Research Article/Engineering Education

http://doi.org/10.15446/ing.investig.109389

An actionable learning path-based model to predict and describe academic dropout

Un modelo accionable basado en el camino de aprendizaje para predecir y describir la deserción académica

Cristian Olivares-Rodríguez[1], Pedro Manuel Moreno-Marcos[2], Eliana Scheihing Garcia[3], Pedro J. Muñoz-Merino[4], and Carlos Delgado-Kloos[5]

ABSTRACT

The prediction and explainability of student dropout in degree programs is an important issue, as it impacts students, families, and institutions. Nevertheless, the main efforts in this regard have focused on predictive power, even though explainability is more relevant to decision-makers. The objectives of this work were to propose a novel explainability model to predict dropout, to analyze its descriptive power to provide explanations regarding key configurations in academic trajectories, and to compare the model against other well-known approaches in the literature, including the analysis of the key factors in student dropout. To this effect, academic data from a Computer Science Engineering program was used, as well as three models: (i) a traditional model based on overall indicators of student performance, (ii) a normalized model with overall indicators separated by semester, and (iii) a novel configuration model, which considered the students' performance in specific sets of courses. The results showed that the configuration model, despite not being the most powerful, could provide accurate early predictions, as well as actionable information through the discovery of critical configurations, which could be considered by program directors could consider when counseling students and designing curricula. Furthermore, it was found that the average grade and rate of passed courses were the most relevant variables in the literature-reported models, and that they could characterize configurations. Finally, it is noteworthy that the development of this new method can be very useful for making predictions, and that it can provide new insights when analyzing curricula and and making better counseling and innovation decisions.

Keywords: academic trajectory, student model, dropout, explainability, curriculum analysis

RESUMEN

La predicción y explicabilidad de la deserción estudiantil en programas académicos es un asunto importante, pues impacta a estudiantes, familias e instituciones. Sin embargo, los principales esfuerzos en este sentido se han centrado en el poder predictivo, aunque la explicabilidad es más relevante para los tomadores de decisiones. Los objetivos de este trabajo fueron proponer un modelo novedoso de explicabilidad para predecir la deserción, analizar su poder descriptivo para proporcionar explicaciones sobre configuraciones clave en trayectorias académicas y comparar el modelo con otros enfoques bien conocidos en la literatura, incluyendo el análisis de los factores clave en la deserción estudiantil. Para ello, se utilizaron datos académicos de un programa de Ingeniería en Informática, así como tres modelos: (i) un modelo tradicional basado en indicadores generales de rendimiento estudiantil, (ii) un modelo normalizado con indicadores generales separados por semestre y (iii) un modelo de configuración novedoso que considera el rendimiento de los estudiantes en conjuntos específicos de cursos. Los resultados mostraron que el modelo de configuración, a pesar de no ser el más poderoso, podría proporcionar predicciones tempranas precisas, así como información accionable a través del descubrimiento de configuraciones críticas, las cuales podrían ser consideradas por los directores de programa al asesorar a los estudiantes y diseñar planes de estudio. Además, se encontró que la nota promedio y la tasa de cursos aprobados fueron las variables más relevantes en los modelos reportados en la literatura, y que estas podrían caracterizar configuraciones. Finalmente, es notable que el desarrollo de este nuevo método puede ser muy útil para hacer predicciones y que puede proporcionar nuevas perspectivas al analizar planes de estudio y al tomar mejores decisiones de asesoramiento e innovación.

Palabras clave: trayectoria académica, modelo de estudiantes, deserción, explicabilidad, análisis curricular

Received: June 7th 2023
Accepted: September 30th 2023

Introduction

Year by year, university decision-makers counsel thousands of students who are not able to manage the contents of their courses, with some of them even dropping out of courses and/or the whole program. Statistics report that about 30-35% of the students drop out of a degree (Paura and Arhipova, 2014; Vidal *et al.*, 2022).

Particularly in Chile, many students drop out, and many of those who do not spend a lot of time finishing their degree, especially their first two years (commonly known as *bachillerato*) (Donoso *et al.*, 2013). There are many reasons for dropping out, such as prior academic preparedness (Smith and Naylor, 2001; Carvajal *et al.*, 2018), poor academic results, lack of funding, and loss of interest

[1] Facultad de Ingeniería, Universidad Alberto Hurtado and Member of the Centro de Estudios en Ciencia, Tecnología y Sociedad, CECTS, Chile, colivares@uahurtado.cl
[2] Universidad Carlos III de Madrid, Spain, pemoreno@it.uc3m.es
[3] Universidad Austral de Chile, Chile, escheihi@uach.cl
[4] Universidad Carlos III de Madrid, Spain, pedmume@it.uc3m.es
[5] Universidad Carlos III de Madrid, Spain, cdk@it.uc3m.es

(Breier, 2010; Li and Carroll, 2020). Therefore, decision-makers require an early understanding { not just a numerical prediction { of why students fail during the program in order to enhance their learning in the long term. They also need clear guidelines to act in these situations.

To understand students' academic behavior, data mining techniques can be used (Palacios *et al.*, 2021; Chung and Lee, 2019). The development of predictive models can provide significant benefits to different stakeholders. Students improve in self-reflection regarding their learning, which facilitates their decision-making. Teachers focus on the students at risk and try to adapt the course and methodology. Finally, university decision-makers can redesign curricula to balance workload, improve personalized counseling, and potentially enhance the learning process.

Some works have attempted to analyze dropout as an academic behavior. Bottcher *et al.* (2020) found that the number of credits passed in the first semester and the frequent changes made to bachelor's degree programs are significant indicators. Hutt *et al.* (2018) analyzed other variables such as personal factors, work experience, and academic tests. However, one of the less studied factors that might influence dropout is the set of courses that a student takes, hereafter called *configurations* (Hutt *et al.*, 2018). If predictions are made with regard to these configurations, a lot of benefits can be reaped. Moreover, when managers decide whether a learner can take a course, they can use dropout information to drive their counseling, leading students into an enhanced learning environment.

