University Dropout: A Prediction Model for an Engineering Program in Bogotá, Colombia

Conference Paper · July 2019

CITATIONS
0

READS
191

3 authors, including:

Andres Acero
Universidad de Monterrey
26 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Juan Morales Piñero
Sergio Arboleda University
22 PUBLICATIONS **19** CITATIONS
SEE PROFILE

University Dropout: A Prediction Model for an Engineering Program in Bogotá, Colombia

Andrés Acero
Universidad Sergio Arboleda, Bogotá, Colombia
andres.acerol@usa.edu.co

Juan Camilo Achury
Universidad Sergio Arboleda, Bogotá, Colombia
juan.achury@usa.edu.co

Juan Carlos Morales
Universidad Sergio Arboleda, Bogotá, Colombia
juan.morales@usa.edu.co

***Abstract**: In Colombia, the desertion average rate shows that only around half of the students that begin an undergraduate program are finishing their studies. Although several models have been developed, the results of the implementation of strategies for student retention have not been sufficiently effective. Therefore, this study focuses on the creation of robust predictive models that allow timely anticipation of the risk that an engineering student will retire prematurely from the program. This phenomenon is analyzed in an empirical way through a methodology of Knowledge Discovery in Databases (KDD) using different machine learning techniques. Results show that the academic cost (in terms of the number of subjects viewed) and the semester the student entered have a significant impact on the probability of dropout occurring, especially when considering the dropout in the firsts semester. Then, acting on the students predicted by the model might reduce the number of dropouts.*

Introduction

In the pursuit of growth and equity, no country can afford to ignore higher education. Through higher education, a country forms skilled labor and builds the capacity to generate knowledge and innovation, which in turn drives productivity and economic growth (Gitto, Minervini, & Monaco, 2016). Given that the acquisition of skills increases productivity and the expected income of people, a good education system is the basis for achieving greater equity and shared prosperity at the social level. Education allows us to update our potential, develop skills that allow us to be better (Patrick & Borrego, 2016).

However, the fact that a person abandons education has severe consequences. The event acquires greater relevance when occurs at higher education if you think about the preparation necessary to achieve this educational level. Furthermore, it is common to attribute the dropout to the failure of the student. A critical look at this position allows to notice that the problem is crossed by multiple factors; some are nested in the learning process itself, while others are alien to it and even out of their control (Lacave, Molina, & Cruz-Lemus, 2018).

In fact, in the Colombian context, one of the main problems the higher education system faces is the high levels of academic dropout from undergraduate programs. Although during the last five years have been characterized by increases in coverage and income of new students, the number of university students who complete their higher education is about 50%, suggesting that a large part drop out, mainly in the first semesters(Morales, Cordero, & Ramírez, 2017). According to statistics from the Ministry of Education of Colombia (2009), of

every hundred students who enter a higher education institution, nearly half do not complete their academic cycle and obtain graduation. Thus, knowledge of the factors that affect this phenomenon is the basis for developing effective policies in order to increase student retention (Esteban, Bernardo, Tuero, Cervero, & Casanova, 2017; Geisinger & Raman, 2013). Then, the measurement and study of student attrition should be part of the evaluation of the efficiency of the education system and of the quality of their processes.

Research in other countries provides evidence of a relationship between social or academic factors and the decision to leave the university without graduating (Munizaga, Cifuentes, & Beltrán, 2018). Likewise, (Fan & Wolters, 2014) suggest that explanations for student dropout without considering their motivation are incomplete. Official data from the United Kingdom indicate that dropping out of college is more likely for students from lower classes. Longitudinal studies conducted in the USA, using both national data and institutional data and differences in academic and social integration, resulted in different levels of institutional and educational commitment (Dicovskiy Riobóo & Pedroza Pacheco, 2018). However, the little research on high school dropout in Colombia opens a spectrum of possibilities to generate research. In fact, the few studies that have included, for example, social class in their investigation of the phenomenon of school dropout have found little evidence of class differences. In contrast, attrition decisions were more likely to be related to the averages of the high school aptitude tests with which the students entered university (Flórez-Nisperuza & Carrascal-Padilla, 2016; Torres Guevara, 2012).

Research Question

Given the background presented above, this study wants to explore the following research question:

- How can a university anticipate that a specific student is going to drop out from an engineering undergraduate program?

The purpose of this question is to develop a predictive model, based on four machine learning models and using a top layer model for the ensemble, to elucidate the social and academic factors that affect the student's decision to drop out from the Industrial Engineering program at Sergio Arboleda University. Even more, based on the model results, formulate strategies and institutional policies that increase student retention for engineer students. This model is important for engineering education because administrative staff and academic divisions could understand the relevant factors in the decision of dropout in engineer students and provide appropriate support to reduce dropout rates.

