Prediction of Students' Grades Based on Non-Academic Data

Beatriz Lacerda[1][0009−0000−1556−2346], Francisco S. Marcondes[1][0000−0002−2221−2261], Henrique Lima[2] Dalila Durães[1][0000−0002−8313−7023], and Paulo Novais[1][0000−0002−3549−0754]

[1] LASI/ ALGORITMI Centre, University of Minho, Guimarães, Portugal
[2] Codevision, S.A., Braga, Portugal
a89535@alunos.uminho.pt, francisco.marcondes@algoritmi.uminho.pt, ,henrique.lima@e-schooling.com, dad@di.uminho.pt, pjon@di.uminho.pt

Abstract. This study examines the use of machine learning techniques to predict Math and Portuguese grades based on student demographics and survey data regarding their school experiences. Using a sample of 53 middle school students, an accuracy rate of 93% was achieved with a support vector machine model. This paper's findings suggest that non-academic factors such as school climate and student engagement can have a significant impact on academic performance.

Keywords: Educational Data Mining · Academic Performance · Machine Learning

1 Introduction

The consequences of dropping out of school are significant, both for the individual and for the society in which he or she is inserted [1]. That being said, predicting academic performance has been one of the focus points of educational data mining. It is beneficial to detect declines in academic performance as early as possible, especially in the compulsory education domain. Unexpectedly, most studies focus on higher education. In addition, the vast majority of studies use academic data such as exam grades or papers eventually neglecting demographic or social data [2].

Nonetheless, studies have found that the school environment in which students are inserted as well as family support, socioeconomic situation, and motivation are also important and can explain the variation between expected academic performance and actual performance [3–6].

Taking this into consideration, this paper aims to investigate the relation between students' perception of school environment, their demographic data, their relationship with family with their overall study motivation and school performance. Thus it wants to answer the following research questions: RQ1) Is it possible to predict a student's academic performance without using academic data? RQ2) Which features are the most relevant within this domain?

Ethical Statement. Taking into account that the children are underaged, their parents signed an informed consent form so they could participate in the study. The collection of the data was controlled by the class director and a psychologist that were always present in the classroom.

2 Literature Review

This review based its steps on Kitchenham's Systematic Literature Review (SLR) Methodology guidelines [7] due to it be suited for research on the software engineering field. This literature review took place in March 2023 upon SCOPUS (www.scopus.com). The query aimed to identify studies related to the intersection of machine learning/predictive analysis, educational data mining/learning analytics, and academic achievements/student academic performance as follows:

```
TITLE-ABS-KEY(("Machine Learning" OR "Predictive Analysis")
    AND ("Educational Data Mining" OR "Learning Analytics")
    AND ("Academic Achievements" OR "Student Academic
    Performance"))
```

To screen the articles obtained, some exclusion criteria were established. As such, any documents that meet the following criteria were excluded/included:
AC1 Considered relevant considering the analysis being made - The paper in question was not written in the time frame previously mentioned. However, it is still recent and constitutes a literary review in the scope of this paper as it mentions the most relevant models and best features to consider.
EC1 Are not open access - The papers chosen had to be available so the authors could fully read it.
EC2 Are not written in Portuguese or English - That are the languages that the authors speak fluently.
EC3 Articles not written in 2023 or 2022 - The area of machine learning is in constant change and so this paper chose to focus on the most recent developments in this specif sector.
EC4 Are not from the field of computer science - The field of computer science is the main focus of this paper as it investigates the use of machine learning in the field of education.
EC5 Do not focus on the context of the study/explain clearly the dataset used - It was important for this literary review that the dataset used and its composition was discussed.

After searching the SCOPUS repository, 120 articles were identified, and the inclusion and exclusion criteria were later applied to them. Refer to Figure 1 for the filtering precedure.

