An Explainable Machine Learning Approach for Student Dropout Prediction

João Gabriel Corrêa Krüger[a] (kruger.joao@ppgia.pucpr.br), Alceu de Souza Britto Jr.[a] (alceu@ppgia.pucpr.br), Jean Paul Barddal[a] (jean.barddal@ppgia.pucpr.br)

[a] Graduate Program in Informatics (PPGIa), Pontifícia Universidade Católica do Paraná (PUCPR), Rua Imaculada Conceição, 1155, Curitiba, Paraná, Brazil

Corresponding Author:
Jean Paul Barddal
Graduate Program in Informatics (PPGIa)
Pontifícia Universidade Católica do Paraná (PUCPR)
R. Imaculada Conceição, 1155, Curitiba, Paraná, Brazil
Tel: (+55) 41 99610-1446
Email: jean.barddal@ppgia.pucpr.br

An Explainable Machine Learning Approach for Student Dropout Prediction

João Gabriel Corrêa Krüger Author[a], Alceu de Souza Britto Jr.[a], Jean Paul Barddal[a,*]

[a] *Graduate Program in Informatics (PPGIa), Pontifícia Universidade Católica do Paraná (PUCPR), Rua Imaculada Conceição, 1155, Curitiba, Paraná, Brazil*

Abstract

School dropout is a relevant socio-economic problem across the globe. Predictive models have been developed to determine the likelihood of students dropping out of their studies precociously in an attempt to overcome such a problem. Academic systems, which gather data from many students, are potential sources for datasets that feed dropout prediction algorithms, thus leading to general improvements in education quality. Despite successful past attempts to predict dropout, several works depict small datasets with features that are hard to reproduce. Furthermore, predicting whether a student will drop out is not enough to diagnose and prevent the problem as it is also necessary to provide potential justifications for the dropout. This paper proposes an approach for creating and enriching a dataset for dropout prediction, which has been applied for dropout prediction using data from 19 schools in Brazil. With this dataset and using classifiers and model explaining techniques, our experiments achieved Area Under the Precision-Recall Curve (AUC-PR) scores of up to 89.5% when predicting dropout at different year moments. This study also shows differences when predicting dropouts in different educational stages, such as preschool and secondary education, with the former being more complex than the latter. In addition to the high recognition rates, our proposal identifies potential reasons

[*] Corresponding author.

Email addresses: `kruger.joao@ppgia.pucpr.br` (João Gabriel Corrêa Krüger Author), `alceu@ppgia.pucpr.br` (Alceu de Souza Britto Jr.), `jean.barddal@ppgia.pucpr.br` (Jean Paul Barddal)

for student dropout, which are relevant for educational institutions to take preemptive actions.

Keywords: Dropout prediction, Machine Learning, Explainable model

1. Introduction

Student dropout is a severe problem of utmost importance in education and society. Schooling is known to be highly correlated with the student's future success due to its impacts on future wages, income distribution, possible employment options, and quality of life (Adelman & Szekely, 2016; Snyder et al., 2019; Rumberger, 2020). Making sure students finish their studies can lead to a better society overall.

Students may quit studies for many different reasons: financial or economic issues (Adelman & Szekely, 2016; Snyder et al., 2019; Yao et al., 2017), struggle with various classes and subjects (Yao et al., 2017), or lack of interest (Adelman & Szekely, 2016). The subjectivity and difficulty of measuring such issues renders predicting evasion as a challenging task. Therefore, this scenario is a sensitive issue, especially considering the school's large number of students.

Attempts have been made to predict dropout in the past, with most of them conducting extensive surveys of a small number of students (Márquez et al., 2016; Lykourentzou et al., 2009), an approach that generates a small yet rich dataset that is hard to reproduce. Another approach, not as common as the former, is using educational systems and data collected automatically to generate larger datasets (Sales et al., 2016). A common characteristic in such studies and approaches is the highly imbalanced profile of the problem since most students finish their studies.

Many educational stages can benefit from dropout prediction, such as secondary (Márquez et al., 2016) and higher education (Sales et al., 2016). There has also been a history of e-courses (Lykourentzou et al., 2009) using these techniques. However, the successful approaches in one will not necessarily be successful in another given their differences in student profile and legal requirements.

This paper describes an end-to-end approach to predict student dropout in 19 Brazilian schools. First, we create a dataset merging data available in the school's educational system with derived temporal features and external socio-economic information related to the school's local region. Next, monolithic and ensemble-based classifiers are applied to each educational stage and trimester of the year of the schools. Finally, model explanation and interpretability techniques generate insights into why a student drops out. We emphasize that the proposed pipeline does not exhibit major technical novelties, yet, its application is relevant for society as it allows education institutions to preemptively act on dropouts. In this paper, we are interested in determining (i) whether temporal and publicly available features, i.e., socio-economic data, contribute to predicting whether students will drop out or not, and (ii) which features play essential roles in predicting dropout in different school stages.

This paper is organized as follows. Section 2 describes related works on student dropout prediction and model interpretability. Section 3 describes the proposed method, including details on data extraction, feature engineering, missing data preprocessing, classification algorithms, and model explaining techniques used. Next, Section 4 depicts and discusses the results obtained. Finally, Section 5 concludes this work.

