ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

A Machine Learning Approach for School Dropout Prediction in Brazil

João Gabriel Corrêa Krüger[1], Jean Paul Barddal[1], and Alceu de Souza Britto Junior[1]

Graduate Program in Informatics (PPGIa)
Pontifícia Universidade Católica do Paraná (PUCPR)
R. Imaculada Conceição, 1155, Curitiba, PR 80215-901 - Brazil

Abstract. School dropout is a problem that impacts many socio-economic aspects, including inequality. Dropout prediction algorithms can help remediate this problem, although several past attempts in the literature did so using small datasets. This paper brings forward an experimental approach of machine learning for school dropout prediction in Brazilian schools. The data used for this study was first retrieved from the academic systems of a group of Brazilian private schools, which was later enriched with socio-economic data extracted from governmental sources. Using the dataset to train different types of classifiers, we obtained up to 95.2% precision rates when predicting dropout at different year and educational stages, thus allowing schools to plan and apply retention strategies.

1 Introduction

Schooling is an essential part of a citizen's life due to the high correlation of quality education with future wages, possible employment options, and available life quality [1, 2]. This correlation implies that a student finishing their studies leads to a better overall society.

The impact of schooling encourages the research of diagnosing and potentially preventing student dropout. Studies show that many reasons could lead to a student dropping out: financial or economic issues [3, 1, 4], lack of interest [3], or struggle with different classes and subjects [4]. Although many factors can lead to a dropout, many are subjective and difficult to measure. Therefore, constantly evaluating which students are susceptible to dropping out is cumbersome, given the number of students enrolled in a school. Different attempts have been made to predict dropout using machine learning. A reasonable portion of such approaches uses extensive and complex surveys [5, 6] that generate datasets with a large number of features yet, are limited to a small number of students. While not as common as the former, approaches that rely on data extracted automatically from academic systems also exist [7], thus resulting in datasets that represent a larger number of students, yet, with a smaller number of features.

In this paper, we describe a machine learning approach for dropout prediction in 19 Brazilian schools. First, we present the dataset creation process that accounted for data from the schools' educational system and the addition of external socio-economic information from the school's region obtained from governmental sources. An additional feature engineering process was performed to derive temporal features, demonstrating the student's evolution. Lastly, classifiers are applied to each educational stage and quarter of the year of the schools.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

2 Related Works

Before advancing on the related student dropout works, it is relevant to further elaborate on the different segment educational stages as they pose various challenges. According to the International Standard Classification of Education (ISCED) [8] the educational levels can be categorized according to the following stages: Level 0 - Preschool, Level 1 - Basic Education (elementary school), Level 2 - Basic education (middle school), and Level 3 - Secondary education (high school). This is relevant as not all education stages are the same, and thus, the techniques applied to one scenario do not mean they will be equally successful when applied to a different stage. Regardless of the level, the socio-economic consequences of student dropout are high. Thus, studies have been made in the past to diagnose the causes behind students quitting their studies [6, 5, 7].

In [6], the authors used the Moodle Platform data to predict whether a student will quit their computer science e-course. Using different classifiers based on student's activity on the platform, submission dates, and grades, 95% recall and 82% precision rates were achieved for a pool of 193 students.

A Mexican study brought forward in [5] used survey data obtained from secondary education students to predict student dropout. With the data collection taking place in different stages of the school year, the model achieved recall scores of up to 98.8% when predicting whether a student would quit. This study contemplated 419 students, a small number for secondary education. Finally, another relevant work is brought forward in [7], where authors applied machine learning approaches in higher education, where recall scores up to 82.4% were achieved, yet, encompassed 32,348 students in 72 different degrees.

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, partly because of the datasets adopted. In contrast, other studies tried to account for a large number of students, yet, the results were not particularly meaningful. Consequently, one of our hypotheses is that enriching the data available in academic systems with external data and new derived features would improve the results in the literature.