The analysis of students in the field of education can be carried out at the course level (*e.g.*, predicting student dropout from a course) or at the program level (*e.g.*, predicting student dropout from a degree). At the course level, Jiang and Li (2017) predicted dropout by using ensemble learning methods. Meanwhile, Moreno-Marcos *et al.* (2019) predicted student success in an admission test based on interactions in an edX-based blended program. There are also many contributions to prediction in online programs (Gardner and Brooks, 2018; Moreno-Marcos *et al.*, 2018; Kang and Wang, 2018; Jin, 2021; Mubarak *et al.*, 2021), as well as in blended contexts, such as the in-session model proposed by Rzepka *et al.* (2022).

Regarding the program level, there have also been plenty of relevant works. Yu *et al.* (2021) analyzed how important it is to include sensitive attributes (*e.g.*, gender, underrepresented minorities, *etc.*) in student modeling. They concluded that these attributes can only provide a marginal improvement and do not significantly impact predictive power. Dekker *et al.* (2009) predicted dropout after the first semester using several algorithms, concluding that academic data offer higher predictive power than pre-university features. Furthermore, Delen (2011) found that, while the most important variables are academic in nature, financial variables are also relevant predictors. Similarly, Quadri and Kalyankar (2010) found that low-income levels have a strong influence on dropout, while demographics are irrelevant factors.

Lázaro Alvarez *et al.* (2020) found that accuracy significantly improves after the first semester, suggesting that early predictions can be made. Berens *et al.* (2018) also analyzed this issue, reporting that accuracy could reach 79-85% after the first semester and 90-95% after the fourth one, which also suggests the possibility of early prediction. In another study (Gašević *et al.*, 2016), different values were obtained for different degrees, indicating the importance of course context and generalizability analysis. Meanwhile, Wagner *et al.* (2020) presented a preliminary study and reported good cross-program models, and Panagiotakopoulos *et al.* (2021) provide an early prediction method for massive open online courses (MOOC). Nevertheless, further research is needed in this direction, as is described in Moreno-Marcos *et al.* (2019).

This work proposes an innovative method to analyze dropout as academic behavior via variables related each set of courses taken simultaneously by a student. The aim is for this model to offer information about curricula, *i.e.*, which configurations can be taken by a student so that dropout can be reduced. This is an aspect that has not been considered in previous works. This paper contributes by analyzing how critical configurations can be obtained and how they can be used by decision-makers to analyze student behavior. Finally, this paper addresses relevant topics in the literature, such as temporal analysis to discover the moment at which early predictions are accurate enough to be used for causing a positive effect on students' learning.

This paper aims to analyze different models for predicting dropout, including a novel proposal based on configurations. To this effect, several specific objectives have been defined:

- Objective 1: To propose a prediction model based on course configurations
- Objective 2: To compare the capabilities of different models for predicting dropout
- Objective 3: To analyze the suitability of the configurations-based model to detect a set of courses that are likely or unlikely to produce or prevent dropout
- Objective 4: To analyze the key factors that affect dropout in different models

Methodology

Dropout models

This section describes the models used to analyze dropout. First, a novel model is proposed, aming to describe student trajectories while considering the set of courses (configurations) taken by students throughout their degree (objective 1). Additionally, we outline two traditional models based on student performance variables for comparison purposes.

Configuration model

Decision-makers analyze student behavior by exploring their academic trajectories, in order to support their advice and innovations. This work proposes a novel dropout analysis model based on the temporal course configurations that students take simultaneously, rather than considering opaque student behavior in integrated variables, as in traditional models.

The model is operationalized through some mathematical expressions. A degree is implemented via a program

curriculum (P), which is composed of a set of semester-based courses (α). Every student draws their academic trajectory (t) with these sets of courses, also defined as *configurations* (c). For example, a student i takes a first configuration c_1^i during their first semester. Afterwards, he takes a second configuration c_2^i, and so on. The resulting trajectory (t^i) is presented in Equation 1, which contains the ordered sequence of configurations. It is noteworthy that, when a student drops out, it implies a trajectory that does not lead to program completion, which must be analyzed nonetheless.

$$t^i = c_1^i \Rightarrow c_2^i \Rightarrow \cdots \Rightarrow c_p^i \tag{1}$$

Every academic program P has a finite set of configurations c, and each student in a program can select only one configuration per semester. Thus, it is possible to model a program through a trajectory matrix (Mt), which relates students with configurations, as shown in Equation 2.

$$M_t^p = \begin{bmatrix} W_{1,1} \cdots W_{1,N} \\ \cdots \\ W_{M,1} \cdots W_{M,N} \end{bmatrix} \tag{2}$$

In this matrix, each column represents every possible configuration c, and each row represents a student i who is enrolled in the program P. The content of each cell represents a measurement based on several weights ($w_(i, j)$), indicating the performance of the student i in the configuration c_j. The value of this measurement is 0 when the student does not select a configuration. Each row is a vector that describes the performance of a student throughout their academic trajectory, considering all configurations. The measurement w, which captures most of a student's behavior, is presented in the subsection called *Variables and techniques*.

Traditional model

This model considers a set of global variables $[X1, X2, \cdots, Xn]$, which are obtained by integrating data over the entire analysis period. Regardless of the length of the student's trajectory, the number and nature of the variables will be the same, even though they will be computed with more or less data. If there are N independent variables, one dependent variable, and M students, the matrix with the training set will always have $(N + 1) * M$ elements. An example of this matrix is presented in Equation 3, where $X_{i,j}$ represents the variable j of student i, and Y_i denotes their dependent variable (dropout).

$$Trad = \begin{bmatrix} X_{1,1} \cdots X_{1,N} Y_1 \\ \cdots \\ X_{M,1} \cdots X_{M,N} Y_M \end{bmatrix} \tag{3}$$

This is regarded as a traditional model because most contributions use a similar approach when developing predictive models. Although this approach can be effective in many scenarios, one of its drawbacks is that academic trajectories become opaque.