2. Related Work

Due to the issues that come from student dropout, different attempts have been made to diagnose possible causes for students to quit their studies (Lykourentzou et al., 2009; Márquez et al., 2016; Sales et al., 2016). However, few studies have explored model explaining techniques for dropout prediction. Model explainability is relevant since dropout prediction may be useless unless academic staff understands why better. Consequently, it allows the academic staff to take assertive and preemptive action to avoid such a dropout.

Similar areas, such as churn prediction, have succeeded and made significant advancements in the literature applying these techniques to student dropout (DUMITRACHE et al., 2020; Villarreal et al., 2020). This section presents an overview of existing approaches for student dropout prediction and works related to applying model interpretability techniques in different customer evasion prediction scenarios.

2.1. Dropout prediction

There have been many attempts to model student factors into features for predicting dropout. In (Lykourentzou et al., 2009), authors used Moodle's data to predict whether a student will drop out of computer science e-courses. The authors used different classification algorithms coupled with basic information, activity on the platform, submission dates, and students' grades to achieve their dropout predictions. The recall and precision rates achieved 95% and 82%, respectively. A recall of 82%, for instance, means that 82% of the potential dropouts were detected, while a precision of 92% means that 92% of the students predicted as dropouts did, in fact, quit their studies. A significant a limitation of the study is that it encompassed only 193 students.

In a study made in Mexico (Márquez et al., 2016), authors conducted surveys on secondary education students with the goal of student dropout prediction. The study evaluated the prediction obtained using data collected in different stages of the school year and achieved recall scores of up to 98.8%. However, it is noteworthy that the study contemplated 419 students, also considered low for secondary education. Furthermore, the data acquired was subjective or hard to replicate, such as the number of hours spent in many activities, the level of difficulty of tasks assigned to the student, and the number of friends.

The approach taken by (Sales et al., 2016) handles the problem differently, using the educational system adopted by the subject higher education institution. The study used academic record data from 32,342 students, with a period of twelve and a half years. Results depicted that the number of semesters students spend in the university makes dropout less predictable. Furthermore, results were volatile, as null precision rates were observed in specific scenarios, while recall scores up to 82.4% were achieved.

2.2. Model interpretability and explaining in similar problems

The ability to correctly interpret a prediction model's output is relevant and, in some applications, as critical as the output itself (Lundberg & Lee, 2017; Ribeiro et al., 2016). In cases where explainability is relevant, simpler models may be preferred over complex ones, even when coupled with worse results (Lundberg & Lee, 2017). However, the results provided by simpler models may not be enough for more intricate problems. The use of model explaining and interpretability techniques, such as SHAP (Lundberg & Lee, 2017), and LIME (Ribeiro et al., 2016), is not widely used in dropout prediction. However, similar areas like churn prediction have successfully applied these techniques to provide explainable predictions. It is also relevant to highlight that this kind of problem is also tackled in a function-behavior-state approach, where the goal is to predict the state (dropout or not) of a student according to its context and behavior (Zhang et al., 2018).

The principle leads to a new view of the dimension that accounts for why and how the structure along with the state results in the behavior. The context leads to another view of the dimension that accounts for why and how the structure along with the state and behavior plays the function or exhibits the information or makes up the meaning.

For instance, in (DUMITRACHE et al., 2020), authors developed a churn prediction model for the telecommunication industry using demographic data, payment information, acquisition power, discount history, and support interactions to predict churn. Their goal was to identify the possible factors that led to a customer ceasing their services. With the aid of SHAP (Lundberg & Lee, 2017), they derived both visualization and scores for the specific customers, which they used to analyze and generate possible churning profiles.

In the banking sector, a bank chain had a series of hypotheses on why customers closed their accounts they wanted to validate (Villarreal et al., 2020). The authors used agency location and resources information, support speed and effectivity, demographic information, and general account movement information to conclude that some of the services provided were not up to the expectations of their clients.

2.3. Discussion

Despite some cases in the literature having great success when predicting whether a student will quit their studies, not all of them cover a reasonable amount of students, primarily because of the data acquisition process. On the other hand, specific studies tried to account for a large number of students, yet, the results were not compelling. Consequently, one of our hypotheses is that enriching the data available in academic systems with external data and newly derived features would improve the results observed in the literature. Furthermore, model-explaining techniques are not widely applied in dropout prediction. Similarly to churn prediction, we believe that explaining why a student is likely to drop out allows a better understanding of the problem and allows the school to take preventive actions rather than reactive ones.

3. Proposed Pipeline

This section describes the pipeline built and followed for developing an explainable student dropout prediction system. This pipeline encompasses the six steps given in Figure 1, and each step is detailed in a specific subsection. First, Section 3.1 describes the data extraction section process, i.e., how the data were extracted from the academic system. Next, Section 3.2 explains the process used to enrich the student's data with their region's socioeconomic data. Next, Section 3.3 describes a feature engineering process in which data from different trimesters of the year are adequately represented. Also, regarding data preprocessing, Section 3.4 describes the process adopted for handling missing data. Given that the data have been properly extracted, enriched, and treated, Section 3.5 describes the process of using it to train classifiers to predict dropout. Finally, Section 3.6 explains the techniques used to explain model outcomes.