Theoretical Framework

Student dropout models

Among the researches carried out around the world, the Tinto model (1975) is the most accepted model in the literature on student retention in higher education. Tinto affirms that the students' decision to persist or drop out of their studies is strongly related to their degree of academic integration and social integration in the university. On the other hand, within the studies on student retention using quantitative tools, the classification algorithms (Kumar & Verma, 2012) are the most applied data mining technique to predict school dropout. A good example of the use of these methods is the research of Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, & Loumos (2009), in which several classification techniques were applied (advanced neural networks, support vector machines (SVM), fuzzy schemes with probabilistic models and decision plans) for the prediction of dropout in e-learning courses at the University of Athens. The most successful technique to promptly and accurately predict the students who could defect was the decision plan. In other studies, for example, the authors used another comparative analysis of several classification methods (artificial neural networks, decision trees, SVM and logistic regression) to develop early models of freshmen

who are more likely to drop out (Delen, 2010). However, current schemes have focused only on predicting using a single model, not parallel models, which is the objective of this study.

In Colombia, studies for follow-up, early diagnosis, and strategies for the prevention of student desertion are focused on survival analysis are the ones that have been used the most and with which the desertion models are created in the Colombia Ministry of Education (Ministerio de Educación Nacional, 2009). This type of dynamic models focuses on the study of this phenomenon according to survival functions, considering that the population is heterogeneous and assuming that the risk is equally probable for each person. This model also includes personal, academic, socioeconomic and institutional information, as suggested by Tinto's models (Tinto, 1975). However, the results obtained, to the opinion of the authors, are vague and do not allow to capture the phenomenon entirely.

Methodology

As shown in the background, the traditional methodologies for predicting student dropout uses only the information from an academic background with single, classical and well-known classification algorithms. Therefore, we propose both a new methodology and the use of an ensemble model that attempts to produce better results and detect students' dropout as early as possible. Three different models of student dropout were developed starting from the data gathered at different steps of the (Figure 1).

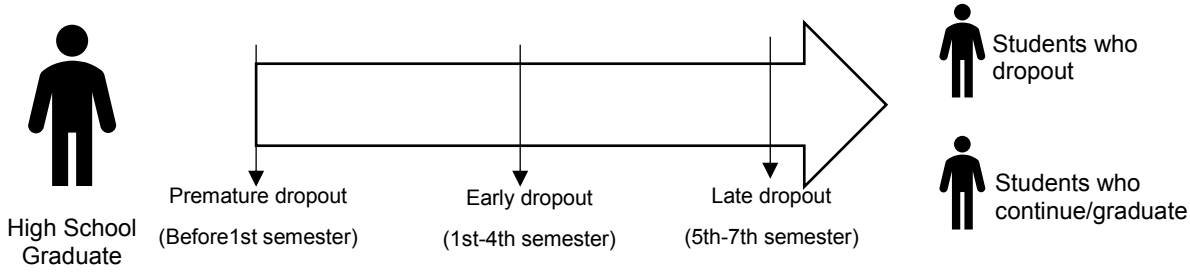

Figure 1: Proposed dropout prediction methodology.

As you can see in Figure 1, even at the beginning of the program, a premature dropout prediction can be made by using only the data available from personal and high school aptitude test (Saber 11) information about the student. As the program progresses, more information progressively becomes available about the performance of students. Therefore, two more models (early and late dropout) were developed to understand the phenomenon at different stages of the career.

Data Set

The dataset used in this study was built from the information of 422 students from the Industrial Engineering program from Sergio Arboleda University in the period 2016-2019. The main variable for examination on each one of the models is the dropout rate, which in the first model represents the students who did not obtain credits during their first semester, and the other models represent students who did not enroll by two or more consecutive semesters. The socioeconomic factors include social stratum, age, payment type (family resources, scholarships, or institutional agreements), and gender. About the academic performance, the data includes grade point average (GPA) each period, the number of lost courses and the average GPA before dropout. Furthermore, 23% of the students drop out, as you can see in Figure 2, being early dropout the most common case.