Normalized model

This model aims to enhance traditional approaches by considering the evolution of a student throughout the degree. To this effect, this model calculates the independent variables for each semester. If there are N variables and the student has taken six semesters, there will be $6 * N$ independent variables. As some students may have taken more semesters than others, the variables of the academic terms that the student has not yet taken are set as 0, in order to make the matrix consistent.

If there are N independent variables, one dependent variable, S possible semesters, and M students, the number of elements in the matrix will be $(N * S + 1) * M$. This number increases with the number of semesters, providing temporal information about a student's performance/behavior. An example of the matrix used for the training set is presented below. Independent variables are named $X_{i,j,k}$, where j indicates the number of the variable (from 0 to N), k the semester (from 0 to S), and i the number of the student (from 0 to M). The dependent variables are named Y_k, as they only depend on the student, as shown in Equation 4.

$$Norm = \begin{bmatrix} X_{1,1,1} \cdots X_{1,2,1} \cdots X_{1,N,S} Y_1 \\ \cdots \\ X_{M,1,1} \cdots X_{M,2,1} \cdots X_{M,N,S} Y_M \end{bmatrix} \tag{4}$$

By following this approach, it is possible to observe how a student evolves throughout a degree program, but their decisions regarding the configurations taken are still opaque and differ from one student to another. For example, student A in semester 6 might have taken a different configuration from student B in the same semester. In this case, the model is said to be *normalized*, since the variables for each semester are computed independently of the courses taken; they are normalized to the number of semesters.

Dataset

This study has been conducted using academic data from one degree program (Computer Science Engineering) at Universidad Austral de Chile (UACh), in the context of the Erasmus+ LALA project. The data collection period goes from 2011 to 2017. This period defines a complete cohort of students attending the same academic program (without changes) and sets of courses. The dataset is composed of records of 479 students, with 50.73% of student dropout. The mean number of semesters per student is 5.54 (SD: 3.76) and the mean number of courses per semester is 5,77 (SD: 2.78). At the end of every semester, each student must decide which set of courses they will take during the next term while considering several program constraints and their trajectory. Trajectories become more and more diverse as students advance through the program, as more decisions regarding courses become necessary.

The data were obtained only from the SIS (Student Information System). Only academic data were considered (*e.g.*, grades in different courses). This fact, despite being a limitation, can showcase the potential of dropout analysis without using many different data types (*e.g.*, without interaction logs of learning platforms). If models can work accurately enough with SIS data, institutional adoption will be easier. In this case, the data are in the form of a table with students' grades for different courses.

Table 1. Example of the data collected from UACh

ID	Course	Year	Semester	Concept	Grade
ID1	Course A	2011	1	TAKEN	5.2
ID1	Course B	2011	1	TAKEN	2.5
ID1	Course C	2011	2	CANCELED	0.0
ID2	Course A	2012	1	VALIDATED	6.1

Source: Authors

As shown in Table 1, there is one column for the student (ID), another for the course name (course), and two columns to identify the period during which the course was taken (year and semester, which can only be 1 or 2). The concept of a course indicates whether the student took the course (taken) or if it was canceled (taken and not finished, withdrawn), validated (taken in another degree program), recognized (taken at another institution), or retrieved from another curriculum (when the student has changed their curriculum). An enrolled course comprises all these situations. The last column indicates the student grade from 0 to 7, with 4,0 the passing score.

Dropout setup

To develop the predictive model, it was necessary to define how to determine dropout. It is worth mentioning that a ground truth was not available to train the models. As only data from 2011 to 2017 were available, a first criterion was defined to filter out students who started long before 2011 or just before the end of the data collection period.

A degree is composed of 11 semesters (including the final project). However, the first four semesters (*bachillerato*) only contain basic training courses (*e.g.*, Algebra, Programming, Communication Skills, *etc.*). In general, according to historical data, students who finish the *bachillerato* have a high probability of finishing the degree. To increase the number of cohorts and obtain sufficient data to train the model, we considered that a student has not dropped out when they have a non-*bachillerato* configuration, *i.e.*, a set of courses that do not contain any course from the first two years. This means that the students have finished the first two years in their program; otherwise, they will report at least one course from these years in their configuration. In contrast, if a student only has *bachillerato* configurations (a set of courses with at least one belonging to the first two years), it means that they have not finished the first four semesters. In some cases, these students were discarded. Note that they might not have the four *bachillerato* configurations because they might have enrolled in the program recently.

Several filters to classify and discard students were considered. First, the students who enrolled before 2011 were removed because their initial configurations were missing, as this was the start of the analyzed period. Afterwards, the students with at least one non-*bachillerato* configuration were labeled as *no dropout* since, as explained above, those who finish *bachillerato* are very likely to finish the degree. Finally, for students who did not have non-*bachillerato* configurations, it was necessary to distinguish those who were currently attending the *bachillerato* from those who had dropped out. In this vein, students were considered dropouts if they were not enrolled in the last year of the data collection period (2017), since they had started the *bachillerato* but had not finished it; otherwise, they would have been enrolled in the program and would later take non-*bachillerato* configurations. Those who were enrolled in 2017 were discarded, as they were attending the *bachillerato* at that moment and there was no dropout information. Figure 1 shows the flowchart of the process for determining dropout with the above-mentioned criteria. A sample of N = 270 students was obtained (after discarding students) for the analysis.

