Education and Information Technologies (2023) 28:12169–12189
https://doi.org/10.1007/s10639-023-11682-z

A stacking ensemble machine learning method for early identification of students at risk of dropout

Juan Andrés Talamás-Carvajal[1] · Héctor G. Ceballos[2]

Received: 10 October 2022 / Accepted: 20 February 2023 / Published online: 7 March 2023

Abstract

Early dropout of students is one of the bigger problems that universities face currently. Several machine learning techniques have been used for detecting students at risk of dropout. By using sociodemographic data and qualifications of the previous level, the accuracy of these predictive models is good enough for implementing retention programs. In addition, by using grades of the first semesters, the accuracy of these models increases. Nevertheless, the classification errors produced by these models cause undetected students to be discarded from the retention programs, whereas students with no actual risk consume additional resources. In order to provide more accurate models, we propose the use of a stacking ensemble technique to obtain an improved combined dropout model, while using relatively few variables. The model results show values on the expected ranges for an early dropout model, but with considerably fewer features and historical information, and we show that deploying the models would be cost-efficient for the institution if applied towards an intervention program.

Keywords Educational Innovation · Dropout · Educational Data Mining · Higher education

✉ Juan Andrés Talamás-Carvajal
juan.talamas@tec.mx

Héctor G. Ceballos
ceballos@tec.mx

[1] School of Engineering and Science, Tecnológico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, Tecnológico, Monterrey, Nuevo Leon 64849, Mexico

[2] Institute for the Future of Education, Tecnológico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, Tecnológico, Monterrey, Nuevo Leon, Mexico

1 Introduction

Education has been shown to be one of the most important factors for the development and advancement of countries. It is one of the most powerful tools for alleviating poverty and can be directly linked to the overall growth of a society (Latif et al., 2015). Going into a more individual perspective, education (specifically the completion of a degree) has been shown to be directly linked to earning potential, self-worth, and future opportunities (Liem et al., 2001).

It is then apparent that we should aim to improve on education as much as possible, as its effects can be seen from the individual up to society as a whole. One of the bigger problems that universities face presently is dropout. Based on the OECD Education at a Glance 2022 report, only 44.3% of women and 33.1% of men finish their higher education degrees on time, and overall, only 68% end up finishing at all (OECD, 2022). This information is worrisome as there tends to be a significant difference between the best and worst performing Higher Education Institutions and countries. If we go by the OECD report again, the difference between the United Kingdom with 69%, and Colombia with 12%, it is clear that there is work to be done regarding dropout and degree completion.

While the national averages can be somewhat misleading due to the differences between public and private institutions, and prestigious and more accessible ones, it is a problem that all institutions will face in some manner or another. For some, it may be a problem of reputation and tuitions recollection, for others, one of societal growth and fighting poverty. In either case, the ability to identify students at risk of dropout is the main obstacle towards fighting it.

The term dropout has been used to describe any situation where a student leaves their studies before they can obtain their degrees (Larsen et al., 2013). However, it is important to distinguish between two main types of student dropout: early dropout and late dropout. Dropout diminishes greatly with time spent at the institution, with most of it happening in the first year of studies. Early dropout also differs from late dropout on its causes: while late dropout usually stems from financial difficulties or unexpected emergencies, early dropout usually occurs due to student difficulties at school, either personal or in terms of performance, and occurs in the first few semesters of higher education.

Ulrich Heublein (2013) proposed a model of the dropout process in his article regarding dropout in German higher education, which was later translated and used in Larsen's Evidence on dropout phenomena at universities. This model states that socio-demographic factors, study prerequisites, expectations, living conditions and financial situation act as external factors (to the university) that could affect the possibility of dropout, while achievement potential, study motivations, mental and emotional resources, university integration and interactions, and study conditions are internal factors (Larsen et al., 2013).

Some of the internal factors that affect dropout are in constant development, like the curriculums and study conditions. As for the rest of them, universities usually employ staff that is capable of helping them with tutoring and mentoring, as well as counsel when needed. It would then seem that education institutions have all the tools needed for fighting dropout. However, there is still the problem of identifying the students at risk of dropping out and doing so quickly, as these students tend to leave in the first two to three semesters. This specific issue is the target of this paper.

We propose the use of a stacking ensemble, machine learning method for early detection of students at risk of dropping out using a classification approach. The ensemble method combines the results of several different models to obtain an improved, combined model. Some examples of ensemble models are bagging, boosting, and stacking. A stacking method is different from bagging or boosting in that the base models are different from each other, and that a single meta model is used to combine the predictions of base models (Wolpert, 1992). The base models will include Logistic regression (LR), k-Nearest Neighbors (KNN), Decision Trees, and Naïve Bayes; while the ensemble or meta-model will be logistic regression. As with any ensemble approach, different combinations of the models will be tested to find the best performing model.

In general, stacking models is useful when different techniques are all good for tackling the same problem, but in different ways. We know that classification, random trees, naïve bayes and other have been used to try and predict dropout in previous works, so it stands to reason that an ensemble model could use the best traits of those models and improve upon their results (Casanova et al., 2018; Ozay & Vural, 2012; Viloria et al., 2019).

We believe that the use of an ensemble method for classification, specifically stacking, will produce improved results compared to individual models when attempting to develop an early-detection model for student dropout. Our research questions are as follows:

R1. Can a stacking ensemble method for classification improve upon the results of other classification methods when working towards early dropout detection?
R2. What is the earliest point at which the dropout model can be deployed viable for the institution? (After admission test, after first semester)

2 Background

Identification of students at risk of dropping out is a complex problem, as it entails combinations of factors, including economic situation, socio-demographic data, academic performance, as well as mental health factors, among others. As such, several different approaches that use machine learning have been applied to this problem, usually in the form of classification algorithms. Random Forest, Neural Networks, Support Vector Machines and Logistic Regression are the most common ones, even to the point that all of them have been compared to each other, as is the case of a team at "Instituto Tecnológico de Costa Rica" (Solis et al., 2018). Decision trees and logistic regression were also used at Karlsruhe Institute of Technology (Kemper et al., 2020). Another example can be seen in the Berens et al. (2019) paper, in which an early detection system was implemented for dropout detection using logistic regression, Random Forest with bagging, and the AdaBoost algorithm.

