COVID-19 habits, such studies have gained even greater relevance. Educational institutions and faculty have a responsibility to support students, especially those in need of guidance and assistance. By continuously monitoring academic performance using predefined criteria, institutions can identify students who are at risk of dropping out and implement early interventions (Kara et al., 2019; Musso et al., 2020; Ortiz-Lozano et al., 2020; Segovia et al., 2020).

Previous research has shown that early identification of at-risk students is crucial for implementing effective measures such as counselling, mentoring, and tutoring, which significantly prevents dropout rates (Cabrera et al., 2006; Herzog, 2004; Lowis & Castley, 2008; Soto & Amores, 2020). Focusing on early intervention, educators aim to address the challenges faced by freshmen (Thomas, 2012; Vinson et al., 2010; Wilson et al., 2016), minimizing the duration and impact of crises that may lead to dropout (Pérez et al., 2018).

The issue of student dropout is not limited to a specific region and has been studied globally. In Spain, for example, the Conference of Rectors of Spanish Universities (CRUE) has conducted comprehensive research on university dropout rates. These reports highlight the substantial percentage of students who either drop out or extend their studies beyond the expected duration, with a rate of 14.5% in public face-to-face universities. They emphasize the need for policies and initiatives to reduce dropout rates as a priority objective (Constante-Amores et al., 2021; Lassibille & Navarro, 2007, 2009). However, institutions must balance the urgency of their retention

strategies with the resources available (Brooman & Darwent, 2014; Lowis & Castley, 2008).

While various variables influence the decision to drop out, not all have the same impact. This study focuses on the data collected up to 2021. The study aims to identify potential predictive indicators of student engagement or attrition at the university level. By examining these variables, the goal is to establish an "early warning system" that can identify students at risk of dropping out and enable the implementation of supportive interventions.

This study recognizes the importance of university education and the challenges students face. By identifying early dropout predictors and implementing timely interventions, educational institutions can effectively support students and reduce dropout rates, ensuring their academic success and contributing to societal advancement.

2. Factors affecting student dropout.

The field of understanding student dropout from university began with Tinto's model (Tinto, 1975), which identified various factors that influence academic dropout, including demographic, cultural, social, family, socioeconomic, psychological, and academic progress variables. Bharadawaj and Pal (2012) found that factors such as upper secondary school exam results, family income, place of residence, and family status significantly influence student dropout. Other studies have utilized enrolment data (Kovacic, 2010), maternal educational level with family income (Devasia et al., 2016), and course grade performance to predict dropout. While numerous variables have been studied, academic performance consistently emerges as a strong predictor of dropout (Casanova et al.,2018).

The variables involved in a dropout decision can be categorized into three groups: personal, organizational, and goal achievement variables. Personal variables encompass characteristics specific to each student, such as gender, ethnic background, place of residence, ease of access to university studies, year of study (particularly freshmen), early marriage responsibilities, financial concerns, and the perceived loss of individual autonomy due to a structured schedule (Aparicio-Chueca et al., 2021; Ashour, 2020; McGhie, 2017; Tentsho et al., 2019; Triadó et al., 2020). Lassibille & Navarro (2007) confirmed that factors such as gender (only in technical schools), age of enrolment, type of pre-university education, type of financial support, paternal level of education, and not residing in the university city can impact attrition. Some of these factors appear to be stronger than others, but there is no consistent agreement across different studies (Hergoz, 2006; Lassibille & Navarro, 2007; Musso et al., 2020).

Organizational variables pertain to personal plans, such as poorly planned coursework, perception of exams as overly challenging, inadequate class scheduling, coursework beginning at an advanced level (Salas-Morera et al., 2019) or well-paid job opportunities (Ashour, 2020).

Academic literature consistently links goal achievement variables to academic success (Aina, 2013; Demeter et al., 2022; Lowis & Castley, 2008; Schmitt et al., 2020; Stoessel et al., 2015; Von Hippel & Hofflinger, 2021). Interestingly, these factors are more potent predictors of dropout than sociodemographic data, challenging the conventional view of sociodemographics as a reliable dropout indicator (Esteban-García et al., 2016; Ortiz-Lozano et al., 2020). Factors such as poor pre-college preparation (Tentsho et al., 2019; Toomsalu-Stefanova & Fokina, 2020), low first-semester grade point average, poor performance in specific subjects, and failing subjects at the end of the first year can contribute to dropout (McGhie, 2017).

Identifying at-risk students early on is crucial, with the first year of study being the optimal time for detection. Early identification allows for timely intervention and support, reducing dropout rates (Herzog, 2006; Lowis & Castley, 2008; Vivian, 2005). Maintaining consistent contact with students and providing tutorials (Al-Shabandar et al., 2017; Jacobsen, 2019) and engagement opportunities are effective strategies to mitigate dropout (Thomas, 2012; Vinson et al., 2010; Wilson et al., 2016, Chemers et al., 2001; Horstmanshof & Zimitat, 2007; Kovacic, 2010; Strayhorn, 2009). By focusing on predictive factors immediately after the end of the first semester, this research aims to confirm the hypothesis that academic results during the first semester have the highest predictive capacity for dropout compared to personal and organizational variables.

The objective of this research is twofold. Firstly, it aims to confirm the following research hypothesis: Out of the three broad categories of variables previously described (personal, organizational, and goal achievement), it is the latter, specifically academic results during the first semester, that has the most predictive capacity for dropout. The second objective, if the hypothesis is confirmed, is to develop a predictive model of dropout based on readily available academic performance data. The research aims to provide universities with a practical tool for identifying at-risk students and implementing timely interventions to reduce dropout rates.