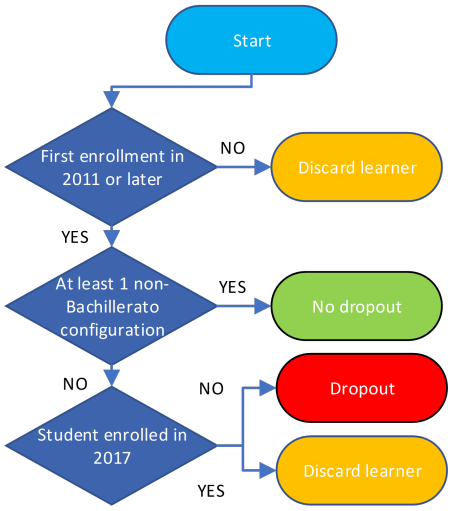

Figure 1. Flowchart with the criteria to flag students as dropouts
Source: Authors

Variables and techniques

To carry out the analyses, it was important to define the set of variables that could be obtained from the previously described dataset. Table 2 presents the full list of variables considered in this study. Some variables were directly taken from the indicators used in the LALA project (Muñoz-Merino *et al.*, 2020). Some variables were used in all models, although others were only tested in the configuration model, which required further tuning given the need to define a weight function.

Table 2. Variables used for the traditional (T), normalized (N), and configurations (C) models

Variable	Model	Description
avg_grade	T, N, C	It indicates the average grade of the enrolled courses.
r_passed	T, N, C	It indicates the relationship between passed and enrolled courses.
r_taken	T, N	It indicates the relationship between the courses taken and enrolled.
r_takreg	C	It indicates the relationship between taken and recognized courses and those enrolled.
r_cancel	T, N	It indicates the relationship between canceled and enrolled courses.
repeat	T, N	It indicates the relationship between the number of courses a student takes and the total number of attempts. If a student takes three courses for the first time and one for the second, the value is 4/(1+1+1+2)=0,8.

Source: Authors

The T and N models use these variables directly. Meanwhile, the C model aggregates such variables via a performance function w, assessing every student's behavior in every particular configuration. To compute this function,we used the formula presented in Equation 5.

$$w^{i,j} = (avg\_grade + r\_takreg) * r\_passed \quad (5)$$

In this work, the *caret* library of the R open-source software was used to develop the predictive models. Particularly, the following algorithms were used: i) random forests (RF), ii) the generalized linear model (GLM), iii) support vector machines (SVMs), iv) decision trees (DT), and (5) single-hidden-layer neural networks (NNs). The hyperparameters of these models were discovered during the training phase. In addition, ten-fold cross-validation was used to validate the results (five-fold cross-validation was also used to compare the results, although no significant changes were observed, so only those related to 10-fold cross-validation are presented), and the area under the curve (AUC) was the metric computed to evaluate the predictive power of the models, as has been done in previous works (Pelánek, 2015; Jeni *et al.*, 2013).

Results

First, this section presents an analysis of dropout behavior, comparing the three models and including a temporal analysis (related to objective 2). Next, we delve into the configuration model to determine whether it provides descriptive value regarding the academic trajectories and configurations that are more or less likely to enhance the learning context (related to objective 3).

Predictive power

We analyzed how early it was possible to predict dropout and how the aforementioned predictive models behaved with small amounts of data. The dropout models by only using data related to the *bachillerato* (four semesters). The data context was named *xS* (*e.g.*, 1S, 2S, *etc.*), where x indicates the number of semesters considered for developing the model.

Table 3. Temporal analysis of dropout prediction (results expressed in AUC)

Alg	Traditional				Normalized				Configurations			
SEM.	1S	2S	3S	4S	1S	2S	3S	4S	1S	2S	3S	4S
LR	**0.92**	0.92	**0.93**	**0.95**	**0.92**	**0.94**	0.95	0.96	0.73	0.69	0.75	0.54
DT	0.78	0.85	0.84	0.89	0.78	0.79	0.90	0.86	0.77	0.77	0.77	0.78
RF	0.90	**0.93**	0.92	0.94	0.90	0.92	0.94	**0.97**	0.77	0.77	0.80	0.81
SVM	0.88	0.90	0.91	0.94	0.88	0.91	0.95	**0.97**	**0.85**	**0.86**	**0.88**	**0.82**
NN	0.91	0.92	**0.93**	**0.95**	**0.92**	**0.94**	**0.96**	**0.97**	0.79	0.71	0.78	0.81

Source: Authors

Table 3 shows the strength of the proposed model, which performs well (AUC≥0.8) for every semester and is useful to analyze critical configurations. This is not possible in the other models. The N and T models provide high predictive power but no actionable information. Particularly, the N model can yield the most accurate results. This implies that separating the variables by semester can provide further information, despite increasing dimensionality, as performance can vary over time. This was checked with the correlation between variables in different semesters. In the case of the average grade, the correlations were between 0.56 and 0.81 (with the latter being the highest correlation between the same variable in different semesters), which shows that there is a high correlation, but that further information can still be added.

Apart from that, a surprising result is that the performance of the C model deteriorates after the third semester, while the other two models improve, which is the usual and expected behavior when adding new data. A possible reason for this is that the number of configurations to be considered in the model increases as new semesters are incorporated. This increases the number of variables, and the model becomes sparser, losing its predictive power as reported in Kohavi and John (1997) for similar situations. This implies that the C model behaves better in the first stages, when the number of configurations is manageable. The positive aspect is that these stages are the most relevant, as most dropouts occur early. Nevertheless, the C model can be used to gather actionable information about critical configurations to help decision-makers and validate the predictions of other algorithms.

In terms of anticipation, the results show that the predictive power of all models is very good since the first semester, which is consistent with Lázaro Alvarez *et al.* (2020). This means that it is possible to obtain accurate early predictions with the first results of the students, which is useful to anticipate performance and support academic decisions.