As shown in Table **??**, the selected papers are presented in a synthesized way. It is also noticiable that ensemble models obtain the best results [9–16, 2] with an emphasis on the Random Forest model. Therefore, this paper opted to mostly choose ensemble models (XGBoost, Random Forest, Gradient Boost and AdaBoost). Aditionally, the SVM model was also used for this paper since it also

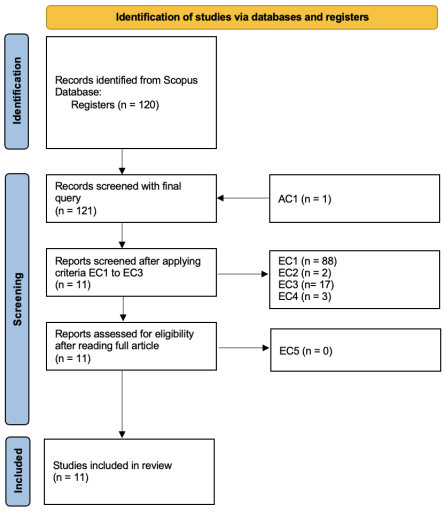

Fig. 1: Flow Diagram

produced positive results [11–14]. Furthermore, in relation to the evaluation of the top features to a model, [18] mentioned the Permutation Feature Importance technique to execute this assessment. Since, through this technique, [18] was able to present clear and good results it was selected to be used in this paper.

3 Data Mining

3.1 Data Preparation

The dataset used for this paper is proprietary and is composed by the answers of 53 Portuguese middle school students. It consisted of 64 questions. Through this questionnaire information was obtained about the students that ranged from demographic data such as their age and their parents' educational level to data related to their relationship with their friends at school, their teachers, family and overall study motivation. In the survey, most of the questions answer's could be one of 5: completely agree, agree, not sure, disagree, and completely disagree. To this there was only one exception. The question "My grades are:" had 5 possible possibilities: 1) way below what I can do; 2) below what I can do; 3) match what I can do; 4) above what I can do; 5) way above what I can do.

Veloso et al. [2] present a systematic review on the prediction of students' failure where students' activities such as raising their hand and visiting resources, were taken into account and considered relevant. In others activities English and Math scores were the most relevant. the results show that Random Forest, and XGBoost stand out. On the other hand, Verma et al. [9] shows the predicting the GPA of the first semester which were composed of 550 students and showed that while academic features were relevant, attributes such as parent's qualifications

or confidence also significantly contributed to the accuracy of the model. These models were Decision trees, Logistic Regression, Support Vector Machine. The best model was Ensemble with 3 Logistic Regression models.

Memon et al. [10] studying the importance of various features when it comes to predicting if students passed their final exams and the dataset was composed of 11814 students. The article showed that demographic data scored low accuracy. However, data such as the relationship of the student with their family affected the student's performance at school. The models used were Decision Tree, Random Forest, Logistic Regression, Naive Bayes, and the best result was the ensemble meta based models. Alboaneen et al. [11] create a web-based machine learning model in order to predict students' academic achievement but also to assess the best features to do it. The dataset contained information on 168 students from a university. In this particular study, it was concluded that test scores are the most relevant alongside other academic features. The models used were SVM, RF, K-NN, ANN, and LR. The best results were obtain by Random Forest.

Gaftandzhiev et al [12] aimed to explore data coming from e-learning platforms, were the data set consists of the final grades for 105 university students who study programming, their activities in the online course and attendance. This study found that there is a strong correlation between attendance and a good academic performance, solidifying the idea that students who are active and participatory in class tend to get better grades. The models used were Random Forest, XGBoost, KNN, and SVM where Random Forest obtain better results. Orrego Granados et al. [13]also studied the best variables and models to predict students' performance using data from students of a Peruvian university. The dataset focused mainly on academic data and it was found that the number of failed classes, grades of the first year and second year were the most crucial. It also recognizes it would be beneficial to add non-academic data to the study.11 models including Random Forest, Gradient Boost, K-NN, Linear SVM, XGBoost and Linear Regression, and the best model wasXGBoost.