The two most common techniques for classification approaches are Random Forest and Naïve Bayes, which has led to them usually being taken as the baseline against which to compare new models and approaches. The same surveys that mention these statistics also include the use, although comparatively smaller, of ensemble models, primarily boosting and bagging (Chung & Lee, 2019; Isphording & Raabe, 2019; Mduma et al., 2019; Silva & Roman, 2021; Zhang et al., 2022).

One of the difficulties when developing models for dropout prediction is that each individual University has its own idiosyncrasies. What may work for one will deliver mediocre results for another. In this particular case, we will be conducting our study on a dataset provided by the Tecnologico de Monterrey which includes several years' worth of students data. The dataset includes anonymized information related to undergraduate students who have enrolled and attended at least one semester at Tecnologico de Monterrey from 2014 to 2020. Among the information categories available in this dataset are: sociodemographic information (age, gender, place of origin), enrollment information (program/school, region), academic information related to the student (previous level average, current average, periods completed), information associated with scores on admission tests (PAA, TOEFL, other initial evaluations), academic history (type of school, region, national/international, Tec or no Tec system), student life (participation in sports, cultural, entrepreneurial activities), and financial information (type of scholarship, percentage of scholarship) (Alvarado-Uribe et al., 2022).

As mentioned before, dropout has heavy repercussions on individuals, educational institutions, and society as a whole (Liem et al., 2001; Latif et al., 2015). Currently, we are at a point where the reasons for dropping out are varied and complex, but understood to some degree: socio-demographic status, mental health, university environment, and several others form the basis of the dropout phenomenon (Heublein, 2013; Larsen et al., 2013). We believe the most urgent step is the identification of these students. Even if we develop an excellent counseling and tutoring program to aid those students, it is meaningless if we do not find them before they leave the institution. The feasibility of early warning systems has been demonstrated before, and a moder example can be seen in Xia and Qi work (2022).

Therefore, any improvement (even relatively small ones) can have a big impact on both the institution and the student. The effect is especially important for the individuals, as it is no exaggeration to say that dropping out could affect their future significatively (Liem et al., 2001).

As ensemble approaches have proven to generally improve upon singular classification algorithms, it is reasonable to believe that the model will result in an improvement compared to single method models that have been used successfully before (Casanova et al., 2018; Ozay & Vural, 2012; Viloria et al., 2019; Wolpert, 1992).

3 Materials and methods

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was followed for this project. Initial research of the issue has already been performed and explained above, so the rest of the steps refer to data understanding and data preparation. Data description of the Dataset can be found on this data paper (Alvarado-Uribe et al., 2022).

The original dataset contains 121,576 unique students with 49 distinct features and an identification number. Approximately half of the entries regard information from high school programs. As for the data preparation and preprocessing, we started by doing data binarization and ordinalization for categorical variables that contained only 2 possible values or were ordered, respectively. Some examples for binarization were gender, school level (level in the dataset), and previous school type (Tec.No_Tec), while ordinalization examples are previous school cost (school.cost), and parents' education level (max.degree.parents). Data points with values like "No information", "Does not apply", or similar were changed for "nan" values after making sure those features were indeed missing and not dependent on another feature. Finally, we merged features regarding the students' participation on Cultural (cultural.diffusion and art.culture), leadership (student.society and student.society.leadership), or Sport related activities (physical.education and athletic.sports), as different generations of students were labeled into separate features. Features with too many missing values were discarded.

After cleaning our dataset, we proceeded to separate it into high school and undergraduate students, as our interest this time deal with early dropout of undergraduates. The final dataset before feature selection consisted of 63,912 students, with 21 features each. Of this original dataset, 8.79% of students ended up dropping out.

It is important to note that this number refers to the dataset before feature selection, and before some reduction happens due to the availability of historical data, as some of the older datapoints are missing some of the newest features.

Feature selection is an important part of the CRISP-DM methodology and of overall machine learning. In our case, we are interested in developing our models with the least amount of historical data possible. The reason for this is so the students at risk of dropping out can be identified earlier. Due to this, historical data will be manually selected for our experiments, while the rest of the data will go through a normal feature selection process, meaning highly correlated features are removed, and a selection methodology is used. In this case, we used Recursive Feature Elimination (RFE) and took the most relevant features.

Finally, due to the inherently imbalanced nature of the problem (the large difference in number of students that stay vs. those that drop out), we used the balancing algorithm Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) in order to reach an equilibrium between the target of interest and the common case. It is important to note that SMOTE creates synthetic datapoints that have been shown to improve classification algorithms,, and as such the datasets tend to change in size to balance the classes. Base models will be trained using k-fold cross validation. As usual, the dataset will be separated into training (80%) and testing (20%) sets, using 5-fold cross validation. The predictions obtained from the testing sets will then be used as inputs for the meta model training.

One of the objectives of our project was to determine the earliest point at which we could identify dropout students. As such, we progressively added historical data to our selected features in order to find the optimal point in time to deploy our model. We trained 2 distinct models for this in the following order:

1. Stacking ensemble model using socio-demographic, and post-admission data.

2. Stacking ensemble model using socio-demographic, post-admission, and 1st semester data.

Evaluation of the models was based on the sensitivity they achieved and done on the testing set of the meta-model (true positive rate). Table 1 contains the variable definitions of the final selected features and the different values they can take, while Table 2 shows the variables used in each iteration of the model.