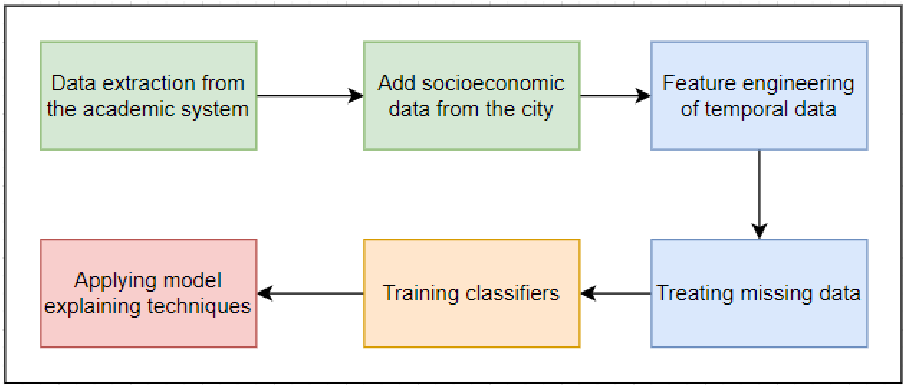
Data extraction from the academic system
Add socioeconomic data from the city
Feature engineering of temporal data
Treating missing data
Training classifiers
Applying model explaining techniques

Figure 1: Different steps taken to the student dropout prediction.

3.1. Data Extraction from the Academic System

Our study was conducted in partnership with a group of Brazilian private schools that chose to remain anonymous. After anonymization, the group allowed access to their student database so that no student could be identified. First, the data present in the academic system was extracted using Structured Query Language (SQL) and comprised the students' grades, tuition fees, age, parents' occupation, and house location.

3.2. Adding Socioeconomic Data

While the student's grades and financial status can be relevant indicators of whether they will quit their studies, external variables such as an economic crisis, the average expected years in school, or a region's financial status can also correlate with dropout. Due to this possibility, there is a motivation to enrich the dataset with these statistics. Since the original dataset extracted from the academic system comprises the students' home location, it was possible to merge it with public data made available by the Brazilian National Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística, IBGE) (IBGE, 2021) on cities' statistics GDP, HDI, life expectancy, and study expectancy.

Table 1: Example of socioeconomic features added to the dataset. Features with an asterisk represent characteristics added in this step of the pipeline.

Student identifier	**ZIP Code**	**GDP per capita***	**HDI***
123456789	12345000	R$ 45318.46	0.823
456456789	45678000	R$ 47683.47	0.751
987123465	78945000	R$ 48615.15	0.763

As an example, we have in Table 1 a fictitious student identified by 123456789, which lives at a neighborhood enumerated with ZIP Code equal to 12345000. Using the ZIP code, the GDP per capita and HDI of such a neighborhood can be extracted from IBGE's dataset (IBGE, 2021) and then used to enrich the dataset. The newly added features are indicated in the table with an asterisk.

3.3. Temporal Data Feature Engineering

Since the school year is divided into three trimesters, a single student in a school year has the potential to be represented once per trimester with the respective grading. However, the student's progress during the year is determined by all the grades up until that point. Consequently, it is intuitive to create new features representing the cumulative sum of the trimesters' grades until that specific moment. In other words, the grading features related to the second trimester should represent the first and second trimester grades, while the features for the third trimester should aggregate the first, second, and third trimesters grades. More formally, this new feature is given by Equation 1, where N_t is the respective trimester student grade for an arbitrary discipline. The value of $N_{cumulative}$ reflects the student's performance in a discipline and school year.

$$N_{cumulative} = \sum_{1}^{t} N_i \tag{1}$$

Despite the overall performance being relevant, a single trimester may significantly impact the prediction of whether a student will stay in school or drop

Table 2: Example containing engineered features for an arbitrary student.

Student Identifier	**School Year**	**Trimester**	**N**	$\mathbf{N_{cumulative}}$	$\mathbf{\Delta N}$
123456789	2015	1	7.9	7.9	0
123456789	2015	2	6.5	14.4	-1.4
123456789	2015	3	3	17.4	-3.5
123456789	2016	1	8.1	8.1	0

out. Therefore, we hypothesize that a sudden grade gain or loss between different school years or trimesters may indicate whether the student's performance worsens with time, causing this individual to drop out. With this rationale in mind, a new feature ΔN was engineered according to Equation 2, where N_t is the respective trimester to each entry in the dataset, representing the gain or loss from the last trimester of that year. In a similar manner, the same principle was applied to the yearly variation of each variable.

$$\Delta N = N_t - N_{t-1} \tag{2}$$

In other words, ΔN measures the difference between the current and previous trimester, thus indicating the student's sudden loss or performance gain in a school year.

Table 2 illustrates how these features are engineered in the dataset. In particular, an arbitrary student with identifier 123456789 has his grades N for different trimesters in 2015 and 2016. Using Equation 1, $N_{\text{cumulative}}$ is computed by summing the N values obtained until that moment, i.e., the grades obtained in previous trimesters. Likewise, the ΔN (cf. Equation 2) represents the difference between the current and previous trimesters in a school year. It is also relevant to highlight that, in the example, ΔN for the first trimester of 2016 is zero since the previous data regards a different school year and it should be unaccounted for.

3.4. Handling Missing Data

Considering that the dataset was extracted from the school's transactional system, which has a table structure that is constantly changing, it is relevant to check for and treat missing data. Table 3 shows the features with the highest missing rate in the dataset. Here, we notice that features related to the student's guardian are the most incomplete. This behavior occurs since certain schools did not register the respective information during enrollment.

