Use of SHAP values for identifying differences in behaviors for subpopulations under intervention

Juan A. Talamás-Carvajal[a], Hector G. Ceballos-Cancino[b]

[a] *Tecnologico de Monterrey, School of Engineering and Science, Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, N.L, Mexico*

[b] *Tecnologico de Monterrey Institute for the Future of Education, Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, N.L, Mexico*

Abstract

The advent of Artificial Intelligence (AI) is currently leading a new industrial revolution on almost all aspects of human life. Adoption of AI in traditional education has been lower than expected due to several reasons, including a lack of understanding of the processes behind it, which is fatal for situations like student dropout. An ideal AI tool for this problem would provide individually tailored interventions towards student retention, but that would require a much deeper understanding of what entails a successful intervention. Using a novel methodology for feature comparison between subpopulations, we found that the features obtained through our machine learning models coincide with both the opinion of interviewed mentors/tutors and with independently performed research with the same dataset origin, that the explanations obtained regarding student dropout match the real-world experiences of mentors and tutors, especially when dealing with highly explanatory features like previous average grades and interventions, and that additional beneficial features would be psychological and emotional well-being information. The results from our proposed methodology were validated directly by practicing mentors and tutors that deal with student dropout on a regular basis.

Keywords

Dropout, XAI, Intervention, Higher Education, Educational Innovation

1. Introduction

With the advent of Artificial Intelligence (AI) into the mainstream of society, it has become clear that we are dealing with a change in our world of the same importance as several of the past industrial revolutions. The potential and implications for the use of AI in almost any discipline are hard to measure, but one area where its effects could have lasting and relevant effects in the future is in the field of education. We are dealing with a tool that could develop perfect individualized learning plans on one side or destroy critical thinking on the other.

Currently, the world seems to be embracing all AI tools and products, but this is not the case in education. The adoption of AI in traditional education appears to be slower than what we might expect [1], and while there is no denying the uses and potential of these tools, it might be the case that two main issues are slowing down adoption: fear of misuse and lack of understanding from the side of the practitioners and final users [2, 3]. While there needs to be a widespread effort from the Learning Analytics community to aid in the adoption of proven methods and tools, we believe that those same tools should be as user friendly as possible. This does not mean that tools should be designed with the general public in mind, but they should definitely be developed for the final user. The most common manner in which this information has been delivered is through the use of dashboards.

One of the many avenues where AI tools could have a large benefit is on individualized advising for students. Currently, advising remain a difficult topic due to many factors: the state of advising inside the institution [4], the difficulty for delivering appropriate and timely advice to students [5, 6], or even the lack of clarity regarding what must be done [7]. This becomes even more complicated when we look into advising for students that are at risk of dropping out or on academic probation.

Joint Proceedings of LAK 2024 Workshops, co-located with the 14th International Conference on Learning Analytics and Knowledge (LAK 2024), Kyoto, Japan, March 18-22, 2024.
EMAIL: juan.talamas@tec.mx (A. 1); ceballos@tec.mx (A. 2)
ORCID: 0000-0002-6140-088X (A. 1); 0000-0002-2460-3442 (A. 2)

Ideally, an AI tool that could deliver individualized recommendations for each student and is capable of outputting specific and achievable counterfactuals for at-risk students would be a dream come true. However, we require a much deeper understanding of what entails a successful intervention (helping the student graduate on a timely basis as an example) before we can make a tool to do it for us.

There are several types of interventions, starting from the Institution-wide programs to what are called targeted interventions. These interventions are called so because they "are theoretically precise and address basic psychological processes that can interfere with optimal academic functioning" [8]. Targeted interventions have been shown to have a definite positive effect when applied at an appropriate time for the students [9, 10]. These characteristics highlight two important aspects of successful interventions: doing so in time and having a concrete plan for the intervention.