Using the variables mentioned before, we developed 2 distinct models: a post-admission test model (before 1st semester), and a post-1st semester model. Each model has the same overall structure: a stacking ensemble classifier that uses as level 0 models (base models) a decision tree classifier, a K-nearest neighbor classifier, a Naïve Bayes classifier, and a logistic regression classifier. The level 1 model (meta model) is also a logistic regression classifier. Stratified K-fold was used to feed the stacking ensemble.

The 2 different models were evaluated by obtaining the accuracy, precision, recall, f1, and the area under the Receiver Operating Characteristic curve (ROC) and Precision Recall Curve (PRC) plots, for the ensemble model and its individual components. While accuracy and ROC Area Under the Curve (AUC) values can be misleading metrics when dealing with imbalanced data (Davis & Goadrich, 2006; Fawcett, 2004; Saito & Rehmsmeier, 2015), they are still relevant as benchmarks and as commonly used values in order to compare our results to other studies.

In order to avoid overly complex systems that might make better predictions but make interpretation significantly harder, we decided to only use the best 5 sociodemographic features available before students start the admission process from our RFE and included historical data progressively (Admission data was classified as historical in this instance). As mentioned before, those were the previous academic level (1-100 scale), the general math (0-100) and English evaluations (1–7), age (13–55), and previous school cost (Ordinalized, 1–5).

4 Results

4.1 Post-admission exam model

The training and testing datasets for this model after cleaning and eliminating cases with missing values was compromised of 33,012 datapoints, and 6,003 datapoints respectively. Each set contained 7 features each.

The scores for the base models, the stacking classifier, and 2 ensemble classifiers (added for comparison) can be seen in Table 3. Contrary to what was expected, the stacking classifier performed lower than the base classifiers. It is important to note that all scores that take into account the target variable (dropout) are below 0.30, and that ROC AUC values didn’t go past 0.70. As was mentioned before, accuracy is not a great indicator of reliability on unbalanced datasets.

Table 1 Variable definitions

Variable name	Definition	Values
PNA	Previous level score (average)	Range from 0 to 100
General.math.eval	Mathematics grade from the admission test or from the school of origin	Range from 0 to 100, Does not apply, No information
Age	Student's age	Range from 13 to 55 years
English.evaluation*	Level of English obtained from a standardized English proficiency test	Level 0: No information, Level 1: Beginner, Level 2: Lower Basic, Level 3: Higher Basic, Level 4: Lower Intermediate, Level 5: Intermediate, Level 6: Upper Intermediate, Level 7: Advanced
School.cost*	Cost level of the student's tuition from the school of origin	Public, Low cost, Medium cost, Medium high cost, High cost, Not defined
Admission.test	Admission test score	Range from 0 to 1600, Does not apply
Admission.rubric	Score generated from the student profile where 50 is outstanding and 0 is average	Range from 0 to 50
Failed.subject.first.period	Number of subjects failed in the first period (Undergraduate) or in the first partial (High School). Only for the AD19 and AD20 generations.	Range from 0 to 8
Average.first.period	Average obtained in the first period (Undergraduate) or in the first partial (High School). This data corresponds only to the AD19 and AD20 generations.	Range from 0 to 100
Dropped.subject.first.period	Number of subjects dropped out in the first period (Undergraduate) or in the first partial (High School). This data corresponds only to the AD19 and AD20 generations.	Range from 0 to 9
Sports (originally physical.education and athletic.sports, merged)	Value that indicates if the student enrolled in any physical education activities during that period	0, 1, Does not apply, No information
Culture (originally cultural.diffusion and art.culture, merged)	Value that indicates if the student enrolled in any Cultural Diffusion activities during that period	0, 1, Does not apply, No information
Leadership (originally student.society and student.society.leadership, merged)	Value that indicates if the student enrolled in any Student Society activities during that period	0, 1, Does not apply, No information

Variables marked with * were ordinalized before use. Ordinalization refers to assigning numbers to categorical variables with sequential characteristics

Table 2 Model variables

Variable name	Post-adm	1st sem
PNA	X	X
General.math.eval	X	X
Age	X	X
English.evaluation	X	X
School.cost	X	X
Admission.test	X	X
Admission.rubric	X	X
Failed.subject.first.period		X
Average.first.period		X
Dropped.subject.first.period		X
Sports		X
Culture		X
Leadership		X

Table 3 Post-admission model scores

	Models						
	Decision trees	KNN	Naïve Bayes	LR	Stacking clasifier	Adaboost	Xgboost
Accuracy	0.7998	0.7566	0.8424	0.8811	0.8496	0.8967	0.9135
Precision	0.1143	0.1052	0.1713	0.2100	0.0996	0.1809	0.2174
Recall	0.2130	0.2617	0.2394	0.1623	0.1034	0.0730	0.0203
F1	0.1487	0.1501	0.1997	0.1831	0.1015	0.1040	0.0371
ROC AUC	0.5326	0.5511	0.6645	0.6675	0.5383	0.6301	0.5952
PRC AUC	0.1959	0.1248	0.1583	0.1786	0.0981	0.1281	0.1204

4.2 Post-first semester model

The training and testing datasets for this model after cleaning and eliminating cases with missing values was compromised of 9,832 datapoints, and 2,459 datapoints respectively. Each set contained 13 features.

Table 4 shows the different scores for the models with all historical data included (admission tests, 1st semester average, failed and dropped courses, and extracurricular participation). Model scores show an overall increase in all metrics following the inclusion of 1st semester information for all models.

The inclusion of historical data affected the models differently: while our stacking model saw increase in the range of 10 points (0.10) for all base metrics, the base models ranged from 3 points at the lowest to 13 points at the highest. The ensemble models saw a notorious increase in precision (~30 points), while other metrics saw more moderate increases.

