JOURNAL OF LEARNING ANALYTICS

Volume 12(2), 211-223. https://doi.org/10.18608/jla.2025.8553

The Facts Behind the Prophecy: Validating a Methodology for Identifying Behavioural Differences in Higher Education Student Subpopulations Under Intervention

Juan Andrés Talamás-Carvajal[1], Héctor G. Ceballos[2] and Isabel Hilliger[3]

Abstract

Artificial intelligence (AI) is currently leading an industrial revolution in most aspects of human life, and education is no exception. With the increasing ratio of students to faculty, AI could be an extremely beneficial tool for individual mentoring; for example, for cases of dropout and for student retention. While many models have already been built, the adoption of AI in education has been lower than expected, and few interventions have emerged from those models. Several factors may be in play, but one is that AI models are not easily explained, and the lack of explanation is fatal for situations like dropout prevention. An ideal AI-based tool for this problem would provide individually tailored interventions, but that would require a much deeper understanding of what a successful intervention entails. Using a novel methodology for feature comparison between student subpopulations, we compared regular students against students under academic guidance on a dataset containing 124,000 unique students and 36 informative features. We found that the explanations obtained regarding student dropout matched the real-world experiences of mentors and tutors, especially when dealing with highly explanatory features like previous average grades and interventions.

Notes for Practice

- Regardless of all the existing models to predict dropout, AI has seen a slower than expected adoption in education, especially to inform interventions towards dropout prevention.
- We developed and validated a methodology that provides individual and global explanations for dropout predictions in higher education.
- Application of this methodology on a dataset of 124,000 unique students resulted in a 90% match with opinions and experiences of practising mentors and tutors.
- Use of this methodology can help mentors and tutors by explaining the reasons behind dropout predictions, which can be used as a head start when developing and applying interventions.

Keywords: Dropout, XAI, higher education, intervention, educational innovation

Submitted: 09/07/2024 **— Accepted:** 29/12/2024 **— Published:** 13/04/2025

Corresponding author [1]Email: juan.talamas@tec.mx Address: School of Engineering and Sciences, Tecnológico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, N.L., Mexico. ORCID iD: https://orcid.org/0000-0002-6140-088X
[2]Email: ceballos@tec.mx Address: Institute for the Future of Education, Tecnológico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, N.L., Mexico. ORCID iD: https://orcid.org/0000-0002-2460-3442
[3]Email: ihillige@uc.cl Address: Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago, Chile. ORCID iD: https://orcid.org/0000-0001-5270-7655

1. Introduction

The only constant in the world is change, and one of the most prominent changes that we can see now is the assimilation of artificial intelligence (AI) into almost every aspect of our lives. A general tendency to include AI-based tools can be seen in almost every discipline, but that does not appear to be the case in education. Adoption of AI-based tools and resources in traditional education has been slower than expected (Rodway & Schepman, 2023): there is a low usage of predictive learning analytics (PLA) in higher education institutions (HEIs). According to Herodotou et al. (2023), only 42% of the interviewed course instructors were actively using tools of this nature, 19% had never even heard of such tools, 18% had heard of but not

used these tools, and 20% had tried to use them and stopped. Any novel approach to AI in education must consider this situation and either attempt to tackle it or resign itself to (most likely) short-term adoption and low usage.

While there are many reasons why this might happen, some of the main issues that appear to be the cause of this slow adoption are fear of misuse (cheating, automated essay writing, plagiarism concerns, etc.), lack of understanding and trust from the side of the practitioners and final users (i.e., the Black Box problem; Cave et al., 2019; Stanton & Jensen, 2021; Rahiman & Kodikal, 2024), and the reliability of predictions obtained from AI tools. There is an argument to be made in favour of increasing data literacy in general as part of the needs of the modern world, but a more focused effort from the learning analytics (LA) community is needed — and from other related research lines (e.g., educational data mining [EDM] and artificial intelligence in education [AIED]) — to make these types of AI-based tools interpretable, accepted, and finally used to their full potential (Khosravi et al., 2022).

In higher education, one of the main problems that AI-based tools could help with is the issue of scale: the ratio of teachers to students has made individual feedback and attention hard at best and impossible at worst, particularly in Latin America, where inequalities regarding accessibility, income, gender, and more are commonplace (Valenzuela & Yáñez, 2022). This issue is aggravated in the case of mentoring and tutoring, as the ratio is even more pronounced. Currently, advising students is a non-trivial issue: there are difficulties arising from the state of advising inside the institution (McMurtrie et al., 2022a), the difficulty in delivering appropriate and timely advice to students (Arin, 2022; McMurtrie et al., 2022b), or even the lack of clarity regarding what must be done (Calhoun-Brown, 2022). Then there are the added complications when dealing with advising students who are at risk of dropping out or on academic probation, and it is easy to see why so many HEIs struggle with this issue.

In most cases where student advising is involved, it is not a case of "if" something should be done but of "what" should be done. In other words, how the intervention should approach the specific problem faced by a specific student. There are several types of interventions mentioned in the literature, everything from institution-wide programs to targeted interventions. These latter interventions are called so because they "are theoretically precise and address basic psychological processes that can interfere with optimal academic functioning" (Harackiewicz & Priniski, 2018). So far, targeted interventions have shown positive results when applied at an appropriate time for the students (Lazowski & Hulleman, 2016; Walton, 2014). We can observe then that both the timing of the intervention and the specific strategy it follows are vital for success.

Additionally, the context of the intervention and the final users of the tools being developed/used can drastically change the necessary inputs and outputs. For example, the research done by Hilliger et al. (2020) mentions three distinct types of dashboards depending on their audiences: student-facing, advisor-facing, and dialogue dashboards. Each one requires distinct information to be presented and comes with its own set of limitations: What types of dashboards are more effective? Should student-facing dashboards show comparisons with other students? What information could be useful for advisors? Are predictions useful or desired in tools available to students? These types of questions have not yet been fully answered.