The reason is practical. In the Spanish university system, it is difficult to obtain certain information about the student. For example, according to the literature, the education of the parents seems to be a relevant variable in predicting both academic failure and dropout (Battin-Pearson et al., 2000; Englund et al., 2008; Pritchard & Wilson, 2003; Opazo et al., 2021; Spady, 1970). However, this variable is not available in many universities, or if it is, it has low reliability, as it is based on voluntary surveys

that few students complete, and those who do may provide false information. Therefore, for practical purposes, it would be desirable to have a predictive model that does not require this variable. The question is to what extent it is possible to develop a sufficiently accurate predictive model for dropout using only variables that any university has available with absolute certainty: the student's academic performance.

With these two objectives, the paper is structured as follows. Firstly, the methodology is described, which consists of three different phases. Subsequently, the results are presented and compared with those obtained in previous research. Finally, the conclusions of the study and future lines of research are exposed.

3. Materials and Methods

The methodology comprises three main parts. First, a predictive model of dropout is developed using data from 3,583 first-year students in the Business Administration (BA) degree at the University of Barcelona. Second, an interpretable neural network methodology (Pizarroso et al., 2022) is employed to evaluate the variable relevance. Finally, a simplified predictive model based solely on academic performance variables is proposed and tested on a total of 10,784 students from three different degree programs (Law, BA, and Economics) at the Complutense University of Madrid. The R programming environment (R Core Team, 2013) and various packages were used for data management and statistical analysis: caret (Kuhn, 2020), dplyr (Wickham, François, Henry and Müller, 2020), ggplot2 (Wickham, 2016), NeuralSens (Pizarroso et al., 2022) and pROC (Robin et al., 2011) packages.

Phase 1: dropout predictive model (University of Barcelona)

The first phase of the analysis focuses on first-year students in the Business Administration (BA) degree at the University of Barcelona. This context is significant

due to the diverse professional opportunities offered by the BA degree, attracting students without a clear vocation but requiring specific skills and knowledge, which contribute to higher dropout rates (Sosu & Pheunpha, 2019; Asian-Chaves et al., 2021). Furthermore, reducing dropout rates is a priority for universities to address unemployment and reshape the economic model (Arce et al., 2015).

The dropout of first-year students was used as the dependent variable. Sociodemographic and academic performance variables were used as predictors of this behaviour. The sociodemographic variables included: sex, parental educational level, parental occupations, student's employment status, and academic origin. The variable parental educational level was categorized into two distinct groups: without university education (non-graduate) and with university education (base level in the model). Regarding parental occupations, we distinguished four categories: Unemployed, Non-qualified employment, Qualified employment, and University-qualified manager or technician (base level in the model). In terms of student employment, three categories were defined: no job, employment involving less than 15 hours per week, and employment involving more than 15 hours per week (base level in the model). Finally, the academic origin was divided into 5 categories: traditional access from high school (University Access), transfer from another university degree which the student entered from high school (Transfer UA), transfer from another university degree to which the student entered from higher vocational training (Transfer HVT), access from higher vocational training (base level in the model), and access for individuals over 25 years old (Over 25).

Regarding academic performance, pre-university variables such as university access mark and university admission mark are considered. Additionally, performance

in the first semester is measured by the percentage of failed subjects, not shown rate (subjects not taken), and average grade in the subjects passed.

After data cleaning, a sample of 3,583 first-year students in the BA degree at the University of Barcelona, spanning eight academic years, was analyzed. The data collection period ranged from 2010/2011 to 2017/2018, excluding students pursuing dual degrees. Overall, this phase of the analysis focuses on understanding the characteristics and performance of first-year BA students at the University of Barcelona to identify potential predictors of dropout.

[Table 1]

Three distinct models were utilized to compare their performance: Neural Networks (NN), Random Forest, and Logistic Regression (Logit). NN are computational models inspired by the way biological brains process information. They are composed of layers containing interconnected nodes, often named "neurons", that transform input data into an output. NN excel at modeling complex non-linear relationships and can automatically learn features from data. They offer adaptability and performance advantages, especially with intricate patterns, compared to conventional econometric models. On the other hand, Random Forest is a learning method that combines multiple decision trees to produce a more accurate and stable prediction, enhancing generalization. Compared to conventional econometric models, Random Forest effectively handles non-linear relationships and interactions between variables and is less sensitive to outliers. Both Random Forest and NN are models that generally exhibit a good performance across a wide variety of problems, and this was the main reason for their selection. However, we believed it was relevant to contrast them with a logit model, given its simplicity and ease of use. For the NN model, hyperparameters were chosen using a grid search method, evaluating one or two neurons in the hidden

layer and decay rates between 10^{-7} and 1. In the case of the Random Forest model, 100 trees were used, with the minimum node size varying among 2, 5, and 10. The number of variables sampled at each split was kept at 1, and the Gini impurity was utilized as the criterion to determine where to split the nodes. For the Logit model, an Elastic Net penalized logistic regression model was utilized, combining both L1 and L2 penalties. In all cases, the optimal cut-off value (threshold) was determined automatically based on the Receiver Operating Characteristic (ROC) curve. To avoid overfitting, in all models a 10-fold cross-validation was employed, and the train/test split was set at 80/20.

Phase 2: analysis of variables relevance (University of Barcelona)

In Phase 2, the methodology proposed by Pizarroso et al. (2022), NeuralSens, was used to evaluate the importance of the variables. It is a methodology that, initially conceived in the field of engineering, has already been successfully applied to studies on education (see Arroyo-Barrigüete et al, 2023). NeuralSens calculates the sensitivity of the dependent variable to each independent variable. For each observational unit, a sensitivity is estimated. Therefore, for the i-th variable, we do not obtain a single sensitivity (as in an OLS model) but a distribution of them. If the distribution is narrow (standard deviation of sensitivities close to zero) and its mean is also close to zero, the variable is considered irrelevant. To avoid biases due to the choice of a specific seed, the NN was fitted 50 times using bootstrap samples, obtaining the sensitivity distribution of each variable based on all of them (179,150 sensitivities for each variable). Based on this distribution, the mean and standard deviation were calculated following the scheme proposed by Pizarroso et al. (2022). The importance of each variable was assessed using the metric proposed by the authors of NeuralSens, the mean of squared sensitivities.

