An early warning dropout model in higher education degree programs: A case study in Ecuador

Vanessa Heredia-Jimenez[1], Alberto Jimenez[1], Margarita Ortiz-Rojas[1], Jon Imaz Marín[2], Pedro Manuel Moreno-Marcos[2], Pedro J. Muñoz-Merino[2], and Carlos Delgado Kloos[2]

[1] Escuela Superior Politecnica del Litoral, ESPOL, Information Technology Center, Guayaquil, Ecuador
{estefania.heredia,alberto.jimenez,margarita.ortiz}@cti.espol.edu.ec
[2] Universidad Carlos III de Madrid, Madrid, España
jimaz@pa.uc3m.es,{pemoreno,pedmume,cdk}@it.uc3m.es

Abstract. Worldwide, a significant concern of universities is to reduce academic dropout rate. Several initiatives have been made to avoid this problem; however, it is essential to recognize at-risk students as soon as possible. In this paper, we propose a new predictive model that can identify the earliest moment of dropping out of a student of any semester in any undergraduate course. Unlike most available models, our solution is based on academic information alone, and our evidence suggests that by ignoring socio-demographics or pre-college entry information, we obtain more reliable predictions, even when a student has only one academic semester finished. Therefore, our prediction can be used as part of an academic counseling tool providing the performance factors that could influence a student to leave the institution. With this, the counselors can identify those students and take better decisions to guide them and finally, minimize the dropout in the institution. As a case study, we used the students' data of all undergraduate programs from 2000 until 2019 from a public high education university in Ecuador.

Keywords: Data mining · Dropout prediction · Early detection · Algorithm · Learning analytics · Higher education

1 Introduction

Worldwide, one main concern in all Higher Education Institutions (HEI) is student dropout [12, 19]. For instance, according to the Organization for Economic Cooperation and Development (OECD), the dropout rate is about 24% [18].

In the specific case of Latin America, the World Bank reports a similar rate than the OECD with a 22% dropout rate [14]. Some proposals to solve this situation have been, among others, offering students an introductory orientation phase [25], offering financial aids or scholarships [11], and use ICT in their University Tutorials Systems [8]. Nonetheless, the overarching strategies and the funding programs on academic success do not sufficiently reduce the academic dropout. It is vital to recognize at-risk students before dropping out.

The available data about students' academic performance, creates an opportunity to identify what factors affect student dropout and take preventive measures to solve this problem. This is the reason why studies have been done using prediction algorithms to solve this situation [6,21]. In the case of Latin America, there are not many studies in this area. However, when found, these are confined to: a) identify attrition through descriptive statistics [17], b) study only the first years of student study [5], c) focus on a specific undergraduate program [15] or a courses [13] and, d) use a different type of variables such as academic and sociodemographic data, which results in over-fitting or biased results due to the large number of variables [23]. Moreover, the studies do not specify the input characteristics, causing that the models can not be replicated. This study addresses the problems above by proposing a predictive module considering the academic history of students throughout their undergraduate programs with a strong emphasis on analyzing which input variables have the most substantial influence on predicting dropout. A case study is presented in a Latin American institution of higher education where the model was tested, scaling it to all undergraduate programs. This model will help as input for the Early Warning Systems (EWS) dashboards to be developed in the Learning Analytics area.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the relevant literature. Section 3 describes the data used in our predictive model and summarises our work. Section 4 shows the experimental results from the study. The discussion about our work is presented in Section 5. Finally, Section 6 draws conclusions from this analysis, and suggests possible directions for future work in this area.

2 Related Work

When it comes to input variables to predict dropout, there are different models and input features found in the literature, some focused on analyzing a single type of data, while others mix different groups of data.

Regarding a single type of data, for example, in studio [22], the authors used psychosocial variables with a statistics model on 690 students, and the results ranged from 11.8% to 22% of the variance. The studio of [24], focused on analyzing academic data such as the first partial grade and the percentage of non-attendance accumulated of 171 students. Unfortunately, the prediction rate is no mentioned. The study of [16] combined intelligence, personality, and motivational predictors in 137 students and the results showed 33% of the variance in GPA and 30% of the variance in time to graduation.