Nonetheless, there is a degree of consensus on the benefits of focused interventions for promoting self-reflection (Harackiewicz & Priniski, 2018), so any advising AI-based tool must ideally inform personalized intervention plans, especially AI-based ones. All tools should at least include individualized suggestions or other information or features to support timely interventions. Task value information could be linked to specific course grades and extracurricular activities, while some additional features could be used to recommend framing and value interventions so appropriate outputs could be obtained for each specific case. In this context, an AI-based tool that could deliver individualized recommendations for each student — besides outputting specific and achievable counterfactuals for at-risk students — would be a dream come true. However, much deeper understanding of what a successful intervention actually entails to help a student graduate in a timely basis is needed before we can make a tool meaningful for their tutors and mentors.

So far, the use of predictive algorithms and early warning systems covers the "acting on time" aspect, but there is still the issue of having "a" plan. The fact remains that, at least in most HEIs, mentors and tutors will not necessarily have information about their student mentees until they are actively working with them, and even then, the intervention is often a general case approach or becomes a targeted one as the case progresses.

In this paper, we present the validation of a methodology developed to identify the differences between the feature effects of two student groups that were fed through a predictive algorithm: one comprised of regular students and another comprised of students in an academic advising program. The predictive models were developed for each group for the first four semesters of their studies, after which both models were fed through an explainable artificial intelligence (XAI) model based on Shapley values. This methodology was first presented by Talamás-Carvajal and Ceballos-Cancino (2024) based on results first published in Talamás-Carvajal (2023). Since then, further work has been invested to provide much needed explanations of predictive models to mentors and tutors who are tasked to make interventions using those predictions. This validation is based on expert opinions and the experience of mentors and tutors belonging to the same institution where the intervention took place.

The research questions we aim to answer in this study are as follows:

RQ1: Do the features obtained through our machine learning models coincide with the opinions of mentors and tutors regarding student dropout?

RQ2: Do the values obtained from SHAP analysis correspond to plausible explanations of individual cases in the opinion of mentors and tutors?

RQ3: What additional features and/or transformations could be obtained to improve upon the models and make them accessible to practitioners?

This methodology was verified with input from practising mentors and tutors with relevant professional training to support students in higher education (e.g., psychology) and/or with training in academic advising, and their input is analyzed below. We aim to show that the models are both coherent with the experiences of the people actually working with students at risk of dropping out, and that these explainability tools could be used to build personalized intervention plans for individual cases, without heavily increasing the practitioner's workload.

1.1. Dropout Prediction

Student dropout and/or retention is an extremely prominent topic in both educational research and policy. It remains an important metric for all types of modern higher education, whether it is private or public institutions looking to improve themselves and help their students achieve their goals, or distance education that aims to reach a broader audience and provide a more flexible alternative, all institutions must wrestle with this issue. While the effect of and reasons for dropping out may be different in every case, there is no denying the general negative connotation of dropping out of education at a social and individual level.

Going as far back as Tinto (1975), reviews on the nature and reasons for dropout can be found, with discussion on how to deal with this problem continuing to today. While the strategies to deal with dropout have evolved together with technology, there is still no clear-cut answer to this problem; however, there have been several attempts to create models to better understand it (Berens et al., 2019; Kemper et al., 2020; Solis et al., 2018).

While there have been many approaches and strategies to combat dropout, the most common are early warning systems (EWSs). According to the United Nations Children's Fund (2018), an EWS is any system that uses a set of rules to try to identify students at risk of dropping out and then respond accordingly. EWSs have repeatedly shown their benefits when applied in experimental cases (Bañeres et al., 2020, 2023; Jokhan et al., 2019; Akçapınar et al., 2019). The main difference between predictive models and an EWS is that, in general, EWSs are based on a set of concrete "rules" or red flags determined by the institution. On the other hand, predictive models use any available information and, depending on the algorithm, might be able to capture more complex relationships instead of simple ones. This can lead to situations where predictive models are better at identifying at-risk students, but the reasons are not as clear as those defined in EWSs.

One problem that arises when trying to fight against dropout is the fact that there are many reasons why students leave their studies, from external factors (economic hardship, culture) to personal/internal ones (achievement potential, mental and emotional health, overall integration, illness, family issues). These features have been previously used to build predictive models that do indeed contribute to assessing dropout risk (Heublein, 2014; Russell et al., 2020), but the issue comes with trying to generalize or transfer those results. Due to the particularities of every institution, there are cultural, educational, and even individual differences that might make a model that worked well at one HEI perform mediocrely at another.

Education is a discipline that requires the application of theory to both validate it and to be of actual use; there is no meaning to a novel educational theory if it is not proven to work "in the field," and it seems that modern learning analytics is failing in this aspect. A systematic review of *Journal of Learning Analytics* articles and Learning Analytics and Knowledge conference papers found that only 11% of all articles/papers attempt to intervene in the learning environment (Motz et al., 2023). When considered together with the results shown in Herodotou et al. (2023), there is no doubt that something needs to change.

Both practitioner literacy with these new tools and their accessibility need to increase. The second point falls to us as researchers, and one way we can provide it is by using explainable and interpretable models.

1.2. Explainable AI

One of the main tools used during this project was the Python SHAP library. SHAP stands for SHapley Additive exPlanations and was first presented by Lundberg and Lee (2017) as a unified approach to what at that time were several different methods for model explanation in artificial intelligence, currently referred to as explainable AI (XAI). XAI refers to a set of artificial intelligence systems or models that can provide a meaningful explanation for their decision-making process, with the objective of helping final users — usually decision makers or stakeholders — make informed decisions instead of blindly trusting the result (Phillips et al., 2021).

As machine learning has advanced into more complex classifiers or predictors, like deep learning and ensemble models, it has also become more difficult to explain the inner working of these systems, to the point that they are commonly referred to as "black box" models. XAI helps solve this problem by delivering a series of explanations that range from global explanations that encompass the whole algorithm to local ones that can be applied to a small sample or even single cases. While the success of black box models cannot be denied, they suffer in areas like education due to their own complexity. For example, suppose our deep learning model identifies one of our students as a high-risk case for dropout. While we could approach the student at that point, what would be our message to them? Black box models do not disclose their inner workings, and even if they do, they are usually hard to interpret for non-experts. This is where XAI shines. By delivering a local explanation of the student's situation, it is possible to both better understand the specific case and help the tutors or mentors approach the student with valuable information.