Phase 3: simplified dropout predictive model

In the final stage, models were adjusted with the three most relevant variables according to the previous phase. Similar to Phase 1, three models were compared: NN, Random Forest, and Logistic Regression. The hyperparameter selection procedure was conducted identically to the process described in Phase 1. A 10-fold cross-validation was also performed, and the train/test split was set at 80/20. Again, to avoid biases due to a specific choice of seeds, the models were fitted ten times, resulting in mean values for accuracy, sensitivity, and specificity. This process ensured the robustness of the results and minimized the impact of random variation in individual model fits.

A critical aspect to consider is evaluating to what extent the results obtained are local in nature, that is, they are the consequence of the peculiarities of the degree studied, the specific university, or the years from which the sample was collected. For this reason, at this stage, we have chosen to evaluate the performance of the model with other datasets, specifically data from 10,784 students from three different degree programs (Law, BA, and Economics), from different cohorts, and from a different university (see Table 2). If the conclusions reached are robust, the model should also demonstrate good predictive capacity in these new datasets.

[Table 2]

4. Results and Discussion

Phase 1: dropout predictive model (University of Barcelona)

Table 3 shows the training and testing results of the model with all variables, using data from the 3,583 first-year students of BA at the University of Barcelona. This table includes three metrics: accuracy (which indicates the percentage of instances correctly

predicted), sensitivity (or recall, which indicates the proportion of actual positives correctly identified as such), and specificity (the proportion of actual negatives correctly identified as such). The metric of most interest is sensitivity, as it indicates the percentage of students correctly identified by the model as at risk of dropping out. However, it is also necessary to maintain high levels of accuracy, as a model with a high percentage of false positives would not be useful. It can be observed that all three models (NN, logit, and random forest) provide good results both in training and test, achieving performance equal to or better than that achieved in previous studies in Spain, such as those by Ortiz-Lozano et al. (2020), Fernández-García et al. (2021), and Segura et al. (2022). Additionally, through an analysis of variable importance, we found that even though there are some differences between the models, Not shown rate and Fail rate are identified as critical by all of them.

[Table 3]

Phase 2: analysis of variables relevance (University of Barcelona)

Table 4 presents the sensitivity mean, the standard deviation, and the mean squared sensitivity of each of the variables included in the model. As stated in the Materials and Methods section, these values indicate the type of relationship between input and output. The results are clear: Not shown rate and Fail rate are by far the two most relevant variables. Other predictors do have an impact, but it is substantially lesser. For instance, a lower likelihood of dropout is observed among women, and among students who access the studies through a transfer from another degree (whether it's an HVT transfer or a UA transfer) and with decreasing access and admission grades. A potential explanation for this latter relationship is that a student with a lower admission grade may have fewer opportunities in other programs. Consequently, dropping out from his

or her current degree could signify a definitive withdrawal from higher education, given the lack of alternative options. Holding a paid job for more than 15 hours a week increases the likelihood of dropout compared to students without a job or with a job of less than 15 hours, and the risk of dropout decreases as the average grade in passed subjects increases. Parents' occupation also seems to have a certain impact, negative in some cases and positive in others. However, as already indicated, these are much smaller effects than in the case of the two critical variables. Particularly interesting is the finding that the parents' educational level appears to have a low (in the case of the mother) or null (in the case of the father) impact, a result that coincides with that of Constante-Amores et al. (2021), and contradicts the findings of Contini et al. (2018), that pointed out that students with university-educated parents are more likely to continue in higher education.

[Table 4]

These results confirm that neither sociodemographic variables, nor variables external to the student, are key discriminant predictor variables of first-year dropout. Indeed, these variables are important, and their inclusion can enhance the predictive capacity of the model, but the improvement is, at best, minimal. This is a result of great interest because these variables, such as the education or occupation of the parents, are often difficult to obtain. The fact that they don’t seem indispensable for developing a predictive model with good precision (as will be shown in the next section) facilitates the use of this type of models. In contrast, the results of this study reveal that academic results are appropriate data for predicting dropout, as stated by Aina (2013), Lin (2015), Lowis & Castley (2008), Stoessel et al. (2015), Schmitt et al. (2020) and Von Hippel, & Hofflinger (2021). So, our first hypothesis is confirmed: out of the three broad categories of variables (personal, organizational, and goal achievement), it is the latter,

specifically academic results during the first semester, that has the most predictive capacity for dropout. Indeed, the models confirm that not only does passing a subject have a substantial impact, but the act of appearing to exams has an even more significant influence: the percentage of subjects not attempted (not shown rate) is the most important variable in all the models.

Phase 3: simplified dropout predictive model

In the third phase, a simplified model was developed with the aim of evaluating its predictive capacity when using only a limited number of variables that are readily available in any university. Based solely on the relative importance in the model outlined in the previous section (see Table 4), it would seem advisable to include the not shown rate and fail rate, followed by other variables. However, adhering to the criterion of simplicity (avoiding variables that may not be easily available in any university), a model with only three variables was chosen: not shown rate, fail rate, and access mark. The results (Table 5) show good performance across the five datasets, with average accuracy and sensitivity of 79% and 74% respectively in training, and 77% and 69% in testing. Indeed, this is a remarkable finding: the results for the BA degree at the University of Barcelona with the simplified model are very similar to those obtained with the complete model (as seen in table 3).