Targeted Intervention can be separated into 3 main distinct types: task value, framing, and personal value interventions. These are all student-centered, all encourage the students to engage in written or spoken reflections, and all focus on a psychological process through the information they provide. The difference between them is the topic of student's attention [8]. Task value interventions explicitly state the importance of a specific theme or topic and are well equipped to address course or field specific challenges. Framing interventions deal with how students frame the challenges they face throughout their academic paths, and common examples include the feeling of not belonging, imposter's syndrome, and adaptation problems, for example. These interventions help students deal with their specific situations by framing these challenges as both common and solvable. Finally, value interventions reinforce students' self-worth and identity, and could be considered the most general of the 3 types.

Any AI tool aiming to provide personalized intervention plans, or even individualized suggestions, should in some way or another include information or features that correspond with the interventions that could be provided. Task value information could be linked to specific course grades, while extracurricular activities and some additional features could be used to recommend framing and value interventions. Additionally, this information should be presented to the non-data experts in an appropriate manner, for which data storytelling could be used a base for the dashboard design.

In this paper we present an analysis of the previously mentioned issues from the perspective of a dropout prediction model by using an explainability tool that was aimed towards the measurement of intervention effects regarding dropout prevention in a Higher Education Institution. We present a set of initial base models developed for early dropout risk detection, which then were fed through the SHAP python library. The models include a special feature that measures a student's performance in academic tutoring and mentoring programs. This is used as an informative feature, and our future work aims to measure the effectiveness of these programs through modeling. While feature importance has existed for some time and is available for several machine learning models, we believe a more in-depth analysis of these features that includes input from active teachers, mentors, and tutors is valuable and could increase our understanding of what truly matters on student interventions.

The questions we aim to answer in this study are:

- Do the features obtained through our machine learning models coincide with the opinion of mentors and tutors regarding student dropout?
- Do the values obtained from SHAP analysis correspond to plausible explanations of individual cases in the opinion of mentors and tutors?
- What missing features and/or transformations could be obtained to improve upon the models and make them accessible to practitioners?

These models have been verified with comments from practicing mentors and tutors of the institution, and their input is analyzed below. We aim to show that the models are both coherent with the experiences of the people 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

Both in public and private institutions, and in distance or traditional learning, dropout remains an important metric for all stakeholders involved. While the overall importance and effect of dropout is very different in each of the cases mentioned before, there is no doubt that it is a topic that remains

relevant to all of them. In the case of Higher Education institutions (HEI), dropout is generally seen as a student's failure to obtain their degree. While not an exact equivalence, some HEIs measure graduation rates and timely graduation rates instead.

Going as far back as 1975 [11], reviews on the nature and reasons for dropout can be found, with discussion on how to deal with this problem continuing up to this day. While the strategies to deal with this problem have evolved together with technology, there is no clear-cut answer to this problem, and there have been several attempts to create models to better understand this phenomenon [12, 13, 14].

Whatever the model may be the reasons for dropout are varied and diverse, from external factors to personal/internal ones. Several models [15, 16] have previously shown these factors to contribute to overall dropout risk but transfer of those models into specific cases is usually not straightforward, as every institution has its particularities, and even the overall culture of the city or country might affect what leads to dropout. This problem is further increased by the relatively low usage of predictive learning analytics in HEI. For example, in [17], only 42% of the interviewed teachers were actively using tools of this nature, 19% had never even heard of such tools, 18% had heard but not used the tools, and 20% had tried to use them and stopped.

1.2. SHAP and Shapley values

One of the main tools used during this project was the Python SHAP library. SHAP stands for SHapley Additive exPlanations and was first presented in [18] as an unified approach to what at that time were several different methods for model explanation in Artificial Intelligence and is currently being referred to as Explainable AI (XAI). XAI refers to a set of Artificial Intelligence systems or models that can provide a meaningful explanation behind their decision-making process, with the objective of helping final users (usually decision makers or stakeholders) make informed decisions instead of blindly trusting a result [19].

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 they are commonly referred as "black box" models. XAI helps solve this problem by delivering a series of explanations, which range from global explanations that encompass the whole algorithm, to local ones that can be applied to a small sample or even singular cases. While the success of black box models can't be denied, they suffer in areas like education due to their own complexity. One example: say our Deep Learning model identifies one of our students as a high-risk case for dropout. While we could approach the students at that point, what would be our message to them? Black box models don't 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, it is possible to both better understand the specific case and help the tutors or mentors approach the student with valuable information for them.