Table 4 Post first semester model scores

	Models						
	Decision trees	KNN	Naïve Bayes	LR	Stacking clasifier	Adaboost	Xgboost
Accuracy	0.8345	0.8081	0.8800	0.9008	0.8841	0.9183	0.9256
Precision	0.1452	0.1550	0.2692	0.3273	0.2118	0.4186	0.5526
Recall	0.2406	0.3422	0.3369	0.2888	0.1925	0.1925	0.1123
F1	0.1811	0.2133	0.2993	0.3068	0.2017	0.2637	0.1867
ROC AUC	0.5620	0.6320	0.6909	0.6944	0.6379	0.6443	0.6464
PRC AUC	0.2218	0.1864	0.2558	0.2655	0.1519	0.2186	0.2144

4.3 Expected value framework

In order to better understand the possible impact of deploying our model, we are using an expected value framework in terms of tuitions and the cost of providing tutoring for identified leaving students (stated as 1/6 of the tuition, the equivalent of one course traditionally) for our post-1st semester model. An expected value framework has its origins in probability theory, and it basically takes into account both the cost/benefit of an outcome, and the actual probability of it happening. As such, we can determine if it would be worthwhile to pursue a particular experiment or model, as is our case. Specifically, we use the frequency of our results to obtain a probability for each of our 4 combinations: real positives, false positives, real negatives, and false negatives; as well as the overall positive and negative frequencies. These probabilities are then multiplied by the costs and benefits of their respective cases, and finally, further multiplied by the overall case (positive rate for true positives and false negatives, and negative rate for true negatives and false positives) before summing everything together and obtaining an average expected value if the experiment was run infinitely. Tables 5 and 6 show this framework.

The final Expected Value amounts to the following equation: $-0.0067*T - 0.0021*T + 0.0126*MT$. If we assume that students leave at around the same time in early dropout (3rd semester), the "lost" and "rescued" semesters amount to normally 5 semesters in an 8-period program. The benefit if the model is applied would be a 0.0542 Tuition benefit per student on average (Assuming that early detection allows the student to stay). This number in reality would be probably higher, as students in academic probation tend to stay longer. Furthermore, this study focuses on early dropout, making it so that even at the latest point a student leaves, the expected value stays on the positives.

Applying this model to our specific dataset would result in the following: 2182 cases with no cost or benefit, 99 cases with a 1/6 T cost and no benefit, 147 cases with no tuition cost or benefit, and 31 cases with 1/6 T cost and 5*T benefit, resulting in an additional 133.33 tuitions per semester for our total of 2459 students in this specific dataset (test dataset regarding post 1st semesters model). This constitutes an additional 5.42% increase in tuitions received after considering cost. While this

Table 5 Cost/benefit table

	Real negative	Real positive
Predicted negative	0 “additional” tuitions, no cost	No direct cost in terms of tuition
Predicted positive	Cost: 1/6 tuition (T) per class	Cost: 1/6 tuition (T) per class Benefit: M tuitions (T) (M being the number of “rescued” semesters.

Table 6 Conditional probabilities

	Real negative	Real positive
Predicted negative	2182 (0.9565)	147 (0.8258)
Predicted positive	99 (0.0434)	31 (0.1741)
Sum	2281 (0.9276)	178 (0.0723)

expected value framework assumes a 100% retention rate for the initial numbers shown before, it is logical to consider that not all students will stay. For the program to at least break even and not cost the university additional resources, it requires at least a 14% success rate for students participating in the interventions. Any result higher than that will result in a net positive regarding tuition cost vs. return.

4.4 Logistic regression coefficients, odds and probability

In order to better understand the models, we extracted the coefficients from the logistic regression classifiers for each iteration. There were 2 main reasons for this: of the base models, logistic regression is the one with a more direct route towards explainability; and since the meta-model for our stacking ensemble was again logistic regression, it can paint a picture of the true values obtained by the “black box” model. From this point on, we will be using a series of identifiers for our variables in order to avoid overly long names on our tables. These identifiers can be seen on Table 7, along with the variable type.

The coefficients for the models are shown in Table 8. It is important to note that these coefficients correspond to the logarithmic odds, and as such, can’t be directly used as probabilities. Larger positive values do have a bigger positive effect, while larger negative values have a bigger negative effect. The standard errors for all estimates ranged from 0.018 to 0.024. Individual errors can be seen in parenthesis in each cell. Using the standard errors and comparing them to the coefficients, we colored Table 8 to show features that flip in sign or go to zero (and therefore would be negligible) in red, and features that do not flip in sign and are more trustworthy in terms of predictive power in green.

Since there is no direct translation from logarithmic odd coefficients into probability coefficients, we instead obtained the average marginal gain for numerical values (given in probability), and the change in probability when present for categorical values. Table 9 shows these numbers. For average marginal gains, the number was obtained using the “divide by 4” rule, as seen in Chap. 5 of (Gelman

Table 7 Variable identifiers

Variable	Identifier	Variable type
PNA	V1	Numerical
General.math.eval	V2	Numerical
Age	V3	Numerical
English.evaluation	V4	Numerical (Ordinalized)
School.cost	V5	Numerical (Ordinalized)
Admission.test	V6	Numerical
Admission.rubric	V7	Numerical
Failed.subject.first.period	V8	Numerical
Average.first.period	V9	Numerical
Dropped.subject.first.period	V10	Numerical
Sports	V11	Categorical
Culture	V12	Categorical
Leadership	V13	Categorical

Table 8 Logistic regression coefficients for each model, and standard errors in parenthesis

	Logistic regression coefficients												
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
Post adm	🟥 0.015 (0.02)	🟥 0.0105 (0.019)	🟩 -0.144 (0.018)	🟩 0.0788 (0.019)	🟩 -0.027 (0.019)	🟥 0.0001 (0.019)	🟥 0.0088 (0.018)						
1st sem	🟩 -0.023 (0.021)	🟥 0.0103 (0.02)	🟩 -0.14 (0.019)	🟥 0.012 (0.019)	🟩 -0.042 (0.02)	🟥 0.0005 (0.019)	🟥 0.011 (0.018)	🟩 -0.037 (0.022)	🟩 0.0359 (0.024)	🟥 -0.022 (0.022)	🟥 0.0139 (0.019)	🟥 0.0055 (0.019)	🟥 0.0129 (0.019)

& Hill, 2006), while the categorical gains were obtained using average values and for all other coefficients and obtaining the probability with and without the target variable. The intercept values were obtained setting all other variables to 0. We can observe that the average marginal values with the highest absolute scores by model in Table 9 correspond with those that were colored in green in Table 8.