The results shown in table 5 confirm the high predictive power of the three selected variables, regardless of the specific characteristics of the degree, the year, the university, or the predictive model used. It underscores their relevance in predicting student dropout and demonstrates that a simplified model using just these variables can be almost as effective as a more complex one that includes additional factors. It's an important insight because it suggests that universities can effectively anticipate dropout risks using readily available data on student academic performance. The recent

systematic review by Cardona et al. (2023) reveals that there is an immense quantity of factors that can be used to predict dropout. However, our results suggest that it is feasible to obtain good predictions with only three variables, at least in the Spanish university context and for degrees in social sciences.

[Table 5]

Figures 1 to 3 present the testing results for each model across the various datasets. In all cases, it is observed that the best performance is achieved for the BA degree at the University of Barcelona and the Economics degree at the Complutense University of Madrid. The least effective performance is seen for the Law degree (Courses 2015/16 to 2020/21) at the Complutense University of Madrid.

[Figure 1]

[Figure 2]

[Figure 3]

In this regard, it is necessary to contextualize the results obtained, as they are unusually high. Indeed, there are studies that have achieved similar accuracy levels in the Spanish university context and using academic data from the first semester. This is the case, for example, of Ortiz-Lozano et al. (2020), who achieved 76% accuracy with a classification tree model. Another research by Fernández-García et al. (2021) reached an accuracy of 80.8% (with a sensitivity of 82.2% and specificity of 79.4%) using an ensemble of Gradient Boosting, Random Forest, and Support Vector Machine models. For comparison purposes, these studies achieve accuracies similar to those obtained with the dataset from the University of Barcelona. However, two very important nuances must be highlighted here.

First, both studies are based on data from an engineering school, not from the social sciences, and the sample size is substantially smaller than that used in this study.

In this sense, for comparison purposes, a much more realistic benchmark would be the work of Segura et al. (2022), whose best model for the field of social sciences and law, achieves a sensitivity of 39.05%. In other words, the model proposed in this study significantly improves the predictions obtained in that research, which among the three mentioned, is the most comparable since it considers the same area of knowledge.

The second nuance is that all three of the mentioned studies use many more variables[1]. In fact, this is the most notable result obtained: achieving such high levels of accuracy with only three variables. This is not a minor issue, as noted in the introduction, but it is a matter of enormous practical importance. In the Spanish university system, it is challenging to obtain certain student information. Therefore, it would be desirable to have a predictive model that requires the fewest variables possible and ideally includes exclusively predictors related to information that every university has available with absolute certainty: the student's academic performance. The results show that it seems feasible to achieve this goal: the three variables identified in Phase 2, which correspond exclusively to the student's academic performance, appear to have a high predictive capacity. Regardless of the model considered, degree, year or university, high levels of accuracy, specificity, and sensitivity are obtained.

Other studies, such as that of Lizarte Simón and Gijón Puerta (2022), in this case using a sample of students from Early Childhood, Primary, and Social Education and Pedagogy degree programs, achieve an accuracy of 91%, using predictors derived from a survey that evaluates various academic dimensions. This means, once again, the model requires access to a series of variables that are challenging to obtain.

5. Conclusion

This study adopted a student's performance-centered action approach. Spain was chosen as an interesting context for this analysis because higher education has emerged as a field of study in the country for more than two decades and, also, due to the greater availability of data sets.

The objective of this research was twofold. Firstly, it aimed to confirm the hypothesis that out of the three broad categories of variables frequently used as predictors of dropout - personal, organizational, and goal achievement - it is the latter, specifically academic results during the first semester, that holds the most predictive capacity for dropout. Using a dataset of 3,583 first-year students on the BA degree at the University of Barcelona, this hypothesis has been confirmed.

The second objective was to develop a simple predictive model of dropout based on easily accessible variables, that is, academic performance variables. For this purpose, five datasets were used, corresponding to the degrees of BA, Economics, and Law across different years and universities, which included a total of 14,367 students. Considering only three variables (not shown rate, fail rate, and access mark), a good performance was achieved across the five datasets, with average accuracy and sensitivity of 79% and 74% respectively in training, and 77% and 69% in test. This result allows universities to develop retention strategies for first-year students. By identifying the academic performance variables that affect dropout, university stakeholders can define strategies to minimize dropout among students at risk.

In summary, the main conclusion of this study is that it is possible to obtain predictive models at the end of the first semester, with good performance and using only three performance variables easily accessible to any university.

However, this research has a significant limitation: in our opinion, the results are confined to the Spanish university context and to degrees within the field of social

sciences. Given that we have considered two different universities and three distinct degrees over several years, we believe that the results are generalizable to other degrees within this field. However, it's possible that in other areas, such as STEM studies (Science, Technology, Engineering, and Mathematics studies), the predictive capacity of the three identified variables may not be as high. In this regard, we consider it a future line of research to verify if these results are applicable to other types of degrees (i.e., STEM and arts degrees) and to expand the sample to include other universities.

References

Aina, C. 2013. "Parental background and university dropout in Italy". *Higher Education*, 65, 437-456. https://doi.org/10.1007/s10734-012-9554-z

Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., & Lunn, J. 2017. "Towards the differentiation of initial and final retention in massive open online courses". In *International Conference on Intelligent Computing* (pp. 26-36). Springer, Cham. http://dx.doi.org/10.1007/978-3-319-63309-1_3

Aparicio-Chueca, P., Domínguez-Amorós, M. & Maestro-Yarza, I. 2021. "Beyond university dropout. An approach to university transfer", *Studies in Higher Education*, http://dx.doi.org/10.1080/03075079.2019.1640671

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, 247–264. http://dx.doi.org/10.5539/ies.v8n5p247

Arroyo-Barrigüete, J. L., Carabias-López, S., Borrás-Pala, F., & Martín-Antón, G. 2023. "Gender Differences in Mathematics Achievement: The Case of a Business School

in Spain", *SAGE Open*, 13 (2), 21582440231166922. https://doi.org/10.1177/21582440231166922