In terms of mixing different groups of data, in [4], the authors used demographic, pre-college entry information, and complete transcript records of 69,116 students, generating a total of more than 700 features and had an accuracy of 66%. In [3], the authors used the Cox proportional hazards model with different groups of variables, such as demography, family history, financial information, high school, college enrollment and semester credits of 11,121 students. The precision of its model ranged between 71% and 82%. In [15], the attributes used were about the student's behavior in the first part of his undergraduate program of 498 students, ignoring classical attributes. The precision of their models ranged between 74% and 78%. Another predictive model, as well as an EWS, is presented by [5], which only uses academic performance and data gathered of LMS, the precision ranges where from 79% to 92%.

While all of the above models are accurate in their contexts, they are not scalable because they use small samples, only use the first academic years, or focus on specific fields or undergraduate programs. Moreover, a mix of different variables, such as psychological, demographic and academic, could cause biased results or less accuracy prediction. Besides, certain models cannot be replicated because the variables are not specified but rather mentioned in a general way. Therefore, a model that can be scalable and replicable is needed to understand the student's path from beginning to end and consider variables that cause an impact on learning performance.

3 Case Study

The study took place in Escuela Superior Politécnica del Litoral, ESPOL, a public engineering-oriented university located in Ecuador. The model was implemented in Spyder[3], an open-source platform, and we used Python V3.7 as programming language and made use of the Scikit-learn library, an open-source library implements many machine learning algorithms. The proposed model and each of its phases is described below.

3.1 Data set

The information was provided by the Information Technology and Systems Management[4]. For the model, we considered the bachelor's degree in Industry Engineering, which has 753 students enrolled from 2000 until the first semester of 2019. This is one of the undergraduate programs with the most significant number of students enrolled during this period. The years selected were used as boundaries to allow five years of full studies from the time the students began their first academic semester until obtaining a professional degree.

[3] https://www.spyder-ide.org/
[4] https://www.serviciosti.espol.edu.ec/

3.2 Pre-processing

This phase focused on the information acquired, with the purpose of generating consistent data that could be used for the algorithm. The data was categorized as follows: socio-demographic data of the student (gender, marital status, employment status, city of residence, among others), data about the undergraduate program (number of total credits, code of the course, among others) and data on the student's performance (courses taken, state of the course, number of times the course has been taken, GPA, credits of the course, among others). During the pre-processing phase, cleaning, transformation and integration procedures were performed due to the origin of the information.

A recognition process was carried out, in which the academic terms were appropriately identified because the period of study varied depending on the term. Therefore, the periods were the first and second term, also called ordinary periods, and third term also called extraordinary period. For the predictive model, ordinary periods were selected as they had the same duration. Subsequently, we enumerated the semesters in which students had been studying with their respective academic record, in order to execute the analysis per semester.

After the pre-processing, the most representative academic performance variables were calculated and used as inputs for the model, which are detailed in Table 1. Besides, Section 4.1 explains the reasons why the model uses academic information alone, and discards socio-demographics information. Other indicators as undergraduate program performance indicators per semester were calculated in order to obtain additional information to compare the results gathered from the prediction.

Table 1. Learning indicators for the model

ID	Description
V1	Average number of subjects taken per semester during the years of study
V2	Number of times of second enrollment in a subject after failing the first time
V3	Number of times of third enrollment in a subject after failing the second time
V4	The period in years that the student takes to return to study
V5	GPA average of all subjects
V6	GPA average of subjects taken in the current undergraduate program
V7	GPA average of subjects with number of credits greater than zero, penalized for the number of times the subjects are taken by the student
V8	Ratio of the approved subjects taken by the student
V9	Ratio of the failed subjects taken by the student
V10	Ratio of the canceled subjects taken by the student

3.3 Desertion alert algorithm solution

For this preliminary work, we define the concept of academic dropout when a student has not been enrolled in any course for some time since his/her last registration. Based on ESPOL's internal guidelines and regulations[5], the period is five years. Therefore, those students have a dropout value equal to one. On the other hand, students who graduated had a dropout value equals to zero. Also, students with 90% of their undergraduate degrees completed are assigned with dropout equal to zero. This percentage based on other studies [2], which indicated that people with this advanced level abandon their undergraduate degree for reasons outside their academic performance. Thus, dropout is considered as a single, binary outcome feature. As a classification problem, we used the Random Forest Classifier (RFC) algorithm [7]. This is a well-known tree-based learning algorithm, known for its low tendency to over-training and its high accuracy.