SHAP values revolve around the computation of close approximations of the Shapley values of the model and a series of characteristics that make it desirable in terms of model explanations. First, we must explain what Shapley values represent: Shapley values is a term from collaborative game theory, where several players (in our case, the model features) interact to obtain a payout (prediction). The individual Shapley values refer to the marginal contribution of each player or feature to the difference between the expected value (average) and the real value. Lloyd Shapley (1953) first described this as a means of fairly estimating how much of an outcome could be attributed to each player if they were co-operating.

2. Materials and Methods

2.1. Comparison Methodology

While SHAP values and library visualizations can provide valuable explainability tools by themselves, their main target remains single-model explainability. In general, visual inspection of the figures, along with a deep understanding of the mechanisms that underlie Shapley values, could lead to informative and useful insights. This becomes a more complicated endeavour if we want everyday users to perform these analyses, as they would require expertise in data science, SHAP/Shapley values, and the specific topic the models were built for. It is unreasonable to expect this mastery for every single final user.

The comparison methodology presented in Talamás-Carvajal and Ceballos-Cancino (2024) is summarized below for clarity. The methodology allows for a model and context-agnostic comparison between two distinct populations from an explainability standpoint, allowing for the identification of changes in the effects of features between these different populations (e.g., regular students versus students under academic guidance). This methodology can be accompanied by a visual analysis so that people without AI expertise can better understand these differences (e.g., mentors and tutors). We present the methodology below and explain its application in our dataset afterwards:

1. Develop a prediction model to compare each of the subpopulations that provides an adequate level of certainty for your situation (i.e., good enough levels of accuracy, precision, recall, f1, etc.).
2. Obtain the Shapley values for the models of the populations of interest.
3. Compute Cohen's d for each set of features of interest between the populations (for example, the Shapley values for GPA for both regular and intervened students).
4. Determine a limit value for the type of change you are looking for. This corresponds to the established Cohen's d values for small, medium, or large effects (0.2 for small effects, 0.5 for a medium effect, and .8 or higher for large effects).
5. Identify the cases where Cohen's d absolute value is above the determined limit defined in the previous step.
6. Each case of a Cohen's d that surpasses the value indicates a change in the population distribution of at least the selected effect.
7. Cohen's d indicates which of the two features tends more towards the target feature. A positive value indicates the first feature in Cohen's d calculation averages values that push the prediction towards the target feature, while a negative sign indicates that the second feature in Cohen's d calculation is the one to do so.
8. Together with the feature ranking of importance from the models (obtained at the same time as the Shapley values), it is possible to interpret the data without plotting the swarm plot: a rank change indicates an importance change between the sets, while Cohen's d indicates how much the population distribution changed, and in which direction.
9. Finally, complement the information from this methodology with a visual inspection of the Shapley values from the swarm plot.

The results section will include a table highlighting the results of the methodology and will serve as an example of how to read and interpret the values obtained by these steps. It is important to note that step 2 is model agnostic, and while computation is faster with tree-based models, the base model does not need to belong to this family of algorithms. That is to say that any model could be used in step 1, from simple regressions to neural networks.

2.2. Model Dataset

The database consists of student data received from the institution's data warehouse. Several datasets were merged into one that was adequate for our objectives. Initially, we requested information regarding variables related to the overall performance of each student through their studies, including both sociodemographic and academic data. Some examples of variables for each period were as follows: age, gender, if the student's primary residence was in the campus city or not, average grade from their previous education level, type of program of their previous school, inscription status, program inside the institution, educational model, campus location, previous semester average (when applicable), failed courses, dropped courses, course load, percentage of financial aid, student progress by period, school period, and final status (i.e., student graduated, is active, or dropped out). A secondary dataset obtained from academic services was used to validate the graduation status when applicable.

All data was provided by the institutional data warehouse, and privacy issues relating to data collection, curation, and publication were validated with the relevant data owners and the Data Security and Information Management departments (Alvarado-Uribe et al., 2022).

The base dataset used for this study was initially comprised of 27 columns, which included all the previously mentioned information and some additional features that were either informative (further explanation of a different feature for case-by-case use) or dropped in the final model due to the data cleaning process. The initial dataset contained 711,846 rows for 125,198 unique students because several semesters worth of information were recorded for each individual student.

2.3. Data Preparation

Table 1 summarizes feature names and a small explanation for each one of them. We started the data preparation by taking all features with only two possible answers and transformed them into binary outputs. Some examples of this type of variable are as follows: 1) if the student was from outside the campus city or not, 2) if the previous school was from the same educational system, 3) if the student was enrolled as a regular student or not, etc.

Table 1. Feature Names and Explanations

Feature name	**Feature details**
*Scholarship**	Indicates if the student had a scholarship during that semester (1: YES, 0: NO)
*FTE **	% of course load the student had during that semester, where 1.0 means full academic load
Conditioned	Indicates whether the student was under academic probation or not during their studies (1: YES, 0: NO)
System_Highschool	Indicates if the student comes from the same family of institutions as the current one (1: YES, 0: NO)
*Foreign**	Indicates if the student's main residence was in a different city from the campus during that semester (1: YES, 0: NO)
Gender	Male: 1, Female: 2
*Sem_Interruption**	Indicates if the student requested a leave of absence during that semester (1: YES, 0: NO)
Highschool_GPA	Average of their previous degree
*Cumulative_GPA**	Average of the previous semester
*Dropped_Courses**	Number of dropped courses in that semester
*Failed_Courses**	Number of failed courses in that semester

Following this binarization, we proceeded to get rid of features with large amounts of missing or redundant data. Examples of this type of variable were cases with large percentages of missing data that was not imputable in any way (admission scores during the pandemic, scholarship descriptors), redundant features (age and year of birth, school descriptor, school code), and informative features that were not discrete categories but comment based (inscription status, motive for dropping out of a semester). Finally, some normalizations were performed in cases where the previous school's grades were on a different scale than the 100-point base (GPA, 10-point base, etc.).