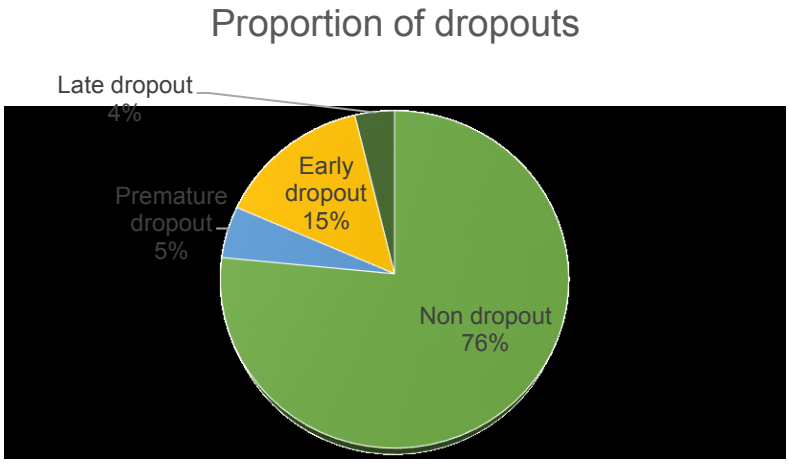

Figure 2: Distribution of student dropout

Data analysis

The models were built using the caret package of R©. The size of the training sets was 70% of the original dataset, according to the standard used by Microsoft (2018). Four models of classification were selected for the bottom level of the model: random forest, Bayesian classifier (using naïve Bayes), logistic regression, and neural networks. On the top level, a gradient boosting model was used, with the data of the previous models as input for the model.

Table 1: Confusion Matrix.

Actual vs. predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

TP: True positive; FN: False negative; FP: False positive; TN: True negative

To compare the models, the chosen measures were specificity (equation 1), recall (equation 2), accuracy (equation 3), and AUC (equation 4), as suggested by Sokolova and Lapalme (2009), from the information obtained from the confusion matrix (Table 1).

$$Specificity: \frac{TN}{TN+FP} \quad (1)$$

$$Recall\ (Sensitivity): \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy: \frac{TP+TN}{N} \quad (3)$$

$$Area\ under\ the\ curve\ (AUC): \frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \quad (4)$$

All classification algorithms were executed using a 10-fold cross-validation procedure in which all executions are repeated 10 times using different train/test partitions of the data set. The data was validated to comply with the assumptions of each one of the individual classifiers, including independence between variables and considerations about the use of categorical variables.

Findings

Model 1: Premature dropout

In this model, the sample size is 422 students from an Industrial Engineering program. The variables used in this model were socio-economic variables (gender, rural or urban, type of funding, age, social stratum and type of admission) and the results of the high school aptitude test (Torres Guevara, 2012). The results of the four models and the two ensemble methods used are presented in table 2.

Table 2: Measures of performance for the premature dropout model.

Model	Specificity	Recall	Accuracy	AUC
Naïve Bayes	0%	95.24%	95.24%	N/A
Logistic Regression	0%	95.24%	95.24%	N/A
Random Forest	0%	95.24%	95.24%	N/A
Neural Network	0%	95.24%	95.24%	N/A
Ensemble model (Voting)	N/A	N/A	N/A	N/A
Ensemble model (Gradient boosting)	0%	95.24%	95.24%	N/A

These models, unfortunately, were not able to predict properly this kind of desertion. A possible explanation for this phenomenon is the high imbalance on the data (21 over 401) that makes difficult to predict. After checking the results, the four models only produce non-dropout results for everybody. This result suggests that, maybe, this kind of desertion is not significative to be considered.

Model 2: Early dropout

In this model, the sample size is 312 students from an Industrial Engineering program due to the only student on their third semester or higher were included. The variables used in this model were socio-economic variables (gender, rural or urban, type of funding, age, social stratum and type of admission), the number of subjects seen, the last available GPA, and the number of subjects lost. The results of the four models and the two ensemble methods used are presented in table 3.

Table 3: Measures of performance for the early dropout model.

Model	Specificity	Recall	Accuracy	AUC
Naïve Bayes	92.86%	93.67%	93.55%	93.26%
Logistic Regression	77.78%	94.67%	91.4%	86.22%
Random Forest	85.71%	100%	96.77%	92.86%
Neural Network	88.24%	96.05%	94.62%	92.14%
Ensemble model (Voting)	87.50%	96.81%	93.55%	91.55%
Ensemble model (Gradient boosting)	85.71%	100%	96.77%	92.86%

As we can see in the results, the four models were capable to produce accurate results that adjust properly to the data. In the level of the bottom models, we can appreciate that naïve Bayes and Random Forest outperform the Logistic Regression and the Neural Network, showing that they can be more specific and more accurate than the other models. In the top layer, both voting and gradient boosting were able to capture a higher performance that some of the models. However, the gradient boosting model shows better results by capturing the recall and accuracy of the random forest model.