Finally, we normalized the marginal gains of the numerical values to have a more intuitive set of probabilities. This normalization was performed in each variable so that they ranged from 0 to 100 so it was easier to compare the values, as some variables had extremely large ranges (admission scores had a maximum of 1600, for example), or very small ranges (0–5 for school cost). These values can be seen on Table 10. Values for variables 11 through 13 and the intercept were omitted due to them being categorical variables, as marginal probabilities are not applicable.

With all values now in the same scale, we are able to compare them more intuitively. We can observe that the most significant positive variable before the 1st semester is the previous academic score followed by the general math

Table 9 Average marginal gains and categorical probabilities

	Average marginal gains and probabilities													
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	Int.
Post adm	0.0037	0.0026	-0.036	0.0197	-0.007	3E-05	0.0022							0.496
1st sem	-0.006	0.0026	-0.035	0.003	-0.01	0.0001	0.0028	-0.009	0.009	-0.005	0.0013	0.0004	0.0012	0.4952

Table 10 Normalized marginal gains

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Post adm	0.0037	0.0026	-0.02	0.0014	-3E-04	0.0004	0.0011			
1st sem	-0.006	0.0026	-0.019	0.0002	-5E-04	0.0021	0.0014	-7E-04	0.009	-5E-04

evaluation, and the most significant negative variable is age followed by school cost. After the 1st semester, the most significative positive variable changes to the average of the 1st semester, while age remains as the most important negative factor. It is important to note that several values are so close to 0 that their effects, negative or positive, are negligible in this context.

5 Discussion

5.1 Feature limitations and context

One of the realities of research regarding dropout in higher education is that it depends heavily on the availability of historical data regarding the student's performance: the more detailed information available, the better the models are, which can be easily seen in the articles by Berens et al. (2019), Chung and Lee (2019), and Kemper et al. (2020). However, there is an inherent cruelty in this need for historical data, as the students that might benefit the most from these prediction models are the ones that tend to leave the earliest. As such, the features that can realistically be used to build these models are severely limited, and usually are compromised by 1st and 2nd semester averages and a few extracurricular activities like sports and student societies.

Our models were built with this in mind, and as such, we used data limited strictly to what would be available at the end of the 1st semester of our students. Even then, the data was limited to very broad indicators of the student's performance (i.e., semester average instead on course grades, number of dropped or failed subjects instead of more detailed information, etc.).

The effect of these limitations can be observed in the True Positive Rates reported. By design, our models deal with limited amounts of data and broad indicators of how a student is performing after the first semester. This initial period of time is usually one of change in which students need to adapt to a new reality quickly, and their overall approach to their new responsibilities is in flux. Translated into feature effects, this limits their general prediction power and, in our case, resulted in lower TP rates.

This is by no means to say that the model is inadequate, as was shown in Section 4.4 with the expected value framework, but it does shed some information on the paradox of dropout: if we are able to make "decent" predictions about it, why is it still categorized as one of the main problems of higher education currently? One answer could be the limitations that a fast prediction imposes on the models. Additional data granularity could improve the early prediction power of our models, as would the inclusion of missing data like academic guidance program participation, emotional and psychological health, among others.

5.2 Related works

While higher scores have been reported before, they have the disadvantage of needing precise academic and/or socioeconomic data, while our model uses only a few specific features. For example: Solis et al. (2018) reported scores as high as 94% true positive rates for their Random Forest model. However, this were obtained using large amounts of historical data, including the semester where dropout happened for students that interrupted their studies. Viloria et al. (2019) reported values around 90% accuracy, but with a low recall (33%), with the distinction that detailed and extensive socio-economic data was used, along with information from the student's first semester.

Another important distinction to make is that it is notoriously hard to compare different studies and models. The differences in sample size, data used, research methods, and even the differences in university culture and "personality" make it hard to transfer discoveries and knowledge from one to another. While one study might focus on one specific class, another attempts to encompass a complete program, while one study uses at least 32 features, another goes for a maximum of 7. Table 11 shows a collection of models from several different authors, highlighting some of these differences for a better understanding of the context of these models and their scores.

The reported scores from the papers mentioned in Table 11, as well as some additional examples, are summarized in Table 12. The scores are accompanied by a small description of the historical academic data used. As each individual article reported their results individually, there is no definite set of scores that was used by every team. In an effort to maintain fairness, we included all scores mentioned in their papers, and calculated the ones we were missing.

It is important to note that accuracy continues to be one of the most reported metrics, even though it is discouraged when dealing with imbalanced datasets. This same problem occurs with ROC AUC, as was mentioned before.

Some of the additional examples include models similar in concept to our selected method: a stacking ensemble method. In the article by (Niyogisubizo et al., 2022), the researchers use a stacking ensemble method, with some key differences with our approach. First, they use data from a single course through several generations of students, and second, they selected a complex set of level 0 and level 1 classifiers. While this resulted in good scores, the model is more expensive to implement regarding resources and time.