After pre-processing, the process started with the dropout assignment equal to one or zero, according to the definition of academic dropout indicated above. Subsequently, the semester selection was made cumulatively. This means, selecting the academic history of all students in their first semester and entering this data into the model; then, selecting the students in their first and second semesters and entering this data into the model; and so on, until the fifth semester, which represents two and a half years of studies. After the fifth semester, the evaluation of all students who had more than five semesters in their undergraduate programs was carried out. In Ecuador, according to [9], the average of student dropout occurs between the second and third year of the studies.

Every group of semesters was used to calculate the academic performance indicators, and consequently, define the set of data that was taken for the training and testing. The training data was built with the group of students whose dropout was previously defined, and the algorithm was trained; while, for testing data, the set of students who did not have a dropout defined was considered.

In general, high probability values mean that the student has a high probability of dropping out of the undergraduate program, while low values indicate a high probability of finishing the degree. Fig. 1 depicts the steps that were carried out to produce meaningful information.

4 Results

4.1 One degree

For data acquired from the bachelor's degree in Industry Engineering, we applied a correlation matrix to analyze how the different variables were affected within the proposed model. As consequence, the socio-demographic variables were discarded, such as gender, married status, school type, and work status, because this variables had a correlation coefficient lower than 0.1; therefore, were

[5] http://reglamentos.espol.edu.ec/

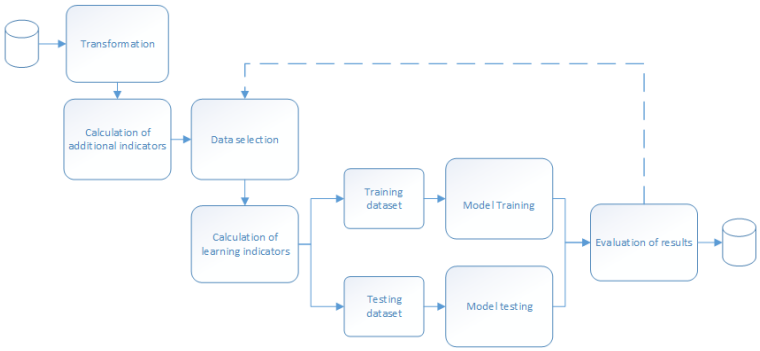

Fig. 1. Predictive model

not included into the model. Additionally, we use the RFC with the following parameters: number of trees equal to 300, depth of trees equal to 6, minimum number of samples required to split equal to 6, minimum number of samples required to be at a leaf node equal to 7. These parameters were chosen based on the results in accuracy when the model was applied to our testing set with an accuracy of more than 95%.

In order to evaluate the effectiveness of the proposed algorithm, accuracy, area under curve (AUC), F1-score, recall and precision were used as evaluation criteria. Accuracy is the proportion of correct prediction of dropout. Precision is the proportion of dropout learners predicted correctly by the classifier. Recall is the proportion of dropout learners predicted correctly by the classifier in all real dropout learners. F1-score is the harmonic mean of precision and recall [10]. Table 2 presents the results of the evaluation criteria per semester of the predictive model based on the variables described in Section 3.2, for the undergraduate program selected.

Table 2. Evaluation criteria per semester

Semester	AUC	Accuracy	F1	Recall	Precision
1	0.997	99%	0.928	0.928	0.928
2	0.995	96%	0.948	0.915	0.984
3	0.994	96%	0.925	0.911	0.939
4	0.972	91%	0.904	0.863	0.95
5	0.996	97%	0.937	0.937	0.937
>5	0.996	98%	0.966	0.96	0.972

4.2 Scalability

It was decided to scale the model for all undergraduate programs that are dictated in ESPOL. The data selected from 2000 until the first semester of 2019

has a total of 29,983 students. Due to the time of the selected data, there were some changes in denomination codes of undergraduate programs. Nevertheless, the essential structure of the undergraduate programs had been preserved. With the purpose of scalability, a process of unifying undergraduate program codes was carried out to avoid the loss of information.

A total of 65 undergraduate programs were predicted within the model. The model predicted the dropout degree by degree, and as explained before, every group of semesters was used to calculate the academic performance indicators, and consequently, define the set of data that was taken for the training and testing. In this way, in parallel to testing our proposal, we proved that socio-demographic data are unnecessary for dropout prediction, and eliminate typical biases from the learning process. A summary of the five evaluation criteria of the model for the 10 undergraduate programs with the largest of number of students enrolled are shown in Table 3.