There are two key factors to consider with this dataset. The first is that each data point corresponds to the performance of a student during a specific course period. In other words, we have our data organized in panel forms. While this could be useful in terms of taking the temporal component of dropout into account, we want to be able to make predictions at any point in the student's path through their studies. Due to this, we decided to transform our dataset into a more common shape, with one-row per student. The second factor is that the dataset contains no direct feature regarding student dropout, requiring us to define it ourselves. For the purposes of this article, we defined dropout as a case where a student has not graduated, is not currently active (enrolled in the latest active term), and has not enrolled for at least one consecutive year. The reason for the last condition is because single-semester sabbaticals are relatively common, either due to personal, emotional, or economic

reasons, and a good percentage of these cases return to the institution. As a quick example, a student who fails to enroll for a year after their first semester would be classified as having dropped out in their second semester. While the data regarding higher semesters was intentionally cut, dropout could happen at any point during their studies.

Following the previously mentioned transformations, mergers, and other necessary procedures, we ended up with a final dataset comprised of 69,066 unique students with 31 informative features and one dependent feature (dropout). This dataset contained 15,285 cases classified as dropout, which represents 22.13% of this sample. It is important to note that this value is likely inflated when compared to that of the institution due to the difference in definitions, the fact that several data points needed to be discarded due to missing data, and because this data does not indicate if the student changed majors or campuses instead of leaving their studies.

2.4. Validation Surveys

To validate the mathematical results obtained from the application of the methodology, we developed a qualitative, anonymous survey to be applied to active mentors and tutors currently advising students at a prestigious private university in Latin America. The objective of this survey was to collect expert opinions from active practitioners, meaning trained psychologists and career advisors.

The survey was designed to enquire about the features that were considered relevant (Cohen d values above 0.20), and the questions developed based on the features that showed those relevant results. As such, this survey consisted of two sections with a total of 20 questions, all to be answered on a 5-point Likert scale. While the phrasing in the two sections differed, all values were from –2 to 2. The first section consisted of questions to validate the choice of features used in the prediction models (were the features important or not?) and two additional questions about whether the importance of those features changed depending on the semester the student was enrolled in, and if there were visible differences in the importance of features between regular students and students in a guidance program. In this section, the lowest value (–2) corresponded to "completely disagree," and the highest value (2) to "completely agree."

The second section consisted of questions in which the importance of the features was ranked from "much more important to students in academic guidance programs" (–2) to "much more important to regular students" (2). Table 2 summarizes the survey questions and possible answers. It is important to note that due to granularity limitations, the questions in the survey do not separate features by semester.

Table 2. Summary of Survey Questions and Possible Answers

Questions	Section	Answers in the Likert Scale
Does the (Feature in the model)* positively affect student retention?	1	Completely disagree (–2) – Completely agree (2)
Does the importance of the previous variables change depending on the semester the student is enrolled in?	1	Completely disagree (–2) – Completely agree (2)
Are there visible differences between the importance of these features for regular students and students under academic guidance programs?	1	Completely disagree (–2) – Completely agree (2)
Do you consider the (Feature in the model)* to be more important for regular students or students under academic guidance programs?	2	Much more important for students on academic guidance programs (–2) – Much more important for regular students (2)

A total of 34 mentors/tutors voluntarily answered our anonymous qualitative survey, with all participants answering all questions. The answers were then taken from the survey and converted into numerical values, both for visualizing results and obtaining the average agreement among tutors regarding the features and their importance between regular students and students in academic guidance programs.

3. Results

3.1. Model Results

Using the cleaned database as described in sections 2.1 and 2.2, we trained three distinct tree-based machine learning algorithms with the objective of obtaining their specific Shapley values, and both compare them between each other and show the results to practitioners. Regardless of whether the methodology is model agnostic, in this study, tree-based algorithms were chosen as they allow for fast and exact Shapley value computations (Lundberg et al., 2018). We decided on using the XGBoost (XGB) algorithm from the library of the same name, Histogram-based Gradient Boosting Classification Tree (HB) from the sklearn library, and the Random Forest Classifier (RF), also from sklearn. All three models were trained with the same dataset, and with the same training and testing splits (80% training, 20% testing), resulting in a training set of 55,252 data points, and

a testing set of 13,814 data points. The scores for all three models can be seen in Table 3 below. We observed very similar scores between the XGB and HB models, with the RF model having better precision but worse recall and F1 scores.

Table 3. Score Summaries for the Tree-Based Models

Model	Accuracy	Precision	Recall	F1	Expected value*
XGB	0.8933	0.8712	0.6146	0. 7207	–1.4196
HB	0.8935	0.8876	0.6003	0.7162	–1.6388
RF	0.8802	0.9205	0.5092	0.6557	−1.2632

** Expected value is given in terms of SHAP values and is a log-odds number. Lower values indicate a lower probability of a student dropping out.*

Having seen the previous results, we decided to focus on the XGBoost model since it offered the overall best scores of the three. For this particular project regarding student dropout, we value recall as being more important than precision.

The specific intervention form applied in the HEI of this study is called an "academic support program" and revolves around framing and personal value interventions: the student is accompanied by mentors and tutors in regular meetings regarding their mental well-being, the challenges they are facing in their studies, and how they could approach them. Finally, they are given specialized courses in order to help them improve their time management and overall study habits. While these are not explicit task-value interventions, they serve as part of the institution's overall intervention program.

We proceeded to separate our student population into three distinct groups: a group containing only students with intervention, a group of regular students, and the full population. Intervened students are enrolled in the academic support program, which is mandatory for them due to institutional rules. Of these groups, the dropout rates were as follows: full population: 3093/13814 (22.39%); intervention population: 912/6641 (13.73%); regular population: 2176/48611 (4.48%). We can observe just from these values that students from the intervened population are more likely to end up dropping out despite their participation in the intervention program. This is consistent with the SHAP results obtained from the different populations as seen in Figure 1, which is a side-by-side comparison of the "bee swarm" plots obtained for each SHAP value set. These plots show how the value of a particular score (high or low) affects a particular student (data point) regarding its impact on the model output (our prediction for dropout or retention).