Yağcı et al. [14] predict final exam grades of the students and the dataset used exclusively academic data. The results were compared with other studies that took demographic and socioeconomic data into account and acknowledged the positive impact this type of data has in the prediction. Models used were Random Forest, SVM, Neural Network, and K-Nearest Neighbour, and the best model was Random Forest. Ramaswami et al. [15] prediction of the student's risk of failing and used both academic and demographic data even though the demographic data used was limited. The model's prediction relies heavily on academic data, since the top three features in the model's reasoning are from this category. CatBoost, Random Forest, Naive Bayes, Logistic Regression, and K-Nearest Neighbours were used and CatBoost obtain the best results.

Alhazmi et al. [16] seeks to study the impact of several features on the final cumulative GPA of students at early stages and the dataset also contained a mix of several types of data. The best results were obtained through the combination of admission scores features and all first-level course scores. the models used

were Random Forest, XGBoost, and SVM, where Random Forest obtain the best combination. Poudyal et al. [17] want to build a hybrid 2D Convolutional Neural Network (CNN) to predict students' academic performance. The OULAD dataset was used that focuses primarily in academic features and the proposed CNN model, k-NN, and linear regression were used, but the CNN obtain the better result. Finally, Realinho et al. [18] aimed to predict whether or not a student was going to drop out. The dataset contained 4424 records from higher education institution and included academic, demographic, socioeconomic, and macroeconomic data. Academic data such as the number of approved curricular activities was significant but so were some non-academic factors such as the parents' occupation or whether the student was a scholarship holder. Models used were Random Forest, Catboost, XGBoost, and LIGHTGBM.

The questionnaire was developed by a team of psychologists and was presented this year in a classroom environment.

Data preparation steps according to CRISP-DM [8] were undertaken upon the dataset. The first step taken was to handle missing values and check for duplicate values. Additionally, there was a need to convert the data into numerical values in order to use it as input for the models. Afterwards, redundant data in the dataset was eliminated. In the survey there was more than one question regarding whether or not they received private tutoring, being one of them if they received it and other how often. That information was accumulated all in one column, where "1" corresponded to students who have no tutoring, "2" to students who have tutoring once a week, and so on. Then columns such as "username" and "class" were dropped. These columns were dropped since their sole purpose was to identify the students. Not only this information was not useful for the model, but this way the student's anonymity is also preserved. After observing the distribution of the grades it was found that the data is not balanced, and so oversampling [19] was performed. The tool used to do this was the RandomOverSampler from the imblearn.over_sampling library.

The resulting dataset then originated two datasets: one to predict Portuguese grades and another to predict Math grades. They have the same demographic information. However, the Portuguese dataset only contains the Portuguese final grades which constitute the target variable while the Math dataset only contains the students' final Math grade.

As seen in Table 1a, the model that worked the best with the Portuguese dataset was SVM, with which 93% of accuracy was obtained. To tune the models' hyperparameters, GridSearch was used which improved its performance. The hyperparameters obtained were:

```
Best Parameters: {'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
```

For the Math dataset, the best model was also SVM as it can be seen in Table 1b with an accuracy of 79%. The hyperparameters selected by GridSearch for this model were:

```
Best Parameters: {'C': 0.1, 'gamma': 0.1, 'kernel': 'poly'}
```

Table 1: Obtained predictions, Acc. means accuracy, Prec. means Precision, F1 refers to F1 Score and CV means Cross-Validation

(a) Model Results - Portuguese Grade Prediction

Model	Acc.	Prec.	F1	Recall	CV
XGBoost	0.89	0.89	0.89	0.89	0.86
AdaBoost	0.48	0.73	0.40	0.48	0.72
SVM	**0.93**	**0.94**	**0.93**	**0.93**	**0.92**
Gradient Boost	0.77	0.78	0.75	0.77	0.90
Random Forest	0.81	0.81	0.81	0.81	0.87