Table 3. Prediction accuracy for the undergraduate programs with largest number of students enrolled between 2000 and 2019

Degrees	**AUC**	**Accuracy**	**F1**	**Recall**	**Precision**
BIBIMBM	0.997	99%	0.929	0.929	0.929
CI004	0.996	97%	0.938	0.938	0.938
CI005	0.997	98%	0.966	0.960	0.973
CI008	0.994	92%	0.929	0.929	0.929
CI009	0.998	98%	0.975	0.975	0.975
CI013	0.993	97%	0.959	0.935	0.983
CI020	0.996	99%	0.989	0.979	1.000
ECCBA	0.973	99%	0.940	0.986	0.899
INACP	0.993	97%	0.934	0.949	0.918
INALL	0.980	98%	0.902	0.974	0.841

5 Discussion

Our proposed model was based on academic data alone and the analysis of the academic semesters of students throughout the student's undergraduate program. The initial results indicated a strong signal in the use of data purely academic for predicting student dropout, with an accuracy of more than 95%. In addition, the experimental results proved the effectiveness and scalability of our algorithm due to the different numbers of students enrolled.

Unfortunately, these results can not be compared with similar studies that have only used academic data as well because of the following reasons. First, the prediction model focuses on the first semesters only [1], while ours studies the student throughout his/her undergraduate program. Second, the variables are not the same. For instance, in [15], the authors used three academic variables, but the studio considered the fees deadline as academic indicator, while we used

in our model data purely academic. Third, we tested our model in different undergraduate programs. Meanwhile in [4] used one undergraduate program.

We also acknowledge the limitation that prevents us from generating a better model. These limitations arise from the data gathered. Although, it is true we only focus on the use of academic data for the model, it should be noted the importance of the students' financial status, because this factor is one of the motivation behind students' decisions to stop their studies [20]. Thus, our limitation is because the students' financial status was calculated in ESPOL using differed range values for students from 2000 until 2014 and for students from 2015 onwards. Another limitation is when students withdraw for a period of time and their undergraduate programs change the assigned code, so when they return to their studies, they are registered into the database with the new undergraduate program code, but with all their academic records as a single semester. The portion of this student population with more than 10 subjects taken was 6%. This limitation was addressed considering as maximum average 7 subjects taken per semester.

6 Conclusion and future work

Dropout prediction is an essential prerequisite to make interventions of at-risk students. To address this issue, we proposed a predictive model, which used academic data alone and analyzed the information semester by semester to obtain more reliable predictions. Besides, it was able to identify the earliest moment of dropping out of a student. To the best of our knowledge, this is the first proposal of a predictive model able to predict even for students who have just finished their first semester.

Our model was also tested with real data of ESPOL and was also applied to other large data sets of the different undergraduate programs. Experimental results showed that our model has better results than existing proposals in terms of input features and accuracy. Therefore, we have proven that our model allows to predict with sufficiently high accuracy but, at the same time, can predict the intention of dropping out early enough, so that some corrective actions can be attempted.

Additionally, the fact that the variables that worked are specifically mentioned can help other universities to replicate this model or adapt it to their needs.

Our future work aims to make technical improvements to our prediction model and use other machine learning algorithms to compare them as other researches do. Furthermore, we plan to use this model in postgraduate programs and in academic data sets from other types of universities where attrition rates tend to be much higher. Additionally, we are currently working on a visualization prototype to integrate the results of the predictive model into the Academic Counseling System. Using visualization techniques with explanatory models, allows counselors to have the opportunity to identify those students who have the possibility of an early dropout, and be able to adapt a recommendation strategy

to minimize the risk of failure of a student. Lastly, we plan to analyze and visualize which input variables have the most substantial influence on predicting dropout.

Acknowledgements

Work partially funded by the LALA project (grant no.586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP). This project has been funded with support from the European Commission. This work has also been partially funded by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307, a project which is co-funded by the European Structural Funds (FSE and FEDER). This work has also been partially funded by the FEDER/Ministerio de Ciencia, Innovación y Universidades–Agencia Estatal de Investigación, through the Smartlet Project under Grant TIN2017-85179-C3-1-R. This publication reflects only the views of the authors, and the funders cannot be held responsible for any use which may be made of the information contained therein.