3 Method

This section describes the protocol adopted and followed for the development of the proposed student dropout prediction process. First, regarding data extraction, our study was conducted with a group of Brazilian private schools that chose to remain anonymous yet allowed access to their student database after anonymization so that no student could be identified. The initial dataset was extracted and encompassed students' grades, tuition fees, age, parent occupation, and house location. Next, while grades and financial status are potential indicators of a dropout, external variables such as an economic crisis, the average expected years in school, or a region's financial status are also factors that can lead to student dropout. Therefore, we chose to enrich the dataset with

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

socioeconomic statistics. Assuming each student's home location, the dataset was enriched with governmental information retrieved from the National Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística, IBGE)[1] on cities' statistics GDP, HDI, life expectancy, and study expectancy. Furthermore, since the school year is divided into three trimesters, a single student in a school year has the potential to generate three entries to the dataset, with each corresponding to a quarter of the school year.

Despite the quarter's information being a potential factor in a student dropping out, the year's result also is. Therefore, we engineered temporal features to encapsulate different aspects of the school year. By calculating the cumulative sum of the student's grades up to each quarter of the school year, it is possible to check how well a student is performing in that school year. This new feature is given by $N_{cumulative} = \sum_1^t N_i$, where N_t is the value respective to the quarter to each entry in the dataset. While the general performance is relevant, a sudden change in a grade may significantly impact whether a student will stay in school or drop out. With this rationale in mind, a new feature ΔN was engineered according to $\Delta N = N_t - N_{t-1}$, where N_t is the value respective to the quarter to each entry in the dataset, representing the gain or loss from the last quarter of that year. In other words, ΔN measures the difference between the current and previous trimesters, thus indicating the student's sudden loss or gain of performance in a school year. It is important to highlight that the computation of these values was made upon each school year and not previous ones, as new students in a school would have detached values for such variables.

Finally, different classifiers were evaluated to determine whether students would quit or finish their studies. The chosen classifiers for this experiment were Decision Tree [9], Logistic Regression [10], Random Forest [11], AdaBoost [12] and XGBoost [13]. All the implementations used, except for XGBoost [13], were the ones present in scikit-learn [14]. The hyperparameters were tuned using an exhaustive search through many possible values coupled with manual fine-tuning. The training protocol adopted was cross-validation (k=5), where the classifiers were trained for the different educational stages and trimesters. This protocol was chosen due to the different behaviors throughout the year and the general characteristics of each educational stage. The evaluation metrics adopted were precision and recall, as these can be explained to non-technical individuals.

4 Results and discussion

Following the proposed methodology, a dataset was built using the school's educational system and enriched using governmental socio-economic data. The dataset was then used to predict the potential dropouts properly.

The generated dataset comprises 299,722 rows and 137 columns of labeled data from three educational stages spanning from 2015 to 2019. The general layout of the dataset is given in Table 1, where 1 marks features that come

[1]Statistics available at `https://www.ibge.gov.br/estatisticas/todos-os-produtos-estatisticas.html`

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

from sources other than the school's academic system and 2 marks features that also include the cumulative sum ($N_{cumulative}$), quarterly and yearly variations (ΔN) per discipline. Each row in the dataset represents a student's grades and information during a quarter of the school year, with Table 2 demonstrating the attendance numbers of each educational stage in the dataset.

Table 1: Dataset layout - categories, number of columns (#Cols), and examples

Category	**#Cols**	**Examples**
Educational stage information	9	School year, educational stage
Student information	11	Absences, years in school
Extra classes	3	Attends extracurricular classes
Financial situation and fees	4	Tuition fee, discounts
Parent information	20	Parents professions, marriage status
Location statistics	14	GDP[1], HDI[1]
Current grades	75	Mathematics[2], Physics[2], Portuguese[2], English[2].
Dropout	1	Student drops out in that school year
Total	137	

Table 2: Educational stage and unique students in the dataset.

Educational stage	**Dropouts**	**Regulars**	**Total**	**Dropout rate**
Preschool	778	16,835	17,613	04.42%
Basic school	1,048	17,027	18,075	05.80%
Secondary school	1,472	12,459	13,931	10.57%

Table 3: Results from predicting dropout for different educational stages.

Best classifier	**Educational stage**	**Precision**	**Recall**
XGBoost	Preschool	74.58%	52.92%
XGBoost	Basic	91.91%	59.88%
XGBoost	Secondary	92.98%	72.78%

After creating the dataset, a series of classifiers were trained to predict dropouts at different moments in the year and educational stages. First, the results demonstrated in Table 3 show that the most challenging stage to predict dropout is preschool, a behavior that can be explained by the non-mandatory status of this educational stage in Brazil and the students' assessment not requiring grading and thus not making use of the new derived features. To analyze these results, we assume the secondary school education stage, in which a recall of 72.78% means that 72.78% of the students who quit their studies were pre-

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

dicted as such by the model; and a precision of 92.98% means that from all the students predicted as dropouts, 92.98% did, in fact, quit their studies.