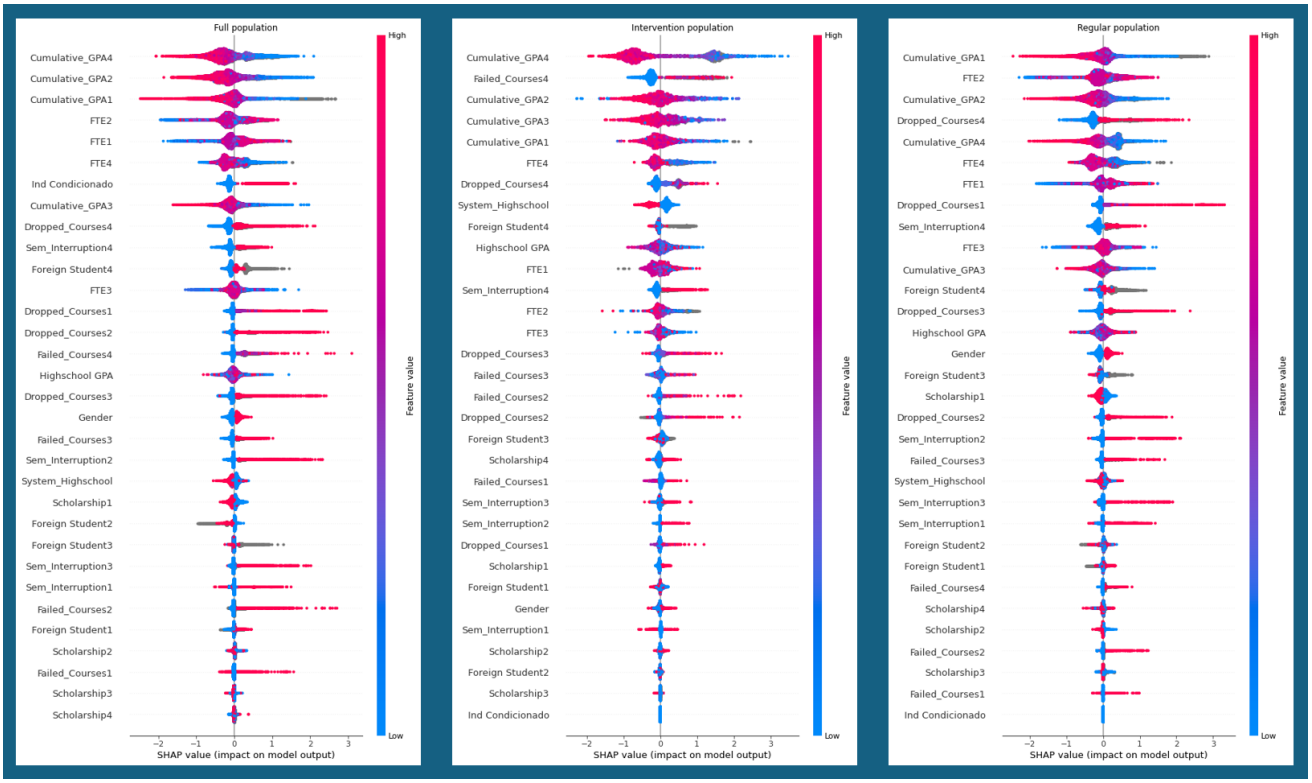


Figure 1. Side-by-side comparison of the full population (left), intervened population (middle), and regular population (right).

Already at this point, we can observe some interesting differences between the populations. In our full population, the intervention variable shows up as the seventh most important feature, but on both the intervention population and in the regular

one, it does not show up at all. The "intervention" feature captures more than just the intended overall performance of the students in their specially assigned courses since those students differ from the regular population by the very fact that they are already under a special status.

3.2. Methodology Results

We applied the methodology described in 2.1 to our models with a selected effect size of 0.20 (we are looking for anything larger than a small effect). We compared the intervened students against the regular ones and found that 17 of our 31 feature sets displayed larger values than our cut-off number. We summarize these results in Table 4.

Table 4. Summary of Features with Significative Differences in the Intervened Population

Features	Cohen's d (Int-Reg)	Ranking change (Int vs. Reg)	Interpretation
FTE2	0.206927042	–11	FTE in the second semester has greatly lower importance for intervened students; the average intervened student has a slightly higher risk of dropout than the average regular student.
FTE4	0.388835956	0	FTE in the fourth semester has the same importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Foreign Student2	-0.272460918	–6	Being a foreign student in their second semester has lower importance for intervened students; the average intervened student has slightly lower risk of dropout than the average regular student.
Foreign Student3	0.354731106	–3	Being a foreign student in their third semester has slightly lower importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Foreign Student4	0.379938664	3	Being a foreign student in their fourth semester has slightly higher importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
System_Highschool	0.205199963	13	Coming from a school from the same system has greatly higher importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Scholarship3	-0.208973997	–1	Scholarship in the third semester has slightly lower importance for intervened students; the average intervened student has slightly lower risk of dropout than the average regular student.
Scholarship4	-0.281130123	7	Scholarship in the fourth semester has higher importance for intervened students; the average intervened student has slightly lower risk of dropout than the average regular student.
Sem_Interruption4	0.295888705	–3	A semester interruption in the fourth semester has slightly lower importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Cumulative_GPA1	0.218999481	–4	GPA in the first semester has slightly lower importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Cumulative_GPA2	0.401176553	0	GPA in the second semester has the same importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Cumulative_GPA3	0.348769231	7	GPA in the third semester has higher importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Cumulative_GPA4	0.204378116	4	GPA in the fourth semester has slightly higher importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Dropped_Courses4	0.757115459	–3	Number of dropped courses on the fourth semester has slightly lower importance for intervened students; the average intervened student has higher risk of dropout than the average regular student.
Failed_Courses2	0.272823741	12	Number of failed courses in the second semester has greatly higher importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Failed_Courses3	0.34520925	3	Number of failed courses in the third semester has slightly higher importance for intervened students; the average intervened student has slightly higher risk of dropout than the average regular student.
Failed_Courses4	0.627525772	24	Number of failed courses in the fourth semester has much higher importance for intervened students; the average intervened student has higher risk of dropout than the average regular student.