(b) Model Results - Math Grade Prediction

Model	Acc.	Prec.	F1	Recall	CV
XGBoost	0.71	0.73	0.70	0.71	0.82
AdaBoost	0.58	0.63	0.59	0.58	0.70
SVM	**0.79**	**0.85**	**0.78**	**0.79**	**0.826**
Gradient Boost	0.66	0.72	0.66	0.66	0.81
Random Forest	0.71	0.73	0.70	0.71	0.80

Table 2: Best Features for this prediction

(a) Questions corresponding to best features - Portuguese prediction

Question	Importance
What is your father's education level?	0.34814815
What is your mother's education level?	0.13333333
Most of what is important to know is learned in school.	0.11851852
Do you have tutoring? How many times a week?	0.11111111
I strive to understand the subjects, even when they are difficult.	0.11111111
I usually participate actively in class	0.0962963
My family/parents help me with schoolwork	0.0962963
My family/parents are interested in knowing if something good happens at school.	0.07407407
I turn in my homework on time.	0.06666667
After I finish my schoolwork, I review everything to see if it is correct.	0.06666667

(b) Questions corresponding to best features - Math prediction

Question	Importance
My family/parents help me with schoolwork.	0.05833333
The adults in my school listen to the students.	0.04166667
I do unexpected things without thinking about the long-term consequences.	0.04166667
When I have a project that takes some time to accomplish, I can keep working on it.	0.03333333
My friends respect what I have to say.	0.025
My family/parents talk to me about my problems.	0.025
My teachers listen to me when I want to say something.	0.01666667
My teachers believe that I will succeed.	0.01666667
I feel that my friends are good friends.	0.01666667
My grades are:	0.01666667

3.2 Discussion

RQ1 - *Is it possible to predict a student's academic performance without using academic data?* The results presented in this paper suggest that it is possible to obtain a prediction of a student's academic performance based on the relationship they have with their school and home environment and their study motivation. While academic performance factors such as previous grades may have the greatest correlation with academic performance, these results show that the perception of the students' environment is very relevant and has a direct impact on a student's ability to learn. This provides new insight to teachers and other faculty members and raises new questions regarding the importance of setting up a safe space for students.

RQ2 - *Which features are the most relevant within this domain?* This paper plotted the 10 most important features to the model that obtained the best results in predicting both the students' Portuguese and Math grades in order to assess which features were the most significant in obtaining these results and if some pattern could be found through the Permutation Feature Importance technique. Table 2a and 2b present the outcome of the use of that technique. As it can be observed, in the case of the Portuguese prediction, the questions selected were all related to the students' perception of their family environment, their participation and dedication when it comes to school work. This suggests that the student's own motivation and the support of the family in their academic pursuits are some of the most influential factors when it comes to their academic success. For this dataset, results suggests that friend's support influences a student's Math score. They also corroborate the previous result obtained that indicated that their relationship with authoritarian figures in their life is an influential factor.

4 Conclusion

In this paper it was assessed whether it is possible or not to predict a student's academic achievement without using any academic factors in the data. To do this, it was extracted a dataset from a survey that was distributed to several students and from which it was gathered demographic information as well as data regarding their relationship with the people in their lives, their own perception of the environment in which they are inserted, and their view on school and studying. After preprocessing the dataset, various classification models were used to obtain a prediction. The best result came from SVM with an accuracy that exceeded the 90% mark when predicting students' Portuguese grades.

These results suggest that the student's demographic data as well as their opinions on their school environment, their study motivation and the support from family are strong predictors of their academic performance and can provide valuable insights into predicting their grades.

However, there are some limitations in this study. The data used in the analysis was limited to a specific private school and may not be representative of the entire population. Secondly, the number of participants is limited which resulted in a relatively small dataset. With this being said, the study can be extended to include a larger sample size and data from different geographical regions to increase the generalization of the results.