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 together 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. They were first described by Lloyd Shapley in [20] as a means of fairly estimating how much of an outcome could be attributed to each player if they were cooperating. Shapley values come with a series of characteristics that make them desirable as fair representations of cooperative games:

- Efficiency: The sum of all contributions for one game results in the difference between the expected value of the game and the real value (average of the model vs the predicted value or probability of the model).
- Symmetry: If two players contribute the same, their Shapley values will be the same.
- Null players: A player with no contribution will have a Shapley value of 0. This is especially important in machine learning because of the common use of "Dummy" variables in some models.
- Linearity/Additive properties: For a collaborative game that is made of other games, the Shapley values of the different games add up to the values of the combined game.

With this information, we can now proceed to its use in XAI. One of the main strategies for explainability is additive feature attribution: for each individual prediction, how much did each feature

contribute to the final decision? This is based on the notion that linear explanations are both easier to understand and valid for local points. The desirable qualities for an additive feature attribution method are the following: local accuracy is maintained, meaning the explanation's result matches the model's original one; the absence of a value should have no impact over the result; an increase on the contribution of an input should never decrease its model contribution. Following those characteristics, we can observe that Shapley values follow both the additive feature attribution method definition and the desirable qualities of the method. From a mathematical standpoint, Shapley values offer an unique solution to the feature additive attribution problem, and their characteristics allow for a more intuitive understanding of individual predictions.

It is from this notion that the SHAP library was built. The exact computation of Shapley values is computationally expensive and grows exponentially with the number of informative features, therefore approximations are necessary for efficiency's sake in complex models. The original SHAP paper [18] shows that the approximations used on the library closely resemble the true Shapley values, while more specialized ones were developed shortly after, as is the case of their "Tree SHAP" algorithm [21]. This algorithm was shown to be capable of computing exact Shapley values in low order polynomial time.

One of the main benefits of using the SHAP library is the ability to create visualizations regarding the explanations obtained from a model. These can range from individualized explanations (figure 1) to global explanations (figure 2), and they vary in shape, style, and information provided. In our case, we used waterfall plots for individual visualizations, and beeswarm plots for global ones. Waterfall plots like the one in figure 1 below present how feature effects push a single prediction towards the model outcome, with the values of the features causing either a positive or negative effect, the sum of which adds up to the final prediction value. For waterfall plots like the one presented in figure 1, each prediction starts at the expected value (prediction average) of all data points, and the effect of each feature is represented as a positive or negative value. The sum of all these effects is equal to the final prediction value, and the waterfall plot represents this by using blue and red bars. A general reading of these plots is performed from the bottom towards the top, as the effect and direction of each feature can be followed as they sum up to the final value.

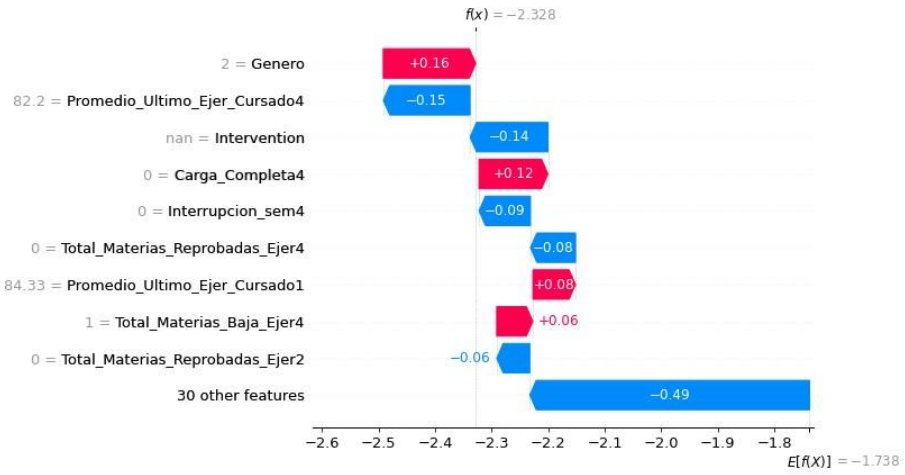

Figure 1. Example of a waterfall plot for individual explanation of a model output