It is important to mention that, due to the nature of Cohen's d, it is possible for this methodology to capture changes in features with low overall importance, and as mentioned in the steps above, the analysis should be accompanied by the feature importance rankings as well. A visual inspection of the Shapley swarm plots confirmed the distribution changes between populations for the mentioned features. A close-up comparison of some of these features can be seen in Figure 2.

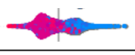
Feature	Intervention students	Regular students
FTE4		
Cumulative GPA 4		
Failed courses 4		
Cumulative GPA 2		

Figure 2. Side-by-side comparison of the intervened and regular student populations.

We can observe clear changes for all four sets shown above, from clusters moving or appearing to move for the population regarding the 0–Shapley value (vertical line in all images). An example interpretation of the Cumulative_GPA4 feature (taking into account all the information mentioned in the article so far) would be that for both regular and intervened students, high GPA scores in the fourth semester aid in student retention, and low scores push students towards dropout, but scores in the medium ranges are not as negative for intervened students as they are for regular ones, as can be seen by the blue-red-purple cluster on the left side of the axis in the intervened population.

3.3. Survey Results

Regarding Section 1 of the survey, we obtained values ranging from 0.9394 at the lowest and 1.4242 at the maximum regarding feature importance. As a value of 1 corresponds to "Agree," the mentors and tutors agree that those features are important regarding student dropout. The average result for feature importance changes per semester was 1.2121 and for differences between regular students versus students on academic guidance the value was 1.1818, meaning that experts agree that feature importance changes both by semester and by group.

Section 2 of the survey is summarized in Table 5. We found that the opinions of the surveyed experts matched our mathematical results in four out of the six variables of the survey. The methodology values were taken from the ranking changes when the feature had a Cohen score magnitude above the threshold and assigned as 0 when they were not.

Table 5. Comparison of Methodology Results versus Expert Survey Results

Feature	Methodology values	Survey values	Match
Scholarship	1	0.424242	✓
FTE	11	0.393939	✓
Cumulative GPA	4	0.424242	✓
Failed Courses	–24	–0.48485	✓
Dropped Courses	3	–0.48485	X
Semester Interruption	3	–0.39394	X

4. Discussion

The purpose of this research is to validate the results of the methodology presented in Talamás-Carvajal and Ceballos-Cancino (2024) with the experience and expertise of mentors and tutors who currently work in the HEI where the data was obtained. It is important to us that the results are both useful to the people who will ultimately use the information and consistent with their specific contexts. For this to be the case, we needed to make sure that the results obtained via our methodology made sense to the experts who could become the final users in future. AI tools with low levels of trust or confidence are generally at a disadvantage when looking for long-term adoption and use (Cave et al., 2019; Stanton & Jensen, 2021; Rahiman & Kodikal, 2024). As we mentioned in the introduction, the validation step by its potential users is a crucial one. From our findings, there are several insights that can be extracted from both Figure 1 and Tables 4 and 5. The first is that the intervention effect seen in the total population could be capturing the background propensity of these students of dropping out, as that population is already at a significant dropout risk according to the HEI indicators. The effect captured in the "conditioned" feature in the full population should still be present in some shape in the intervened population and removed completely from the regular

one. We believe that many of these changes are "absorbed" by the other features present in the dataset. By using our methodology, it is possible to identify these relevance changes, both in overall importance compared to the rest of the features, and the individual feature changes between populations. Consequently, the use of this methodology can be beneficial to mentors and tutors, as it will allow them to identify features that become more relevant for specific populations and how.

Thus, one of the most important innovations and contributions of this work is the identification of actionable insights that can be directly tied to the differences between groups. We can identify cases where the effects of features for regular and intervened students are different to the point that they are sometimes even opposite (one such example can be seen on the effect of sports between the intervention and regular populations). Even recently published articles that deal with dropout predictions do not consider the possible differences between student subgroups or how these might be leveraged, which in turn cycles back to the problems mentioned before regarding long-term adoption and AI tool use. Some examples can be seen in Roslan et al. (2025), Peng et al. (2024), and Jiménez-Gutiérrez et al. (2024). These three publications used different techniques to predict dropout or dropout intentions, and all of them report high levels of accuracy and AUC scores. However, they do not necessarily report differences among students or potential interventions derived from their results. Our methodology could be used as a follow-up to those studies, aiming to find the difference between the groups and provide academic advisors with valuable information to support specific students. It is also possible to isolate the actionable features from non-actionable ones, so that information can be used to produce highly effective interventions by the mentors and tutors.

From the survey results and their comparison against our methodology results, we found a consensus between expert opinions and our mathematical results (4/6, 66.66%). The notion that the feature importance could change between semesters is something that we also observed in our results, and this same concept could explain cases where there was no match between the survey and our method. One possible interpretation is that the model is capturing this effect while the tutors are not focusing on the semester as a possible explanatory feature. The results from both the methodology and the mentor/tutor opinions are consistent with Heublein (2014) and Russell et al. (2020). A systematic review from Liz-Domínguez et al. (2019) also mentions that the progress of a student in their studies (i.e., their year) has been shown to affect learning analytics predictions. That same review mentions how EWS have both been sparsely used outside of experimental setups when combined with AI and that those results commonly go unexploited as they require people to interpret and act upon them. At this point, our results could make a difference, as we offer explanations regarding the predictions, which we hope will, in turn, inform interventions.

Still, there are some limitations to consider, as with every research project. Regarding missing features and transformations that the mentors/tutors believe could be of use to these prediction models, the unanimous response was psychological and emotional well-being information. All the interviewed mentors quickly mentioned that this information could be extremely useful for the models, either in the form of categories or made into a number in some manner. However, all mentors also commented that this information could prove difficult to obtain even in its most basic form, as it deals with deeply personal and protected information in the institution. Additionally, validation was performed in a mainly quantitative manner using a Likert scale, and on a limited case study. While valuable, education as a discipline often deals with significant differences between contexts and even cohorts under the same HEI and part of the future work is focused on additional validations.