Another instance of an ensemble classifier being used can be seen in the paper by (Zeineddine et al., 2021). This team used Automated Machine Learning in order to select the best performing model for their dataset. The Automated Machine Learning methodology is easy to understand, as it entails a brute force approach where a "grid-search" (all possible combinations) is performed and the best model selected. While more computationally intensive and lengthier regarding time, it allows optimal model selection. The team reported that the search algorithm arrived at an ensemble model based on a voting scheme of its components.

Readers more acquainted with machine learning algorithms might realize that there are few cases mentioned of more powerful methods such as Deep Learning

Table 11 Model characteristic summary

Models	Features	Application period	Data
Chung and Lee (2019)	12	All cohorts	Absences and grades, specific period,all levels
Viloria et al. (2019)	18	All cohorts	Socio-economic and grade data, specific period, all levels
Solis et al. (2018)	21	End of program	All students from 5 years, all historical data
Berens et al. (2019) admission	> 30	At admission	All students from 10 years, all historical and demographic data
Berens et al. (2019) 1st sem	> 30	1st semester	All students from 10 years, all historical and demographic data
Berens et al. (2019) 2nd sem	> 30	2nd semester	All students from 10 years, all historical and demographic data
Berens et al. (2019) 3rd sem	> 30	3rd semester	All students from 10 years, all historical and demographic data
Berens et al. (2019) 4th sem	> 30	4th semester	All students from 10 years, all historical and demographic data
Kemper et al. (2020)	> 32	1st semester	All students from 5 years, all historical and demographic data
Kemper et al. (2020)	> 32	2nd semester	All students from 5 years, all historical and demographic data
Kemper et al. (2020)	> 32	3rd semester	All students from 5 years, all historical and demographic data
Niyogisubizo et al. (2022)	12	Course specific	One course through 5 years, all data from course
Zeineddine et al. (2021)	13	All cohorts	Demographics, math and english data, average (to determine pass or fail)
Borrella et al. (2022)	25	Course specific (MOOC)	Data from 3 courses, log data and grade information
Mubarak et al. (2021)	> 100	Database specific (several courses [MOOCS])	Log data and grades from 2 MOOC databases
Stacking Ensemble (post exam)	7	At admission	Demographic and admission data, all levels
Stacking Ensemble (1st semester)	13	1st semester	Demographic and admission data, 1st semester data, all levels

Table 12 Score comparison between different models

	Scores						
Models	Accuracy	Precision	Recall	ROC AUC	PRC AUC	TP rate	TN rate
Chung and Lee (2019) All cohorts	0.95		0.85	0.97		0.85	0.95
Viloria et al. (2019) All cohorts	0.91		0.33			0.33	0.98
Solis et al. (2018) End of program			0.71			0.71	
Berens et al. (2019) admission	0.67		0.49			0.48	0.73
Berens et al. (2019) 1st sem	0.84		0.72	0.86		0.69	0.86
Berens et al. (2019) 2nd sem	0.89		0.76	0.91		0.72	0.90
Berens et al. (2019) 3rd sem	0.81		0.93	0.92		0.75	0.92
Berens et al. (2019) 4th sem	0.95		0.83	0.94		0.74	0.94
Kemper et al. (2020) 1st sem	0.865	0.695	0.764			0.76	0.90
Kemper et al. (2020) 2nd sem	0.933	0.887	0.83			0.83	0.97
Kemper et al. (2020) 3rd sem	0.918	0.741	0.90			0.90	0.92
Niyogisubizo et al. (2022) Single course	0.92	0.93	0.93	0.98		0.93	
Zeineddine et al. (2021) All cohorts	0.75						
Borrella et al. (2022) MOOC	0.78		0.64			0.64	0.98
Mubarak et al. (2021) MOOC	0.94	0.94	0.85	0.86	0.8	0.85	
Stacking Ensemble (post exam)	0.84958	0.09961	0.103	0.538	0.098	0.10	0.92
Stacking Ensemble (1st semester)	0.8841	0.21176	0.193	0.638	0.152	0.19	0.94

for example. Generally, such “black box” methods are not preferred in educational research as they offer little insight as to “what” is affecting the target variable. However, these models do generally show high scores. Such an example is the article by (Mubarak et al., 2021). The authors used a convolutional neural network with a custom cos-sensitive loss function. They achieved high levels of precision, recall, and f1 scores using only log data from a series of courses and grades. Finally, as a way of having a more visual comparison, we developed bar charts regarding what we believe are the most comparable models, with Fig. 1 representing the two models implemented at admission, and Fig. 2 summarizing the scores for a post 1st semester application of the models.

We can observe that the models shown in Berens et al. (2019) and Kemper et al. (2020) are the closest ones in scope and timing to our own. Regarding this comparison, even though these two models are closest to ours in terms of scope and timing, both models had more than twice the number of features than our model: 7 and 13 for our admission and 1st semester models, against their (at least) 30 for Berens et al. (2019), and 32 for Kemper et al. (2020).