Missing data imputation was conducted by replacing missing values with the attribute-wise median and mode for numeric and categorical data, respectively. Furthermore, a new boolean feature was added to indicate whether a specific row has been imputed or not. Table 4 exemplifies the imputation process. Assuming an arbitrary student identified by 123456789, categorical features (e.g., extracurricular classes), have been imputed using the mode, whereas the median was used for numerical features (e.g., tuition fee). It is also relevant to highlight the *Participation imputed* and *fee imputed* features, which are flags to depict whether imputation has been made in the respective features.

Table 3: Top missing features.

Ranking	**Feature**	**Missing percentage**	**Feature Type**
1	Male guardian connection	49.28%	Categorical
2	Female guardian connection	50.16%	Categorical
3	Total tuition discount	56.78%	Numerical
4	Male guardian's education level	60.22%	Categorical
5	Female guardian's education level	62.69%	Categorical

Table 4: Example containing imputed missing features.

Student Identifier	**Extracurricular classes**	**Participation imputed**	**Tuition fee**	**Fee imputed**
111111111	Yes	No	50000	No
123456789	*No*	Yes	*52000*	Yes
555555555	No	No	54000	No
999999999	No	No	52000	No

3.5. Classifier Training

Different classifiers were evaluated to determine whether a student will drop out or finish the school year. The chosen classifiers for this experiment were included Decision Tree, Logistic Regression, Random Forest, AdaBoost, and XGBoost.

The Decision Tree is a structure that represents a set of sequential decisions created using criteria, such as entropy, based on the different examples on the training dataset (Quinlan, 2014). Whenever a prediction is required, an instance and its attribute values are used to traverse the tree until a decision node is reached. The model can use a larger tree to better fit the training data, but it can also be pruned, making it less likely to overfit. Logistic Regression algorithm learns a hyperplane using the logit function to divide the different instances of each class (Anzai, 2012). Due to its propensity to overfitting, implementations often use regularization methods to prevent this problem. Random Forest (Breiman, 2001) uses the average prediction of a bagging ensemble of randomized Decision Trees (Quinlan, 2014) trained on sub-samples of the original training dataset. Random Forest is known to be a strong algorithm as it reduces both bias and variance, thus it often exhibits positive results against both under and overfitting. AdaBoost uses a large number of weak learners over sub-samples of the original training set, assigning higher weights to incorrectly classified instances so that future learners can learn the most difficult cases (Freund & Schapire, 1997). Similarly to Random Forest, AdaBoost is a strong ensemble-based algorithm to overcome underfitting, yet, it is known to overfit if the dataset has noisy instances. Finally, XGBoost (Chen & Guestrin, 2016) is an implementation of the gradient tree boosting algorithm, which uses a growing set of decision tree variants to generate predictions while trying to minimize a loss function using gradient boosting during the training process. The algorithms counts with different base learners, and the adopted tree-based boosting algorithm counts with both tree based methods and regularizers to avoid overfitting while not underfitting to the data.

All the implementations used, except for XGBoost (Chen & Guestrin, 2016), were present in scikit-learn (Pedregosa et al., 2011). The hyperparameters were tuned using an exhaustive search through many possible values coupled with manual fine-tuning and are presented in Table 5.

With the proper parameters in hand, the algorithm with the best results is further used to generate explanations for the classifier results.

The training protocol adopted was cross-validation (k=5), where the classifiers were trained for the different educational stages and trimesters in the year. This protocol was chosen due to the different behaviors and general characteristics of each educational stage and moment in the year, leading to different reasons behind a dropout.

The metrics chosen for evaluation were Area under Precision-Recall Curve (AUC-PR), Kolmogorov-Smirnov statistic (KS score), precision, and recall. The reasoning behind the selection of these metrics is that (i) AUC-PR and KS are

Table 5: Evaluated algorithms and parameters tuned for better results.

Algorithm	**Tuned parameters**	**Best found value**
Decision Tree	criterion, max_depth, max_leaf_nodes	criterion: gini, max_depth: 25, max_leaf_nodes: 64
Logistic Regression	solver, penalty, C	solver: saga, penalty: l2, C: 0.01
Random Forest	n_estimators, criterion, max_depth	n_estimators: 700, criterion: gini, max_depth: 5
AdaBoost	n_estimators, learning_rate	n_estimators: 500, learning_rate: 0.05
XGBoost	colsample_bytree, gamma, max_depth, max_leaves, n_estimators, reg_alpha	colsample_bytree: 0.6, gamma: 1, max_depth: 6, max_leaves: 15, n_estimators: 600, reg_alpha: 2.4

robust to class imbalance, and (ii) the ease to explain the results and performance of the classifiers to non-technical and educational teams in the case of precision and recall.

The precision of a predictive model is given by Equation 3, which quantifies the ratio of actual positive cases that have been predicted as such (actual dropouts that were predicted as dropouts, or true positives) and positive predictions (predicted dropouts, regardless of being actual dropouts or not), i.e., true positives and false positives. On the other hand, recall is given by Equation 4 and it quantifies the ratio of dropouts identified correctly (true positives) and the amount of actual dropouts, regardless of them being predicted as such or not (true positives and false negatives).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

Another relevant approach to evaluate predictive models focuses on plotting the precision of a model against its recall over different decision thresholds. The area under the generated curve is called *Area under Precision-Recall Curve*, or AUC-PR. AUC-PR is a metric that is better suited for imbalanced datasets (see Section 4.1) due to its characteristic of evaluating the trade-off between precision and recall over different thresholds.