Overall, this study has the potential to contribute to the development of effective and anticipated interventions that can improve the academic performance of students.

Acknowledgements This work is supported by: FCT - Fundação para a Ciência e Tecnologia within the RD Units Project Scope: UIDB/00319/2020 and the Northern Regional Operational Programme (NORTE 2020), under Portugal 2020 within the scope of the project "Hello: Plataforma inteligente para o combate ao insucesso escolar", Ref. NORTE- 01-0247-FEDER-047004 and by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

1. Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. Applied Sciences, 9(15), 3093.
2. Veloso, B., Barbosa, M. A., Faria, H., Marcondes, F. S., Durães, D., & Novais, P. (2023). A Systematic Review on Student Failure Prediction. In International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning (pp. 43-52). Springer, Cham.
3. Ekwochi, U., Osuorah, D. C., Ohayi, S. A., Nevo, A. C., Ndu, I. K., & Onah, S. K. (2019). Determinants of academic performance in medical students: evidence from a medical school in south-east Nigeria. Advances in Medical Education and Practice, 737-747.
4. Edgerton, E., & McKechnie, J. (2023). The relationship between student's perceptions of their school environment and academic achievement. Frontiers in Psychology, 13, 959259.
5. Durães, D., Carneiro, D., Jiménez, A., & Novais, P. (2018). Characterizing attentive behavior in intelligent environments. Neurocomputing, 272, 46-54.
6. Muntean, L. M., Nireștean, A., Sima-Comaniciu, A., Mărușteri, M., Zăgan, C. A., & Lukacs, E. (2022). The relationship between personality, motivation and academic performance at medical students from Romania. International Journal of Environmental Research and Public Health, 19(15), 8993.
7. B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," vol. 2, 2007
8. Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM.
9. Verma, S., Yadav, R. K., & Kholiya, K. (2022). A scalable machine learning-based ensemble approach to enhance the prediction accuracy for identifying students at-risk. International Journal of Advanced Computer Science and Applications, 13(8) doi:https://doi.org/10.14569/IJACSA.2022.0130822

10. Memon, M. Q., Lu, Y., Yu, S., Memon, A., & Memon, A. R. (2022). The Critical Feature Selection Approach using Ensemble Meta-Based Models to Predict Academic Performances. INTERNATIONAL ARAB JOURNAL OF INFORMATION TECHNOLOGY, 19(3 A), 523-529.
11. Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L., & Alharthi, R. (2022). Development of a Web-Based Prediction System for Students' Academic Performance. Data 2022, 7, 21.
12. Gaftandzhieva, Silvia & Talukder, Ashis & Gohain, Nisha & Hussain, Sadiq & Theodorou, Paraskevi & Salal, Yass & Doneva, Rositsa. (2022). Exploring Online Activities to Predict the Final Grade of Student. Mathematics. 10. 3758. 10.3390/Math10203758.
13. Orrego Granados, D., Ugalde, J., Salas, R., Torres, R., & López- Gonzales, J. L. (2022). Visual-Predictive Data Analysis Approach for the Academic Performance of Students from a Peruvian University. Applied Sciences, 12(21), 11251.
14. Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1), 11.
15. Ramaswami, G., Susnjak, T., & Mathrani, A. (2022). Supporting Students' Academic Performance Using Explainable Machine Learning with Automated Prescriptive Analytics. Big Data and Cognitive Computing, 6(4), 105.
16. Alhazmi, E., & Sheneamer, A. (2023). Early Predicting of Students Performance in Higher Education. IEEE Access.
17. Poudyal, S., Mohammadi-Aragh, M. J., & Ball, J. E. (2022). Prediction of student academic performance using a hybrid 2D CNN model. Electronics, 11(7), 1005
18. Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Succesè constitus. Data, 7(11), 146.
19. R. Mohammed, J. Rawashdeh and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 243-248, doi: 10.1109/ICICS49469.2020.239556.