Ashour, S. 2020. "Analysis of the attrition phenomenon through the lens of university dropouts in the United Arab Emirates", *Journal of applied research in higher education*, 12 (2), 357-374. https://doi.org/10.1108/JARHE-05-2019-0110

Asian-Chaves, R., Buitrago, E. M., Masero-Moreno, I., & Yñiguez, R. 2021. "Advanced mathematics: An advantage for business and management administration students". *The International Journal of Management Education*, *19*(2), 100498. https://doi.org/10.1016/j.ijme.2021.100498

Battin-Pearson, S., Newcomb, M. D., Abbott, R. D., Hill, K. G., Catalano, R. F., & Hawkins, J. D. (2000). Predictors of early high school dropout: A test of five theories. *Journal of educational psychology*, 92(3), 568

Bhardwaj, B. K., & Pal, S. 2012. "Data Mining: A prediction for performance improvement using classification". *arXiv preprint arXiv:1201.3418*. https://doi.org/10.48550/arXiv.1201.3418

Brooman, S., & Darwent, S. 2014. "Measuring the beginning: A quantitative study of the transition to higher education". *Studies in Higher Education*,39, 1523–1541. https://doi.org/10.1080/03075079.2013.801428

Cabrera, L., Bethencourt, J. T., Pérez, P. Á., & Alfonso, M. G. 2006. "El problema del abandono de los estudios universitarios". *Relieve*,12, 171–203.

Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J. 2023. "Data mining and machine learning retention models in higher education". *Journal of College Student Retention: Research, Theory & Practice, 25*(1), 51-75.

Casanova, J. R., Cervero Fernández-Castañón, A., Núñez Pérez, J. C., Almeida, L. S., & Bernardo Gutiérrez, A. B. 2018. "Factors that determine the persistence and dropout of university students". *Psicothema, 30*. http://dx.doi.org/10.7334/psicothema2018.155

Chemers, M. M., Hu, L. T., & Garcia, B. F. 2001. "Academic self-efficacy and first year college student performance and adjustment". *Journal of Educational psychology*, *93*(1), 55. DOI: 10.1037//0022-0663.93.1.55

Constante-Amores, A., Martínez, E. F., Asencio, E. N., & Fernández-Mellizo, M. 2021. "Factores asociados al abandono universitario." *Educación XX1*, *24*(1), 17-44. https://doi.org/10.5944/educxx1.26889

Contini, D., Cugnata, F., & Scagni, A. 2018. "Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy". *Higher Education, 75*(5), 785-808. https://doi.org/10.1007/s10734-017-0170-9

CRUE, 2019, https://www.crue.org/wp-content/uploads/2020/02/UEC-1718_FINAL_DIGITAL.pdf)

Demeter, E., Dorodchi, M., Al-Hossami, E., Benedict, A., Slattery Walker, L., & Smail, J. 2022. "Predicting first-time-in-college students' degree completion outcomes". *Higher Education*, 1-21. https://doi.org/10.1007/s10734-021-00790-9

Devasia, T., Vinushree, T. P., & Hegde, V. 2016. *Prediction of students' performance using Educational Data Mining*. In International Conference on Data Mining and Advanced Computing (SAPIENCE). IEEE

Englund, M. M., Egeland, B., & Collins, W. A. (2008). Exceptions to high school dropout predictions in a low-income sample: Do adults make a difference? *Journal of social issues*, 64(1), 77-94.

Esteban-García, M., Bernardo Gutiérrez, A. B., & Rodríguez-Muñiz, L. J. 2016. "Persistence in university studies: The importance of a good start". *Aula Abierta*, 44(1), 1–6. https://doi.org/10.1016/j.aula.2015.04.001

Fernández-García, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sánchez-Figueroa, F. 2021. "A real-life machine learning experience for predicting university dropout at different stages using academic data". *IEEE Access, 9*, 133076-133090.

García-Ros, R. & Pérez-González, F. 2009. "Una aplicación web para la identificación de sujetos de nuevo acceso a la universidad en situación de riesgo académico". *@tic. Revista d'innovació Educativa*, 2, 11-17.

Herzog, S. 2004. *Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen.* Procceding of 44th Annual Forum of the Association for Institutional Research (AIR).

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. http://dx.doi.org/10.1002/ir.185

Horstmanshof, L., & Zimitat, C. 2007. "Future time orientation predicts academic engagement among first-year university students". *British Journal of Educational Psychology*, 77(3), 703–718. https://doi.org/10.1348/000709906X160778

Jacobsen, D. Y. 2019. "Dropping out or dropping in? A connectivism approach to understanding participants' strategies in an e-learning MOOC pilot". *Technology, Knowledge and Learning*, *24*(1), 1-21. http://dx.doi.org/10.1007/s10758-017-9298-z

Kara, M., Erdoğdu, F., Kokoç, M., & Cagiltay, K. 2019. “Challenges faced by adult learners in online distance education: A literature review”. *Open Praxis*, *11*(1), 5-22. https://doi.org/10.5944/openpraxis.11.1.929

Kovacic, Z. 2010. *Early prediction of student success: Mining students' enrolment data*.