As for the algorithms, NN was the most consistent for both the T and N models, although SVM stood out in the C model, as happened in Tekin (2014). Nevertheless, the differences were not great. There was a predominance of SVMs in the C model, perhaps due to their capability to handle problems with high dimensionality (Bersimis and Varlamis, 2019).

In summary, the proposed models can provide accurate predictions from the beginning, so it is possible to anticipate student behavior with regard to the issue under study. Nevertheless, the accuracy of the N model stands out, and the C model provides promising results and new insights, since it does not only analyze overall performance, but also that in specific sets of courses.

Critical configurations

It is particularly relevant to provide actionable information about students, and the C model was designed to track both temporal relations and performance in course configurations at any given moment. Thus, this model aids in determining the most critical course configurations.

The importance of the configurations was computed through the Mean Decrease Gini, which is commonly used to evaluate the importance of features (Louppe *et al.*, 2013). This metric was computed with data from one to four semesters. The three most relevant configurations (among 2127 possible configurations) for each semester are shown in Table 4.

In light of the above, decision-makers must pay attention to the most relevant configurations and, more precisely, to the sets of courses appearing at the top in more than one of the rankings provided by the classifiers. Particularly, configurations 29, 30, 31, 46, and 97 are the most critical.

Table 4. Most critical configurations

SEM	Most relevant	2nd most relevant	3rd most relevant
1S	S29	S97	S46
2S	S29	S30	S97
3S	S29	S30	S31
4S	S30	S29	S31

Source: Authors

These configurations have the following in common: (i) they contain initial programming courses (five out of five) or initial mathematics courses, such as Algebra or Calculus (four out of five); and (ii) they have a large number of courses per semester (six or more). The negative results of the students taking these configurations implies that programming and mathematics courses are becoming harder to pass. This is consistent with other works, such as that of Figueiredo and García-Peñalvo (2021), who highlighted the difficulty of programming courses. Moreover, this could imply that students who take many courses can have a very high workload and may have difficulties in dealing with it, as was also reported in other works (Radovanović *et al.*, 2021). To prevent this, decision-makers (*e.g.*, program directors) should not allow students to take configurations with such a high workload, in addition to preventing them from taking specific courses in the same semester, as these configurations may lead to dropout. In any case, it is important to note that decision-makers should interpret this actionable information and consider possible special cases or circumstances to support their decisions.

Key factors affecting dropout

The models can be further analyzed to discover what features are more relevant or actionable for program directors. The importance of the variables was initially computed for the T and N models using the same methodology presented in the previous section. The results shown in Figures 2 and 3 indicate that the most important variables for the T model are the average grade, as reported by Delen (2011), followed by the ratio of passed courses. These two variables are very clear indicators of performance, since students who fail repeatedly or have low grades are more likely to drop out. It is also interesting that the repeat variable becomes relevant after the second semester. In the first two semesters, it is not important, as students do not repeat courses, but once they start to do so, this variable takes on great significance.

In the N, similar variables are relevant, as the average grade and the ratio of passed courses stand out. Nevertheless, additional patterns can be found when considering individual variables. For example, the average grades of the first semesters became very relevant over time. The importance of the average grade of the first semester does not significantly differ from that of the average grade of the second semester, also showing high values in the third semester. In contrast, variables from previous semesters become less relevant in the fourth semester. Like the average grade, the ratio of passed courses shows a similar behavior, although it is surprisingly the most relevant variable in the fourth semester. The repeat variable shows a similar behavior to the T model, except in the fourth semester.

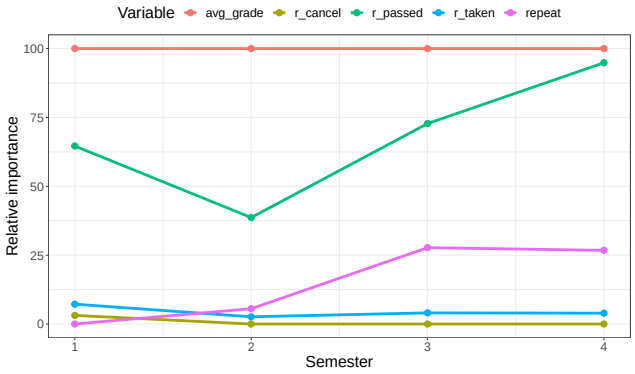

Figure 2. Evolution of the importance of variables in the T model
Source: Authors

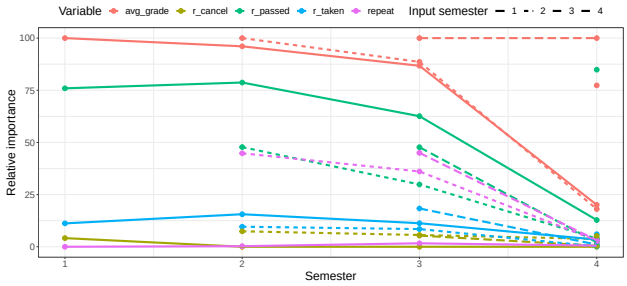

Figure 3. Evolution of the importance of variables in the N model
Source: Authors

Furthermore, it is interesting to see how the values of these relevant features are related to critical configurations. To analyze this issue, the top three features in the aforementioned models were computed at the configuration level, averaging the results of all the students who took each configuration. In this case, as knowing the dropout level was not necessary, all 695 students in the dataset were considered (we removed the filters to flag dropout presented in Figure 1). The results for the top 20 critical configurations of the first four semesters (there are 32 in total) are shown in Figure 4.