Finally, Kolmogorov-Smirnov (KS) is a separability statistic that quantifies how well a predictive model separates two classes. KS indicates the maximum distance between the cumulative probability distribution functions (*cdfs*) obtained by students that dropped out and those who did not. Assuming n as the number of students who dropped out ($y_i = 1$) and m the number of students who did not ($y_i = 0$), the empirical cumulative distribution function of the respective subsets of students are given by Equations 5 and 6, respectively, where $L = \min s_i, 1 \leq i \leq n + m$, $H = \max s_i, 1 \leq i \leq n + m$ and $a \in [L, H]$.

$$F_{\text{dropout}} = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 1, & \text{if } s_i \leq a \text{ and } y_i = 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$F_{\text{no dropout}} = \frac{1}{m} \sum_{i=1}^{m} \begin{cases} 1, & \text{if } s_i \leq a \text{ and } y_i = 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

KS is given by $KS = \max_{a \in [L,H]} |F_{\text{dropout}} - F_{\text{no dropout}}|$, which is the maximum difference between the *cdfs* that describe the two classes. The larger the KS, the higher the class separability achived by a predictive model.

3.6. Model Explanation

While it is important to achieve compelling classification results, in the case of dropout prediction, it is also relevant to identify the reasons behind a dropout and potentially improve the school's retention rate in the future. The best model obtained for each school year had SHAP (Lundberg & Lee, 2017) applied to it to enlighten the reason behind the dropouts predictions. While other techniques such as LIME (Ribeiro et al., 2016) were evaluated, we observed that the visualization provided by SHAP better illustrates the results to educational teams.

The SHAP algorithm (Lundberg & Lee, 2017) uses a game theory approach for measuring the impact of the different features on the model's output. The technique does it by varying all different subsets of features and the resulting value produced by the model. Applying the different results to the formula described by the technique (Lundberg & Lee, 2017) and adopted kernel, SHAP can predict the contributions of the different features in the model.

The kernel adopted when applying SHAP was either `TreeExplainer` or `KernelExplainer`, the first being adequate for tree-based models such XGBoost and Random Forest, while the latter being suitable for the remainder of the models tested.

4. Results and discussion

As an initial result, a dataset was built from the data present in the school's transactional system, which was then enriched and used for posterior prediction of possible students who will drop out.

4.1. Dataset

The generated dataset comprises 299,722 rows and 137 columns of labeled data. The general layout of the dataset is given in Table 6, where an asterisk marks features that come from sources other than the school's academic system.

Each row in the dataset represents a student's grades and information during a trimester of a school year. The data was collected for the first three educational stages a student can pass through. Table 7 demonstrates the number of students in each educational stage in the dataset. Here, we observe that the segment with the highest number of students is the basic stage, followed by the preschool segment. The secondary school, despite having the smallest number of students, has a bigger dropout rate than its counterparts. A reasonable explanation for this behavior is that students drop out of their studies since they are old enough to work in the informal market.

The data was collected using the year 2015 to 2019. Table 8 shows the general number of students in these years.

Table 8 shows that the number of students remains nearly constant between the different years in the dataset, as does the dropout rate.

During a school year, a single student comprises up to three rows in the dataset, with each row corresponding to their grades over the trimesters in the school year. Table 9 demonstrates the number of rows of each school year in the dataset.

It is important to note that the dropout rate is lower for each year in Table 9 in comparison to Table 8 primarily because of students dropping out in the first

[1] These features were added from an external dataset, each using the student house location to assign its value.

Table 6: Dataset layout.

Category	**# of columns**	**Examples**
Educational stage information	9	School, grade, school year, class, educational stage, expected age at current grade
Student information	11	Age, sex, grade, absences, years in that school,
Extra classes	3	Involved in religious classes, involved in extracurricular classes, full-time study.
Financial situation and Fees	4	Tuition fee, an increase of tuition fee compared to last trimester and last year, discounts
Parent information	20	Parents professions, marriage status, raised by parents
Location statistics*	14	GDP[1], HDI[1], life expectancy[1], study expectancy[1]
Current grades	21	Arts, Mathematics, Physics, Portuguese, English, Spanish. It also includes all-time high, low and average grades.
Cumulative grades	18	Sum of all the grades in the year up to that point to each discipline
Grade difference against last trimester	18	Difference of the grades compared to the last trimester in that year (if available)
Grade difference against last year	18	Difference of the grades compared to the last year (if available)
Dropout	1	Student drops out in that school year
Total	137	

Table 7: Educational stage and unique students in the dataset.

Educational stage	Dropouts	Regulars	Total	Dropout rate
Preschool	778	16835	17613	04.42%
Basic school	1048	17027	18075	05.80%
Secondary school	1472	12459	13931	10.57%

Table 8: Years and unique students in the dataset.

School year	Dropouts	Regulars	Total	Dropout rate
2015	738	20143	20881	3.53%
2016	672	19636	20308	3.31%
2017	628	19248	19876	3.16%
2018	639	19129	19768	3.23%
2019	648	19482	20130	3.22%

Table 9: Years and row count in the dataset.