Model 3: Late dropout

In the third model, the sample size is 158 students from an Industrial Engineering program due to the only student on their third semester or higher were included. The variables used in this model were socio-economic variables (gender, rural or urban, type of funding, age, social stratum and type of admission), the number of subjects seen, the last available GPA, and the number of subjects lost. The results of the four models and the two ensemble methods used are presented in table 4.

Table 4: Measures of performance for the late dropout model.

Model	Specificity	Recall	Accuracy	AUC
Naïve Bayes	25%	92.86%	86.96%	58.93%
Logistic Regression	50%	95.24%	91.3%	72.62%
Random Forest	100%	93.33%	93.48%	96.67%
Neural Network	50%	95.24%	91.3%	72.62%
Ensemble model (voting)	100%	93.33%	93.48%	96.67%
Ensemble model (Gradient boosting)	N/A	91.3%	91.3%	N/A

Finally, as shown in the previous table, there is a high variance between the model in terms of the specificity and the area under the curve (AUC). Specifically, the higher level on most of the parameters was reached by using the random forest technique. However, if we compare the models based on the sensitivity (or recall), both logistic regression and neural networks outperform the other models. In the top layer of the ensemble model, only the voting model was able to really capture the performance of the other models. The gradient boosting model was not even able to predict a single one case of late dropout.

Discussion and Future Research

Higher education dropout is being a big concern for academic institutions all over the world. The risk of student attrition has been studied broadly and deeply (Hällsten, 2017; Truta, Parv, & Topala, 2018; Vallejos & Steel, 2017), and the amount of data available about this phenomenon keeps growing every day. This data contains hidden knowledge about the features that could be used to predict the risk of the students to drop out in different moments of their studies. This paper proposes a student attrition model based on machine learning methods on different levels and using both social information and academic performance features. In general, the individual models and the ensemble models used by the authors produces good results in terms of the accuracy and AUC (more than 80% in model 2 and 3), which shows that the model strongly adjusts to the real data obtained. Compared with similar studies, this level of accuracy in the literature can be only found in Zhang's study (2010), where using naïve Bayes models, 89% of accuracy was reached. However, due to access limitations to more complete data, our model can be improved to be more robust for the early detection of students who can drop out.

In terms of the use of the machine learning techniques, the key result from these models is that none of the individual methods used can be assured to be completely superior to the others. The objective of these models is to construct a discriminative classifier, which tries to approximate to the binary decision to drop out from an engineering program. Due to the size of the sample used in this study, the individual method can produce different results with different performance on each of the measurements checked, looking for local optima in the individual search strategy. Thus, using a combination of the classifiers with an ensemble method can provide better results by outperforming the individual searches and combining the outputs. In fact, ensemble machine learning has been used for several years for classification models, but their uses on the desertion of engineering students are relatively new because most of the models use a single learning approach (Amrieh, Hamtini, & Aljarah, 2016). From the results of the models, we can see that the use of gradient boosting method in the early dropout model, and the voting method for the late dropout model, let us obtain better capture the highest performance than using any individual model. These results support our hypothesis by showing that we can anticipate a certain pattern in our data to provide early alerts about students with a high risk of dropout. However, in this study, the cost in terms of computational time was ignored. The next step of this research will include a comparison of the models in terms of the costs associated with the performance level reached.

In terms of our results, even if our first model (premature dropout) does not provide enough information to anticipate if students will drop out before entering the program, our best estimation is every student will enter in the program and study at least one semester. Because of the little number of students who drop out prematurely (only 5%), this model can be improved by obtaining more information from other engineering programs. The second model, early dropout, and the third model, late dropout, reaches around 93% of accuracy on predicting drop out. A deep look in the model shows us that are two features that help us predict desertion. First, the number of subjects seen by the students throughout their careers is a great predictor of non-dropout. In other words, when a student takes more subjects over time is less likely that wants to drop out. This is related to some theories that support that exist both a social and economic cost associated with the decision to drop out (Hällsten, 2017). Further research about the social, emotional and economic cost of dropping out will improve the results of this model. Second, there are differences between the groups of the students who start their studies in January (Group 1) than those who start their studies in July (group 2). Even this effect has not been completely understood by the authors, the model predicts more frequently than group 2 is more likely to drop out from an engineering program. This requires strategies to understand the differences between these groups and create policies to prevent the desertion. These two features, even more, will be studied by using more and complete data in future research.