Kuh, G.D., Cruce, T.M., Shoup, R., Kinzie, J., & Gonyea, R.M. 2008. “Unmasking the Effects of Student Engagement on First-Year College Grades and Persistence”. *The Journal of Higher Education 79*(5), 540-563. http://dx.doi.org/10.1353/jhe.0.0019

Kuhn, M. 2020. *caret: Classification and Regression Training*. R package version 6.0-86. https://CRAN.R-project.org/package=caret

Lassibille, G., Navarro Gomez, L. 2007. “Why do higher education students drop out? Evidence from Spain”. *Education Economics* 16(1), p. 89-105. https://doi.org/10.1080/09645290701523267

Lassibille, G., Navarro Gómez, L., 2009. “Tracking students’ progress through the Spanish university school sector”. *Higher Education,* 58**,** 82. https://doi.org/10.1007/s10734-009-9227-8

Lin, S.P. 2015. “Using EDM for developing EWS to predict university students drop out”. *International Journal of Intelligent Technologies and Applied Statistics*,8, 365–388. https://doi.org/10.6148/IJITAS.2015.0804.04

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, 333–343. https://doi.org/10.1080/14703290802377232

McGhie, V. 2017. “Entering university studies: identifying enabling factors for a successful transition from school to university”. *Higher Education*, *73*(3), 407-422. https://doi.org/10.1007/s10734-016-0100-2

Musso, M.F., Hernández, C.F.R. & Cascallar, E.C. 2020. “Predicting key educational outcomes in academic trajectories: a machine-learning approach”. *Higher Education* 80**,** 875–894 (2020). https://doi.org/10.1007/s10734-020-00520-7

Opazo, D., Moreno, S., Álvarez-Miranda, E., & Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. *Mathematics*, 9(20), 2599.

Ortiz-Lozano, J.M., Rua-Vieites, A., Bilbao-Calabuig, P., & Casadesús-Fa, M. 2020. “University student retention: Best time and data to identify undergraduate students at risk of dropout”, *Innovations in Education and Teaching International*, 57:1, 74-85. https://10.1080/14703297.2018.1502090

Pérez, B., Castellanos, C., & Correal, D. 2018. “Predicting student drop-out rates using data mining techniques: A case study”. In *IEEE Colombian Conference on Applications in Computational Intelligence* (pp. 111-125). Springer, Cham.

Pizarroso, J., Portela, J., & Muñoz, A. 2022. “NeuralSens: sensitivity analysis of neural networks”. *Journal of Statistical Software, 102*(7), 1-36. https://doi.org/10.18637/jss.v102.i07

Pritchard, M. E., & Wilson, G. S. (2003). Using emotional and social factors to predict student success. *Journal of college student development*, 44(1), 18-28.

R Core Team 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. 2011. "pROC: an open-source package for R and S+ to analyze and compare ROC curves". *BMC Bioinformatics*, 12, 77. https://doi.org/10.1186/1471-2105-12-77

Salas-Morera, L., Molina, A. C., Olmedilla, J. L. O., García-Hernández, L., & Palomo-Romero, J. M. 2019. "Factors affecting engineering student's dropout: A case study". *The International journal of engineering education*, *35*(1), 156-167.

Sarfraz, M., Khawaja, K. F., & Ivascu, L. (2022). Factors affecting business school students' performance during the COVID-19 pandemic: A moderated and mediated model. *The International Journal of Management Education*, 20(2), 100630.

Schmitt, J., Fini, M. I., Bailer, C., Fritsch, R., & de Andrade, D. F. 2020. "WWH-dropout scale: when, why and how to measure propensity to drop out of undergraduate courses". *Journal of Applied Research in Higher Education*. 13 (2), 540-560. https://doi.org/10.1108/JARHE-01-2020-0019

Segovia, N., Orellana, D., & Rincón, A. G. 2020. "La evaluación del rendimiento del aprendizaje como elemento de la predicción del abandono en programas de educación superior virtual". *Redes de Investigación e Innovación en Docencia Universitaria, 2020*, 605-614

Segura, M., Mello, J., & Hernández, A. 2022. "Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?". *Mathematics, 10*(18), 3359. https://doi.org/10.3390/math10183359

Sosu, E. M., & Pheunpha, P. 2019. "Trajectory of university dropout: investigating the cumulative effect of academic vulnerability and proximity to family support." *Frontiers in Education*, 4. https://doi.org/10.3389/feduc.2019.00006

Soto, M. F. M., & Amores, I. A. C. 2020. “Determinantes del rendimiento académico de los estudiantes de nuevo acceso a la Universidad Complutense de Madrid”. *Revista de educación*, (387), 213-240. http://dx.doi.org/10.4438/1988-592X-RE-2020-387-433.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1), 64-85.

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. http://doi.org/10.1007/s11162-014-9343-x

Strayhorn, T. L. 2009. “Different folks, different hopes: The educational aspirations of Black males in urban, suburban, and rural high schools”. *Urban Education*, 44(6), 710-731. https://doi.org/10.1177/0042085908322705

Su, C. Y., & Guo, Y. (2021). Factors impacting university students' online learning experiences during the COVID-19 epidemic. *Journal of computer assisted learning,* 37(6), 1578-1590.

Tentsho, K., McNeil, N., & Tongkumchum, P. 2019. “Examining timely graduation rates of undergraduate students”. *Journal of Applied Research in Higher Education*, 11(2), 199-209. 10.1108/JARHE-10-2017-0124

Teodorescu, D., Aivaz, K. A., & Amalfi, A. (2022). Factors affecting motivation in online courses during the COVID-19 pandemic: the experiences of students at a Romanian public university. *European Journal of Higher Education*, 12(3), 332-349.

Terenzini, P. T., Lorang, W. G., Pascarella, E. T. 1981. "Predicting freshman persistence and voluntary dropout decisions: a replication". *Research in Higher Education* 15(2), 109-127. https://doi.org/10.1007/BF00979592

Thomas, L. 2012. "Building student engagement and belonging in Higher Education at a time of change". *Paul Hamlyn Foundation*, *100*, 1-99.

Tinto, V. 1975. "Dropout from higher education: A theoretical synthesis of recent research". *Review of educational research*, *45*(1), 89-125.

Tinto, V. 1993. *Learning college: Rethinking the causes and cures of student attrition*. (2nd ed.) Chicago: University of Chicago Press.