Readers with more experience in machine learning might be wondering why Shapley values would be needed at all if the models we selected belong to the tree-based family of algorithms, as those models are already established explainability tools and/or transparent or white box models. It is important to note that one of the main issues HEIs have with the adoption of AI tools is the lack of reliability in practice, as many models are not entirely generalizable across cohorts. In this sense, one possible solution is the use of more complex machine learning models. However, the increase in predictive power comes with an important increase in model complexity or size, to the point that even if the models are explainable in theory (like tree-based algorithms), sufficiently complex or large models lose that characteristic. This can be easily explained by a thought experiment: imagine a decision tree with three branches. The results from such an algorithm are easily interpretable by a human without much issue. Now imagine instead of three branches, we have 10. While still interpretable, it has a much higher complexity. Finally, imagine a system that computes hundreds or thousands of such trees, and makes its final decision by majority vote. Such a system would be impossible to interpret by a person, which is exactly what happens in Random Forest models. This specific example and the previous argument were given by Petch et al. (2022), along with a series of recommendations on when to use interpretable models and when to go for black box models with additional explanations. The choice of models for this validation was made based on some of the currently commonly used and popular algorithms, but other researchers could as easily use the methodology on results obtained from neural networks, for example.

Future work will include both an independent validation in a distinct context to the one presented here and comparisons of dropout or viable academic phenomena between distinct HEIs. We plan to initially compare two and then extend the methodology to a larger group. We also plan to perform in-depth interviews with the more experienced mentors of the universities in order to improve upon the methodology and guide its possible use to feed academic-advising dashboards, either for students, mentors/teachers, or both.

5. Conclusions

Going back to our research questions, we found the following: 1) the features obtained through our machine learning models coincide with the opinions of interviewed mentors and tutors; 2) the explanatory values obtained through the methodology match the real-world experiences of mentors and tutors, especially when dealing with highly explanatory features like previous average grades and interventions; 3) the additional features that would be most beneficial to these models are psychological and emotional well-being information from students, which could prove difficult to obtain.

Thus, these results could be used to build a holistic system that predicts student dropout and delivers context and user-specific explanations, which in turn can be leveraged to best avoid the dropout prediction by allowing the design of achievable individualized interventions from mentors and tutors.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability

The data that support the findings of this study are available from the Institute for the Future of Education (IFE)'s Educational Innovation collection of the Tecnologico de Monterrey's Research Data Hub, but restrictions apply to the availability of these data, which were used under a signed Terms of Use document for the current study, and so are not publicly available. Data are, however, available from the IFE Data Hub upon reasonable request at https://datahub.tec.mx/dataverse/datahubprivate (accessed on 6 October 2024).

Funding

Tecnológico de Monterrey's authors are grateful for the institutional funding provided via seed funds from "TEC-CSIC Ethical Challenges of the Use of Artificial Intelligence Towards 2030 (DesafIA2030)" and "TEC-PUC Identifying Changes in Feature Effects Between Regular and At-Risk Students in Higher Education across Latin America." The work of Isabel Hilliger was funded by the Chilean National Agency for Research and Development (ANID) under the Millennium Nucleus NCS2021_083 grant for a project entitled "Student Experience in Higher Education in Chile: Expectations and Realities."

Acknowledgments

The authors would like to thank Tecnológico de Monterrey and the Data Hub of the Institute for the Future of Education for the data provided for this research.

References

Akçapınar, G., Altun, A. & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education, 16*, 40. https://doi.org/10.1186/s41239-019-0172-z

Alvarado-Uribe, J., Mejía-Almada, P., Masetto Herrera, A. L., Molontay, R., Hilliger, I., Hegde, V., Montemayor Gallegos, J. E., Ramírez Díaz, R. A., & Ceballos, H. G. (2022). Student dataset from Tecnológico de Monterrey in Mexico to predict dropout in higher education. *Data, 7*(9), 119. https://doi.org/10.3390/data7090119

Arin, J. (2022). The missing link in academic advising: The faculty perspective. In B. McMurtrie & B. Supiano (Eds.), *The future of advising: Strategies to support student success* (pp. 40–43). Chronicle of Higher Education.

Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A. E., & Karadeniz, A. (2020). An early warning system to detect at-risk students in online higher education. *Applied Sciences, 10*(13), 4427. https://doi.org/10.3390/app10134427

Bañeres, D., Rodríguez-González, M. E., Guerrero-Roldán, A.-E., & Cortadas, P. (2023). An early warning system to identify and intervene online dropout learners. *International Journal of Educational Technology in Higher Education, 20*, 3. https://doi.org/10.1186/s41239-022-00371-5

Berens, J., Schneider, K., Görtz, 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

Calhoun-Brown, A. (2022). How data and technology can improve advising and equity. In B. McMurtrie & B. Supiano (Eds.), *The future of advising: Strategies to support student success* (pp. 36–39). Chronicle of Higher Education.

Cave, S., Coughlan, K., & Dihal, K. (2019). "Scary robots": Examining public responses to AI. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '19), 27–28 January 2019, Honolulu, HI, USA (pp. 331–337). ACM Press. https://doi.org/10.1145/3306618.3314232

Harackiewicz, J. M., & Priniski, S. J. (2018). Improving student outcomes in higher education: The science of targeted intervention. *Annual Review of Psychology, 69*, 409–435. https://doi.org/10.1146/annurev-psych-122216-011725

Herodotou, C., Maguire, C., Hlosta, M., & Mulholland, P. (2023). Predictive learning analytics and university teachers: Usage and perceptions three years post-implementation. *Proceedings of the 13th International Learning Analytics and Knowledge Conference* (LAK '23), 13–17 March 2023, Arlington, TX, USA (pp. 68–78). ACM Press. https://doi.org/10.1145/3576050.3576061

Heublein, U. (2014). Student drop-out from German higher education institutions. *European Journal of Education, 49*(4), 497–513. https://doi.org/10.1111/ejed.12097

Hilliger, I., De Laet, T., Henríquez, V., Guerra, J., Ortiz-Rojas, M., Zuñiga, M. Á., Baier, J., & Pérez-Sanagustín, M. (2020). For learners, with learners: Identifying indicators for an academic advising dashboard for students. In C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, & S. M. Dennerlein (Eds.), *Addressing global challenges and quality education: 15th European Conference on Technology Enhanced Learning, EC-TEL 2020, Heidelberg, Germany, September 14–18, 2020, Proceedings* (pp. 117–130). Springer, Cham. https://doi.org/10.1007/978-3-030-57717-9_9