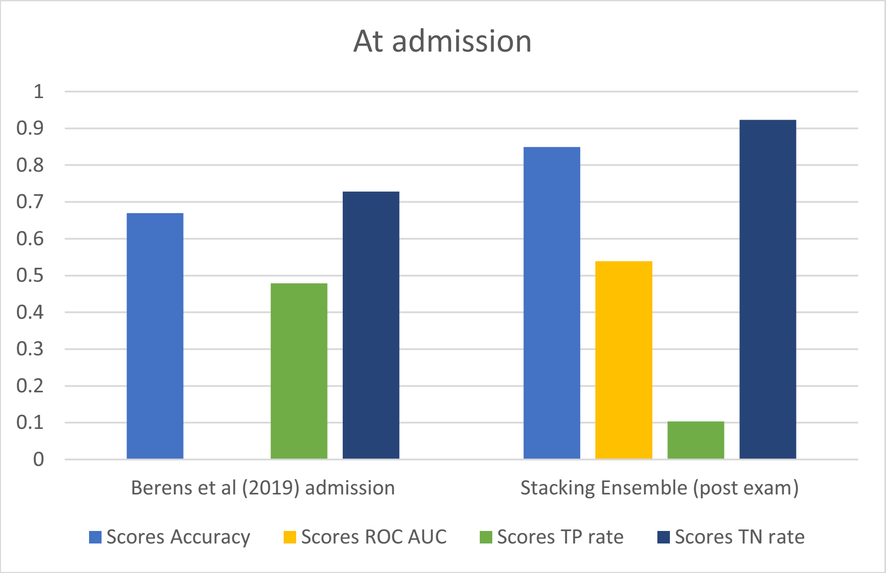

Fig. 1 Scores comparison for admission models

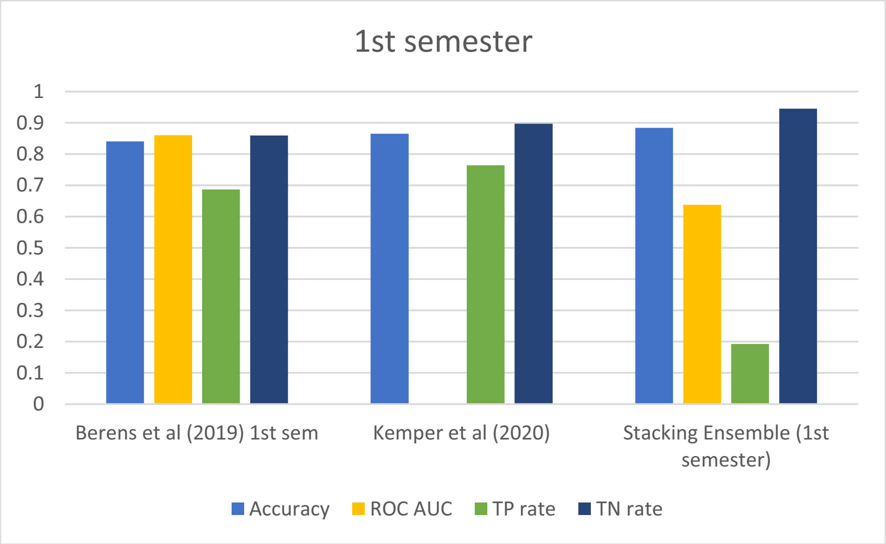

Fig. 2 Scores comparison for 1st semester models

5.3 Feature effects

Regarding the logistic regression coefficients, there are some possible explanations as to their values and effects. Regarding positive effects, the overall most recent grade appears as the biggest positive variable. In pre-1st semester models, this refers to the PNA variable. It is reasonable to see grades as an overall indicator of several different student characteristics: for example, their adaptation to a different school model, preparedness, enthusiasm and learning capabilities all affect their grades. The effect of these unseen variables can be interpreted by changes in time. For example, a large drop in grades from admission to the 1st semester might indicate problems adapting, stress, and more. And as grades also determine passing or failing, a lower grade also has a direct negative impact on retention probability by means of the dropped or failed subjects variables.

These unseen variables can also help understand why age is seen as an important negative variable on average: for example, older students might be working and studying at the same time, they might be going through a career change, or even have children of their own that require their attention. In a vacuum, age by itself more than likely has no real effect on undergraduate performance, but instead, it becomes a proxy variable for more complex characteristics.

Traditionally, school cost could be thought of as a positive indicator of future student performance. After all, a more expensive school is a better school, is it not? While this may come off as “common sense”, it is not necessarily true. A much better indicator would be the student’s performance at their previous school, regardless of the tuition cost. In our case, we believe that school cost shows another indirect feature that has little to do with the quality of the school: a family’s economic capacity. A family with the economic freedom to move their kids to another school, take the hit for a lost semester, or encourage less traditional education is more likely to leave the school than one with economic difficulties and/or a child with a scholarship.

It is important to note that these feature effects are an initial approximation towards understanding feature effects (positive or negative), and future work will include more specialized explainability tools. (SHAP values, for example).

6 Conclusion and future work

This paper contributes a series of early detection models that can be used to either assign regularization programs for students that might be enrolling with lower scores than ideal, or for possible interventions for struggling students. The models presented have the advantage of requiring little to no academic history data, allowing for early detection of dropout-likely students, which in turn can be used to deploy interventions and retention strategies.

A general expected value framework was also discussed regarding a possible cost-benefit analysis, showing that the deployment of the model would be self-sufficient (cost of deployment is less than benefit) if we could ensure at least a 14% student retention after the intervention. While this has its own set of difficulties, the benefits for individual students, school reputation, and economic health outweigh the hardships presented. Having said that, the current True Positive rate of our models is not optimal, and improvements upon this would lead to both increased confidence in the model predictions, as well as a more beneficial deployment for both students and institutions.

Taking these points into account, future work will revolve around possible interventions and their effect on the dropout or retention prediction, as well as the inclusion of emotional and phycological variables in order to improve the model and consider these important aspects. A set of variables that is not present in the current dataset is academic guidance and psychological data for students. It is difficult to observe and control emotional and psychological backgrounds, so one alternative is mentoring and guidance information. Several institutions, including Tecnologico de Monterrey, assign mandatory guidance and counseling for students that are underperforming or missing too many classes. Data from these interventions could be used to improve the models and assign interventions appropriately.

Abbreviations *CRISP-DM*: Cross Industry Standard Process for Data Mining; *RFE*: Recursive Feature Elimination; *SMOTE*: Synthetic Minority Oversampling Technique; *ROC*: Receiver Operating Characteristic curve; *PRC*: Precision Recall Curve; *AUC*: Area Under the Curve; *LR*: Logistic regression; *KNN*: k-Nearest Neighbors

Acknowledgements The authors would like to acknowledge the Living Lab & Data Hub of the Institute for the Future of Education, Tecnológico de Monterrey, Mexico, for the data published through the Call "Bringing New Solutions to the Challenges of Predicting and Countering Student Dropout in Higher Education" used in the production of this work.