Toomsalu-Stefanova, L. M., & Fokina, E. N. 2020. "Study of the Quality of Applicants' Admission to Universities Based on the Results of the Unified State Exam in Russia". *International journal of instruction*, *13*(2), 73-88. https://doi.org/10.29333/iji.2020.1326a

Triadó-Ivern, X. Aparicio-Chueca, P. Elasri-Ejjaberi, A. Maestro-Yarza, I. Bernardo M. & Presas Maynegre, P. 2020. "A factorial structure of university absenteeism in higher education: A student perspective". *Innovations in Education and Teaching International*, 57:2, 136-147, https://doi.org/10.1080/14703297.2018.1538896

Vinson, D., Nixon, S., Walsh, B., Walker, C., Mitchell, E., & Zaitseva, E. 2010. "Investigating the relationship between student engagement and transition". *Active Learning in Higher Education*, 11, 131–143. https://doi.org/10.1177/1469787410365658

Vivian, C. 2005. "Advising the at-risk college student". *The Educational Forum*, 69, 336–351. https://doi.org/10.1080/00131720508984707

Von Hippel, P. T., & Hofflinger, A. 2021. “The data revolution comes to higher education: Identifying students at risk of dropout in Chile”. *Journal of Higher Education Policy and Management*, *43*(1), 2-23. https://doi.org/10.1080/1360080X.2020.1739800

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L. & Müller, K. 2020. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

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, 1023–1040. https://doi.org/10.1080/03075079.2014.966070

Table 1. Description of the variables of the model

Dependent Variable		**Frequency**	**%**
Dropout	Yes	567	15.82%
	No	3,016	84.18%
Independent Variables			
Quantitative Variables		***Mean***	***SD***
Access mark		6.53	1.31
Admission mark		8.55	1.05
Not shown rate		0.06	0.18
Fail rate		0.19	0.23
Mean pass		5.99	1.52
Dichotomic and qualitative Variables		***Frequency***	***%***
Sex	Female	2,029	56.63%
	Male	1,554	43.37%
Father educational level	Graduate	1,291	36.03%
	Non-Graduate	2,292	63.97%
Mother educational level	Graduate	1,329	37.09%
	Non-Graduate	2,254	62.91%
Father Occupation	Non-qualified	200	5.58%
	Qualified	2,084	58.16%
	University-qualified	1,283	35.81%
	Unemployed	16	0.45%
Mother Occupation	Non-qualified	421	11.75%
	Qualified	2,014	56.21%
	University-qualified	1,027	28.66%
	Unemployed	121	3.38%
Type of access to university	Over 25	40	1.12%
	Transfer HVT	30	0.84%
	Transfer UA	197	5.50%
	University access (UA)	2,858	79.77%
	HVT	458	12.78%
Student work	No job	2,321	64.78%
	Less 15 hours	190	5.30%
	More 15 hours	1,072	29.92%

Table 2. Sample used to evaluate the model's generalization capacity.

University	Degree	Years	Students
U. of Barcelona	Business Administration	2010/11 - 2017/18	3,583
U. Complutense of Madrid	Business Administration	2015/16 - 2021/22 (excluding 2019/20 due to Covid)	2,896
U. Complutense of Madrid	Law	2010/11 & 2011/12	2,003
U. Complutense of Madrid	Law	2015/16 - 2021/22 (excluding 2019/20 due to Covid)	4,688
U. Complutense of Madrid	Economics	2015/16 - 2021/22 (excluding 2019/20 due to Covid)	1,197

Table 3. Results obtained with all variables (NN, RF, and logit models), both in training and test sets. Highlighted in bold are the cells with the best accuracy metrics in each set

Model	Train set			Test set		
	Accuracy	**Sensitivity**	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**
NN	**0.83**	0.82	**0.83**	**0.83**	**0.78**	**0.84**
Logit	0.82	0.83	0.82	0.81	**0.78**	0.82
RF	0.82	**0.86**	0.81	0.79	0.77	0.80

Table 4. Mean,standard deviation and mean squared metrics for sensitivities of each variable (NN using 50 bootstrap samples)

		Mean Sens(*)	Std Sens(**)	Mean sq Sens (***)
Access mark		0.22	0.26	0.34
Admission mark		0.19	0.22	0.29
Female		-0.18	0.72	0.74
Father educ. Level: non-Graduate		0.05	0.33	0.33
Mother educ. Level: non-Graduate		-0.12	0.41	0.42
Father Occupation	Non-qualified	0.32	0.69	0.76
	Qualified	-0.05	0.41	0.42
	Unemployed	1.06	1.95	2.22
Mother Occupation	Non-qualified	0.04	0.63	0.64
	Qualified	0.18	0.44	0.48
	Unemployed	-1.04	1.32	1.69
Type of access to university	Over 25	0.50	1.74	1.81
	Transfer HVT	-0.93	0.95	1.34
	Transfer UA	-1.08	1.20	1.61
	University access (UA)	-0.09	0.57	0.57
Student work	Less 15 hours	-0.17	0.90	0.91
	No job	-0.13	0.49	0.51
Not shown rate		9.76	4.76	10.86
Fail rate		4.00	1.49	4.27
Mean pass		-0.14	0.22	0.26

(*) Mean sensitivity of each variable included in the model.

(**) Sensitivity standard deviation of each variable included in the model.

(***) Mean squared sensitivity of each variable included in the model.