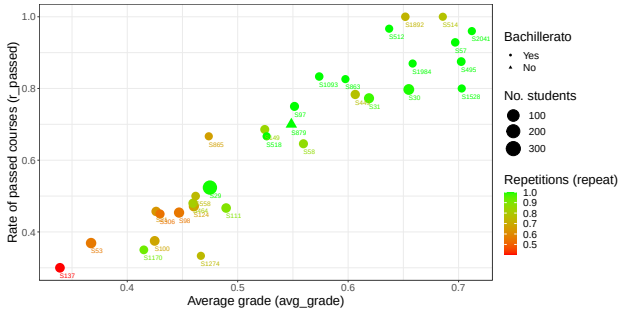

Figure 4. Characterization of the top critical configurations in the C model
Source: Authors

This Figure indicates that not all critical configurations are negative, but actionable. There are some configurations where most of the students obtain low grades and fail most of the courses (*e.g.*, S137) while there are others with very good results. This is an important finding, since program directors can not only detect configurations where students

may struggle, but also those where it might be easier to succeed.

Moreover, S29 is the most frequent configuration (320 students took it), and, although it does not contain repeated courses (repeat=1 means that all courses are taken for the first time, as shown in Table 2), many students get poor results. This actionable information serves to provide insights to program directors regarding the workload or the possible reasons behind taking this configuration.

Additionally, a hierarchical clustering analysis was conducted to discover whether there were any patterns or distinctions between the configurations. This was done by using the variables avg_grade, r_passed, r_taken, r_cancel, and repeat at the configuration level. The optimal number of clusters was computed via Silhouette analysis. Figure 5 shows the clusters obtained.

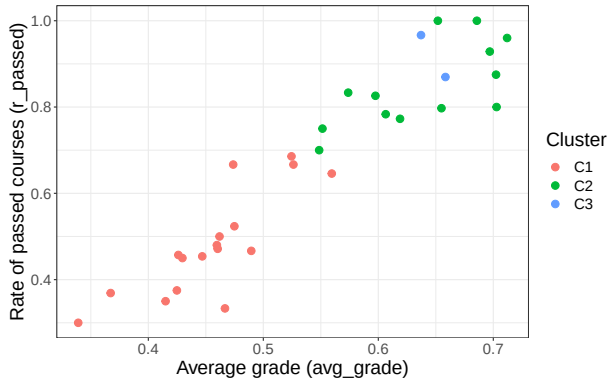

Figure 5. Hierarchical clustering of the top critical configurations
Source: Authors

Two clusters are relevant: the red one and the green one. The red cluster shows configurations with poor results and low average grades and rates of passed courses. Meanwhile, in the green cluster, students usually obtain good results and high average grades and rates of passed courses. The repeat variable is also usually high (students do not repeat courses). Therefore, red configurations should be avoided, and green ones may be encouraged by program directors. Finally, there is a special cluster (the blue one) with two configurations where all students validate all courses, which is not relevant in terms of student performance. An analysis of the the 17 red configurations reveals interesting facts. In six configurations (29, 149, 306, 464, 518, and 1170), all courses belong to the same semester in the program, but the average grade in most of the courses is below 4, implying that these configurations include difficult courses. In addition, there are seven configurations (53, 58, 98, 100, 111, 137, and 1274) where all courses belong to the same semester except one. In all those cases except one, most of the courses show an average grade below 4, implying that they are difficult configurations and the students may not perform as expected. Finally, there are four configurations (21, 124, 558, and 865) where students have courses from different semesters. In those cases, it is likely that student performance also affects the configuration. Further analysis should be conducted in this regard.

Finally, considering all the results and the actionable information provided by the C model, the flow of actions in an institution could be as follows:

1. Making predictions using the different models (as all of them are quite accurate).
2. Identifying the most critical configurations with the C model.
3. Computing the main features and identifying the cluster they belong to.
4. Program directors could review the predictions and support the students' decisions. To this effect, a dashboard could be implemented to provide this information, as in Guerra *et al.* (2019).

If these steps are taken, both descriptive and predictive information based on configurations could be used to provide better counseling to students, which may in turn result in better graduation rates.

Conclusions

In this article, a novel dropout analysis model was proposed as a method to support decision-makers during student counseling and/or curriculum innovations. Two models were compared against our proposal, upon the basis of students' performance in each specific set of courses. All models exhibited a high predictive power, although the normalized model stood out. Nevertheless, the configuration model (the proposed approach) implies a reasonable trade-off between prediction power and actionable information about critical configurations. The main findings are summarized below:

- The configuration model has a strong predictive power, particularly in the first semesters, so it can be used for early dropout prediction. However, it may face issues in later stages due to the diversity of configurations, as reported in other works (Kohavi and John, 1997), which implies too many features).
- Predictive power is very high for all studied models since the first semester, as was previously reported by Lázaro Alvarez *et al.* (2020) and by Berens *et al.* (2018) to a lesser extent. This means that it is possible to obtain accurate early predictions once the first results of the students are available.
- The configuration model aids in identifying critical configurations that program directors can consider in their counseling sessions. Program directors may warn students about negative configurations and recommend courses, providing another perspective, different from other traditional recommender systems, as is the case of Denley (2014)).
- An analysis of critical configurations shows that initial programming and mathematics courses may cause difficulties for students and configurations with more courses than expected (heavy workload). This is consistent with other works highlighting the difficulties of programming courses (Figueiredo and García-Peñalvo, 2021) or the students' issues with heavy workloads (Radovanović *et al.*, 2021).
- The most relevant variables in the models are the average grade of the courses, which is consistent with

Delen (2011); the rate of passed courses; and, to a lesser extent, a variable indicating if the student is repeating many courses. In addition, relevant variables are usually also relevant even in subsequent semesters (*e.g.*, the average grade in semester 1 was highly important in the first three semesters).

- The most relevant variables enabled a cluster analysis to identify positive and negative configurations.