Author contribution JAT performed the literature search, data analysis, and developed the 1st draft of the document. HC critically reviewed the work, provided commentary, supervised, and guided the final development of the article. All authors read and approved the final manuscript.

Funding The authors would like to thank the financial support from Tecnológico de Monterrey through the "Challenge-Based Research Funding Program 2022". Project ID # I004 - IFE001 - C2-T3 – T.

Data availability The datasets analyzed during the current study are available in the Institute for the Future of Education's Educational Innovation collection of the Tecnologico de Monterrey's Data Hub repository, https://doi.org/10.57687/FK2/PWJRSJ.

Declarations

Institutional review board statement Privacy issues related to the collection, curation, and publication of student data were validated with Tecnológico de Monterrey's Data Owners and the Data Security and Information Management Departments.

Competing interests Juan Talamás and Héctor Ceballos declare that they have no competing interests.

References

Alvarado-Uribe, J., Mejía-Almada, P., Masetto-Herrera, A., Molontay, R., Hilliger, I., Hegde, V., Montemayor-Gallegos, J., Ramírez-Díaz, R., Ceballos, H. (2022). Student dataset from Tecnologico de Monterrey in Mexico to Predict Dropout in Higher Education. *Data*.

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk - Predicting student dropouts using administrative student data from german universities and machine learning methods. *Journal of Educational Data Mining, 11*(3), 1–41. https://doi.org/10.5281/zenodo.3594771

Borrella, I., Caballero-Caballero, S., & Ponce-Cueto, E. (2022). Taking action to reduce dropout in MOOCs: tested interventions. *Computers & Education, 179*, 104412. https://doi.org/10.1016/J.COMPEDU.2021.104412

Casanova, J. R., Cervero, A., Núñez, J. C., Almeida, L. S., & Bernardo, A. (2018). Factors that determine the persistence and dropout of university students. *Psicothema, 30*(4), 408–414. https://doi.org/10.7334/psicothema2018.155

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chung, J. Y., & 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

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *ACM International Conference Proceeding Series*, *148*, 233–240. https://doi.org/10.1145/1143844.1143874

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine learning*, *31*(1), 1–38. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:ROC+Graphs:+Notes+and+Practical+Considerations+for+Researchers#0. Accessed 24 Aug 2022

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Heublein, U. (2013). Student drop-out from german Higher Education Institutions. *European Journal of Education, 49*(4), 497–513. https://doi.org/10.1111/EJED.12097

Isphording, I. E., & Raabe, T. (2019). *Early Identification of College Dropouts Using Machine-Learning* (IZA Research Reports 89). Institute of Labor Economics (IZA). https://ftp.iza.org/report_pdfs/iza_report_89.pdf. Accessed 1/11/2022

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: a machine learning approach. *European Journal of Higher Education, 10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Larsen, M., Sommersel, H., & Larsen, M. (2013). *Evidence on dropout phenomena at universities* (1). Danish Clearinghouse for Educational Research. 1–53. http://edu.au.dk/fileadmin/edu/Udgivelser/Clearinghouse/Review/Evidence_on_dropout_from_universities_brief_version.pdf. Accessed 1/6/2022

Latif, A., Ai, C., & Aa, H. (2015). Economic effects of student dropouts: a comparative study. *Journal of Global Economics, 3*(2), 2–5. https://doi.org/10.4172/2375-4389.1000137

Liem, J., Dillon, C., & Gore, S. (2001). Mental health consequences associated with dropping out of high school. *Annual Conference of the American Psychological Association*, 109. https://eric.ed.gov/?id=ED457502. Accessed 10/04/2022

Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal, 18*(1), 14. https://doi.org/10.5334/dsj-2019-014

Mubarak, A. A., Cao, H., & 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

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization. *Computers and Education: Artificial Intelligence, 3*, 100066. https://doi.org/10.1016/J.CAEAI.2022.100066

OECD. (2022). *Education at a glance 2022: OECD Indicators*. OECD Publishing. https://doi.org/10.1787/3197152b-en

Ozay, M., & Vural, F. T. Y. (2012). A new fuzzy stacked generalization technique and analysis of its performance. *arXiv: Learning*. http://arxiv.org/abs/1204.0171. Accessed 1/6/2022

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One, 10*(3), e0118432. https://doi.org/10.1371/JOURNAL.PONE.0118432

Silva, J., & Roman, N. (2021). Predicting dropout in Higher Education: a systematic review. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*. *Porto Alegre: SBC*, 1107–1117. https://doi.org/10.5753/sbie.2021.21743.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. *2018 IEEE International Work Conference on Bioinspired Intelligence*, IWOBI 2018 - Proceedings, September. https://doi.org/10.1109/IWOBI.2018.8464191

Viloria, A., Lezama, O. B. P., & Varela, N. (2019). Bayesian classifier applied to Higher Education dropout. *Procedia Computer Science, 160*, 573–577. https://doi.org/10.1016/J.PROCS.2019.11.045

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Xia, X., & Qi, W. (2022). Early warning mechanism of interactive learning process based on temporal memory enhancement model. *Education and Information Technologies, 28*, 1019–1040. https://doi.org/10.1007/s10639-022-11206-1

Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: automated machine learning approach. *Computers and Electrical Engineering, 89*, 106903. https://doi.org/10.1016/j.compeleceng.2020.106903

Zhang, W., Wang, Y., & Wang, S. (2022). Predicting academic performance using tree-based machine learning models: a case study of bachelor students in an engineering department in China. *Education and Information Technologies, 27*(9), 13051–13066. https://doi.org/10.1007/s10639-022-11170-w

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.