School year	Dropouts	Regulars	Total	Dropout rate
2015	1602	59724	61326	2.61%
2016	1180	58673	59853	1.97%
2017	1175	57825	59000	1.99%
2018	1218	57805	59023	2.06%
2019	1283	59237	60520	2.12%

part of the year and not having grades for the last and second to last trimesters.

4.2. Classification

After creating the dataset, a series of classifiers were trained for each combination for predicting dropouts in different educational stages and moments in the year. Because of that, the results are split into Tables 10 and 11, respectively.

As shown in Table 10, the results tend to get better the later in the school year the classification is made. However, the greatest increase in results comes from the first and second trimesters. This means that it is a reasonable moment to assess the dropout situation and avoid potential losses with greater success. The classifier that achieved the best results, independent of all the moments in the year, was XGBoost. This is similar when analyzing the results over the different educational stages, as depicted in Table 11. The results demonstrated in Table 11 show that the most challenging stage to predict dropout is preschool. We believe that this behavior relies on the fact that preschool is not mandatory in Brazil and is a time period in which students' assessment does not result in grades. Consequently, distinguishing between regular students and dropouts becomes more cumbersome. The basic and secondary stages both showed better results when predicting if a student will drop out, indicating that the lack of these features could be affecting the predictions in the preschool stage.

An example of how these results could be interpreted is as follows. For instance, a recall of 72.78% means that 72.78% of the students who did, in fact, evade were predicted as possible dropouts by the model; and a precision of 92.98% means that of all students predicted as dropouts 92.98% did, in fact,

Table 10: Results from predicting dropout in different moments in the year.

Best classifier	**Trimester**	**AUC-PR**	**Precision**	**Recall**	**KS score**
XGBoost	1st	0.3822	0.8325	0.4388	0.7031
XGBoost	2nd	0.7198	0.9423	0.7587	0.8550
XGBoost	3rd	0.8950	0.9523	0.9393	0.9703

quit their studies.

4.3. Explaining model results and visualization

After training the classifiers, it is possible to generate explanations for the student's dropout, one of this paper's objectives. One of the approaches to evaluating which factors contribute the most to a model is checking the importance of the model's features. Table 12 demonstrates the top features behind the classification according to the model created in the previous step of the framework.

For the preschool stage we verify that the annual cost, and its yearly variation are correlated with the permanence of a student during the school year. We hypothesize that this factor, coupled with the non-obligatory characteristic of this educational stage in the school's country of origin, may play a role in the student's permanence. Alongside the financial situation, variables such as the city's HDI and school life expectancy also are correlated with the permanence of a student in school. This leads to the hypothesis that students that live in more well developed cities have better conditions to stay in school. This is an interesting result since families' incomes are not available and these variable serve as proxies. Finally, the individual contributions can also be seen in the

Table 11: Results from predicting dropout for different educational stages.

Best classifier	**Educational stage**	**AUC-PR**	**Precision**	**Recall**	**KS score**
XGBoost	Preschool	0.4404	0.7458	0.5292	0.7003
XGBoost	Basic	0.5583	0.9191	0.5988	0.8417
XGBoost	Secondary	0.6878	0.9298	0.7278	0.8668

Table 12: Top features in each educational stage.

Ranking	**Preschool**	**Basic Education**	**Secondary Education**
1	Tuition fee	$N_{\text{cumulative}}$ Physical Education	$N_{\text{cumulative}}$ Portuguese
2	City's HDI	$N_{\text{cumulative}}$ Arts	$N_{\text{cumulative}}$ Geography
3	City's school life expectancy	ΔN Portuguese	Average grade
4	Absences	$N_{\text{cumulative}}$ Portuguese	Portuguese
5	ΔN_{yearly} Tuition fee	Average in Portuguese	Lowest grade

Figure 2.

As shown by Figure 3 and Table 12, the Basic Education scenarios differs from the Preschool stage. In Basic Education the model indicates that the student's grade in Human Sciences disciplines are of great importance, with features such as ΔN, $N_{cumulative}$ and average grade in Portuguese appearing as important features. Besides Portuguese, the overall performance in disciplines generally perceived as more ludic, such as Physical Education and Arts, are shown to be the most impactful in this stage. We hypothesize that this correlation comes from the idea if the student is not performing satisfactorily in these disciplines, which are generally perceived as important, could show signs of difficulty in class, problems with parents or even bullying (e.g., during sports activities). However, it's important to note that this analysis should be made in a case-to-case basis by the responsible people in school to better approach each individual problem.