In light of the above, the proposed model provides actionable information for decision-makers to enhance the learning context. If a dashboard with all this information is given, directors could easily identify students at risk and analyze their trajectories to better guide them in their learning path, which may in turn help to reduce dropouts. In addition, directors can further reflect on how the program is designed, so that the configurations for each semester are better balanced and the workload can be better managed by students.

One limitation of this proposal is that, as completing a degree takes several years and information from only the last seven cohorts (years) is available, data on completion are not available for some students. This means that the dropout definition (which is not unique, unlike a numerical grade, for example) had to be adapted, likely affecting the results. Furthermore, only data for one degree program were used, and more variability would be needed to compare different contexts and analyze the generalizability of the models. Nevertheless, these models can be easily adapted to other programs, although they may require training.

In future work, the addition of new variables, such as the student's financial situation, mental health, change of interest, personal reasons, and specific grades in critical courses, as well as the general features of the configurations (*e.g.*, popularity/frequency), may be relevant to the understanding of dropout and the improvement of the models. Further analyses involving interpretable machine learning tools such as those presented in Nagy and Molontay (2023) could be performed in order to better understand the models. Moreover, putting predictions in a dashboard and evaluating how programs use them would be a relevant contribution. In this vein, a scheme could be developed to ensure that directors get proper feedback and provide adequate counsel to students. This scheme would also allow students to provide feedback if they eventually drop out, and it would help to detect special situations that are useful for enhancing the models and identifying critical configurations. With regard to these configurations, more research is needed in order to obtain a better understanding and to determine whether poor results are related to course difficulty or student performance.

Acknowledgements

This work was partially funded by the LALA project (grant 586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), by FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación through project H2O Learn (grant PID2020-112584RB-C31), and by the Spanish Ministry of Science, Innovation, and Universities under an FPU fellowship (FPU016/00526). The LALA project has been funded with support from the European Commission. This publication only reflects the views of the authors, and the Commission and the Agency, as well as other funding entities, cannot be held responsible for any use of the information contained therein.

CRediT author statement

Cristian Olivares-RodrIguez: conceptualization, formal analysis, investigation, methodology, software, supervision, and writing (original draft). *Pedro Manuel Moreno-Marcos*: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, and writing (original draft). *Pedro J. Muñoz-Merino*: conceptualization, formal analysis, funding acquisition, investigation, supervision, and writing (review and editing). *Eliana Scheihing Garcia*: conceptualization, investigation, methodology, and writing (review and editing). *Carlos Delgado Kloos* provided critical feedback.

Conflicts of interest

The authors declare no conflict of interest.

References

Berens, J., Schneider, K., Gortz, S., Oster, S., and Burghoff, J. (2018). *Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods*. CESifo Working Paper. Retrieved from https://ssrn.com/abstract=3275433

Bersimis, F. G., and Varlamis, I. (2019). Use of health-related indices and classification methods in medical data. In N. Dey (Ed.), *Classification techniques for medical image analysis and computer aided diagnosis* (pp. 31–66). Elsevier. https://doi.org/10.1016/B978-0-12-818004-4.00002-9.

Bottcher, A., Thurner, V., and Hafner, T. (2020, April 27-30). *Applying data analysis to identify early indicators for potential risk of dropout in cs students*. [Conference paper]. 2020 IEEE Global Engineering Education Conference, Porto, Portugal. https://doi.org/10.1109/EDUCON45650.2020.9125378.

Breier, M. (2010). *Student retention and graduate destination: Higher education and labour market access and success*. HSRC Press Cape Town.

Carvajal, C. M., González, J. A., and Sarzoza, S. J. (2018). Variables sociodemográficas y académicas explicativas de la deserción de estudiantes en la facultad de ciencias naturales de la universidad de playa ancha (chile). *Formación universitaria*, *11*(2), 3–12. http://dx.doi.org/10.4067/S0718-50062018000200003.

Chung, J. Y., and Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, *96*, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030.

Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009, July 1-3). *Predicting students drop out: A case study*. [Conference paper]. 2nd International Conference on Educational Data Mining, Córdoba, Spain. https://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf.

Delen, D. (2011). Predicting Student Attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 17–35. https://doi.org/10.2190/CS.13.1.b.

Denley, T. (2014). How predictive analytics and choice architecture can improve student success. *Research & Practice in Assessment*, *9*, 61–69.

Donoso, S., Donoso, G., and Frites, C. (2013). La experiencia chilena de retención de estudiantes en la universidad. *Revista Ciencia y Cultura*, *17*(30), 141–171.

Figueiredo, J., and García-Peñalvo, F. (2021, October 26-29). *A tool help for introductory programming courses.* [Conference paper]. 9th International Conference on Technological Ecosystems for Enhancing Multiculturality, Barcelona, Spain. https://doi.org/10.1145/3486011.3486413.

Gardner, J., and Brooks, C. (2018). Student success prediction in moocs. *User Modeling and User-Adapted Interaction*, *28*, 127–203. https://doi.org/10.1007/s11257-018-9203-z.

Gašević, D., Dawson, S., Rogers, T., and Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002.

Guerra, J., Scheihing, E., Henríquez, V., Olivares-Rodríguez, C., and Chevreux, H. (2019, September 16-19). *TrAC: Visualizing students academic trajectories.* [Conference paper]. 14th European Conference on Technology Enhanced Learning, Delft, The Netherlands. https://doi.org/10.1007/978-3-030-29736-7_84.

Hutt, S., Gardener, M., Kamentz, D., Duckworth, A. L., and D'Mello, S. K. (2018, March 7-9). *Prospectively predicting 4-year college graduation from student applications.* [Conference paper]. 8th International Conference on Learning Analytics and Knowledge, Sydney, New South Wales, Australia. https://doi.org/10.1145/3170358.3170395.