Global explanations are visualized using beeswarm plots. These plots show all data points, and how the specific features affected the model outcome for each case. Figure 2 is a close-up of an example of a beeswarm plot. In this figure it is possible to visualize several characteristics of the model effects: first, every data point corresponds to an unique case. The color of each points indicated the feature value, with red being high and blue being low (depending on the feature ranges). The position of the points indicates the effect of the feature towards the target outcome. The vertical line going through all features is the point where the features have no effect, so any point to the right indicates a higher propensity towards the target, and points to the left indicate a lower one. Finally, the features are vertically arranged from higher average effect to lower average effect. Using these visualizations, it becomes possible to approach practitioners with much more than just a series of numbers and predictions, but also feature effects, tendencies, and even possible targets for interventions.

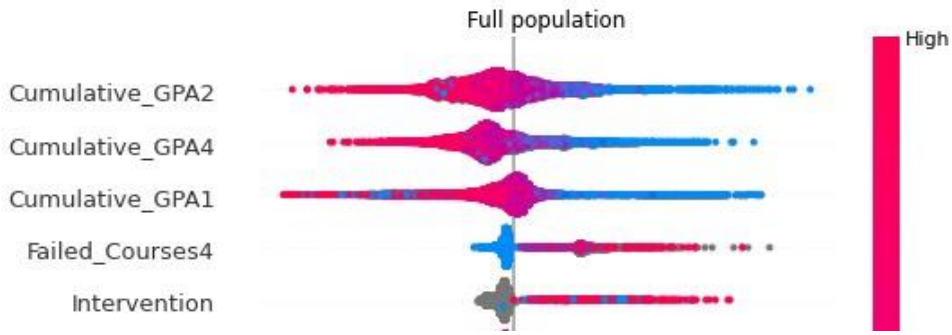

Figure 2. Close-up example of a beeswarm plot for global explanation of a model output

Given these characteristics, SHAP values can be used to deliver visual and numerical explanations regarding black box or hard to interpret models, and together with concepts from data storytelling, could be used to deliver high quality information to non-data experts, which are ultimately the ones who need to act on the information, specially in setting like education in which mentors and tutors are the main contact with struggling students.

2. MATERIALS AND METHODS

2.1. 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 the overall performance of each student through their studies, including both socio-demographic and academic data. Some examples for each period were as follows: age, gender, the student's primary residence, average grade from their previous education level, type of program of their previous school, educational model, previous semester average (where applicable), failed courses, dropped courses, course load, scholarship percentage, student progress by period, school period, and final status (student graduated, is active, or dropped out). A secondary dataset obtained from academic services was used to validate the graduation status where applicable. A third dataset included the student's extracurricular participation inside the school (participation in sports, and cultural or leadership activities).

To merge the relevant data together, the data warehouse provided unique identification numbers for each student. These numbers are randomized and anonymized, to avoid data triangulation. 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. The base dataset used for this study was initially comprised of 36 rows 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 dataset contained 708,266 rows for 124,507 unique students. This is because several semesters worth of information were recorded for each individual student.

2.2. Data preparation

A summary table that includes our feature names and a small explanation for each one is provided in Table 1. We started the data preparation by taking all features with only 2 possible answers and transformed them into binary outputs. Some examples of this were gender, if the student was from outside the Campus city or not, if the previous school was from the same educational system, 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
*Conditioned**	Indicates if the student was under academic probation or not during that semester (1: YES, 0: NO)
culture	Indicates if the student was participating in cultural extracurricular activities at any point during the semesters

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 on 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)
Intervention	For student's that attended academic improvement courses, how they performed in those courses
leadership	Indicates if the student was participating in leadership-based extracurricular activities at any point during the semesters
Highschool_GPA	Average of their previous degree
*Cumulative_GPA**	Average of the previous semester
sports	Indicates if the student was participating in sports-based extracurricular activities at any point during the semesters
*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 were cases with large percentages of missing data that was not imputable in any way, redundant features, and informative features that were not discrete categories, but comment based. Finally, some normalizations were performed in cases where the previous schools' grades were on a different scale than the 100-point base (GPA, 10-point base, etc.).