References

1. Abu-Oda, G.S., El-Halees, A.M.: Data mining in higher education: university student dropout case study. Data mining in higher education: university student dropout case study **5**(1) (2015)
2. Alvarracín, P., Daniel, J.: Identificación del perfil de egreso correspondiente a la licenciatura de la carrera de laboratorio clínico e histotecnológico de la universidad central del ecuador periodo 2017-2022 (2016)
3. Ameri, S., Fard, M.J., Chinnam, R.B., Reddy, C.K.: Survival analysis based framework for early prediction of student dropouts. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 903–912. ACM (2016)
4. Aulck, L., Velagapudi, N., Blumenstock, J., West, J.: Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364 (2016)
5. Baneres, D., Rodriguez-Gonzalez, M.E., Serra, M.: An early feedback prediction system for learners at-risk within a first-year higher education course. IEEE Transactions on Learning Technologies (2019)
6. Barbu, M., Vilanova, R., Lopez Vicario, J., Pereira, M.J., Alves, P., Podpora, M., Ángel Prada, M., Morán, A., Torreburno, A., Marin, S., et al.: Data mining tool for academic data exploitation: literature review and first architecture proposal. Projecto SPEET-Student Profile for Enhancing Engineering Tutoring (2017)
7. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
8. Catalano, V., Murcia Pérez, D.E., Escudero, F., Gerlo, P.D., Pantoja, A.: Influence of a tutorial system based on the use of ict in the decrease of dropout and academic failureof first-year students of the veterinary career of the juan agustín maza university. In: I Jornadas de Inclusión de Tecnologías Digitales en la Educación Veterinaria (La Plata, 2018) (2018)
9. Cevallos, T.: Cuadernos del contrato social por la educación. Cuaderno **10**, 34–46 (2014)

10. Chen, J., Feng, J., Sun, X., Wu, N., Yang, Z., Chen, S.: Mooc dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine. Mathematical Problems in Engineering **2019** (2019)
11. Chen, R.: Financial aid and student dropout in higher education: A heterogeneous research approach. In: Higher education, pp. 209–239. Springer (2008)
12. Chen, R.: Institutional characteristics and college student dropout risks: A multilevel event history analysis. Research in Higher education **53**(5), 487–505 (2012)
13. Dalipi, F., Imran, A.S., Kastrati, Z.: Mooc dropout prediction using machine learning techniques: Review and research challenges. In: 2018 IEEE Global Engineering Education Conference (EDUCON). pp. 1007–1014. IEEE (2018)
14. Ferreyra, M.M., Avitabile, C., Botero Álvarez, J., Haimovich Paz, F., Urzúa, S.: At a crossroads : Higher education in latin america and the caribbean (2017)
15. Jimenez, F., Paoletti, A., Sanchez, G., Sciavicco, G.: Predicting the risk of academic dropout with temporal multi-objective optimization. IEEE Transactions on Learning Technologies (2019)
16. Kappe, R., van der Flier, H.: Predicting academic success in higher education: what's more important than being smart? European Journal of Psychology of Education **27**(4), 605–619 (2012)
17. Murphy, J.P., Murphy, S.A.: Get ready, get in, get through: Factors that influence latino college student success. Journal of Latinos and Education **17**(1), 3–17 (2018)
18. OECD: Education at a Glance 2019 (2019). https://doi.org/https://doi.org/https://doi.org/10.1787/f8d7880d-en, https://www.oecd-ilibrary.org/content/publication/f8d7880d-en
19. Paura, L.: Cause analysis of students' dropout rate in higher education study program. Procedia - Social and Behavioral Sciences **109**, 1282–1286 (01 2014). https://doi.org/10.1016/j.sbspro.2013.12.625
20. Sinchi Nacipucha, E.R., Gómez Ceballos, G.P.: Access and desertion in universities. financing alternatives. ALTERIDAD. Revista de Educación **13**(2), 274–287 (2018)
21. Suganya, S., Narayani, V.: Analysis of students dropout forecasting using data mining. In: 3rd Internaational Conference on Lastest Trends in Engineering, Science, Humanities and Management (2017)
22. Ting, S.M.R., Man, R.: Predicting academic success of first-year engineering students from standardized test scores and psychosocial variables. International Journal of Engineering Education **17**(1), 75–80 (2001)
23. Viloria, A., Lezama, O.B.P.: Mixture structural equation models for classifying university student dropout in latin america. Procedia Computer Science **160**, 629–634 (2019)
24. Viloria Silva, A.J., Parody, A.: Methodology for obtaining a predictive model academic performance of students from first partial note and percentage of absence (2016)
25. Vossensteyn, J.J., Kottmann, A., Jongbloed, B.W., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., Wollscheid, S.: Dropout and completion in higher education in europe: Main report (2015)