The classifier that achieved the best results, independent of all the moments in the year, was XGBoost. This result is similar when analyzing the results over the different educational stages, as shown in Table 3.

Table 4: Results from predicting dropout in different moments in the year.

Best classifier	**Quarter**	**Precision**	**Recall**
XGBoost	1st quarter	83.25%	43.88%
XGBoost	2nd quarter	94.23%	75.87%
XGBoost	3rd quarter	95.23%	93.93%

The results in Table 4 show that dropout recognition rates improve later in the school year, which can be justified by the derived features which are present in the later sections of the school year. Nonetheless, the most significant improvement appears from the first to the second quarter.

To better understand the key factors for predicting dropout, the feature importance of each educational stage's classsifiers where evaluated. The results shown in Table 5 points that the both the $N_{cumulative}$ value and the overall Average Grade are of great importance in the Basic and Secondary stages. When analyzing the Preschool results the tuition fee, the city's HDI and the student's absence count were the features with the most importance when predicting dropout in this stage. It is relevant to highlight at this point that a high importance does not imply causation, yet, important features observed here serve as proxies for evasion.

Table 5: Most important features for predicting dropout in different educational stages.

Importance	**Preschool**	**Basic**	**Secondary**
1st	Tuition fee	$N_{cumulative}$ Phys. Ed.	$N_{cumulative}$ Portuguese
2nd	HDI	$N_{cumulative}$ Arts	$N_{cumulative}$ Geography
3rd	Absences	Avg. Grade	Avg. Grade

5 Conclusion

Student dropout is a serious problem that directly impacts students and society. While predicting student dropout is not a novel idea, the methodology adopted can be improved in the data acquisition or engineering process.

Predicting whether a student will drop out is in constant evolution, with two common approaches used: a rich dataset with a low number of students and a dataset with a higher number of students but less detailed information. We successfully predicted dropouts by adopting the latter approach coupled with feature engineering techniques. The proposed algorithms achieved precision

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

scores ranging from 83.25% to 95.23% when predicting dropouts in different moments (trimesters) of the school year. This study also demonstrates that the behaviors presented by each educational stage are not the same, with results showing that it is harder to identify whether a preschool student will drop out than their counterparts in basic and secondary stages. Potential future studies can evaluate the impact of the different features when predicting dropouts and elaborate on potential visualizations of these results to non-technical teams.

References

[1] Thomas D Snyder, Cristobal De Brey, and Sally A Dillow. Digest of education statistics 2017, nces 2018-070. *National Center for Education Statistics*, 2019.

[2] Russell W Rumberger. The economics of high school dropouts. *The economics of education*, pages 149–158, 2020.

[3] Melissa A Adelman and Miguel Szekely. School dropout in central america: An overview of trends, causes, consequences, and promising interventions. *World Bank Policy Research Working Paper 7561*, 2016.

[4] Yezhou Yao, Hongmei Yi, Linxiu Zhang, Wang Huan, Chu Yang, Yaojiang Shi, James Chu, Prashant Kumar Loyalka, and Scott Rozelle. Exploring dropout rates and causes of dropout in upper-secondary vocational schools. *Available at SSRN 2938383*, 2017.

[5] Carlos Márquez, Alberto Cano, Cristóbal Romero, Amin Mohammad, Habib Fardoun, and Sebastian Ventura. Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, 33:107–124, 02 2016.

[6] Ioanna Lykourentzou, Ioannis Giannoukos, Vassilis Nikolopoulos, George Mpardis, and Vassili Loumos. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education*, 53(3):950–965, 2009.

[7] Allan Sales, Leandro Balby, and Adalberto Cajueiro. Exploiting academic records for predicting student drop out: A case study in brazilian higher education. *Journal of Information and Data Management*, 7(2):166–166, 2016.

[8] UNESCO Institute for Statistics. International standard classification of education: Isced 2011. *Int. Stand. Classif. Educ.(ISCED) 2011*, 2012.

[9] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[10] Yuichiro Anzai. *Pattern recognition and machine learning*. Elsevier, 2012.

[11] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[12] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[13] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

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