Another approach to getting individual explanations behind each student's classification is by using model explaining techniques, such as SHAP. The SHAP values of a model can check the individual contributions to each classification,

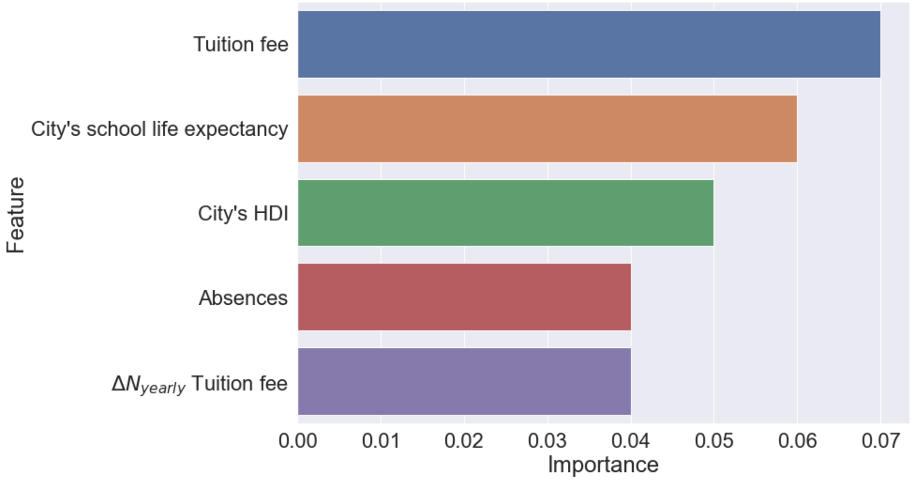

Figure 2: Feature importance for the Preschool stage.

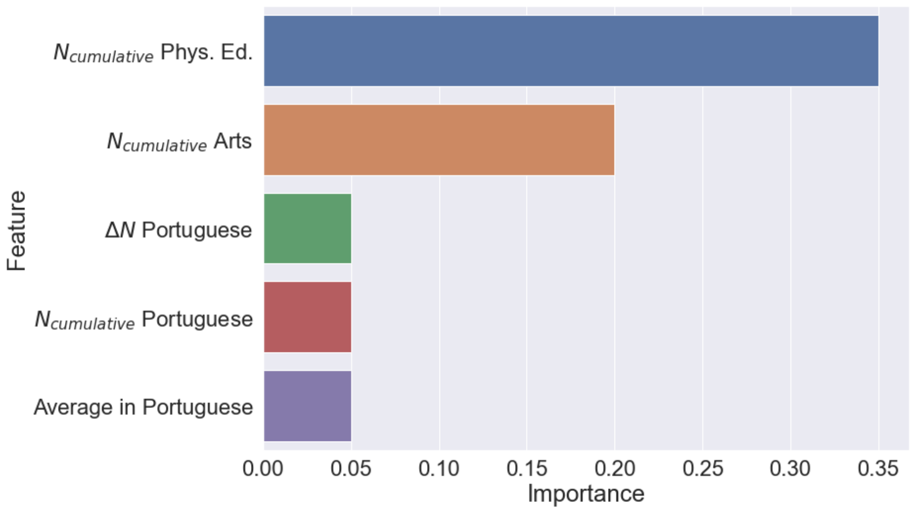

Figure 3: Feature importance for the Basic Education stage.

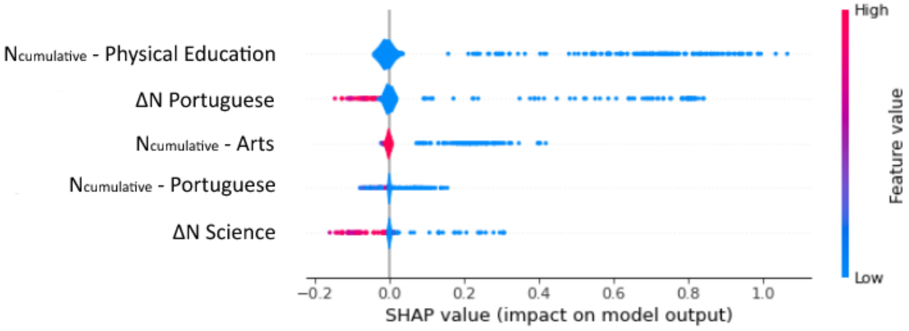

Figure 4: Explanations provided by SHAP for the Basic Education stage.

like check the overall feature importance in the model. Figure 4 illustrates the top contributors appointed by SHAP.

In Figure 4, each individual dot represents a student. A feature's value is illustrated by its color, and the higher its SHAP value, the most influence it has in the final classification. As shown by Figure 4, the SHAP values of features such as $N_{cumulative}$ in Physical Education, Arts and Portuguese, as well as the

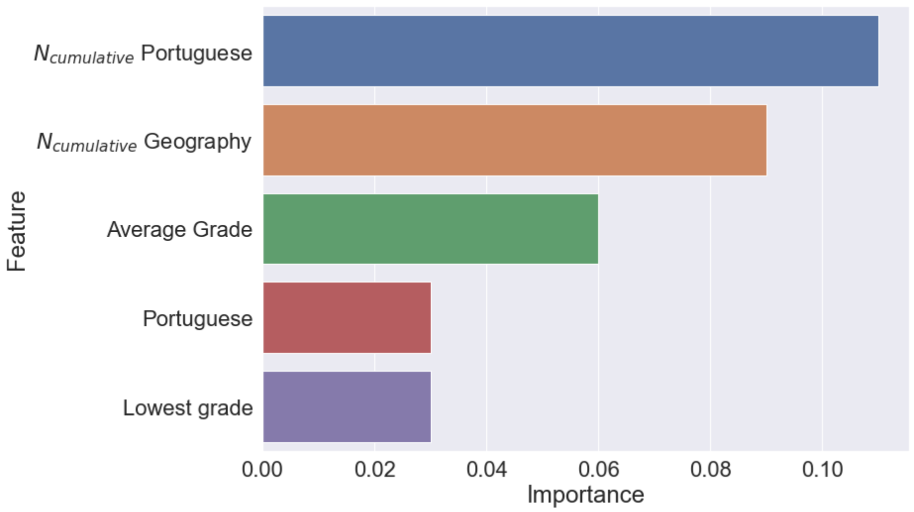

Figure 5: Feature importance for the Secondary Education stage.