The dataset contains no direct feature regarding student dropout, requiring us to define it ourselves. For 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 in at least a 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 that fails to enroll for a year after their 1st semester would be classified as having dropped out in their 2nd semester. While the data regarding higher semesters was intentionally cut, dropout could happen at any point during their studies.

Finally, we included a feature named "Intervention", which summarizes different student's recorded performance in the institution's Academic Advising Program. This program requires students to take specific courses aimed at providing guidance and aid towards their academic life, either in a preventive or corrective manner. Following the previously mentioned transformations, merges, and other necessary procedures, we ended up with a final dataset comprised of 69,732 unique students, with 39 informative features (including our Intervention column) and one dependent feature (dropout). This dataset contained 13,763 cases classified as dropout, which represents 19.7% of this sample.

3. RESULTS

Using the now cleaned database, we trained 3 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. Tree-based algorithms were chosen as they allow for exact Shapley value computations [21] instead of close approximations, which would be useful but less reliable for this project. 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 tree 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 55,785 datapoints, and a testing set of 13,947 data points. The scores for all three models can be seen in Table 2 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 2. Score Summaries for the tree-based models

Model	Accuracy	Precision	Recall	F1	**Expected value***
XGB	0.9158	0.8850	0.6448	0.7461	-1.7378
HB	0.9150	0.8945	0.6314	0.7403	-1.9060
RF	0.9056	0.9403	0.5424	0.6880	-1.3943

*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.

We compared the SHAP values obtained from the three models, and found out by using Pearson's correlation and Cohen's d that there was little difference between the overall SHAP values obtained from the different models, that the SHAP values were highly correlated (Pearson's correlation over 0.5 for 75.3% of the paired comparisons), and with a small distance between their means in the majority of the cases (Cohen's d below 0.5 for 91.5% of the pairs). As SHAP values depend on the overall quality of the model for their won, we can say that the small variability of the values could come from the model's own variability. Having seen the previous results, we decided to focus on the XGBoost model, as it offered the overall best scores of the three. For this project regarding student dropout, we value recall as having more importance 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, and 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 3 distinct groups: full population, intervened students, and regular students. 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: 2675/13947 (19.17%); Intervention population: 697/1995 (34.93%); Regular population: 1978/11952 (16.54%). We can observe from these values that students from the intervened population are more likely to drop out despite their participation in the intervention program. This is consistent with the SHAP results obtained from the different populations.

Already at this point we can observe some interesting differences between the populations. In our full population the intervention variable shows up as the 5th most important feature, but on both the intervention population and in the regular one, it shows either almost at the bottom, or with no effect at all. The "Intervention" feature captures more than just the intended overall performance of the students in their specially assigned courses, as those students differ from the regular population by the very fact that they are already under a special status. As this program deals with the overall situation of the students, we were interested in the effect of intervention on the different features, and not only a "difference between groups" like what would be available through a statistical analysis.

Visual inspection of the swarms can allow for informative insights regarding feature importance by a direct comparison in the order of appearance of the features, but a more detailed analysis might prove complicated from the standpoint of everyday users, as a complete interpretation would require expertise both in data science and SHAP values, and education from the standpoint of a mentor/tutor.

To make this process more palatable, we propose a generalized methodology for feature comparison between two distinct populations from an explainability standpoint. This method allows for the identification of changes in the distribution of Shapley values between populations using already stablished mathematical properties, and it can be accompanied by a visual analysis of the bee-swarm plots to determine differences in what is it that truly matters for distinct populations. We present the methodology below and show its application in our dataset afterwards.