References

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, *9*(8), 119–136. https://doi.org/10.14257/ijdta.2016.9.8.13

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2010.06.003

Dicovskiy Riobóo, L. M., & Pedroza Pacheco, M. E. (2018). Predicción De Deserción Y Éxito En Estudiantes. Caso de Estudio: Ingeniería Agroindustrial de la UniNorte, Nicaragua, 2011-2015. *Nexo Revista Científica*, *31*(01), 16–27.

Esteban, M., Bernardo, A., Tuero, E., Cervero, A., & Casanova, J. (2017). Variables influyentes en progreso académico y permanencia en la universidad. *European Journal of Education and Psychology*, *10*(2), 75–81. https://doi.org/10.1016/j.ejeps.2017.07.003

Fan, W., & Wolters, C. A. (2014). School motivation and high school dropout: The mediating role of educational expectation. *British Journal of Educational Psychology*, *84*(1), 22–39. https://doi.org/10.1111/bjep.12002

Flórez-Nisperuza, E. P., & Carrascal-Padilla, J. J. (2016). Estudio de la deserción estudiantil de la Licenciatura en Ciencias Naturales y Educación Ambiental de la Universidad de Córdoba-

Colombia- 2011 – 2015. - Study of student desertion of the Degree in Natural Sciences and Environmental Education of the Univer. *Revista Científica*, *4*(27), 340. https://doi.org/10.14483/udistrital.jour.RC.2016.27.a4
Geisinger, B. N., & Raman, D. R. (2013). Why they leave: Understanding student attrition from engineering majors. *International Journal of Engineering Education*, *29*(4), 914–925.
Gitto, L., Minervini, L. F., & Monaco, L. (2016). University dropouts in Italy: Are supply side characteristics part of the problem? *Economic Analysis and Policy*, *49*, 108–116. https://doi.org/10.1016/j.eap.2015.12.004
Hällsten, M. (2017). Is education a risky investment? The scarring effect of university dropout in Sweden. *European Sociological Review*, *33*(2), 169–181. https://doi.org/10.1093/esr/jcw053
Kumar, R., & Verma, D. (2012). Classification Algorithms for Data Mining: A Survey. *International Journal of Innovations in Engineering …*. https://doi.org/10.1002/ppul.23134
Lacave, C., Molina, A. I., & Cruz-Lemus, J. A. (2018). Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks. *Behaviour & Information Technology*, *0*(0), 1–15. https://doi.org/10.1080/0144929X.2018.1485053
Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education*. https://doi.org/10.1016/j.compedu.2009.05.010
Microsoft. (2018). Training and Testing Data Sets.
Ministerio de Educación Nacional. (2009). *Deserción estudiantil en la educación superior colombiana: Metodología de seguimiento, diagnóstico y elementos para su prevención*.
Morales, J., Cordero, N., & Ramírez, J. (2017). Influence of economic expectation on choosing a university: A case study in Industrial Engineering. *Espacios*, *38*(35).
Munizaga, F. R., Cifuentes, M. B., & Beltrán, A. J. (2018). Retención y Abandono Estudiantil en la Educación Superior Universitaria en América Latina y el Caribe: Una Revisión Sistemática. *Education Policy Analysis Archives*, *26*(0), 61. https://doi.org/10.14507/epaa.26.3348
Patrick, A. D., & Borrego, M. (2016). A Review of the Literature Relevant to Engineering Identity. *2016 ASEE Annual Conference & Exposition*, *June*. https://doi.org/10.18260/p.26428
Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*. https://doi.org/10.3102/00346543045001089
Torres Guevara, L. E. (2012). *Retención Estudiantil en la Educación Superior: Revisión de la literatura y elementos de un modelo para el contexto colombiano*.
Truta, C., Parv, L., & Topala, I. (2018). Academic Engagement and Intention to Drop Out: Levers for Sustainability in Higher Education. *Sustainability*, *10*(12), 1–11. https://doi.org/10.3390/su10124637
Vallejos, C. A., & Steel, M. F. J. (2017). Bayesian Survival Modelling of University Outcomes. *Statistics in Society Series A*, 1–19.
Zhang, Y., Oussena, S., Clark, T., & Hyensook, K. (2010). Use data mining to improve student retention in HE: a case study. *ICEIS - 12th International Conference on Enterprise Information Systems*.

Acknowledgments

The authors wish to acknowledge the financial assistance of Universidad Sergio Arboleda. The authors also acknowledge the support of Juan David Arboleda and María Paula Flórez during the data gathering and research.

Copyright statement