ΔN in Portuguese and Science are shown as features that have great impact in the permanence of a student, confirming the feature importances appointed by Table 12

For the Secondary educational stage, a similar scenario to the Basic educational stage, as show by Table 12 and Figure 5. The Portuguese discipline, which appeared in the Basic educational stage, also is shown to impact the permanence of a student in Secondary education in its average and $N_{cumulative}$ forms. The average and lowest overall grades are also shown to having an impact in the dropout in this educational stage.

With the factors that contributed the most to the classification of each student as regular or dropout, a private dashboard was created to visualize the reasons and dropout probability. The implemented dashboard shows the students with a higher probability of dropping out of school and, when choosing a specific student, the reason behind the predicted outcome. This makes it easier for educational teams and teachers to reach out to students with potential problems.

It is possible to note that while it was possible to note which factors have impact in the permanence of a student in school, both in individual and general cases, grades were shown to be the most correlated factors with dropout. While grades are important, this result can also be explained by grades being the great majority of the collected features. When taking individual measures against school dropout, it's necessary to take this into consideration when evaluating each individual case. However, we hypothesize that general policies to improve the quality of education and student understanding of the disciplines can be effective to decrease dropout rates.

5. Conclusion

Student dropout is a serious problem, impacting the current and future socioeconomic status of a region. Predicting whether students will quit their studies using machine learning techniques is not a novel idea. However, due to differences between student assessment techniques and country regulations, the difficulty of obtaining rich and updated student data makes predicting dropout a dynamic task with many variables and space for exploration.

Predicting whether a student will quit his studies is in constant evolution, with two common approaches: a rich dataset with a low number of students and a dataset with a higher number of students but less detailed information. This paper uses the latter approach and feature engineering techniques to predict dropouts successfully. The algorithms achieved AUC-PR scores ranging from 38.22% to 89.50% when predicting dropouts in different moments (trimesters) of the school year.

This paper also demonstrates that not all education stages have the same behavior when predicting dropout, with students in the preschool stage being harder to identify as dropouts than their counterparts in the basic and secondary stages of education. The non-obligatory characteristic and different evaluation forms in Brazil's preschool education could explain these differences.

The use of model explaining and interpretability techniques is common in similar problems, like churn prediction, but has yet to be explored in dropout prediction. Using techniques to generate explanations and interpretability of black-box models provided a better way to approach students who would drop out by providing the features that contributed the most to each prediction.

The explanations for each prediction also allowed the creation of a dashboard that can be used to prevent the students from quitting. However, further improvements can be made, such as a more graphical way of showing which students are at a higher risk of evasion.

References

Adelman, M. A., & Szekely, M. (2016). School dropout in central america: An overview of trends, causes, consequences, and promising interventions. *World Bank Policy Research Working Paper 7561*, .

Anzai, Y. (2012). *Pattern recognition and machine learning*. Elsevier.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16 (pp. 785–794). New York, NY, USA: ACM. doi:10.1145/2939672.2939785.

DUMITRACHE, A., NASTU, A. A., & STANCU, S. (2020). Churn prediction in telecommunication industry: Model interpretability, .

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*, 119–139.

IBGE (2021). Indicadores IBGE. URL: https://www.ibge.gov.br/estatisticas/todos-os-produtos-estatisticas.html.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education*, *53*, 950–965. doi:`https://doi.org/10.1016/j.compedu.2009.05.010`.

Márquez, C., Cano, A., Romero, C., Mohammad, A., Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*, 107–124. doi:`10.1111/exsy.12135`.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). ”why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Rumberger, R. W. (2020). The economics of high school dropouts. *The economics of education*, (pp. 149–158).

Sales, A., Balby, L., & Cajueiro, A. (2016). Exploiting academic records for predicting student drop out: A case study in brazilian higher education. *Journal of Information and Data Management*, *7*, 166–166.

Snyder, T. D., De Brey, C., & Dillow, S. A. (2019). Digest of education statistics 2017, nces 2018-070. *National Center for Education Statistics*, .

Villarreal, D., Zhao, L., Hill, T., & Chung, L. (2020). Validating goal-oriented hypotheses of business problems using machine learning: An exploratory study of customer churn. In *Big Data–BigData 2020: 9th International Conference, Held as Part of the Services Conference Federation, SCF 2020, Honolulu, HI, USA, September 18-20, 2020, Proceedings* (p. 144). Springer Nature volume 12402.

Yao, Y., Yi, H., Zhang, L., Huan, W., Yang, C., Shi, Y., Chu, J., Loyalka, P. K., & Rozelle, S. (2017). Exploring dropout rates and causes of dropout in upper-secondary vocational schools. *Available at SSRN 2938383*, .

Zhang, W., Yang, G., Lin, Y., Ji, C., & Gupta, M. M. (2018). On definition of deep learning. In *2018 World Automation Congress (WAC)* (pp. 1–5). doi:10.23919/WAC.2018.8430387.