Jeni, L. A., Cohn, J. F., and De La Torre, F. (2013, September 2-5). *Facing imbalanced data–recommendations for the use of performance metrics.* [Conference paper]. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland. https://doi.org/10.1109/ACII.2013.47.

Jiang, F., and Li, W. (2017). Who will be the next to drop out? anticipating dropouts in moocs with multi-view features. *International Journal of Performability Engineering*, *13*(2), 201–210. https://doi.org/10.23940/ijpe.17.2.p201.mag.

Jin, C. (2021). Dropout prediction model in MOOC based on clickstream data and student sample weight. *Soft Computing*, *25*, 8971–8988. https://doi.org/10.1007/s00500-021-05795-1.

Kang, K., and Wang, S. (2018, March 23-25). *Analyze and predict student dropout from online programs.* [Conference paper]. 2nd International Conference on Compute and Data Analysis, DeKalb, Illinois, USA. https://doi.org/10.1145/3193077.3193090.

Kohavi, R., and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X.

Lázaro Alvarez, N., Callejas, Z., and Griol, D. (2020). Predicting computer engineering students' dropout in cuban higher education with pre-enrollment and early performance data. *JOTSE: Journal of Technology and Science Education*, *10*(2), 241–258. https://doi.org/10.3926/jotse.922.

Li, I. W., and Carroll, D. R. (2020). Factors influencing dropout and academic performance: an australian higher education equity perspective. *Journal of Higher Education Policy and Management*, *42*(1), 14–30. https://doi.org/10.1080/1360080X.2019.1649993.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013, December 5-10). *Understanding variable importances in forests of randomized trees* (Vol. 1). [Conference paper]. 26th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, USA.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., and Kloos, C. D. (2018). Prediction in moocs: A review and future research directions. *IEEE transactions on Learning Technologies*, *12*(3), 384–401. https://doi.org/10.1109/TLT.2018.2856808.

Moreno-Marcos, P. M., De Laet, T., Muñoz-Merino, P. J., Van Soom, C., Broos, T., Verbert, K., and Delgado Kloos, C. (2019). Generalizing predictive models of admission test success based on online interactions. *Sustainability*, *11*(18), 4940. https://doi.org/10.3390/su11184940.

Mubarak, A. A., Cao, H., and Hezam, I. M. (2021). Deep analytic model for student dropout prediction in massive open online courses. *Computers & Electrical Engineering*, *93*, 107271. https://doi.org/10.1016/j.compeleceng.2021.107271.

Muñoz-Merino, P. J., Kloos, C. D., Tsai, Y.-S., Gasevic, D., Verbert, K., Pérez-Sanagustín, M., ... Scheihing, E. (2020, September 14-15). *An overview of the LALA project.* [Conference paper]. Workshop on Adoption, Adaptation and Pilots of Learning Analytics in Under-represented Regions co-located with the 15th European Conference on Technology Enhanced Learning 2020, Online. https://ceur-ws.org/Vol-2704/invited1.pdf.

Nagy, M., and Molontay, R. (2023). Interpretable dropout prediction: Towards XAI-based personalized intervention. *International Journal of Artificial Intelligence in Education*, 1–27. https://doi.org/10.1007/s40593-023-00331-8.

Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., and Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile. *Entropy*, *23*(4), 485. https://doi.org/10.3390/e23040485.

Panagiotakopoulos, T., Kotsiantis, S., Kostopoulos, G., Iatrellis, O., and Kameas, A. (2021). Early dropout prediction in MOOCs through supervised learning and hyperparameter optimization. *Electronics*, *10*(14), 1701. https://doi.org/10.3390/electronics10141701.

Paura, L., and Arhipova, I. (2014). Cause analysis of students' dropout rate in higher education study program. *Procedia-Social and Behavioral Sciences*, *109*, 1282–1286. https://doi.org/10.1016/j.sbspro.2013.12.625.

Pelánek, R. (2015). Metrics for Evaluation of Student Models. *Journal of Educational Data Mining*, *7*(2), 1–19.

Quadri, M., and Kalyankar, D. N. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, *10*(2), 2–5.

Radovanović, S., Delibašić, B., and Suknović, M. (2021). Predicting dropout in online learning environments. *Computer Science and Information Systems*, *18*(3), 957–978. https://doi.org/10.2298/CSIS200920053R.

Rzepka, N., Simbeck, K., Muller, H.-G., and Pinkwart, N. (2022, April 22-24). *Keep it up: In-session dropout prediction to support blended classroom scenarios.* [Conference paper]. 14th International Conference on Computer Supported Education, Online. https://doi.org/10.5220/0010969000003182.

Smith, J. P., and Naylor, R. A. (2001). Dropping out of university:

A statistical analysis of the probability of withdrawal for UK university students. *Journal of the Royal Statistical Society: Series A* (*Statistics in Society*), *164*(2), 389–405. https://doi.org/10.1111/1467-985X.00209.

Tekin, A. (2014). Early Prediction of students' grade point averages at graduation: A data mining approach. *Eurasian Journal of Educational Research*, *54*, 207–226.

Vidal, J., Gilar-Corbi, R., Pozo-Rico, T., Castejón, J.-L., and Sánchez-Almeida, T. (2022). Predictors of university attrition: Looking for an equitable and sustainable higher education. *Sustainability*, *14*(17), 10994. https://doi.org/10.3390/su141710994.

Wagner, K., Merceron, A., and Sauer, P. (2020, March 23-27). *Accuracy of a cross-program model for dropout prediction in higher education*. [Conference paper]. 10th International Learning Analytics & Knowledge Conference, Frankfurt, Germany.

Yu, R., Lee, H., and Kizilcec, R. F. (2021, June 22-25). *Should college dropout prediction models include protected attributes?* [Conference paper]. 8th ACM Conference on Learning @ Scale, Virtual Event, Germany. https://doi.org/10.1145/3430895.3460139.