The methodology is defined as follows:

1. Obtain the Shapley values for the populations of interest.
2. Compute Cohen's d for each set of features of interest between the populations (for example, obtain the Shapley values for GPA for both regular and intervened students)
3. Determine a limit value for the type of change that 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 .08 or higher for large effects)
4. Identify the cases where Cohen's d absolute value is above the determined limit defined in the previous step.
5. Each case of a Cohen's d that surpasses the value indicates that there is a change in the population distribution of at least the selected effect.
6. Cohen's d sign 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.
7. 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.

8. Finally, complement the information from this methodology with a visual inspection of the Shapley values from the swarm plot.

We applied this same methodology to our models with a selected effect size of 0.2 (we are looking for anything larger than a small effect). We compared the intervened student against the regular ones and found that 14 out of our 39 feature sets displayed larger values than our cut-off number. We summarize these results in Table 3.

Table 3. Summary of features with significative differences in the intervened population

Features	**Cohen's d** (**Int-Reg**)	Ranking change (Int vs Reg)	Interpretation
Scholarship1	0.30378	+2	Scholarship in the 1st semester has slightly lower importance for intervened students, and the average intervened student has slightly higher risk of dropout due to the feature's effects than the average regular one.
Scholarship2	-0.27094	+1	Scholarship in the 2nd semester has slightly lower importance for intervened students, and the average intervened student has slightly lower risk of dropout due to the feature's effects than the average regular one.
Scholarship3	-0.75839	+8	Scholarship in the 3rd semester has greatly lower importance for intervened students, and the average intervened student has greatly lower risk of dropout due to the feature's effects than the average regular one.
FTE2	0.31874	-4	FTE in the 2nd semester has higher importance for intervened students, and the average intervened student has slightly higher risk of dropout due to the feature's effects than the average regular one.
FTE3	-0.27005	-12	FTE in the 3rd semester has greatly higher importance for intervened students, and the average intervened student has slightly lower risk of dropout due to the feature's effects than the average regular one.
FTE4	0.58707	+1	FTE in the 4th semester has slightly lower importance for intervened students, and the average intervened student has greatly higher risk of dropout due to the feature's effects than the average regular one.
Conditioned4	0.26968	-3	Being conditioned in the 4th semester has slightly higher importance for intervened students, and the average intervened student has slightly higher risk of dropout due to the feature's effects than the average regular one.
Sem_Interruption2	0.23618	-2	A semester interruption in the 2nd semester has slightly higher importance for intervened students, and the average intervened student has slightly higher risk of dropout due to the feature's effects than the average regular one.
*Intervention**	0.27679	-8	Going through the Academic Support program has greatly higher importance for intervened students, and the average intervened student has slightly higher risk of dropout due to the feature's effects than the average regular one.
Cumulative_GPA2	0.31747	+2	GPA on the 2nd semester has slightly lower importance for intervened students, and the average intervened student has slightly higher risk of dropout due to the feature's effects than the average regular one.
Cumulative_GPA4	0.37555	-1	GPA on the 4th semester has slightly higher importance for intervened students, and the average intervened student has slightly higher risk of dropout due to the feature's effects than the average regular one.
Dropped_Courses3	0.49483	-5	Number of dropped courses on the 3rd semester has higher importance for intervened students, and the average intervened student has higher risk of dropout due to the feature's effects than the average regular one.
Dropped_Courses4	0.32241	-1	Number of dropped courses on the 4th semester has slightly higher importance for intervened students, and the average intervened student has slightly higher risk of dropout due to the feature's effects than the average regular one.
Failed_Courses3	0.35890	-3	Number of failed courses on the 3rd semester has slightly higher importance for intervened students, and the average intervened student has slightly higher risk of dropout due to the feature's effects than the average regular one.

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 mentioned above, analysis should be accompanied by the feature importance rankings as well. A visual inspection

of the Shapley swarm plots confirmed the distribution changes between population for the mentioned features. A close-up comparison of some of these features can be seen in figure 3 below.