Jiménez-Gutiérrez, A. L., Mota-Hernández, C. I., Mezura-Montes, E., & Alvarado-Corona, R. (2024). Application of the performance of machine learning techniques as support in the prediction of school dropout. *Scientific Reports, 14*, 3957. https://doi.org/10.1038/s41598-024-53576-1

Jokhan, A., Sharma, B., & Singh, S. (2019). Early warning system as a predictor for student performance in higher education blended courses. *Studies in Higher Education, 44*(11), 1900–1911. https://doi.org/10.1080/03075079.2018.1466872

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

Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research, 86*(2), 602–640. https://doi.org/10.3102/0034654315617832

Liz-Domínguez, M., Caeiro-Rodríguez, M., Llamas-Nistal, M., & Mikic-Fonte, F. (2019). Predictors and early warning systems in higher education: A systematic literature review. *Proceedings of the Learning Analytics Summer Institute Spain 2019: Learning Analytics in Higher Education* (LASI Spain 2019), 27–28 June 2019, Vigo, Spain (pp. 84–99). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-2415/paper08.pdf

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS '17), 4–9 December 2017, Long Beach, CA, USA (pp. 4768–4777). Curran Associates. https://dl.acm.org/doi/10.5555/3295222.3295230

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). *Consistent individualized feature attribution for tree ensembles*. arXiv. https://doi.org/10.48550/arXiv.1802.03888

McMurtrie, B., & Supiano, B. (2022a). Concerns about bias in advising technology. In B. McMurtrie & B. Supiano (Eds.), *The future of advising: Strategies to support student success* (pp. 33–39). Chronicle of Higher Education.

McMurtrie, B., & Supiano, B. (2022b). The barriers to better advising. In B. McMurtrie & B. Supiano (Eds.), *The future of advising: Strategies to support student success* (pp. 7–12). Chronicle of Higher Education.

Motz, B. A., Bergner, Y., Brooks, C. A., Gladden, A., Gray, G., Lang, C., Li, W., Marmolejo-Ramos, F., & Quick, J. D. (2023). A LAK of direction: Misalignment between the goals of learning analytics and its research scholarship. *Journal of Learning Analytics, 10*(2), 1–13. https://doi.org/10.18608/jla.2023.7913

Peng, P., Liu, L., Wu, Q., Tang, Y.-Y., Tang, J., Liu, T., & Liao, Y. (2024). Establishment and validation of a nomogram for dropout intention in Chinese early year medical undergraduates. *BMC Medical Education, 24*, 868. https://doi.org/10.1186/s12909-024-05835-y

Petch, J., Di, S., & Nelson, W. (2022). Opening the black box: The promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology, 38*(2), 204–213. https://doi.org/10.1016/j.cjca.2021.09.004

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence: National Institute of Standards and Technology, interagency or internal report* (NISTIR 8312). U.S. Department of Commerce. https://doi.org/10.6028/NIST.IR.8312

Rahiman, H. U., & Kodikal, R. (2024). Revolutionizing education: Artificial intelligence empowered learning in higher education. *Cogent Education, 11*(1), 2293431. https://doi.org/10.1080/2331186X.2023.2293431

Rodway, P., & Schepman, A. (2023). The impact of adopting AI educational technologies on projected course satisfaction in university students. *Computers and Education: Artificial Intelligence, 5*, 100150. https://doi.org/10.1016/j.caeai.2023.100150

Roslan, N., Mohd Jamil, J., Mohd Shaharanee, I. N., & Sultan Alawi, S. J. (2025). Prediction of student dropout in Malaysian's private higher education institutes using data mining applications. *Journal of Advanced Research in Applied Sciences and Engineering Technology, 45*(2), 168–176. https://doi.org/10.37934/araset.45.2.168176

Russell, J.-E., Smith, A., & Larsen, R. (2020). Elements of success: Supporting at-risk student resilience through learning analytics. *Computers & Education, 152*, 103890. https://doi.org/10.1016/j.compedu.2020.103890

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games (AM-28), Volume II* (pp. 307–318). Princeton University Press. http://www.jstor.org/stable/j.ctt1b9x1zv.24

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. *Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence* (IWOBI 2018), 18–20 July 2018, San Carlos, Costa Rica (pp. 1–6). IEEE. https://doi.org/10.1109/IWOBI.2018.8464191

Stanton, B., & Jensen, T. (2021). *Trust and artificial intelligence: National Institute of Standards and Technology interagency/internal report* (NISTIR 8332). U.S. Department of Commerce. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087

Talamás-Carvajal, J. A. (2023). The middle-man between models and mentors: Using SHAP values to explain dropout prediction models in higher education. *Companion Proceedings of the 13th International Conference on Learning Analytics & Knowledge* (LAK '23), 13–17 March 2023, Arlington, TX, USA (pp. 68–70). SoLAR.

Talamás-Carvajal, J. A., & Ceballos-Cancino, H. G. (2024). Use of SHAP values for identifying differences in behaviors for subpopulations under intervention. *Joint Proceedings of LAK 2024 Workshops* (LAK-WS 2024), 18–22 March 2024, Kyoto, Japan (pp. 139–149). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3667/DS-LAK24_paper_4.pdf

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89–125. https://doi.org/10.3102/00346543045001089

United Nations Children's Fund. (2018). *Early warning systems for students at risk of dropping out: Policy and practice pointers for enrolling all children and adolescents in school and preventing dropout*. https://www.unicef.org/eca/sites/unicef.org.eca/files/2018-11/Early%20warning%20systems%20for%20students%20at%20risk%20of%20dropping%20out_0.pdf

Valenzuela, J. P., & Yáñez, N. (2022). *Trajectory and policies for inclusion in higher education in Latin America and the Caribbean in the context of the pandemic: Two decades of progress and challenges.* Economic Commission for Latin America and the Caribbean (ECLAC).

Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science, 23*(1), 73–82. https://doi.org/10.1177/0963721413512856