Table 5. Results obtained with three variables (NN, RF and logit models), both in training and test sets, considering different degrees, years, and Universities. Highlighted in bold are the cells with the best accuracy metrics in each dataset

Dataset	Model	Train set			Test set		
		Accuracy	**Sensitivity**	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**
Business Degree (U. of Barcelona) n: 3,583 Train set: 2,867 Test set: 716	NN	0.82	0.82	0.82	0.81	**0.82**	0.81
	Logit	0.82	0.81	0.82	0.81	0.80	0.82
	RF	**0.83**	**0.83**	**0.84**	**0.82**	0.80	**0.82**
Business Degree (2015 - 2021) (U. C. of Madrid) n: 2,896 Train set: 2,318 Test set: 578	NN	0.75	0.66	0.77	0.75	0.65	0.77
	Logit	**0.77**	0.63	**0.80**	**0.77**	0.62	**0.80**
	RF	0.77	**0.77**	0.77	0.73	**0.67**	0.75
Law Degree (2010/11 & 2011/12) (U. C. of Madrid) n: 2,003 Train set: 1,603 Test set: 400	NN	0.79	0.69	0.81	0.78	0.63	0.82
	Logit	0.79	0.69	0.81	**0.79**	0.65	**0.82**
	RF	**0.83**	**0.86**	**0.82**	0.77	**0.68**	0.79
Law Degree (2015 - 2021) (U. C. of Madrid) n: 4,688 Train set: 3,751 Test set: 937	NN	0.76	0.59	0.79	0.76	0.56	0.79
	Logit	**0.79**	0.54	**0.84**	**0.79**	0.52	**0.84**
	RF	0.76	**0.86**	0.75	0.69	**0.64**	0.70
Economics Degree (2015 - 2021) (U. C. of Madrid) n: 1,197 Train set: 958 Test set: 239	NN	0.77	0.77	0.77	0.75	**0.76**	0.75
	Logit	0.77	0.73	**0.79**	**0.76**	0.73	**0.77**
	RF	**0.81**	**0.88**	0.79	0.73	0.75	0.72

Figure 1. Accuracy and Sensitivity (test set) for the NN model in the 5 datasets.

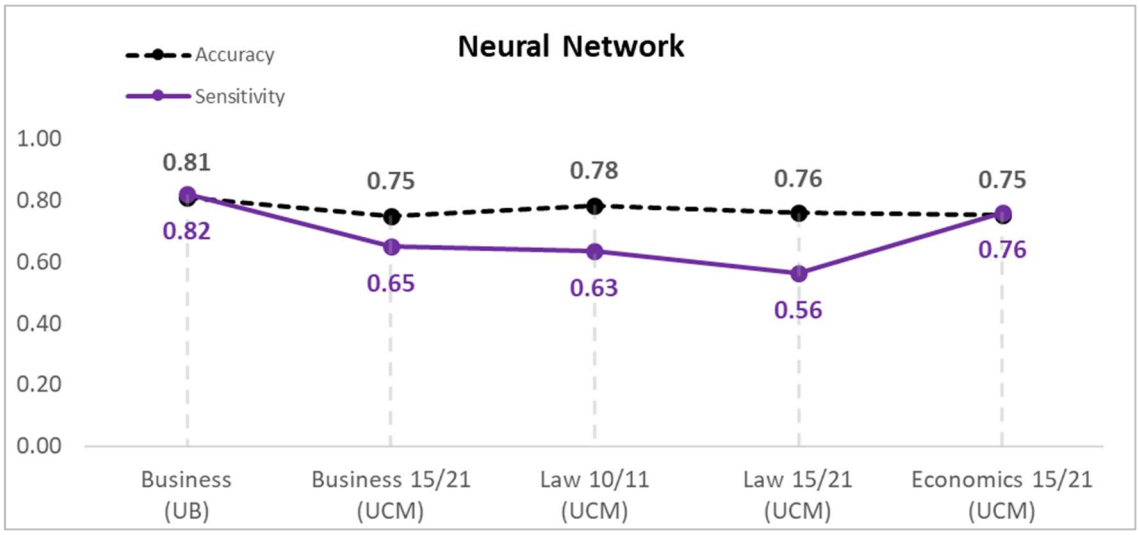

Figure 2. Accuracy and Sensitivity (test set) for the logit model in the 5 datasets.

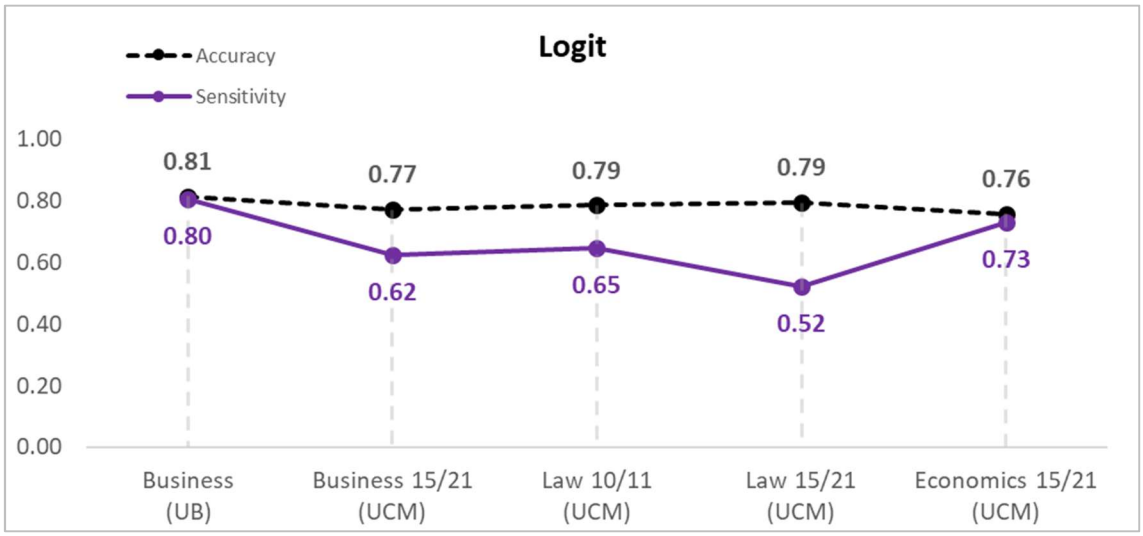

Figure 3. Accuracy and Sensitivity (test set) for the RF model in the 5 datasets.