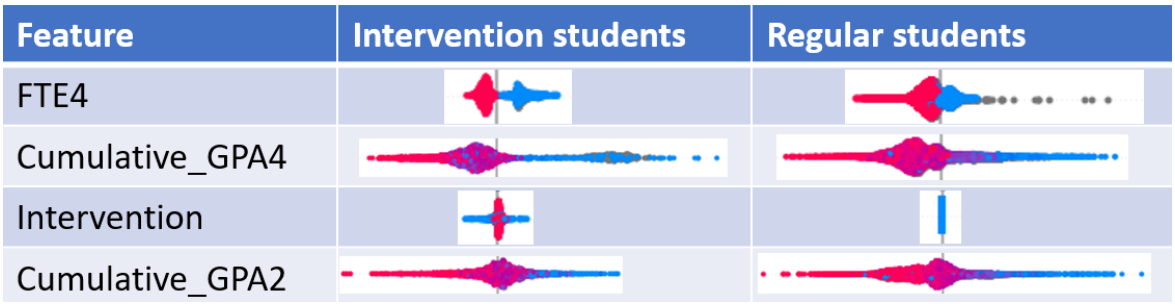

Figure 3. Side by side comparison of the Intervened and regular student populations.

In the case of figure 3, we can observe how the overall distribution of students changes in each feature. For FTE4, it is possible to see that there is a more distinct separation between the red and blue clusters between the groups, with the intervened students having a much clearer separation between the highest densities of positive and negative cases, showing that there is a clearer effect on that group regarding a full academic load against a reduced one. We can observe clear changes for all four sets shown above, from clusters moving or appearing (density), to movement of the population regarding the 0 Shapley value vertical line in all images (overall effect), and even changes regarding the effect of a feature's value (movement of colored clusters). 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 4th semester aid in student retention and low scores push towards student 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.

4. DISCUSSION

To validate our results against other models, we refer to [22]. The researchers in this project used the same dataset as us [23] but used Artificial Neural Networks instead of more traditional machine learning methods. In their research, they present a set of categories that grouped the dataset variables regarding their predictive contributions by generation and found that "University Background" (what they called the academic features such as averages, dropped and missed classes, etc.), Student's characteristics, and Financial aid were the three most informative features of the dataset, which matches the results of our models, although ours reported lower model scores.

There are several insights that we can extract from both the figures and the results of our proposed methodology regarding student dropout between intervened and regular students. 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 state according to the HEI indicators. When we move towards the intervened population, we can observe that the intervention courses themselves have little to no overall effect, ranking 29th in importance instead of the original 5th place. Looking into the regular population, we can observe that the "Intervention" variable has no effect on them, which is consistent with the reality that they did not receive an intervention in any form.

The effect captured in the "Intervention" feature in the full population should still be present in some shape or form in the Intervened population and removed completely from the regular one. We believe that a large amount of these changes are "absorbed" by the other features present in the dataset, and 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. We believe the use of this methodology will be beneficial to mentors and tutors, as it will allow them to identify features that become more relevant for specific populations, and in which ways.

As a final step towards the validation of our work, we consulted with a series of mentors inside the institution. In this initial approach, we presented figures and plots much like the ones used here in a meeting with mentors, tutors, and other accompanying figures for academic guidance. The presentation, and especially the different plots (waterfall and beeswarm) were well received, with several mentors informing us that the model does match their individual experiences. Of importance is the fact that the seemingly counter-intuitive result of the "Intervention" feature was mentioned to be accurate, as they mentioned that "several of the student's that end up in academic probation do tend to leave even after

going through the intervention program". After this initial approach, we are already in talks regarding several case studies for methodological validation, as well as possible application and development of a dashboard using Shapley values with the early alerts team of our institution. A survey was also developed to gather more objective data regarding the agreement of our model results and mentor experiences, in which we expect around 40 to 50 respondents.

Regarding missing features and transformations that the mentors believe could be of use to these prediction models, the unanimous response was psychological and emotional well-being information. All of 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.

5. CONCLUSIONS AND FUTURE WORK

Going back to the research question provided at the beginning of this article, we found that: a) the features obtained through our machine learning models coincide with both the opinion of interviewed mentors and tutors, and with independently performed research with the same dataset origin [24]; b) the explanatory values obtained through the use of the SHAP library do not differ in a relevant amount between each other, that this difference could be attributed to model variability, and that the explanations applied to student dropout match the real-world experiences of mentors and tutors, especially when dealing with highly explanatory features like previous average grades and interventions; c) the additional features that would be most beneficial to these models are psychological and emotional well-being information from students, but it could prove to be difficult to obtain these.

Aside from the answers to our research questions, we developed a methodology based on the use of already stablished mathematical concepts (Cohen's d and feature importance) that can be used towards feature comparison between any two distinct populations from an explainability standpoint. While initially developed for the explanation of educational features, the methodology can be directly applied to any similarly structured data science problem, greatly increasing its potential benefit.

We believe that these results could be used towards building an AI system for mentors/tutors based on the use of Shapley values that could allow for the identification and design of achievable, individualized counterfactuals/interventions, both for student retention and overall well-being. Future work will include a much more ample validation of the features and explanatory values mentioned above by using a series of instruments to measure the level of agreement for different features and explanatory distributions, validation of the presented methodology in contexts different than education and with non tree-based models, and a qualitative analysis regarding tutor and mentor interpretation of the methodology outcomes, along with iterative improvement on the proposed steps and application of the same.

6. Acknowledgements

The authors would like to acknowledge the Living Lab & Data Hub of the Institute for the Future of Education, Tecnologico 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. We would also like to acknowledge the financial support from the "Fondo de Apoyo a Publicaciones" of Tecnologico de Monterrey.

REFERENCES

[1] 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. https://doi.org/10.1016/j.caeai.2023.100150

[2] Cave, S., Coughlan, K., & Dihal, K. (2019). 'Scary Robots': Examining public responses to AI. In Proceedings of the 2019 AAAI/ACM conference on AI, Ethics, and society – (AIES '19). https://doi.org/10.1145/3306618.3314232

[3] Stanton, B., & Jensen, T. (2021). Trust and artificial intelligence, NIST interagency/internal report (NISTIR). Gaithersburg, MD: National Institute of Standards and Technology. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087.

[4] McMurtrie, B. & Supiano, B. (2022). The barriers to better advising. The future of advising: strategies to support student success (7-12). The Chronicle of Higher Education.

[5] Arin, J. (2022). The Missing Link in Academic Advising: The Faculty Perspective. The future of advising: strategies to support student success (40-43). The Chronicle of Higher Education.

[6] McMurtrie, B. & Supiano, B. (2022). Concerns About Bias in Advising Technology. The future of advising: strategies to support student success (33-95). The Chronicle of Higher Education.

[7] Calhoun-Brown, A. (2022). How Data and Technology Can Improve Advising and Equity. The future of advising: strategies to support student success (36-39). The Chronicle of Higher Education.

[8] 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

[9] Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. Review of Educational research, 86(2), 602-640.

[10] Walton, G. M. (2014). The new science of wise psychological interventions. Current Directions in Psychological Science, 23(1), 73-82.

[11] 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/ASSET/00346543045001089.FP.PNG_V03

[12] 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).

[13] 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

[14] 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

[15] 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

[16] 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

[17] Herodotou, C., Maguire, C., Hlosta, M., & Mulholland, P. (2023). Predictive Learning Analytics and University Teachers: Usage and perceptions three years post implementation. LAK23: 13th International Learning Analytics and Knowledge Conference, 68–78. https://doi.org/10.1145/3576050.3576061

[18] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30. https://arxiv.org/abs/1705.07874

[19] Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. Sept. 2021. Four Principles of Explainable Artificial Intelligence. National Institute of Standards and Technology, (Sept. 2021). doi: 10.6028/NIST.IR.8312

[20] Shapley, L. S. (1953) A value for n-person games. Contributions to the Theory of Games 2.28, pp.307–317.

[21] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888.

[22] Rodríguez-Hernández, C. F., Musso, M., & Cascallar, E. (2023, March 9-11). An Artificial Neural Network Approach to Analyze Students' Dropout in Higher Education [Poster presentation]. International Convention of Psychological Science (ICPS) 2023, Brussels, Belgium.

[23] 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.